Training Over-parameterized Deep ResNet Is almost as Easy as Training a Two-layer Network

Huishuai Zhang, Da Yu, Wei Chen, and Tie-Yan Liu

March 19, 2019

Abstract

It has been proved that gradient descent converges linearly to the global minima for training deep neural network in the over-parameterized regime. However, according to Allen-Zhu et al. [2018b], the width of each layer should grow at least with the polynomial of the depth (the number of layers) for residual network (ResNet) in order to guarantee the linear convergence of gradient descent, which shows no obvious advantage over feedforward network. In this paper, we successfully remove the dependence of the width on the depth of the network for ResNet and reach a conclusion that training deep residual network can be as easy as training a two-layer network. This theoretically justifies the benefit of skip connection in terms of facilitating the convergence of gradient descent. Our experiments also justify that the width of ResNet to guarantee successful training is much smaller than that of deep feedforward neural network.

1 Introduction

Although deep neural networks have achieved revolutionary success over various tasks, i.e., computer vision [He et al., 2016] and natural language understanding [Hochreiter and Schmidhuber, 1997], they are still in lack of a rigorous theoretical study of the optimization and generalization properties. Specifically for the optimization, because the loss of deep neural network is highly nonconvex, local search algorithms like gradient descent is hard to analyze with performance guarantee. Many recent works [Choromanska et al., 2015, Kawaguchi, 2016, Nguyen and Hein, 2017, Soudry and Hoffer, 2017] have studied the loss surface of the neural networks and a common claim is that (deep) neural networks have

---

*H. Zhang, W. Chen and TY Liu are with Microsoft Research Asia, Beijing, 100080 China (email: {huizhang, wche, tyliu}@microsoft.com); D. Yu is with School of Data and Computer Science at Sun Yat-sen University, Guangzhou, 510275, China (email: yuda3@mail2.sysu.edu.cn). The work of D. Yu was done when visiting MSRA.
relatively no bad local minima. However, the scenarios they study are often under many strict assumptions on the network architectures, i.e., deep linear network, or shallow network with one hidden layer or differentiable nonlinear activation, and on the input data i.e., Gaussian input data or linear separable input data. In fact, Safran and Shamir [2017] have shown that spurious local minima are common in two-layer ReLU neural networks. Overall, the loss surface study is still far from understanding practical models.

As most neural network models are trained with (stochastic) gradient descent, the optimization property of gradient descent in training deep neural network has also been widely studied. Soltanolkotabi et al. [2018], Brutzkus et al. [2017] point out that over-parameterization might play a key role in the convergence analysis of (stochastic) gradient descent. More recently, Li and Liang [2018], Du et al. [2019] prove that (stochastic) gradient descent converges linearly to the global minimum for training two-layer neural networks as long as the network is sufficiently over-parameterized. The high level idea is to show the gradient of the network exhibits good property at initialization and then argue the gradient descent finds global minimum within a neighborhood of the initialization, in which the benign property roughly maintains.

A breakthrough is achieved by Allen-Zhu et al. [2018b], Du et al. [2018], who extend the analysis to deep neural networks (more than two layers). Specifically, Du et al. [2018] prove that gradient descent finds global minimum with the assumption that activation function is smooth and some gram matrix at the last layer has lower bounded singular value. Their result requires that the width of the network grow exponentially with the depth of the network for feedforward network. At the same time, Allen-Zhu et al. [2018b] prove that the width of the network growing polynomially with the depth of the network for feedforward network with ReLU activation is enough to show the linear convergence of gradient descent. The high level idea is first bounding the forward and backward stability for deep networks and apply a similar argument for the convergence result of the two-layer’s case.

From the above theoretical results, it seems that any vanilla feedforward neural network can be successfully trained as long as it is sufficiently over-parameterized. Alternatively, the practical difficulty of training deep network, i.e., exploding or vanishing gradient, is due to that the network is not wide enough. However, in practice, with skip connection we can successfully train deep network with hundreds and thousands of layers without much difficulty. It naturally motivates us to ask

“Does residual network (ResNet) make itself preferable than vanilla feedforward network from the theoretic convergence analysis of gradient descent?”

We note that although Allen-Zhu et al. [2018b], Du et al. [2018] have established the convergence results of gradient descent for ResNet, their results do not clearly answer this question. Du et al. [2018] show that the provable training steps of ResNet is polynomial in the number of layers while vanilla feedforward network is exponential. Nonetheless, Allen-Zhu et al. [2018b] show that the provable training time for feedforward network is still polynomial in the number of layers and that for ResNet is also polynomial, which makes the benefit of ResNet unclear.
In this paper we establish that for ResNet the over-parameterization requirement on the width does not directly depend on the depth, which is the best possible result we can expect for the depth dependence. Our contribution can be summarized as follows.

- We show that the over-parameterization requirement for ResNet is almost independent with the depth of the network.
- We show that the provable training steps do not depend directly on the depth of the network, which recalls that training deep over-parameterized ResNet can be almost as easy as training a two-layer network.

Moreover, the over-parameterization for ResNet does not depend on the optimization accuracy\(^1\). Technically, we make several critical improvements over the proof in Allen-Zhu et al. [2018b] for analyzing the convergence of gradient descent training over-parameterized deep ResNet. Specifically, we exploit the fact that both the output change of each layer and the magnitude of the gradient on the parameters in the residual block become smaller as the depth of the network increases because the output of the parametric mapping in the residual block is scaled by \(\tau = 1/\Omega(L \log m)\) where \(L\) is the depth and \(m\) is the width, which is adopted in both Allen-Zhu et al. [2018b] and Du et al. [2018]. We note that \(\tau\) being small\(^2\) is necessary both for the proof and for the practice for our ResNet model that does not include batch normalization layer. We fully exploit such setting of \(\tau\) and successfully remove the dependence of the width \(m\) on the depth \(L\). Moreover, we also introduce two new proofs on bounding the forward stability and tighten several arguments in Allen-Zhu et al. [2018b]. Our theoretical result reflects that from the optimization perspective, the training deep neural network with skip connection is much easier than training vanilla feedforward network. Extensive experiments corroborate our finding.

### 1.1 Related works

Several papers argue the benefit of ResNet but they are either lack of rigorous theory or study the ResNet without nonlinear activation. Specifically, Veit et al. [2016] interpret ResNet behaves like an ensemble of shallower networks, which is imprecise because the shallower networks are trained jointly, not independently [Xie et al., 2017]. Zhang et al. [2018] argue the benefit of skip connection form the perspective of improving the local Hessian and Hardt and Ma [2016] show that deep linear residual networks have no spurious local optima.

The most related papers are Allen-Zhu et al. [2018b], Zou et al. [2018], Du et al. [2018]. Zou et al. [2018] shares the same high level proof idea as Allen-Zhu et al. [2018b] and studies binary classification problem and shows stochastic gradient descent can find the

\(^1\) The new version of Allen-Zhu et al. [2018b] also achieves this.

\(^2\) Preliminary experiments show that \(\tau\) may be improved to \(\Omega(1/\sqrt{L})\), whose rigorous argument needs further development.
global minimum when training an over-parameterized deep ReLU network. In contrast, we improve the condition guaranteeing that gradient descent finds global minimum for ResNet and achieve an optimal dependence of over-parameterization on the network depth.

People are skeptical about the over-parameterization partially because of the classic wisdom in learning theory: controlling the complexity of the function space leads to good generalization. However, the great success of deep learning urges to reconsider the generalization property in the over-parameterized regime. Recently, some progress has been made along this line. Brutzkus et al. [2017] provide both optimization and generalization guarantees of the SGD solution for over-parameterized two-layer networks given that the data is linear separable. Li and Liang [2018], Allen-Zhu et al. [2018a] show that the over-parameterized neural network provably generalize for two-layer and three-layer networks. Neyshabur et al. [2019] use unit-wise capacity and obtain a bound on the empirical Rademacher complexity, which can provide an explanation (not rigorous argument) of the generalization for over-parameterized two-layer ReLU networks.

Papers studying other over-parameterized models and the local geometry of neural networks are also related. Xu et al. [2018] show that over-parameterization can help Expectation Maximization avoid spurious local optima. A result with similar flavor [Li et al., 2017] has also been obtained for the matrix sensing problem. Chizat and Bach [2018] use optimal transport theory to analyze continuous time gradient descent on over-parameterized neural network with a single hidden layer. Oymak and Soltanolkotabi [2018], Fu et al. [2018], Zhou and Liang [2017] study the local geometry of neural networks that are responsible for the behavior of gradient descent.

1.2 Paper Organization

The rest of this paper is organized as follows. Section 2 introduce the model and notations. Section 3 presents the main results, including the theory and the proof roadmap. Section 4 presents the the proofs for theorems and critical lemmas. Section 5 gives some experiments that support our theory. Finally, we conclude in Section 6.

2 Model and Notations

There are many residual network models since the seminal paper He et al. [2016]. Here we study a very simple ResNet model\(^3\) because we are targeting understanding how skip connection help the optimization rather than achieving good performance. The ResNet model is described as follows,

---

\(^3\) The same ResNet model has been used in Allen-Zhu et al. [2018b] and Du et al. [2018]. Many notations are borrowed from Allen-Zhu et al. [2018b], which may help readers to better compare the results and proofs.
\begin{itemize}
  \item Input layer: \( h_0 = \phi(Ax) \);
  \item \( L - 1 \) residual layers: \( h_l = \phi(h_{l-1} + \tau W_l h_{l-1}) \), for \( l = 1, 2, \ldots, L - 1 \);
  \item A fully-connected layer: \( h_L = \phi(W_L h_{L-1}) \);
  \item Output layer: \( y = Bh_L \);
\end{itemize}

where \( \phi(\cdot) \) is the point-wise activation function, and we use ReLU activation \( \phi(\cdot) := \max\{0, \cdot\} \). Specifically, we assume the input dimension is \( p \) and hence \( x \in \mathbb{R}^p \), the intermediate layers have the same width \( m \), and hence \( h_l \in \mathbb{R}^m \) for \( l = 0, 1, \ldots, L \), and the output has dimension \( d \) and hence \( y \in \mathbb{R}^d \). Denote the values before activation by \( g_0 = Ax, g_l = h_{l-1} + \tau W_l h_{l-1} \) for \( l = 1, 2, \ldots, L - 1 \) and \( g_L = W_L h_{L-1} \). Use \( h_{i,l} \) and \( g_{i,l} \) to denote the value of \( h_l \) and \( g_l \), respectively, when the input vector is \( x_i \), and \( D_{i,l} \) the diagonal sign matrix where \( [D_{i,l}]_{k,k} = 1_{\{g_{i,l} \geq 0\}} \).

We adopt the following initialization scheme:

\begin{itemize}
  \item Each entry of \( A \in \mathbb{R}^{m \times p} \) is sampled independently from \( \mathcal{N}(0, \frac{2}{m}) \);
  \item Each entry of \( W_l \in \mathbb{R}^{m \times m} \) is sampled independently from \( \mathcal{N}(0, \frac{2}{m}) \) for \( l = 1, 2, \ldots, L \);
  \item Each entry of \( B \in \mathbb{R}^{d \times m} \) is sampled independently from \( \mathcal{N}(0, \frac{2}{d}) \).
\end{itemize}

Specifically, we set \( \tau = 1/\Omega(L \log m) \). We note that a small \( \tau \) is necessary both for the proof and for the practice for our ResNet model with the above initialization because there is not batch normalization layer. For example, with \( \tau = 1 \) the output of the ResNet explodes easily as the depth increases, which can be verified by calculating the expected value and by experiment. However, whether \( \tau = 1/\Omega(L \log m) \) can be improved requires further consideration.

The training data set is \( \{(x_i, y^*_i)\}_{i=1}^n \), where \( x_i \) is the feature vector and \( y^*_i \) is the target signal for all \( i = 1, \ldots, n \). We make the following assumption on the training data.

**Assumption 1.** For every pair \( i, j \in [n] \), we have \( \|x_i - x_j\| \geq \delta \).

We consider \( \ell_2 \) regression task and the objective function is

\[
F(\overrightarrow{W}) := \sum_{i=1}^n F_i(\overrightarrow{W}) \text{ where } F_i(\overrightarrow{W}) := \frac{1}{2} \|B h_{i,L} - y^*_i\|^2,
\]

where \( \overrightarrow{W} := (W_1, W_2, \ldots, W_L) \) are the trainable parameters. Specifically, we clarify some notations here. We use \( \|v\| \) to denote the \( \ell_2 \) norm of the vector \( v \). We further use \( \|M\|_2 \) and \( \|M\|_F \) to denote the spectral norm and the Frobenius norm of the matrix \( M \), respectively. Denote \( \|\overrightarrow{W}\|_2 := \max_{t \in [L]} \|W_t\|_2 \) and \( \|W_{[L-1]}\|_2 := \max_{t \in [L-1]} \|W_t\|_2 \).
We note that the initialization scheme, the choice of $\tau$ and the assumption on the data are the same as those in Allen-Zhu et al. [2018b] so that the result is comparable.

The training is conducted by running the gradient descent algorithm. The gradient is computed through back-propagation. Since the layer $L$ and the following layers $l \in [L-1]$ have different forms, we dispose them separately. Specifically for a fixed sample $i \in [n]$, we have

$$\nabla W_L F_i(\mathbf{W}) := D_{i,L} \left( B^T (B h_{i,L} - y_i^*) \right) h_{i,L-1}^T,$$

$$\nabla W_l F_i(\mathbf{W}) := \tau D_{i,l} \left( \text{Back}_{i,l+1} (B h_{i,L} - y_i^*) \right) h_{i,l-1}^T,$$

for layer $l \in [L-1], \quad \text{where } \text{Back}_{i,l+1} \text{ is a back-propagation operator to simplify the expression given by }$$

$$\text{Back}_{i,l} := BD_{i,L} W_L D_{i,L-1} (I + \tau W_{L-1}) \cdots D_{i,l}(I + \tau W_l).$$

For all $l \in [L]$, we define

$$\nabla W_l F(\mathbf{W}) := \sum_{i=1}^{n} \nabla W_l F_i(\mathbf{W}).$$

### 3 Main Result

Given the model introduced in Section 2, our main result for gradient descent is as follows.

**Theorem 1.** For the ResNet defined and initialized as in Section 2, if the network width $m \geq \max\{L, \Omega(n^{24}d^{-8} \log^2 m)\}$, then with probability at least $1 - \exp(-\Omega(\log^2 m))$, gradient descent with learning rate $\eta = \Theta(\frac{d \delta}{n^2 m})$ finds a point $F(\mathbf{W}) \leq \epsilon$ in $T = \Omega(n^6 \delta^{-2} \log \frac{n \log^2 m}{\epsilon})$ iterations.

This implies that gradient descent converges to global minima in linear time. The bound on $m$ does not depend on $L$ and $\epsilon$ directly if the third term in $m$ dominates, which usually should be the case. We have the following two remarks to compare our result with previous works.

**Remark 1.** Under the regime $L < \Omega(n^{24}d^{-8} \log^2 m)$, the network width requirement imposed on $m$ in Theorem 1 does not depend on the depth $L$, sharply in contrast with Allen-Zhu et al. [2018b] and Du et al. [2018].

**Remark 2.** The network width requirement imposed on $m$ in Theorem 1 does not directly depend on the optimization accuracy $\epsilon$.

We can also have a similar result for mini-batch stochastic gradient descent.
Theorem 2. For the ResNet defined and initialized as in Section 2, the network width \( m \geq \max\{L, \Omega(n^{2b-4d-8d \log^2 m})\} \). Suppose we do stochastic gradient descent update starting from \( \vec{W}^{(0)} \) and

\[
\vec{W}^{(t+1)} = \vec{W}^{(t)} - \eta \frac{n}{|S_t|} \sum_{i \in S_t} \nabla F_i(\vec{W}^{(t)}),
\]

where \( S_t \) is a random subset of \([n]\) with \( |S_t| = b \). Then with probability at least \( 1 - \exp(-\Omega(\log^2 m)) \), stochastic gradient descent (1) with learning rate \( \eta = \Theta(\frac{db \delta}{n^7 b^{-1} \delta^{-2} \log n \log^2 m}) \) finds a point \( F(\vec{W}) \leq \epsilon \) in \( T = \Omega(n^7 b^{-1} \delta^{-2} \log n \log^2 m) \) iterations.

In the following, we first present the proof’s high-level idea from a generic perspective of nonconvex optimization. We then give the proof roadmap for Theorem 1 and explain why and how we can achieve stronger result for optimizing over-parameterized ResNet.

3.1 Proof’s High-level Idea

From the generic nonconvex optimization, we understand that in order to build linear convergence to global minima of function value, one needs at least to build a gradient dominance condition. Suppose that \( x^* \) is a global minimizer of a generic function \( f: \mathbb{R}^d \to \mathbb{R} \), and \( B_{x^*}(\rho) \) is a neighborhood around \( x^* \) with radius \( \rho \), then the \( \lambda \)-gradient dominance condition with respect to \( x^* \) is depicted as

\[
\forall x \in B_{x^*}(\rho), \quad \frac{1}{\lambda} \| \nabla f(x) \|^2 \geq f(x) - f(x^*).
\]

Suppose further the gradient of \( f \) satisfies some smoothness condition, e.g., \( \nabla f(\cdot) \) is \( L \)-Lipschitz continuous

\[
f(x_2) \leq f(x_1) + \langle \nabla f(x_1), x_2 - x_1 \rangle + \frac{L}{2} \| x_2 - x_1 \|^2
\]

for all \( x_1, x_2 \in B_{x^*}(\rho) \). The gradient descent update step

\[
x^{(t+1)} \leftarrow x^{(t)} - \eta \cdot \nabla f(x^{(t)})
\]

gives the linear convergence of function value if choosing \( \eta < L \) [Karimi et al., 2016].

3.2 Proof Roadmap

Next one needs only to build similar gradient dominance condition and gradient smooth condition for deep ResNet to show the linear convergence of gradient descent.

We first build the gradient upper bound for deep ResNet.
Theorem 3. With probability at least $1 - \exp(-\Omega(m))$ over the randomness of $\mathbf{W}^{(0)}, A, B$, it satisfies for every $l \in [L - 1]$, every $i \in [n]$, and every $\mathbf{W}$ with $\|\mathbf{W} - \mathbf{W}^{(0)}\|_2 \leq \omega$ for $\omega \in [0, 1]$,

$$\|\nabla_{\mathbf{W}} F_i(\mathbf{W})\|_F^2 \leq O \left( \frac{F_i(\mathbf{W})}{d} \times \tau^2 m \right),$$

(2)

$$\|\nabla_{\mathbf{W}} F(\mathbf{W})\|_F^2 \leq O \left( \frac{F(\mathbf{W})}{d} \times \tau^2 mn \right),$$

(3)

$$\|\nabla_{W_L} F(\mathbf{W})\|_F^2 \leq O \left( \frac{F(\mathbf{W})}{d} \times mn \right).$$

(4)

We establish tighter gradient upper bound than Allen-Zhu et al. [2018b] by involving $\tau$ for the residual layers. Specifically, Theorem 3 treats the top layer $W_L$ and the residual layers $W_l$ for $l \in [L - 1]$ separately. This gives us the freedom to tighten the smoothness property in Theorem 5.

Theorem 4. Let $\omega = O \left( \frac{\delta^{3/2}}{n^{9/2} \log m} \right)$. With probability at least $1 - \exp(-\Omega(m\omega^{2/3}))$ over the randomness of $\mathbf{W}^{(0)}, A, B$, it satisfies for every $\mathbf{W}$ with $\|\mathbf{W} - \mathbf{W}^{(0)}\|_2 \leq \omega$,

$$\|\nabla_{W_L} F(\mathbf{W})\|_F^2 \geq \Omega \left( \frac{\max_{i \in [n]} F_i(\mathbf{W})}{dn/\delta} \times m \right).$$

(5)

This gradient lower bound on $\|\nabla_{W_L} F(\mathbf{W})\|_F^2$ acts like the gradient dominance condition and it is the same as Allen-Zhu et al. [2018b] except that our range on $\omega$ does not depend on the depth $L$.

With the help of Theorem 3 and several improvements, we can obtain a tighter bound on the semi-smoothness condition of the objective function.

Theorem 5. Let $\omega \in \left[ \Omega \left( \frac{d^{1.5}}{m \log^{1.5} m} \right), O(1) \right]$ and $\mathbf{W}^{(0)}, A, B$ be at random initialization. With probability at least $1 - \exp(-\Omega(m\omega^{2/3}))$ over the randomness of $\mathbf{W}^{(0)}, A, B$, we have for every $\mathbf{W} \in (\mathbb{R}^{m \times m})^L$ with $\|\mathbf{W} - \mathbf{W}^{(0)}\|_2 \leq \omega$, and for every $\mathbf{W}' \in (\mathbb{R}^{m \times m})^L$ with $\|\mathbf{W}'\|_2 \leq \omega$, we have

$$F(\mathbf{W} + \mathbf{W}') \leq F(\mathbf{W}) + \langle \nabla F(\mathbf{W}), \mathbf{W}' \rangle + \sqrt{n F(\mathbf{W})} \cdot \frac{\omega^{1/3} \sqrt{m \log m}}{\sqrt{d}} \cdot O \left( \sum_{l=1}^L \|W_l\|_2 \right)$$

$$+ O(\frac{nm}{d}) \|\mathbf{W}'\|_2^2.$$  

(6)

This semi-smoothness condition is stronger than Allen-Zhu et al. [2018b] because it removes the dependence of the right hand side on $L$ and it holds for larger region, i.e., the range of $\omega$ increases.
Our main improvements include the following, which will be more specific in Section 4.

- We provide a tighter bound on $\| h_l \|$, i.e., the representation at layer $l$. Now $\| h_l \|$ can be arbitrarily close to 1 for all depth ResNet, which is critical for downstream bounding tasks e.g., the $\delta$-separateness for proving Theorem 4.
- We enlarge the region where the good properties hold. Now $\omega$ breaks the dependence on the depth $L$.
- We improve the bound on the spectral norm of the perturbed intermediate mappings, which is helpful for downstream bounding task.

Finally, we can prove Theorem 1 with the help of Theorem 4, Theorem 3 and Theorem 5, which together produce a bound on the over-parameterization requirement of $m$.

**Outline Proof of Theorem 1**

We note that we remove the dependence of $m$ on the solution accuracy $\epsilon$ by employing the fact that the gradient norm shrinks to 0 exponentially fast along the path of gradient descent iteration. We also treat $W_L$ and $W_l, l \in [L - 1]$ separately to obtain a $L$-free bound on $m$. The complete proof is relegated to Appendix D.

Based on the forward stability and the randomness of $B$, we can show that $\| B h^{(0)}_{L}\iota; L - y^*_i \|_2^2 \leq O(\log^2 m)$ with probability at least $1 - \exp(-\Omega(\log^2 m))$, and therefore $F(\hat{W}^{(0)}) \leq O(n \log^2 m)$.

Assume that for every $t = 0, 1, \ldots, T - 1$,

$$\| W^{(t)}_L - W_L^{(0)} \|_F \leq \omega \Omega \left( \frac{\delta^3}{n^8} \right),$$

$$\| W^{(t)}_l - W_l^{(0)} \|_F \leq \tau \omega. \quad (8)$$

From Theorem 5 and Theorem 3, we can obtain that for one gradient descent step,

$$F(\hat{W}^{(t+1)}) \leq F(\hat{W}^{(t)}) - \eta \| \nabla F(\hat{W}^{(t)}) \|_F^2 + O \left( \frac{\eta \omega m}{d} + \frac{\eta^2 n^2 m^2}{d^2} \right) \cdot F(\hat{W}^{(t)})$$

$$\leq \left( 1 - \Omega \left( \frac{\eta \delta m}{dn^2} \right) \right) F(\hat{W}^{(t)}), \quad (9)$$

where the last inequality uses the gradient lower bound in Theorem 4 and the choice of $\eta = \frac{\delta \delta}{n^8 m}$ and the assumption on $\omega$. That is, after $T = \Omega \left( \frac{dn^2}{\eta \delta m} \right) \log \frac{n \log^2 m}{\epsilon} = \Omega \left( n^6 \delta^{-2} \right) \log \frac{n \log^2 m}{\epsilon}$ iterations, $F(\hat{W}^{(T)}) \leq \epsilon$.

We need to verify for each $t$, the iterate $\hat{W}^{(t)}$ stays in the region where good properties
hold. Therefore, we calculate

\[
\| W_L^{(t)} - W_L^{(0)} \|_F \leq \sum_{i=1}^{t-1} \| \eta \nabla W_L F(\bar{W}^{(i)}) \|_F
\]

\[
\overset{(a)}{\leq} O(\eta \sqrt{nm/d}) \sum_{i=1}^{t-1} \sqrt{F(\bar{W}^{(i)})} \overset{(b)}{\leq} O\left( \frac{n^3 \sqrt{d}}{\delta \sqrt{m} \log m} \right)
\]

where (a) is due to Theorem 3 and (b) is due to an upper bound of the sum of a geometric sequence. Similarly, we have for \( l \in [L-1] \),

\[
\| W_l^{(t)} - W_l^{(0)} \|_F \leq \sum_{i=1}^{t-1} \| \eta \nabla W_l F(\bar{W}^{(i)}) \|_F
\]

\[
\leq O(\eta \tau \sqrt{nm/d}) \sum_{i=1}^{t-1} \sqrt{F(\bar{W}^{(i)})} \leq O\left( \frac{\tau n^3 \sqrt{d}}{\delta \sqrt{m} \log m} \right).
\]

By combining (10) and the assumption on \( \omega (7) \), we obtain a bound on \( m \).

4 Proofs of Theorems and Critical Lemmas

In this section, we prove the theorems in Section 3 and introduce several lemmas that helps to establish the proofs. First we list several useful bounds on Gaussian distribution.

**Lemma 1.** Suppose \( X \sim N(0, \sigma^2) \), then

\[
\mathbb{P}\{|X| \leq x\} \geq 1 - \exp\left( -\frac{x^2}{2\sigma^2} \right),
\]

\[
\mathbb{P}\{|X| \leq x\} \leq \sqrt{\frac{2x}{\pi \sigma}}.
\]

Another bound is on the spectral norm of random matrix [Vershynin, 2012, Corollary 5.35].

**Lemma 2.** Let \( A \in \mathbb{R}^{N \times n} \), and entries of \( A \) are independent standard Gaussian random variables. Then for every \( t \geq 0 \), with probability at least \( 1 - \exp(-t^2/2) \) one has

\[
s_{\max}(A) \leq \sqrt{N} + \sqrt{n} + t,
\]

where \( s_{\max}(A) \) are the largest singular value of \( A \).

Next we give a useful lemma related to ResNet (slightly different from that in Allen-Zhu et al. [2018b]).
Lemma 3. For ResNet initialized as in Section 2, with probability at least \(1 - O(L) \cdot \exp(-\Omega(m))\), one have
\[
\|(I + \tau W_b)D_{i,b+1} \cdots D_{i,a}(I + \tau W_a)\|_2 \leq 1 + c
\]
for any \(L - 1 \geq b \geq a \geq 1\) and \(c\) can be made arbitrarily small by the choice of \(\tau\).

Next we show the good property at the initialization with the help of randomization and concentration. Then we show that such properties still hold after small perturbation. At last we prove that the perturbation is indeed small for gradient descent update with an appropriate step size.

4.1 Critical Lemmas at Initialization

The main idea is to build the forward and backward stability at the initialization, i.e., the norm and the distance are kept even after many layers’ mapping.

We first bound how the norm changes after layers’ mapping.

Lemma 4. With probability at least \(1 - O(nL) \cdot e^{-\Omega(\log^2 m)}\) over the randomness of \(A \in \mathbb{R}^{m \times p}\) and \(\overrightarrow{W} \in (\mathbb{R}^{m \times m})^L\), we have
\[
\forall i \in [n], l \in \{0, 1, \ldots, L\} : \|h_{i,l}\| \in [1 - c, 1 + c],
\]
where \(c\) can be arbitrarily small for the choice of \(\tau\) and a sufficiently large \(m\).

We note that Lemma 4 achieve stronger result than the argument in Allen-Zhu et al. [2018b] which cannot guarantee \(\|h_{i,l}\|\) arbitrarily close to 1. The property of \(\|h_{i,l}\|\) arbitrarily close to 1 is required for down-streaming bounding tasks. For example, the gradient lower bound (Theorem 4 requires this property and the \(\delta\)-separateness (Lemma 6).

Proof. With property (14), we can derive that \(\|h_{i,l}\| \leq 1.1\) for every \(i\) and \(l\). The lower bound on \(\|h_{i,l}\|\) is argued as follows for a fixed input \(i\).

Note that each coordinate of \(h_0\) follows i.i.d. from a distribution which is 0 with probability \(\frac{1}{2}\), and \(|N(0, \frac{2}{m})|\) with probability \(\frac{1}{2}\) [Allen-Zhu et al., 2018b, Fact 4.2]. Therefore, with probability \(1 - \Omega(m^2e^2), \|h_0\| \geq 1 - \epsilon\).

The event that \(|W_ih_{i,l-1}|_\infty < \log m\) for all input samples \(i \in [n]\) and all \(l \in [L]\) holds with probability at least \(1 - nL \cdot \exp(-\Omega(\log^2 m))\). Condition on the above event, we have \(\tau \sum_{a=1}^{l}(W_ah_{a-1})_k \leq O(\frac{\tau}{\sqrt{m}} \log m) \leq \frac{1}{10\sqrt{m}}\) due to the choice of \(\tau\).

Moreover, since \((h_0)_k\) is Gaussian with probability \(\frac{1}{2}\) and 0 with probability \(\frac{1}{2}\), then
\[
P\{(h_0)_k \geq \xi \} \geq \frac{1}{2} - \frac{1}{2\sqrt{\pi} \xi \sqrt{m}}, \quad \text{and} \quad P\{0 < (h_0)_k < \xi \} \leq \frac{1}{2\sqrt{\pi} \xi \sqrt{m}}.
\]
Let $S_1 = \{k : (h_0)_k \geq \xi\}$, then $\| (h_0)_{S_1} \|^2 = \|h_0\|^2 - \|(h_0)_{[m]}\|_{S_1}^2 \geq 1 - \frac{1}{2\sqrt{\pi}} m^{3/2} \xi^3$. If choosing $\xi = \frac{0.5}{\sqrt{m}}$, then we have $\| (h_0)_{S_1} \| \geq 0.98$. Hence $\| (h_0)_{S_1} \| - \frac{1}{\sqrt{2m}} \geq 0.9$. We note that the above constants 1.1, 0.9 and 0.98 can be made arbitrarily close to 1 by choosing $\tau$ appropriately and $m$ sufficiently large.

We next prove that the norm of a sparse vector through the network mapping.

**Lemma 5.** If $s \geq \Omega(d/ \log m)$, then for all $i \in [n]$ and $a \in [L]$ and for all $s$-sparse vectors $u \in \mathbb{R}^m$ and for all $v \in \mathbb{R}^d$,

$$|v^T B D_{i,l} W_i D_{i-1}(I + \tau W_{l-1}) \cdots D_{i,a}(I + \tau W_a) u| \leq O \left( \frac{\sqrt{s \log m}}{\sqrt{d}} \|u\| \|v\| \right)$$

holds with probability at least $1 - \exp(-\Omega(s \log m))$.

**Proof.** For any fixed vector $u \in \mathbb{R}^m$, $\| D_{i,l} W_i D_{i-1}(I + \tau W_{l-1}) \cdots D_{i,a}(I + \tau W_a) u \| \leq 1.1\|u\|$ holds with probability at least $1 - \exp(-\Omega(m))$ (over the randomness of $W_l, l \in [L]$).

On the above event, for a fixed vector $v \in \mathbb{R}^d$ and any fixed $W_l$ for $l \in [L]$, the randomness only comes from $B$, then $v^T B D_{i,l} W_i D_{i-1}(I + \tau W_{l-1}) \cdots D_{i,a}(I + \tau W_a) u$ is a Gaussian variable with mean 0 and variance no larger than $O(\|u\| \cdot \|v\|/\sqrt{d})$. Hence

$$\mathbb{P}\{|v^T B D_{i,l} W_i D_{i-1}(I + \tau W_{l-1}) \cdots D_{i,a}(I + \tau W_a) u| \geq \sqrt{s \log m} \cdot \Omega(\|u\| \|v\|/\sqrt{d})\}$$

$$= \text{erfc}(\Omega(\sqrt{s \log m})) \leq \exp(-\Omega(s \log m)).$$

Take $\varepsilon$-net over all $s$-sparse vectors of $u$ and all $d$-dimensional vectors of $v$, if $s \geq \Omega(d/ \log m)$ then with probability $1 - \exp(-\Omega(s \log m))$ the claim holds for all $s$-sparse vectors of $u$ and all $d$-dimensional vectors of $v$. \hfill \Box

We next give a bound on the distance of the representations $h_{i,l}$ and $h_{j,l}$ in each layer for two input vectors $x_i, x_j$ with $\|x_i - x_j\| \geq \delta$. In comparison with a similar result in Allen-Zhu et al. [2018b], our distance bound does not depend on the depth $L$.

**Lemma 6.** For any $\delta$ and any pair $(x_i, x_j)$ satisfying $\|x_i - x_j\|_2 \geq \delta$, then with probability at least $1 - O(n^2 L) \cdot \exp(-\Omega(\log^2 m))$,

$$\|h_{i,l} - h_{j,l}\| \geq \frac{\delta}{2},$$

holds for all $l \in [L]$.

**Proof.** The full proof is relegated to Appendix A. \hfill \Box
4.2 Critical Lemmas after Perturbation

Next we establish the forward stability after perturbation. We use \( \overrightarrow{W}^{(0)} = (W_{1}^{(0)}, \ldots, W_{L}^{(0)}) \) to denote the weight matrices at initialization and use \( \overrightarrow{W}^{'} \) to denote the perturbation matrices. Let \( \overrightarrow{W} = \overrightarrow{W}^{(0)} + \overrightarrow{W}^{'} \). Similarly, we define \( h_{i,l}^{(0)} = \phi((I + \tau W_{l}^{(0)} h_{i,l-1}^{(0)}) \) and \( h_{i,l} = \phi((I + \tau W_{l}^{(0)} + \tau W_{l}^{'} h_{i,l-1}) \) for \( l \in [L - 1] \), and \( h_{i,L}^{(0)} = \phi(W_{L}^{(0)} h_{i,L-1}^{(0)}) \) and \( h_{i,L} = \phi(\tau W_{L}^{0} + \tau W_{L}^{'} h_{i,L-1}) \). Furthermore, we let \( h_{i,l}^{'} := h_{i,l}^{(0)} - h_{i,l}^{(0)} \) and \( D_{i,l}^{'} := D_{i,l}^{(0)} - D_{i,l}^{(0)} \).

**Lemma 7.** Suppose for \( \omega \leq O(1) \), \( \|W_{L}^{'}\|_{2} \leq \omega \) and \( \|W_{l}^{'}\|_{2} \leq \tau \omega \) for \( l \in [L - 1] \). Then with probability at least \( 1 - \exp(-\Omega(m \omega^{2}/3)) \), the following bounds on \( h_{i,l}^{'} \) and \( D_{i,l}^{'} \) hold for all \( i \in [n] \) and all \( l \in [L - 1] \),

\[
\|h_{i,l}^{'}\| \leq O(\tau \omega), \quad \|D_{i,l}^{'}\|_{0} \leq m \omega^{2}/3, \tag{17}
\]

\[
\|h_{i,L}^{'}\| \leq O(\omega), \quad \|D_{i,L}^{'}\|_{0} \leq m \omega^{2}/3. \tag{18}
\]

**Proof.** The proof is relegated to Appendix B.

**Lemma 8.** With probability at least \( 1 - (nL) \cdot \exp(-\Omega(m)) \) over the randomness of \( \overrightarrow{W}^{(0)}, A_{i}, \) for every \( i \in [n], 1 \leq a \leq b \leq L - 1 \), for every diagonal matrices \( D_{i,b}^{''}, \ldots, D_{i,L}^{''} \) such that \( |(D_{i,b}^{''})_{k,k} + (D_{i,0}^{(0)})_{k,k}| \leq 1 \) for all \( k \in [m] \), for every perturbation matrices \( W_{1}', \ldots, W_{L}' \in \mathbb{R}^{m \times m} \) with \( \|\overrightarrow{W}^{'}\|_{2} \leq \omega \in [0, 1] \), we have

\[
\|I + \tau W_{b}^{(0)} (D_{i,b-1}^{(0)} + D_{i,b-1}^{''}) \cdots (D_{i,a}^{(0)} + D_{i,a}^{''}) (I + \tau W_{a}^{(0)})\|_{2} \leq O(1), \tag{19}
\]

\[
\|I + \tau W_{b}^{(0)} + \tau W_{b}^{'} (D_{i,b-1}^{(0)} + D_{i,b-1}^{''}) \cdots (D_{i,a}^{(0)} + D_{i,a}^{''}) (I + \tau W_{a}^{(0)} + \tau W_{a}^{''})\|_{2} \leq O(1). \tag{20}
\]

**Proof.** This is a direct result by using the argument as in the proof of Lemma 3.

We note the spectral norm bound in the above lemma does not depend on the depth \( L \) any more, in sharp contrast with the feedforward case.

4.3 Proofs of Theorems

**Proof of Theorem 4 (Gradient Lower Bound)**

Because the gradient is pathological and data-dependent, in order to build bound on the gradient, we need to consider all possible point and all cases of data. Hence we first introduce an arbitrary loss vector and then the gradient bound can be obtained by taking a union bound.
Definition 1 (Definition 6.1 in Allen-Zhu et al. [2018b]). For any vector tuple \( \overrightarrow{v} = (v_1, \ldots, v_n) \in (\mathbb{R}^d)^n \) (viewed as a fake loss vector), we define

\[
\nabla_{\overrightarrow{v}} F_{i,L}(\overrightarrow{W}) := D_{i,L} \left( B^T v_i \right) h_{i,L}^T,
\]

\[
\nabla_{\overrightarrow{v}} F_i(\overrightarrow{W}) := \tau D_{i,l} \left( \text{Back}_i^{T_{l+1}} v_i \right) h_{i,l-1}^T, \forall l \in [L - 1]
\]

\[
\nabla_{\overrightarrow{v}} F_{i,L}(\overrightarrow{W}) := \sum_{i=1}^{n} \nabla_{\overrightarrow{v}} F_i(\overrightarrow{W}), \forall l \in [L].
\]

Proof. The gradient lower-bound at the initialization is given in [Allen-Zhu et al., 2018b, Section 6.2] via the smoothed analysis [Spielman and Teng, 2004]: with high probability the gradient is lower-bounded, although the worst case it might be 0. The proof is the same given two preconditioned results Lemma 4 and Lemma 6. We shall not repeat the proof here.

Now suppose that we have \( \|\nabla_{W_L} F(\overrightarrow{W}(0))\|_F \geq \Omega \left( \frac{\max_{i \in [n]} F_i(\overrightarrow{W}(0))}{d_{n/\delta}} \times m \right) \). We next bound the change of the gradient after perturbing the parameter. Recall that

\[
\nabla_{\overrightarrow{v}} F_{i,L}(\overrightarrow{W}(0)) - \nabla_{\overrightarrow{v}} F_{i,L}(\overrightarrow{W}) = \sum_{i=1}^{n} \left( v_i^T B D_{i,L}^{(0)} \right) h_{i,L-1}^{(0)} T - \left( v_i^T B D_{i,L} \right) h_{i,L-1}^{(0)} T
\]

By Lemma 7 and Lemma 5, we know,

\[
\|v_i^T B D_{i,L}^{(0)} - v_i^T B D_{i,L}\| \leq O(\sqrt{m \omega^{2/3} \log m / \sqrt{d}}) \cdot \|v_i\|.
\]

Furthermore, we know

\[
\|v_i^T B D_{i,L}\| \leq O(\sqrt{m / d}) \cdot \|v_i\|.
\]

By Lemma 4 and Lemma 7, we have

\[
\|h_{i,L-1}^{(0)}\| \leq 1.01 \quad \text{and} \quad \|h_{i,L-1}^{(0)} - h_{i,L-1}\| \leq O(\omega).
\]

Combing the above bounds together, we have

\[
\|\nabla_{\overrightarrow{v}} F_{i,L}(\overrightarrow{W}(0)) - \nabla_{\overrightarrow{v}} F_{i,L}(\overrightarrow{W})\|_F \leq n\|\overrightarrow{v}\|^2 \cdot O(\sqrt{m \omega^{2/3} \log m / \sqrt{d}} + \omega \sqrt{m / d})^2
\]

\[
\leq n\|\overrightarrow{v}\|^2 \cdot O \left( \frac{m \log m}{d} \omega^{2/3} \right)
\]

Hence the gradient lower bound still holds for \( \overrightarrow{W} \) given \( \omega < O \left( \frac{d^{3/2}}{n^{9/2} \log^{3/2} m} \right) \).

Finally, taking \( \epsilon \)-net over all possible vectors \( \overrightarrow{v} = (v_1, \ldots, v_n) \in (\mathbb{R}^d)^n \), we prove that the above gradient lower bound holds for all \( \overrightarrow{v} \). In particular, we can now plug in the choice of \( v_i = B h_{i,L} - y_i^* \) and it implies our desired bounds on the true gradients. \( \square \)
Proof of Theorem 3 (Gradient Upper Bound)

For each $i \in [n]$, we have

$$\| \hat{\nabla}_{\tilde{W}_L} F_i(\tilde{W}) \|_F = \| D_{i,L} (B^T v_i) h_{i,L-1}^T \|_F = \| D_{i,L} (B^T v_i) \| \| h_{i,L-1}^T \| \leq O(\sqrt{m/d}) v_i. $$

Similarly, we have for $l \in [L-1]$

$$\| \hat{\nabla}_{\tilde{W}_L} F_i(\tilde{W}) \|_F = \| \tau D_{i,l} (\text{Back}_{i,l+1}^T v_i) h_{i,l-1}^T \|_F = \tau \| D_{i,l} (\text{Back}_{i,l+1}^T v_i) \| \| h_{i,l-1} \| \leq O(\tau \sqrt{m/d}) v_i. $$

The above upper bounds hold for the initialization $\tilde{W}^{(0)}$ because of Lemma 3 and Lemma 4. They also hold for all the $\tilde{W}$ such that $\| \tilde{W} - \tilde{W}^{(0)} \|_2 \leq \omega$ due to Lemma 7 and Lemma 8.

Finally, taking $\epsilon$–net over all possible vectors $\tilde{v} = (v_1, \ldots, v_n) \in (\mathbb{R}^d)^n$, we prove that the above bounds holds for all $\tilde{v}$. In particular, we can now plug in the choice of $v_i = B h_{i,L} - y_i^*$ and obtain the desired bounds on the true gradients.

Proof of Theorem 5 (Semi-smoothness)

With the above established lemmas, the proof of Theorem 5 can be derived. We relegate it into Appendix C.

5 Empirical study

In this section we show some empirical evidence to support our theoretical claim. Specifically, we train the feedforward fully-connected neural network models and ResNet models independently, and compare their convergence behavior.

5.1 Theory verification

The ResNet model and the initialization scheme is introduced in Section 2 and choose $\tau = 1/L$. The feedforward model is described as follows, The feedforward model is depicted as follows,

- Input layer: $h_0 = \phi(A x)$;
- $L$ feedforward layers: $h_l = \phi(W_l h_{l-1})$, for $l = 1, 2, \ldots, L$;
- Output layer: $y = B h_L$.

The feedforward model adopts the same initialization scheme as the ResNet model (see Section 2). The models are generated by varying the depth $L \in \{3, 10, 30, 100, 500, 1000\}$ and the width $m \in \{16, 128, 1024\}$.  

15
**Data and hyperparameters.** We use MNIST dataset [LeCun et al., 1998] and do the classification task. MNIST contains 60000 training examples with input dimension 784 and 10 labels. The input feature vector is normalized by subtracting feature mean and dividing the feature standard deviation. We train the model with SGD\(^4\) and the size of minibatch is 256. The learning rate \(lr\) is chosen as a function of the model width \(m\): 
\[lr = \frac{0.01}{(m/16)}.\]

**Experiment results.** For a given width, we evaluate the training performances of ResNet and feedforward NN with different depths as shown in Figure 1.

![Figure 1: Training curve comparison for a given width with different model depths. The left column is ResNet and the right column is feedforward NN. The widths from top to bottom are 16, 128, 1024 respectively. Training loss is evaluated on the whole dataset.](image)

Figures 1 shows that for a given width, the convergence of training ResNet does not depend on the depth much while training feedforward network becomes much harder as the depth increases. For example, 500-layer or 1000-layer feedforward NN with 1024 neurons each layer fails to converge while ResNet with only 16 neurons each layer succeeds to converge even with 1000 layers.

The failure of training feedforward network is mainly due to the network is not wide enough because it is able to train deeper networks if simply increasing the width. However, ResNet does not suffer from the training difficulty: training deep network is not essentially

\(^4\) We test GD for small models and observe the same phenomenon. We use SGD for all experiments due to the extremely expensive one-iteration cost of GD.
harder than training a two-layer network. Moreover, we attribute the fact that the wider network the better performance to the increasing capacity of network.

5.2 Influence of $\tau$

We have seen that a small $\tau$ plays a key role in establishing the convergence of gradient descent for training ResNet. It is natural to ask whether the small $\tau$ is necessary in practice and whether it reduces the expressivity or capacity of the network. In this section, we conduct experiments with different settings of $\tau$ and show how it influences the training performance. Our experiment setting is the same as Section 5.1 except that $\tau$ is chosen from $\{\frac{1}{L}, \frac{1}{L^{0.5}}, \frac{1}{L^{0.25}}\}$ for the ResNet model. We plot the training curves of different models with varying depths and widths in Figure 2. We can see that both $\tau = \frac{1}{L}$ and $\tau = \frac{1}{L^{0.5}}$ are able to guarantee the successful training of very deep ResNets. However, for $\tau = \frac{1}{L^{0.25}}$, the training losses explode for models with depth 50 and more. This suggests that a small $\tau$ is necessary to successfully training the deep ResNet$^5$ in practice. Moreover, we can see larger $\tau$ gives relatively better performance. Hence, $\tau$ should be set large to fully exploit the model capacity while it should be set small enough to guarantee that model can be efficiently trained.

We next empirically verify how our theory applies to a practical model ResNet18 He et al. [2016]. ResNet18 contains 10 residual blocks and each residual block contains two parametric convolution layers. We are interested in the influence of $\tau$ to this practical model. The task is to classify the CIFAR10 dataset[Krizhevsky & Hinton, 2009] that contains 60000 training samples with input dimension $32 \times 32 \times 3$ and 10 labels.

We use the standard preact ResNet18 model as the baseline. Then we remove all the batch normalization layers and multiply $\tau$ at each residual block before adding the residual. The $\tau$ is chosen from a set of $\{\frac{1}{L}, \frac{1}{L^{0.5}}, \frac{1}{L^{0.25}}\}$ where $L = 10$ is the number of residual blocks. We use the same hyper-parameters for all these variant models as the baseline model. We plot the performance curves of these variants and the baseline in Figure 5. We can see that by setting a small $\tau$, the real-world model can be trained effectively even without batch normalization, which corroborates the wide applicability of our theory. We note that empirically if $\tau = 1.0$, our ResNet18 cannot learn anything because of the forward explosion, which from the other side indicates that a small $\tau$ is essential to train network with residual connection if without batch normalization layer. Moreover, from Figure 5, we can see that larger $\tau$ leads to better performance, which reflects the influence of $\tau$ on model expressivity. More empirical studies on different dataset and architecture are available at Appendix F.

---

$^5$ There is not batch normalization layer in our setting of ResNet.
Figure 2: Training curve comparison for ResNet with different $\tau$. The columns from left to right are corresponding to ResNet with $\tau = \frac{1}{L}$, $\frac{1}{L^{0.5}}$ and $\frac{1}{L^{0.25}}$, respectively. The rows from top to bottom are corresponding to the width at each layer 16, 128 and 1024, respectively. The training loss is evaluated on the whole dataset.

Figure 3: Comparison of training performances between baseline ResNet18 and the variants our ResNet18 with $\tau = \{1/L, 1/L^{0.5}, 1/L^{0.25}\}$ and $L$ is the number of residual blocks.

6 Conclusion

In this paper, we establish a stronger result on the convergence behavior of gradient descent for training over-parameterized ResNet than recent work, which states training deep ResNet is as easy as training a two-layer network, bridging the gap between the theoretical result
and the practice. We note the theoretical proof relies on choosing a small $\tau$. A small $\tau$ definitely helps the optimization but may hurt the expressiveness of the network. With batch normalization [Ioffe and Szegedy, 2015] this problem can be mitigated but it is not clear how the proof adapts for network with batch normalization. As for future direction, one interesting question is can we achieve the convergence result with a much relaxed constraint on $\tau$, making the theory even closer to practice.

Acknowledgments

The authors would like to thank Prof. Yingbin Liang for many helpful discussions and thank Zeyuan Allen-Zhu for clarifying the proof in Allen-Zhu et al. [2018b] and discussion.

References

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

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

Alon Brutzkus, Amir Globerson, Eran Malach, and Shai Shalev-Shwartz. Sgd learns over-parameterized networks that provably generalize on linearly separable data. *arXiv:1710.10174*, 2017.

Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for overparameterized models using optimal transport. In *Advances in Neural Information Processing Systems 31*. 2018.

Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In *Artificial Intelligence and Statistics*, 2015.

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*, 2018.

Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019.

Haoyu Fu, Yuejie Chi, and Yingbin Liang. Local geometry of one-hidden-layer neural networks for logistic regression. *arXiv preprint arXiv:1802.06463*, 2018.

Moritz Hardt and Tengyu Ma. Identity matters in deep learning. In *International Conference on Learning Representations*, 2016.
Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, pages 448–456, 2015.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, pages 586–594, 2016.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

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

Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix recovery. *arXiv preprint arXiv:1712.09203*, 2017.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. 2011.

Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. The role of over-parametrization in generalization of neural networks. In *International Conference on Learning Representations*, 2019.

Quynh Nguyen and Matthias Hein. The loss surface of deep and wide neural networks. 2017.

Samet Oymak and Mahdi Soltanolkotabi. Overparameterized nonlinear learning: Gradient descent takes the shortest path? *arXiv:1812.10004*, 2018.

Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer relu neural networks. *arXiv:1712.08968*, 2017.
Mahdi Soltanolkotabi, Adel Javanmard, and Jason D Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 2018.

Daniel Soudry and Elad Hoffer. Exponentially vanishing sub-optimal local minima in multilayer neural networks. *arXiv:1702.05777*, 2017.

Daniel A Spielman and Shang-Hua Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *Journal of the ACM (JACM)*, 51(3):385–463, 2004.

Andreas Veit, Michael J Wilber, and Serge Belongie. Residual networks behave like ensembles of relatively shallow networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *Compressed Sensing, Theory and Applications*, 2012.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Ji Xu, Daniel J Hsu, and Arian Maleki. Benefits of over-parameterization with em. In *Advances in Neural Information Processing Systems*. 2018.

Huishuai Zhang, Wei Chen, and Tie-Yan Liu. On the local hessian in back-propagation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Yi Zhou and Yingbin Liang. Characterization of gradient dominance and regularity conditions for neural networks. *arXiv:1710.06910*, 2017.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Stochastic gradient descent optimizes over-parameterized deep relu networks. *arXiv:1811.08888*, 2018.
A Proof of Lemma 6

Proof. In the input layer since \( \|x_i - x_j\| \geq \delta \), we use numerical integral to calculate
\[
\mathbb{E}[|\phi(A_k x_i) - \phi(A_k x_j)|] \geq \sqrt{\frac{1}{\delta}} \delta (1 - \delta). \]
Hence with high probability
\[
\|h_{i,0} - h_{j,0}\|_1 \geq \frac{\delta}{2} \sqrt{m}. \tag{22}
\]

In the following layer \( l \), we have the following four cases.

Case 1: \( (g_{i,l})_k > 0 \) and \( (g_{j,l})_k > 0 \). This case follows Allen-Zhu et al. [2018b], and one has
\[
(h_{i,l} - h_{j,l})_k = (h_{i,l-1})_k - (h_{j,l-1})_k + \tau (W_l(h_{i,l-1} - h_{j,l-1}))_k
\]
Let \( s_k = 1 \) if \( (h_{i,l-1})_k - (h_{j,l-1})_k \geq 0 \) and \( s_k = -1 \) otherwise. Then,
\[
|(h_{i,l} - h_{j,l})_k| \geq |(h_{i,l-1})_k - (h_{j,l-1})_k| + \tau \cdot s_k (W_l(h_{i,l-1} - h_{j,l-1}))_k.
\]

Case 2: \( (g_{i,l})_k > 0 \) and \( (g_{j,l})_k < 0 \). We have
\[
|(h_{i,l} - h_{j,l})_k| = (h_{i,l-1})_k + \tau (W_l h_{i,l-1})_k
\]
\( (a) \)
\[
\geq (h_{i,l-1})_k + (h_{j,l-1})_k + \tau (W_l h_{j,l-1})_k + \tau (W_l h_{i,l-1})_k
\]
\( (b) \)
\[
\geq |(h_{i,l-1})_k - (h_{j,l-1})_k| + \tau (W_l(h_{i,l-1} + h_{j,l-1}))_k,
\]
where (a) is due to the case assumption and (b) is due to the fact \( (h_{j,l-1})_k \geq 0 \).

Case 3: \( (g_{i,l})_k < 0 \) and \( (g_{j,l})_k > 0 \); and Case 4: \( (g_{i,l})_k < 0 \) and \( (g_{j,l})_k < 0 \). Similar to Case 2, we have
\[
|(h_{i,l} - h_{j,l})_k| \geq |(h_{i,l-1})_k - (h_{j,l-1})_k| + \tau (W_l(h_{j,l-1} + h_{i,l-1}))_k.
\]

Hence we have
\[
\sum_{k \in [m]} |(h_{i,l} - h_{j,l})_k| \geq \sum_{k \in [m]} |(h_{i,l-1} - h_{j,l-1})_k|
\]
\[
+ \sum_{k : (g_{i,l})_k > 0 \text{ and } (g_{j,l})_k > 0} \tau \cdot s_k (W_l(h_{i,l-1} - h_{j,l-1}))_k
\]
\[
+ \sum_{k : (g_{i,l})_k \leq 0 \text{ or } (g_{j,l})_k \leq 0} \tau \cdot (W_l(h_{j,l-1} + h_{i,l-1}))_k. \tag{23}
\]
We claim that $s_k(W_l(h_{i,l-1} - h_{j,l-1}))_k$ and $(W_l(h_{i,l-1} + h_{j,l-1}))_k$ are independent Gaussian variables with zero mean. Therefore, with probability at least $1 - \exp(-\Omega(\log^2 m))$, the sum of the last two terms in (23) is at most $O(\log m)$ in absolute value. Hence,

$$\|h_{i,l} - h_{j,l}\|_1 \geq \|h_{i,l-1} - h_{j,l-1}\|_1 - O(\tau \log m).$$

Continuing this argument for $l = 2, 3, \ldots, L - 1$, we have that every time we move from layer $l - 1$ to layer $l$, the $\ell_1$ norm of $h_l$ decreases by $O(\tau \log m)$. Putting this together, we know $\|h_l\|_1 \geq \|h_0\|_1 - L \cdot O(\tau \log m) \geq \frac{1}{2} \delta \sqrt{m} - O(\log m)$ for all $l = 1, 2, \ldots, L - 1$ with the choice of $\tau = \Omega(L \log m)$. Therefore, $\|h_{i,l} - h_{j,l}\| \geq \frac{1}{2} \delta$. The probability that the above events hold for all pair $(i, j)$ and all layers is $1 - O(n^2 L) \cdot \exp(-\Omega(\log^2 m))$. 

\[\Box\]

## B Proof of Lemma 7

**Proof.** Fixing $i$ and ignoring the subscript in $i$, by Claim 8.2 in Allen-Zhu et al. [2018b], for $l \in [L - 1]$, there exists $D''_l$ such that $|(D''_l)_k,k| \leq 1$ and

$$h'_l = D''_l ((I + \tau W_l + \tau W'_l)h_{l-1} - (I + \tau W_l)h^{(0)}_{l-1})$$

$$= D''_l ((I + \tau W_l + \tau W'_l)h'_{l-1} + \tau W'_lh^{(0)}_{l-1})$$

$$= D''_l (I + \tau W_l + \tau W'_l)D'_{l-1}(I + \tau W_{l-1} + \tau W'_{l-1})h'_{l-2}$$

$$+ \tau D''_l (I + \tau W_l + \tau W'_l)D''_{l-1}W_{l-1}h^{(0)}_{l-2} + \tau D''_l W'_lh^{(0)}_{l-1}$$

$$= \cdots$$

$$= \sum_{a=1}^l \tau D''_a (I + \tau W_l + \tau W'_l) \cdots D''_{a+1}(I + \tau W_{a+1} + \tau W'_{a+1})D''_a W'_a h^{(0)}_a.$$  \hspace{3cm} (24)

We claim that

$$\|h'_l\| \leq O(\tau \omega)$$  \hspace{3cm} (25)

due to the choice of $\tau$, the fact $\|D''_l\|_2 \leq 1$ and the assumption $\|W'_l\|_2 \leq \tau \omega$ for $l \in [L - 1]$. This implies that $\|h'_l\|, \|g'_l\| \leq O(\tau \omega)$ for all $l \in [L - 1]$ and for all $i$ with probability at least $1 - O(nL) \cdot \exp(-\Omega(m))$. One step further, we have $\|h''_l\|, \|g''_l\| \leq O(\omega)$.

As for the sparsity $\|D'_l\|_0$, we have $\|D'_l\|_0 \leq O(m\omega^{2/3})$ for every $l = [L]$.

The argument as follows (adapt from the Claim 5.3 in Allen-Zhu et al. [2018b]).

We first study the case for $l \in [L - 1]$. We observe that if $(D'_l)_{j,j} \neq 0$ one must have

$$|(g'_l)_{j,j}| > |(g^{(0)}_l)_{j,j}|.$$
We note that $(g_i^{(0)})_j = (h_i^{(0)} + \tau W_i h_i^{(0)})_j \sim N \left( (h_i^{(0)})_j, \frac{2\tau^2 \| h_i^{(0)} \|^2}{m} \right)$. Let $\xi \leq \frac{1}{\sqrt{m}}$ be a parameter to be chosen later. Let $S_1 \subseteq [m]$ be a index set satisfying $S_1 := \{ j : |(g_i^{(0)})_j | \leq \xi \tau \}$. We have $\mathbb{P} \{ |(g_i^{(0)})_j | \leq \xi \tau \} \leq O(\xi \sqrt{m})$ for each $j \in [m]$. By Chernoff bound, with probability at least $1 - \exp(-\Omega(m^{3/2} \xi))$ we have

$$|S_1| \leq O(\xi m^{3/2}).$$

Let $S_2 := \{ j : j \notin S_1, \text{ and } (D'_i)_{j,j} \neq 0 \}$. Then for $j \in S_2$, we have $|(g'_{ij})_j | > \xi \tau$. As we have proved that $\| g'_{ij} \| \leq O(\tau \omega)$, we have

$$|S_2| \leq \frac{\| g'_{ij} \|^2}{(\xi \tau)^2} = O(\omega^2 / \xi).$$

Choosing $\xi$ to minimize $|S_1| + |S_2|$, we have $\xi = \omega^{2/3} / \sqrt{m}$ and consequently, $\| D'_i \|_0 \leq O(m \omega^{2/3})$. Similarly, we have $\| D'_L \|_0 \leq O(m \omega^{2/3})$. □

C Proof of Theorem 5

Before going to the proof of the theorem, we introduce a lemma.

**Lemma 9.** There exist diagonal matrices $D''_{i,l} \in \mathbb{R}^{m \times m}$ with entries in [-1,1] such that $\forall i \in [n], \forall l \in [L - 1],$

$$h_{i,l} - \tilde{h}_{i,l} = \sum_{a=1}^{l} (\tilde{D}_{i,l} + D''_{i,l}) (I + \tau W_l) \cdots (I + \tau W_{a+1}) (\tilde{D}_{i,a} + D''_{i,a}) \tau W'_{a} h_{i,a-1},$$

and

$$h_{i,L} - \tilde{h}_{i,L} = (\tilde{D}_{i,L} + D''_{i,L}) W'_L h_{i,L-1}$$

$$+ \sum_{a=1}^{L-1} (\tilde{D}_{i,L} + D''_{i,L}) W_L \cdots (I + \tau W_{a+1}) (\tilde{D}_{i,a} + D''_{i,a}) \tau W'_{a} h_{i,a-1}.\tag{27}$$

Furthermore, we have $\forall l \in [L-1], \| h_{i,l} - \tilde{h}_{i,l} \| \leq O(1) \| W' \|_2, \| Bh_{i,l} - B\tilde{h}_{i,l} \| \leq O(\sqrt{m/d}) \| W' \|_2, \| D''_{i,l} \|_0 \leq O(m \omega^{2/3})$, and $\| h_{i,L} - \tilde{h}_{i,L} \| \leq O(1) \| W'_L \|_2, \| Bh_{i,L} - B\tilde{h}_{i,L} \| \leq O(\sqrt{m/d}) \| W'_{L} \|_2, \| D''_{i,L} \|_0 \leq O(m \omega^{2/3})$ hold with probability at least $1 - \exp(-\Omega(m \omega^{2/3}))$ given $\omega \leq O(1)$.

**Proof.** The proof can adapt from the proof of Claim 8.2 in Allen-Zhu et al. [2018b] and the proof of Lemma 7. □
Proof of Theorem 5. First of all, we know that 
\[ \text{loss}_i := B\tilde{h}_{i,L} - y_i^* \]
\[
\frac{1}{2}\|B\tilde{h}_{i,L} - y_i^*\|^2 = \frac{1}{2}\|\text{loss}_i + B(h_{i,L} - \tilde{h}_{i,L})\|^2 \\
= \frac{1}{2}\|\text{loss}_i\|^2 + \text{loss}_i^T B(h_{i,L} - \tilde{h}_{i,L}) + \frac{1}{2}\|B(h_{i,L} - \tilde{h}_{i,L})\|^2, \quad (28)
\]
and
\[
\nabla_{\bar{W}_i} F(\bar{W}) = \sum_{i=1}^n (\text{loss}_i^T B D_{i,L} W_L \cdots D_{i,L} (I + \tau W_i) D_{i,L})^T (\tau h_{i,L-1})^T. \quad (29)
\]
\[
\nabla_{\bar{W}_L} F(\bar{W}) = \sum_{i=1}^n (\text{loss}_i^T B D_{i,L})^T (h_{i,L-1})^T. \quad (30)
\]

Then,
\[
\begin{multline}
F(\tilde{W} + \hat{W}^*) - F(\tilde{W}) - \langle \nabla F(\tilde{W}), \hat{W}^* \rangle \\
= -\langle \nabla F(\tilde{W}), \hat{W}^* \rangle + \frac{1}{2} \sum_{i=1}^n \|B\tilde{h}_{i,L} - y_i^*\|^2 - \|B\tilde{h}_{i,L} - y_i^*\|^2 \\
= -\sum_{i=1}^n (\nabla_{\bar{W}_i} F(\bar{W}), W_i') + \sum_{i=1}^n \text{loss}_i^T B(h_{i,L} - \tilde{h}_{i,L}) + \frac{1}{2}\|B(h_{i,L} - \tilde{h}_{i,L})\|^2 \\
= \frac{1}{2} \sum_{i=1}^n \|B(h_{i,L} - \tilde{h}_{i,L})\|^2 + \sum_{i=1}^n \text{loss}_i^T B((\tilde{D}_{i,L} + D''_{i,L}) W_L h_{i,L-1} - (\tilde{D}_{i,L}) W_L \tilde{h}_{i,L-1}) \\
+ \sum_{i=1}^n \sum_{l=1}^{L-1} \text{loss}_i^T B((\tilde{D}_{i,L} + D''_{i,L}) W_L \cdots (I + \tau \tilde{W}_{l+1}) (\tilde{D}_{i,L} + D''_{i,L}) \tau W_i h_{i,L-1} \\
- \tilde{D}_{i,L} \tilde{W}_L \cdots (I + \tau \tilde{W}_{l+1}) \tilde{D}_{i,L} W_i ((\tau \tilde{h}_{i,L-1}))) \quad (31),
\end{multline}
\]
where (a) is due to Lemma 9.

We next bound the RHS of (31). We first use Lemma 9 to get
\[
\|B(h_{i,L} - \tilde{h}_{i,L})\| \leq O(\sqrt{m/d}) \|\hat{W}^*\|_2. \quad (32)
\]
Next we calculate that for \( l = L \),
\[
\left| \text{loss}_i^T B((\tilde{D}_{i,L} + D''_{i,L}) W_L h_{i,L-1} - (\tilde{D}_{i,L}) W_L \tilde{h}_{i,L-1}) \right| \\
\leq \left| \text{loss}_i^T B(D''_{i,L} W_L h_{i,L-1}) \right| + \left| \text{loss}_i^T B((\tilde{D}_{i,L} W_L (h_{i,L-1} - \tilde{h}_{i,L-1})) \right|. \quad (33)
\]
For the first term, by Lemma 5 and Lemma 9, we have
\[
\left| \text{loss}_i^T B(D''_{i,L} W_L h_{i,L-1}) \right| \leq O\left( \frac{\sqrt{m\omega^{2/3}}}{}^{\sqrt{d}} \right) \|\text{loss}_i\| \cdot \|W'_L h_{i,L-1}\| \\
\leq O\left( \frac{\sqrt{m\omega^{2/3}}}{}^{\sqrt{d}} \right) \|\text{loss}_i\| \cdot O(\|W'_L\|_2). \quad (34)
\]

25
where the last inequality is due to \( \|h_{i,L-1}\| \leq O(1) \). For the second term, by Lemma 9 we have

\[
\begin{align*}
\left| \tilde{\text{loss}}_i^T \mathbf{B} \left( \tilde{\mathcal{D}}_{i,L} \mathbf{W}_L'(h_{i,L-1} - \tilde{h}_{i,L-1}) \right) \right| \\
\leq \|\tilde{\text{loss}}_i\| \cdot \left\| \mathbf{B} \tilde{\mathcal{D}}_{i,L} \right\|_2 \cdot \|\mathbf{W}_L'\|_2 \|h_{i,L-1} - \tilde{h}_{i,L-1}\| \\
\leq \|\tilde{\text{loss}}_i\| \cdot O \left( \frac{\sqrt{m}}{\sqrt{d}} \right) \cdot \|\mathbf{W}_L'\|_2 \cdot O(\|\mathbf{W}_L'\|_2) \\
\leq \|\tilde{\text{loss}}_i\| \cdot O \left( \frac{\omega \sqrt{m}}{\sqrt{d}} \right) \cdot \|\mathbf{W}_L'\|_2,
\end{align*}
\]

(35)

where the last inequality is due to the assumption \( \|\mathbf{W}_L'\|_2 \leq \omega \). Similarly for \( l \in [L-1] \), we ignore the index \( i \) for simplicity.

\[
\begin{align*}
\left| \tilde{\text{loss}}_i^T \mathbf{B} (\tilde{\mathcal{D}}_L + D''_{L-1}) \tilde{\mathbf{W}}_L \cdots (I + \tau \tilde{\mathbf{W}}_{l+1}) (\tilde{\mathcal{D}}_l + D''_l) \mathbf{D}_l \right| \\
= \left| \tilde{\text{loss}}_i^T \mathbf{B} D''_L \tilde{\mathbf{W}}_L (D_L - 1 + D''_{L-1}) (I + \tau \tilde{\mathbf{W}}_{L-1}) \cdots (D_l + D''_l)(\tau \mathbf{W}_l') (\tau h_{l-1}) \right| \\
+ \sum_{a=1}^{L-1} \left| \tilde{\text{loss}}_i^T \mathbf{B} D''_L \tilde{\mathbf{W}}_L \cdots (I + \tau \tilde{\mathbf{W}}_{a+1}) (\tau \mathbf{W}_a) \cdots (D_l + D''_l)(\tau \mathbf{W}_l') (\tau h_{l-1}) \right| \\
+ \sum_{a=1}^{L-1} \left| \tilde{\text{loss}}_i^T \mathbf{B} D''_L \tilde{\mathbf{W}}_L \cdots (I + \tau \tilde{\mathbf{W}}_{a+1}) (\tau \mathbf{W}_a) \cdots (D_l + D''_l)(\tau \mathbf{W}_l') (\tau h_{l-1}) \right| \\
\leq \left| \tilde{\text{loss}}_i^T \mathbf{B} D''_L \tilde{\mathbf{W}}_L (D_L - 1 + D''_{L-1}) (I + \tau \tilde{\mathbf{W}}_{L-1}) \cdots (D_l + D''_l)(\tau \mathbf{W}_l') (\tau h_{l-1}) \right| \\
+ \sum_{a=1}^{L-1} \left| \tilde{\text{loss}}_i^T \mathbf{B} D''_L \tilde{\mathbf{W}}_L \cdots (I + \tau \tilde{\mathbf{W}}_{a+1}) (\tau \mathbf{W}_a) \cdots (D_l + D''_l)(\tau \mathbf{W}_l') (\tau h_{l-1}) \right| \\
+ \sum_{a=1}^{L-1} \left| \tilde{\text{loss}}_i^T \mathbf{B} D''_L \tilde{\mathbf{W}}_L \cdots (I + \tau \tilde{\mathbf{W}}_{a+1}) (\tau \mathbf{W}_a) \cdots (D_l + D''_l)(\tau \mathbf{W}_l') (\tau h_{l-1}) \right| \\
\end{align*}
\]

(36)

We next bound the terms in (36) one by one. For the first term, by Lemma 5 and Lemma 9, we have

\[
\begin{align*}
\left| \tilde{\text{loss}}_i^T \mathbf{B} D''_L \tilde{\mathbf{W}}_L (D_L - 1 + D''_{L-1}) (I + \tau \tilde{\mathbf{W}}_{L-1}) \cdots (D_l + D''_l)(\tau \mathbf{W}_l') (\tau h_{l-1}) \right| \\
\leq O \left( \frac{\sqrt{m \omega^{2/3} \log m}}{\sqrt{d}} \right) \cdot \left| \tilde{\text{loss}}_i^T \tilde{\mathbf{W}}_L (D_L - 1 + D''_{L-1}) (I + \tau \tilde{\mathbf{W}}_{L-1}) \cdots (D_l + D''_l)(\tau \mathbf{W}_l') (\tau h_{l-1}) \right| \\
\leq O \left( \frac{\sqrt{m \omega^{2/3} \log m}}{\sqrt{d}} \right) \cdot \|\tilde{\text{loss}}_i\| \cdot \|\mathbf{W}_l'\|_2
\end{align*}
\]

(37)

where (a) is due to the fact \( \|\tilde{\mathbf{W}}_L (D_{L-1} + D''_{L-1}) (I + \tau \tilde{\mathbf{W}}_{L-1}) \cdots (D_l + D''_l)\| = O(1) \) and \( \|h_{l-1}\| = O(1) \) holds with high probability.
We have similar bound for every summand in the second term of (36)

\[
\left| \text{loss}^T \mathbf{B} \mathbf{D}_l \mathbf{W}_l \cdots (I + \tau \mathbf{W}_{a+1}) D''_a (I + \tau \mathbf{W}_a) \cdots (D_l + D''_l) (\tau \mathbf{W}_l' h_{l-1}) \right|
\leq O \left( \sqrt{m \omega^{2/3} \log m \over \sqrt{d}} \right) \cdot ||\text{loss}|| \cdot \tau ||\mathbf{W}'_l||_2.
\]

(38)

For the last term in (36), we have

\[
\left| \text{loss}^T \mathbf{B} \mathbf{D}_l \mathbf{W}_l \cdots (I + \tau \mathbf{W}_{l+1}) \mathbf{D}_l \mathbf{W}'_l (h_{l-1} - \hat{h}_{l-1}) \right|
\leq ||\text{loss}|| \cdot O \left( \sqrt{m \over d} \right) \cdot ||\mathbf{W}'_l||_2 \cdot \tau ||h_{l-1} - \hat{h}_{l-1}||
\leq O(\omega \sqrt{m \over d}) \cdot ||\text{loss}|| \cdot \tau ||\mathbf{W}'_l||_2,
\]

(39)

where is the last inequality is due to the bound on \(||h_{i,l-1} - \hat{h}_{i,l-1}||_2\) in Lemma 9.

Hence

\[
(36) \leq O \left( L \sqrt{m \omega^{2/3} \log m \over \sqrt{d}} \right) \cdot ||\text{loss}|| \cdot \tau ||\mathbf{W}'_l||_2
\]

\[
\leq O \left( \sqrt{m \omega^{2/3} \log m \over \sqrt{d}} \right) \cdot ||\text{loss}|| \cdot ||\mathbf{W}'_l||_2,
\]

(40)

where the last inequality is due to the choice of \(\tau\).

Having all the above together and using triangle inequality, we have the result. \(\square\)

**Proposition 1** (Proposition 8.3 in in Allen-Zhu et al. [2018b]). Given vectors \(a, b \in \mathbb{R}^m\) and \(D \in \mathbb{R}^{m \times m}\) the diagonal matrix where \(D_{k,k} = 1_{a_k \geq 0}\). Then, there exists a diagonal matrix \(D'' \in \mathbb{R}^{m \times m}\) with

- \(|D_{k,k} + D''_{k,k}| \leq 1\) and \(|D''_{k,k}| \leq 1\) for every \(k \in [m]\),
- \(D''_{k,k} \neq 0\) only when \(1_{a_k \geq 0} \neq 1_{b_k \geq 0}\),
- \(\phi(a) - \phi(b) = (D + D'')(a - b)\).

**Proof of Lemma 9.** Fixing index \(i\) and ignoring the subscript in \(i\) for simplicity, by Proposition 1, for each \(l \in [L - 1]\) there exists a \(D''_l\) such that \(|(D''_l)_{k,k}| \leq 1\) and

\[
\begin{align*}
    h_l - \hat{h}_l &= \phi((I + \tau \mathbf{W}_l + \tau \mathbf{W}_l') h_{l-1}) - \phi((I + \tau \mathbf{W}_l) \hat{h}_{l-1}) \\
    &= (\mathbf{D}_l + D''_l) (I + \tau \mathbf{W}_l + \tau \mathbf{W}_l') h_{l-1} - (I + \tau \mathbf{W}_l) \hat{h}_{l-1} \\
    &= (\mathbf{D}_l + D''_l)(I + \tau \mathbf{W}_l)(h_{l-1} - \hat{h}_{l-1}) + (\mathbf{D}_l + D''_l) \tau \mathbf{W}_l' h_{l-1} \\
    &= \sum_{a=1}^l (\mathbf{D}_l + D''_a)(I + \tau \mathbf{W}_l) \cdots (I + \tau \mathbf{W}_{a+1}) \hat{D}_a + D''_a) \tau \mathbf{W}_a' h_{a-1}.
\end{align*}
\]

(41)
For \( l = L \), we similarly have
\[
\hat{h}_L - \hat{h}_L = (\hat{D}_L + D''_L)W'_Lh_{L-1} + \sum_{a=1}^{L-1}(\hat{D}_a + D''_a)(\hat{W}_a + \tau D''_a)\tau W'_ah_{a-1}.
\]

(42)

Then we have following properties.

For \( l \in [L-1], \|h_l - \hat{h}_l\| \leq O(1)\|W'\|_2 \), where \( \|W'\|_2 = \max\{\|W'_L\|_2, \|W'_2\|_2, \|W'_{L-1}\|_2\} \).
This is because \( (\hat{D}_l + D''_l)(I + \tau \hat{W}_l) \cdots (I + \tau \hat{W}_{a+1})(\hat{D}_a + D''_a)\| \leq 1.1 \) from Lemma 3; \( \|h_{a-1}\| \leq O(1) \) from Lemma 4; and \( \|W'_ah_{a-1}\| \leq \|W'_a\|_2\|h_{a-1}\| \) and \( \tau = \Omega(L \log m) \).

For \( l = L \), \( \|h_L - \hat{h}_L\| \leq O(1)(\|W'_L\|_2 + \|W'\|_2) = O(\|\hat{W}'\|_2) \).

For \( l \in [L], \|D''_l\|_0 \leq O(m\omega^{2/3}) \). This is because \( (D''_l)_{k,k} \) is non-zero only at coordinates \( k \) where \((\hat{g}_l)_k\) and \((g_l)_k\) have opposite signs, where it holds either \((D''_l)_{k,k} \neq (\hat{D}_l)_{k,k}\) or \((D''_l)_{k,k} \neq (\hat{D}_l)_{k,k}\). Therefore by Lemma 7, we have \( \|D''_l\|_0 \leq O(m\omega^{2/3}) \) if \( \|\hat{W}'\|_2 \leq \omega \).

\[\square\]

D Proof of Theorem 1

**Theorem 1.** For the ResNet defined and initialized as in Section 2, if the network width \( m \geq \max\{L, \Omega(n^2d^{-8}d \log^2 m)\} \), then with probability at least \( 1 - \exp(\Omega(\log^2 m)) \), gradient descent with learning rate \( \eta = \Theta(\frac{d\delta}{n^2m}) \) finds a point \( F(\hat{W}) \leq \epsilon \) in \( T = \Omega(n^2d^{-2}n\log^2 m) \) iterations.

**Proof.** Using Lemma 4 we have \( \|h_{i,L}^{(0)}\|_2 \leq 1.1 \) and then using the randomness of \( B \), it is easy to show that \( \|Bh_{i,L}^{(0)} - y_i^*\|_2 \leq O(\log^2 m) \) with probability at least \( 1 - \exp(\Omega(\log^2 m)) \), and therefore
\[
F(\hat{W}^{(0)}) \leq O(n \log^2 m).
\]

(43)

Assume that for every \( t = 0, 1, \ldots, T - 1 \), the following holds,
\[
\|W_L^{(t)} - W_L^{(0)}\|_F \leq \omega = O\left(\frac{\delta^3}{n^9}\right) \tag{44}
\]
\[
\|W_t^{(t)} - W_t^{(0)}\|_F \leq \tau \omega. \tag{45}
\]

We shall prove the convergence of GD under the assumption (44) holds, so that previous statements can be applied. At the end, we shall verify that (44) is indeed satisfied.
Letting $\nabla_t = \nabla F(\tilde{W}^{(t)})$, we calculate that

$$F(\tilde{W}^{(t+1)}) \leq F(\tilde{W}^{(t)}) - \eta \| \nabla_t \|^2_F + \eta \sqrt{nF(\tilde{W}^{(t)})} \cdot O \left( \frac{\sqrt{m\omega^{2/5}}}{\sqrt{d}} \right)^2 \cdot O \left( \sqrt{\sum_{i=1}^{L-1} \| W^{(t)}_i \|_2^2} + \eta^2 (\eta^2 n^2/d) \| \nabla_t \|^2_F + \eta \sqrt{nF(\tilde{W}^{(t)})} \cdot O \left( \frac{\sqrt{m\omega^{1/3}}}{d} + \eta^2 n^2 m^2 \right) \cdot F(\tilde{W}^{(t)}) \right) \leq \left( 1 - \Omega \left( \frac{\eta \delta m}{dn^2} \right) \right) F(\tilde{W}^{(t)}),$$

where the first inequality uses Theorem 4, the second inequality uses the gradient upper bound in Theorem 3 and the last inequality uses the gradient lower bound in Theorem 4 and the choice of $\eta$ and the assumption on $\omega$ (44). That is, after $T = \Omega(\frac{dn^2}{\eta \delta m}) \log \frac{n \log^2 m}{\epsilon}$ iterations $F(\tilde{W}^{(T)}) \leq \epsilon$.

We need to verify for each $t$, (44) holds. By Theorem 3,

$$\| W_L^{(t)} - W_L^{(0)} \|_F \leq \sum_{i=1}^{t-1} \| \eta \nabla W_L F(\tilde{W}^{(i)}) \|_F \leq O(\eta \sqrt{nm/d}) \cdot \sum_{i=1}^{t-1} \sqrt{F(\tilde{W}^{(i)})} \leq O \left( \frac{\eta \delta m}{dn^2} \cdot \sqrt{n \log m \cdot \frac{1}{1 - \Omega \left( \frac{\eta \delta m}{dn^2} \right)}} \right) \leq O \left( \frac{n^3 \sqrt{\delta}}{\delta \sqrt{m} \log m} \right),$$

where (a) is due to the relation (46) and (b) is due to the fact that $1 - \sqrt{1 - \Omega \left( \frac{\eta \delta m}{dn^2} \right)} \geq \frac{1}{2} \Omega \left( \frac{\eta \delta m}{dn^2} \right)$. Similarly, we have for $l \in [L-1]$,

$$\| W_L^{(t)} - W_L^{(0)} \|_F \leq \sum_{i=1}^{t-1} \| \eta \nabla W_l F(\tilde{W}^{(i)}) \|_F \leq O(\eta \sqrt{nm/d}) \cdot \sum_{i=1}^{t-1} \sqrt{F(\tilde{W}^{(i)})} \leq O \left( \frac{\tau n^3 \sqrt{\delta}}{\delta \sqrt{m} \log m} \right).$$

By combining (47) and the assumption on $\omega$ (44), we obtain a bound on $m$.

\[ \square \]

E Proof of Theorem 2

**Theorem 2.** For the ResNet defined and initialized as in Section 2, the network width $m \geq \max\{L, \Omega(n^{28b^{-4}} d^{-8} d \log^2 m)\}$. Suppose we do stochastic gradient descent update
starting from $\mathbf{W}^{(0)}$ and

\[
\mathbf{W}^{(t+1)} = \mathbf{W}^{(t)} - \eta \frac{n}{|S_t|} \sum_{i \in S_t} \nabla F_i(\mathbf{W}^{(t)}),
\]

(1)

where $S_t$ is a random subset of $[n]$ with $|S_t| = b$. Then with probability at least $1 - \exp(-\Omega(\log^2 m))$, stochastic gradient descent (1) with learning rate $\eta = \Theta(\frac{db}{n^5m})$ finds a point $F(\mathbf{W}) \leq \epsilon$ in $T = \Omega(n^7b^{-1}d^{-2} \log \frac{n \log^2 m}{\epsilon})$ iterations.

Proof. Using Lemma 4 we have $\|h_{i,L}\|_2 \leq 1.1$ and then using the randomness of $B$, it is easy to show that $\|B h_{i,L}^{(0)} - y_i\|_2 \leq O(\log^2 m)$ with probability at least $1 - \exp(-\Omega(\log^2 m))$, and therefore

\[
F(\mathbf{W}^{(0)}) \leq O(n \log^2 m).
\]

(48)

Assume that for every $t = 0, 1, \ldots, T - 1$, the following holds,

\[
\|\mathbf{W}_L^{(t)} - \mathbf{W}_L^{(0)}\|_F \leq \omega \overset{\Delta}{=} O\left(\frac{\delta^3 b^{1.5}}{n^{10.5}}\right)
\]

(49)

\[
\|\mathbf{W}_i^{(t)} - \mathbf{W}_i^{(0)}\|_F \leq \tau \omega.
\]

(50)

We shall prove the convergence of SGD under the assumption (44) holds, so that previous statements can be applied. At the end, we shall verify that (44) is indeed satisfied.

Letting $\nabla_t = \frac{n}{|S_t|} \sum_{i \in S_t} \nabla F_i(\mathbf{W}^{(t)})$, we calculate that

\[
\mathbb{E}F(\mathbf{W}^{(t+1)}) \leq F(\mathbf{W}^{(t)}) - \eta \|\nabla F(\mathbf{W}^{(t)})\|_F^2 + \eta \sqrt{nF(\mathbf{W}^{(t)})} \cdot O\left(\frac{\sqrt{m \omega^{2/3}}}{\sqrt{d}}\right)
\]

\[
\cdot O\left(\mathbb{E}\|\nabla \mathbf{W}_L^{(t)}\|_2 + \mathbb{E} \sum_{i=1}^{L-1} \|\nabla \mathbf{W}_i^{(t)}\|_2\right) + O(\eta^2 nm/d)\mathbb{E}\|\nabla_t\|_2^2
\]

\[
\leq F(\mathbf{W}^{(t)}) - \eta \|\nabla F(\mathbf{W}^{(t)})\|_F^2 + O\left(\frac{\eta m^{3/2} m^{1/3}}{d \sqrt{b}} + \frac{\eta n^2 m^2}{bd^2}\right) \cdot F(\mathbf{W}^{(t)})
\]

\[
\leq \left(1 - \Omega\left(\frac{\eta \delta m}{dn^2}\right)\right) F(\mathbf{W}^{(t)}),
\]

(51)

where the first inequality uses Theorem 4, the second inequality uses the gradient upper bound in Theorem 3 and the following fact

\[
\mathbb{E}\left[\|\nabla_t\|_2^2\right] \leq O\left(\frac{n^2 mf(\mathbf{W}^{(t)})}{db}\right),
\]

(52)

and the last inequality uses the gradient lower bound in Theorem 4 and the choice of $\eta$ and the assumption on $\omega$ (44). That is, after $T = \Omega(\frac{dn^2}{\eta \delta m}) \log \frac{n \log^2 m}{\epsilon}$ iterations $F(\mathbf{W}^{(T)}) \leq \epsilon$. 30
We need to verify for each $t$, (44) holds. By Theorem 3,
\[
\|W_L^{(t)} - W_L^{(0)}\|_F \leq \sum_{i=1}^{t-1} \|\eta \nabla W_L^{(i)}\|_F \leq O(\eta \sqrt{n^2 m/(bd)}) \cdot \sum_{i=1}^{t-1} \sqrt{F(W_L^{(i)})}
\]
\[
\overset{(a)}{\leq} O(\eta \sqrt{n^2 m/(bd)}) \cdot O\left(\sqrt{n \log m} \cdot \frac{1}{1 - \sqrt{1 - \Omega(\frac{\eta m}{dn^2})}}\right)
\]
\[
\overset{(b)}{\leq} O\left(\frac{n^{3.5} \sqrt{d}}{\delta \sqrt{mb} \log m}\right), \quad (53)
\]
where (a) is due to the relation (51) and (b) is due to the fact that $1 - \sqrt{1 - \Omega(\frac{\eta m}{dn^2})} \geq \frac{1}{2} \Omega(\frac{\eta m}{dn^2})$. Similarly, we have for $l \in [L - 1]$,
\[
\|W_L^{(l)} - W_L^{(0)}\|_F \leq \sum_{i=1}^{t-1} \|\eta \nabla W_L^{(i)}\|_F \leq O(\eta \tau \sqrt{n^2 m/(bd)}) \cdot \sum_{i=1}^{t-1} \sqrt{F(W_L^{(i)})} \leq O\left(\frac{\tau n^{3.5} \sqrt{d}}{\delta \sqrt{mb} \log m}\right).
\]

By combining (53) and the assumption on $\omega$ (49), we obtain a bound on $m$.

\[\square\]

\section*{F More Empirical Studies}

In this section we train the feedforward neural network and ResNet models on the Street View House Numbers (SVHN) dataset [Netzer et al., 2011], and compare their convergence behaviors. The model architectures is the same as Section 5. We run our experiments on both fully-connected and convolutional models. The fully-connected model zoos are generated by varying the depth $L \in \{30, 100, 500\}$ and the width $m \in \{128, 1024\}$. The convolution model zoos are generated by varying the depth $L \in \{30, 50, 100\}$ and the number of channels $m \in \{16, 32\}$. The width of convolution model is the number of convolution kernels of each hidden layer. We choose $\tau \in \{\frac{1}{L}, \frac{1}{L^{0.5}}\}$ and test its influence.\footnote{We do not include the case of $\tau = \frac{1}{L^{0.25}}$ into comparison because the training fails to converge.}

\textbf{Data and hyperparameters.} SVHN contains more than 70000 training examples with input dimension $32 \times 32 \times 3$ and 10 labels. The input feature vector are normalized. We use the standard SGD optimizer. Learning rate is chosen as a function of the model width $lr = \frac{0.1}{(m/16)}$ and minibatch size is 64. There is no pooling layer of our convolution model.

\textbf{Experiments results.} For a given width, we evaluate the training performances of ResNet and feedforward NN with different depths. Figure 4 and 5 show the results of fully connected models and convolutional models, respectively. $\tau = \frac{1}{L}$ and $\tau = \frac{1}{L^{0.5}}$ respectively. The results show that deep ResNet with a small $\tau$ is much easier to train than feedforward
NN. However, small $\tau$ hurts the expressivity of the network, i.e., when the width is large enough ($m = 1024$ for fully connected models or $m = 32$ for convolutional models) to train a feedforward NN, ResNet with small $\tau$ performs worse than feedforward NN.

Figure 4: Training performances of fully-connected models on SVHN dataset. The columns from left to right are corresponding to feedforward NN, ResNet with $\tau = \frac{1}{L}$ and $\frac{1}{L^{0.5}}$, respectively. The rows from top to bottom are corresponding to the width at each layer $m = 16, 128$ and $1024$, respectively. Training loss is evaluated on 10% training random samples.
Figure 5: Training performances of convolutional models on SVHN dataset. The columns from left to right are corresponding to feedforward NN, ResNet with $\tau = \frac{1}{L}$ and $\frac{1}{L^0\tau}$, respectively. The rows from top to bottom are corresponding to the width at each layer $m = 8, 16$ and $32$, respectively. Training loss is evaluated on 10% random training samples.