Machine Learning
https://doi.org/10.1007/s10994-019-05839-6

Gradient descent optimizes over-parameterized deep ReLU networks

Difan Zou[1] · Yuan Cao[1] · Dongruo Zhou[1] · Quanquan Gu[1]

Received: 4 May 2019 / Revised: 6 July 2019 / Accepted: 6 September 2019

Abstract
We study the problem of training deep fully connected neural networks with Rectified Linear Unit (ReLU) activation function and cross entropy loss function for binary classification using gradient descent. We show that with proper random weight initialization, gradient descent can find the global minima of the training loss for an over-parameterized deep ReLU network, under certain assumption on the training data. The key idea of our proof is that Gaussian random initialization followed by gradient descent produces a sequence of iterates that stay inside a small perturbation region centered at the initial weights, in which the training loss function of the deep ReLU networks enjoys nice local curvature properties that ensure the global convergence of gradient descent. At the core of our proof technique is (1) a milder assumption on the training data; (2) a sharp analysis of the trajectory length for gradient descent; and (3) a finer characterization of the size of the perturbation region. Compared with the concurrent work (Allen-Zhu et al. in A convergence theory for deep learning via over-parameterization, 2018a; Du et al. in Gradient descent finds global minima of deep neural networks, 2018a) along this line, our result relies on milder over-parameterization condition on the neural network width, and enjoys faster global convergence rate of gradient descent for training deep neural networks.

Keywords Deep neural networks · Gradient descent · Over-parameterization · Random initialization · Global convergence

D. Zou, Y. Cao: Equal contribution.

Editors: Kee-Eung Kim and Jun Zhu.

✉ Quanquan Gu
qgu@cs.ucla.edu

Difan Zou
knowzou@cs.ucla.edu

Yuan Cao
yuancao@cs.ucla.edu

Dongruo Zhou
drzhou@cs.ucla.edu

[1] Department of Computer Science, University of California, Los Angeles, CA 90095, USA

Published online: 23 October 2019

1 Introduction

Deep neural networks have achieved great success in many applications like image processing (Krizhevsky et al. 2012), speech recognition (Hinton et al. 2012) and Go games (Silver et al. 2016). However, the reason why deep networks work well in these fields remains a mystery for long time. Different lines of research try to understand the mechanism of deep neural networks from different aspects. For example, a series of work tries to understand how the expressive power of deep neural networks are related to their architecture, including the width of each layer and depth of the network (Telgarsky 2015, 2016; Lu et al. 2017; Liang and Srikant 2016; Yarotsky 2017, 2018; Hanin 2017; Hanin and Sellke 2017). These work shows that multi-layer networks with wide layers can approximate arbitrary continuous function.

Very recently, there emerges a large body of work that study the global convergence of gradient descent (GD) for training neural networks (Li and Liang 2018; Du et al. 2018b; Allen-Zhu et al. 2018a; Du et al. 2018a). In particular, Li and Liang (2018) showed that for a one-hidden-layer network with ReLU activation function using over-parameterization and random initialization, GD and stochastic gradient descent (SGD) can find the global near-optimal solution in polynomial time. Du et al. (2018b) showed that under the assumption that the ReLU Gram matrix is positive definite, randomly initialized GD converges to a globally optimal solution of a one-hidden-layer network with ReLU activation function and quadratic loss function. Beyond shallow neural network, Du et al. (2018a) considered regression problem with square loss function, and proved that under certain assumptions on the initialization and training data, gradient descent is able to converge to the global optimal solution for training deep neural networks. However, Du et al. (2018a) only investigated DNNs with smooth activation functions, which exclude the widely-used ReLU activation function. Moreover, the theoretical results in Du et al. (2018a) heavily rely on the assumption that the smallest eigenvalue of certain deep compositional Gram matrix is bounded below from zero, which does not explicitly tell the dependency on the problem parameter such as the number of training examples n and the number of hidden layers L, and this assumption cannot be verified in practice. Allen-Zhu et al. (2018a) studied the same problem under a different assumption on the training data, and proved that random initialization followed by gradient descent is able to converge to the global optimal solution for training deep neural networks. Besides, Allen-Zhu et al. (2018a) studied the convergence rate of SGD for training deep ReLU network and discussed various extensions to classification problem and various loss functions. However, the assumption on the training data made in Allen-Zhu et al. (2018a) is very stringent, because they require that any two training data points are separated by some constant, but in practice the data from the same class can be arbitrarily close (e.g., due to data augmentation in deep learning). Our work is independent and concurrent to Du et al. (2018a), Allen-Zhu et al. (2018a).[1]

In this paper, we study the optimization properties of gradient-based methods for deep ReLU neural networks, with more realistic assumption on the training data, milder over-parameterization condition and faster convergence rate. In specific, we consider an L-hidden-layer fully-connected neural network with ReLU activation function. Similar to the one-hidden-layer case studied in Li and Liang (2018) and Du et al. (2018b), we study binary classification problem and show that GD can achieve the global minima of the training loss for any $L \geq 1$, with the aid of over-parameterization and random initialization. The high-level idea of our proof technique is to show that Gaussian random initialization followed by gradient descent generates a sequence of iterates within a small perturbation region centering

[1] The first versions of all these three papers were all posted on arXiv in November 2018.

around the initial weights. In addition, we will show that the empirical loss function of deep ReLU networks has very good local curvature properties inside the perturbation region, which guarantees the global convergence of gradient descent. Compared with the proof technique in Allen-Zhu et al. (2018a), we provide a sharper analysis on the GD algorithm and prove that GD can be guaranteed to have sufficient descent in a larger perturbation region with a larger step size. This leads to a faster convergence rate and a milder condition on the over-paramterization. More specifically, our main contributions are summarized as follows:

- We establish the global convergence guarantee for training deep ReLU networks in terms of classification problems. Compared with Li and Liang (2018), Allen-Zhu et al. (2018a) our assumption on training data is more reasonable and is often satisfied by real training data. Specifically, we only require that any two data points from different classes are separated by some constant, while Li and Liang (2018) assumes that the data from different classes are sampled from small balls separated by a constant margin, and Allen-Zhu et al. (2018a) requires that any two data points are well separated, even though they belong to the same class.
- We show that with Gaussian random initialization on each layer, when the number of hidden nodes per layer is at least $\widetilde{\Omega}\big(n^{14}L^{16}/\phi^4\big)$, GD can achieve zero training error within $\widetilde{O}\big(n^5L^3/\phi\big)$ iterations, where ϕ is the data separation distance,[2] n is the number of training examples, and L is the number of hidden layers. This significantly improves the state-of-the-art results by Allen-Zhu et al. (2018a), where the authors proved that GD can converge within $\widetilde{O}\big(n^6L^2/\phi^2\big)$ iterations if the number of hidden nodes per layer is at least $\widetilde{\Omega}(n^{24}L^{12}/\phi^8)$. Compared with Du et al. (2018a), our result only has a polynomial dependency on the number of hidden layers, which is much better than their result that has an exponential dependency on the depth for fully connected deep neural networks.

2 Additional related work

Due to the huge amount of literature on deep learning theory, we are not able to include all papers in this big vein here. Instead, we review the following two additional lines of research, which are also related to our work.

One-hidden-layer neural networks with ground truth parameters Recently a series of work (Tian 2017; Brutzkus and Globerson 2017; Li and Yuan 2017; Du et al. 2017; Zhang et al. 2018) studied a specific class of shallow two-layer (one-hidden-layer) neural networks, whose training data are generated by a ground truth network called "teacher network". This series of work aim to provide recovery guarantee for gradient-based methods to learn the teacher networks based on either the population or empirical loss functions. More specifically, Tian (2017) proved that for two-layer ReLU networks with only one hidden neuron, GD with arbitrary initialization on the population loss is able to recover the hidden teacher network. Brutzkus and Globerson (2017) proved that GD can learn the true parameters of a two-layer network with a convolution filter. Li and Yuan (2017) proved that SGD can recover the underlying parameters of a two-layer residual network in polynomial time. Moreover, Du et al. (2017) proved that both GD and SGD can recover the teacher network of a two-layer CNN with ReLU activation function. Zhang et al. (2018) showed that GD on the empirical loss function can recover the ground truth parameters of one-hidden-layer ReLU networks at a linear rate.

[2] We will define the data separation distance, training sample size n and number of hidden layers L formally in Sects. 3 and 4.

Deep linear networks Beyond shallow one-hidden-layer neural networks, a series of recent work (Hardt and Ma 2016; Kawaguchi 2016; Bartlett et al. 2018; Gunasekar et al. 2018; Arora et al. 2018a, b) focused on the optimization landscape of deep linear networks. More specifically, Hardt and Ma (2016) showed that deep linear residual networks have no spurious local minima. Kawaguchi (2016) proved that all local minima are global minima in deep linear networks. Arora et al. (2018b) showed that depth can accelerate the optimization of deep linear networks. Bartlett et al. (2018) proved that with identity initialization and proper regularizer, GD can converge to the least square solution on a residual linear network with quadratic loss function, while Arora et al. (2018a) proved the same properties for general deep linear networks.

3 Preliminaries

3.1 Notation

We use lower case, lower case bold face, and upper case bold face letters to denote scalars, vectors and matrices respectively. For a positive integer n, we denote $[n] = \{1, \dots, n\}$. For a vector $\mathbf{x} = (x_1, \dots, x_d)^\top$, we denote by $\|\mathbf{x}\|_p = \big(\sum_{i=1}^d |x_i|^p\big)^{1/p}$ the ℓ_p norm of $\mathbf{x}$, $\|\mathbf{x}\|_\infty = \max_{i=1,\dots,d} |x_i|$ the ℓ_∞ norm of $\mathbf{x}$, and $\|\mathbf{x}\|_0 = |\{x_i : x_i \neq 0, i = 1, \dots, d\}|$ the number of non-zero entries of $\mathbf{x}$. We use $\mathrm{Diag}(\mathbf{x})$ to denote a square diagonal matrix with the elements of vector $\mathbf{x}$ on the main diagonal. For a matrix $\mathbf{A} = (A_{ij}) \in \mathbb{R}^{m\times n}$, we use $\|\mathbf{A}\|_F$ to denote the Frobenius norm of $\mathbf{A}$, $\|\mathbf{A}\|_2$ to denote the spectral norm (maximum singular value), and $\|\mathbf{A}\|_0$ to denote the number of nonzero entries. We denote by $S^{d-1} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 = 1\}$ the unit sphere in $\mathbb{R}^d$.

For two sequences $\{a_n\}$ and $\{b_n\}$, we use $a_n = O(b_n)$ to denote that $a_n \le C_1 b_n$ for some absolute constant $C_1 > 0$, and use $a_n = \Omega(b_n)$ to denote that $a_n \ge C_2 b_n$ for some absolute constant $C_2 > 0$. In addition, we also use $\widetilde{O}(\cdot)$ and $\widetilde{\Omega}(\cdot)$ to hide logarithmic terms in Big-O and Big-Omega notations. We also use the following matrix product notation. For indices l_1, l_2 and a collection of matrices $\{\mathbf{A}_r\}_{r\in\mathbb{Z}_+}$, we denote

$$\prod_{r=l_1}^{l_2} \mathbf{A}_r := \begin{cases} \mathbf{A}_{l_2}\mathbf{A}_{l_2-1}\cdots\mathbf{A}_{l_1} & \text{if } l_1 \le l_2 \\ \mathbf{I} & \text{otherwise.} \end{cases} \tag{3.1}$$

3.2 Problem setup

Let $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\} \in (\mathbb{R}^d \times \{-1, 1\})^n$ be a set of n training examples. Let $m_0 = d$. We consider L-hidden-layer neural networks as follows:

$$f_{\mathbf{W}}(\mathbf{x}) = \mathbf{v}^\top \sigma(\mathbf{W}_L^\top \sigma(\mathbf{W}_{L-1}^\top \cdots \sigma(\mathbf{W}_1^\top \mathbf{x}) \cdots)),$$

where $\sigma(x) = \max\{0, x\}$ is the entry-wise ReLU activation function, $\mathbf{W}_l = (\mathbf{w}_{l,1}, \dots, \mathbf{w}_{l,m_l}) \in \mathbb{R}^{m_{l-1}\times m_l}$, $l = 1, \dots, L$ are the weight matrices, and $\mathbf{v} \in \{-1, +1\}^{m_L}$ is the fixed output layer weight vector with half 1 and half -1 entries. Let $\mathbf{W} = \{\mathbf{W}_l\}_{l=1,\dots,L}$ be the collection of matrices $\mathbf{W}_1, \dots, \mathbf{W}_L$, we consider solving the following empirical risk minimization problem:

$$L_S(\mathbf{W}) = \frac{1}{n}\sum_{i=1}^n \ell(y_i \widehat{y}_i) = \frac{1}{n}\sum_{i=1}^n \ell\big(y_i \mathbf{v}^\top \sigma(\mathbf{W}_L^\top \sigma(\mathbf{W}_{L-1}^\top \cdots \sigma(\mathbf{W}_1^\top \mathbf{x}_i) \cdots))\big) \tag{3.2}$$

where $\widehat{y}_i = f_{\mathbf{W}}(\mathbf{x}_i)$ denotes the output of neural network and $\ell(x) = \log(1 + \exp(-x))$ is the cross-entropy loss for binary classification.

3.3 Optimization algorithms

In this paper, we consider training a deep neural network with Gaussian initialization followed by gradient descent.

Gaussian initialization We say that the weight matrices $\mathbf{W}_1, \ldots, \mathbf{W}_L$ are generated from Gaussian initialization if each column of $\mathbf{W}_l$ is generated independently from the Gaussian distribution $N(\mathbf{0}, 2/m_l\mathbf{I})$ for all $l = 1, \ldots, L$. This initialization mechanism is called He-initialization, which was proposed in He et al. (2015).

Gradient descent We consider solving the empirical risk minimization problem (3.2) with gradient descent with Gaussian initialization: let $\mathbf{W}_1^{(0)}, \ldots, \mathbf{W}_L^{(0)}$ be weight matrices generated from Gaussian initialization, we consider the following gradient descent update rule:

$$\mathbf{W}_l^{(k)} = \mathbf{W}_l^{(k-1)} - \eta \nabla_{\mathbf{W}_l} L_S(\mathbf{W}^{(k-1)}),\ l = 1, \ldots, L,$$

where $\nabla_{\mathbf{W}_l} L_S(\cdot)$ is the partial gradient of $L_S(\cdot)$ with respect to the l-th layer parameters $\mathbf{W}_l$, and $\eta > 0$ is the step size (a.k.a., learning rate).

3.4 Calculations for neural network functions

Here we briefly introduce some useful notations and provide some basic calculations regarding the neural network in our setting.

- *Output after the l-th layer:* Given an input $\mathbf{x}_i$, the output of the neural network after the l-th layer is

$$\begin{aligned}\mathbf{x}_{l,i} &= \sigma(\mathbf{W}_l^\top \sigma(\mathbf{W}_{l-1}^\top \cdots \sigma(\mathbf{W}_1^\top \mathbf{x}_i)\cdots))\\ &= \left(\prod_{r=1}^{l} \mathbf{\Sigma}_{r,i}\mathbf{W}_r^\top\right)\mathbf{x}_i,\end{aligned}$$

  where $\mathbf{\Sigma}_{1,i} = \mathrm{Diag}\big(\mathbb{1}\{\mathbf{W}_1^\top \mathbf{x}_i > 0\}\big)$,[3] and $\mathbf{\Sigma}_{l,i} = \mathrm{Diag}[\mathbb{1}\{\mathbf{W}_l^\top(\prod_{r=1}^{l-1}\mathbf{\Sigma}_{r,i}\mathbf{W}_r^\top)\mathbf{x}_i > 0\}]$ for $l = 2, \ldots, L$.
- *Output of the neural network:* The output of the neural network with input $\mathbf{x}_i$ is as follows:

$$\begin{aligned}f_{\mathbf{W}}(\mathbf{x}_i) &= \mathbf{v}^\top \sigma(\mathbf{W}_L^\top \sigma(\mathbf{W}_{L-1}^\top \cdots \sigma(\mathbf{W}_1^\top \mathbf{x}_i)\cdots))\\ &= \mathbf{v}^\top\left(\prod_{r=l}^{L} \mathbf{\Sigma}_{r,i}\mathbf{W}_r^\top\right)\mathbf{x}_{l-1,i},\end{aligned}$$

  where we define $\mathbf{x}_{0,i} = \mathbf{x}_i$ and the last equality holds for any $l \geq 1$.
- *Gradient of the neural network:* The partial gradient of the training loss $L_S(\mathbf{W})$ with respect to $\mathbf{W}_l$ is as follows:

$$\nabla_{\mathbf{W}_l} L_S(\mathbf{W}) = \frac{1}{n}\sum_{i=1}^{n} \ell'(y_i\widehat{y}_i)\cdot y_i \cdot \nabla_{\mathbf{W}_l}[f_{\mathbf{W}}(\mathbf{x}_i)],$$

[3] Here we slightly abuse the notation and denote $\mathbb{1}\{\mathbf{a} > 0\} = (\mathbb{1}\{\mathbf{a}_1 > 0\}, \ldots, \mathbb{1}\{\mathbf{a}_m > 0\})^\top$ for a vector $\mathbf{a} \in \mathbb{R}^m$.

where the gradient of the neural network function is defined as

$$\nabla_{\mathbf{W}_l}[f_{\mathbf{W}}(\mathbf{x}_i)] = \mathbf{x}_{l-1,i}\mathbf{v}^\top \left(\prod_{r=l+1}^{L} \mathbf{\Sigma}_{r,i}\mathbf{W}_r^\top \right) \mathbf{\Sigma}_{l,i}.$$

In the remaining of this paper, we define the gradient $\nabla L_S(\mathbf{W})$ as the collection of partial gradients with respect to all $\mathbf{W}_l$'s, i.e.,

$$\nabla L_S(\mathbf{W}) = \{\nabla_{\mathbf{W}_1} L_S(\mathbf{W}), \nabla_{\mathbf{W}_2} L_S(\mathbf{W}), \ldots, \nabla_{\mathbf{W}_L} L_S(\mathbf{W})\}.$$

We also define the Frobenius norm of $\nabla L_S(\mathbf{W})$ as

$$\|\nabla_{\mathbf{W}_l} L_S(\mathbf{W})\|_F = \left[\sum_{l=1}^{L} \|\nabla_{\mathbf{W}_l} L_S(\mathbf{W})\|_F^2 \right]^{1/2}.$$

4 Main theory

In this section, we show that with random Gaussian initialization, over-parameterization helps gradient descent converge to the global minimum, i.e., find a point in the parameter space with arbitrary small training loss. We start with assumptions on the training data,

Assumption 4.1 $\|\mathbf{x}_i\|_2 = 1$ and $(\mathbf{x}_i)_d = \mu$ for all $i \in \{1, \ldots, n\}$, where $\mu \in (0, 1)$ is a constant.

As is shown in the assumption above, the last entry of input $\mathbf{x}$ is considered to be a constant μ. This assumption is natural because it can be seen as adding a bias term in the input layer, and learning both weight vector and bias is equivalent to adding an additional dummy variable $((\mathbf{x}_i)_d = \mu)$ to all input vectors and learning the weight vector only. The same assumption has been made in Allen-Zhu et al. (2018a). In addition, we emphasize that Assumption 4.1 is made in order to simplify the proof. Actually, rather than restricting the norm of all training examples to be 1, this assumption can be relaxed to be that $\|\mathbf{x}_i\|_2$ is lower and upper bounded by some constants.

Assumption 4.2 For all $i, i' \in \{1, \ldots, n\}$, if $y_i \neq y_{i'}$, then $\|\mathbf{x}_i - \mathbf{x}_{i'}\|_2 \geq \phi$ for some $\phi > 0$.

Assumption 4.2 basically requires that inputs with different labels in the training data are separated from each other by at least a constant. This assumption is often satisfied in practice. In contrast, Allen-Zhu et al. (2018a) assumes that every two different data points in the training data are separated by a constant, which is much stronger and cannot be satisfied since in classification it is allowed that the data with the same label can be arbitrarily close.

Furthermore, Assumption 4.2 can be easily verified based on the training data. As a comparison, the assumption made in Du et al. (2018a) assumes that certain deep compositional Gram matrix defined on the training data is strictly positive definite, which is not easy to verify, since the definition of their special Gram matrix is based on integration.

Then we have the following assumption on the structure of neural network.

Assumption 4.3 Define $M = \max\{m_1, \ldots, m_L\}$, $m = \min\{m_1, \ldots, m_L\}$. We assume that $M \leq 2m$.

Assumption 4.3 states that the number of nodes at all layers are of the same order. The constant 2 is not essential and can be replaced with an arbitrary constant greater than or equal to 1.

Under Assumptions 4.1–4.3, we are able to establish the global convergence of gradient descent for training deep ReLU networks. Specifically, we provide the following theorem which characterizes the required numbers of hidden nodes and iterations such that the gradient descent can attain the global minimum of the training loss function.

Theorem 4.4 *Suppose* $\mathbf{W}_1^{(0)}, \ldots, \mathbf{W}_L^{(0)}$ *are generated by Gaussian initialization. Then under Assumptions 4.1–4.3, if the step size* $\eta = O(M^{-1}L^{-3})$*, the number of hidden nodes per layer satisfies*

$$m = \widetilde{\Omega}\big(n^{14}L^{16}\phi^{-4} + n^{12}L^{16}\phi^{-4}\epsilon^{-1}\big)$$

and the maximum number of iteration satisfies

$$K = \widetilde{O}\big(n^5L^3/\phi + n^3L^3\epsilon^{-1}/\phi\big),$$

then with high probability, the last iterate of gradient descent $\mathbf{W}^{(K)}$ *satisfies* $L_S(\mathbf{W}^{(K)}) \le \epsilon$.

Remark 4.1 Note that our bound on the required number of hidden nodes per layer, i.e., m, depends on the target accuracy ϵ. However, in practical classification tasks, we are more interested in finding some points with zero training error. In specific, the cross-entropy loss $\ell(x) = \log(1+\exp(-x))$ is strictly decreasing in x, thus $\ell(y_i\widehat{y}_i) \le \ell(0) = \log(2)$ implies $y_i\widehat{y}_i \ge 0$. If we set $L_S(\mathbf{W}) \le \ell(0)/n = \log(2)/n$, it holds that $\ell(y_i\widehat{y}_i) \le nL_S(\mathbf{W}) \le \ell(0)$ for all $i \in [n]$, which further implies that $y_i\widehat{y}_i \ge 0$ for all $i \in [n]$, i.e., all training data are correctly classified. Therefore, Theorem 4.4 implies that gradient descent can find a point with zero training error if the number of hidden nodes per layer is at least $m = \widetilde{\Omega}(n^{14}L^{16}\phi^{-4})$.

Remark 4.2 Here we compare our theoretical results with those in Allen-Zhu et al. (2018a) and Du et al. (2018a). Specifically, Allen-Zhu et al. (2018a) proved that gradient descent can achieve zero training error within $O(n^6L^2/\phi^2)$ iterations under the condition that the neural network width is at least $m = \widetilde{\Omega}(n^{24}L^{12}/\phi^8)$. As a clear comparison, our result on m is significantly better by a factor of $\widetilde{\Omega}(n^{10}L^{-4}/\phi^4)$, and our convergence rate is faster by a factor of $O(nL^{-1})$.[4] On the other hand, Du et al. (2018a) proved similar global convergence result when the neural network width is at least $\widetilde{\Omega}\big(2^{O(L)} \cdot n^4/\lambda_0^4\big)$, where λ_0 is the smallest eigenvalue of the deep compositional Gram matrix defined in their paper. Compared with their result, our condition on m has significantly better dependency in L. In addition, for real training data, λ_0 can have high degree dependency on the reciprocal of the sample size n, which makes the dependency of their result on n much worse.

5 Proof of the main theory

In this section, we provide the proof of the main theory. In specific, we decompose the proof into three steps:

Step 1: We characterize a perturbation region at the initialization, and prove that the neural network attains good properties within such region.

[4] It is worth noting that in practice we usually have $n \gg L$, thus our improvements in terms of the over-parameterization condition and convergence rate are indeed significant.

Step 2: Based on the assumption that all iterates are staying inside the region $\mathcal{B}(\mathbf{W}^{(0)}, \tau)$, we establish the convergence results of gradient descent.
Step 3: We verify that with our choice of m, until convergence all iterates of gradient descent would not escape from the perturbation region $\mathcal{B}(\mathbf{W}^{(0)}, \tau)$, which justifies the derived convergence guarantee.

Now we characterize the perturbation as follows. Given the initialization generated by Gaussian distribution $\mathbf{W}^{(0)} := \{\mathbf{W}_l^{(0)}\}_{l=1,\ldots,L}$, we define by $\mathcal{B}(\mathbf{W}^{(0)}, \tau) = \{\mathbf{W} : \|\mathbf{W}_l - \mathbf{W}_l^{(0)}\|_2 \le \tau$ for all $l \in [L]\}$ the perturbation region centered at $\mathbf{W}^{(0)}$. Then we provide the following Lemmas that provides key results which are essential to establish the convergence guarantees for (stochastic) gradient descent.

Lemma 5.1 (Bounded initial training loss) *Under Assumptions 4.1 and 4.3, with probability at least $1 - \delta$, at the initialization the training loss satisfies $L_S(\mathbf{W}^{(0)}) \le C\sqrt{\log(n/\delta)}$.*

Next we are going to state the following key lemmas that characterize some essential properties of the neural network when its weight parameters satisfy $\mathbf{W} \in \mathcal{B}(\mathbf{W}^{(0)}, \tau)$. Firstly, the following lemma provides the lower and upper bounds of the Frobenious norm of the partial gradient $\nabla_{\mathbf{W}_l}[L_S(\mathbf{W})]$.

Lemma 5.2 (Gradient lower and upper bound) *Under Assumptions 4.1, 4.2, and 4.3, if $\tau = O\big(\phi^{3/2}n^{-3}L^{-2}\big)$ and $m = \widetilde{\Omega}(n^2\phi^{-1})$, then for all $\widetilde{\mathbf{W}} \in \mathcal{B}(\mathbf{W}^{(0)}, \tau)$, with probability at least $1 - \exp\big(-O(m\phi/n)\big)$, there exist positive constants C and C' such that*

$$\|\nabla_{\mathbf{W}_L}[L_S(\widetilde{\mathbf{W}})]\|_F^2 \ge C\frac{m\phi}{n^5}\left(\sum_{i=1}^n \ell'(y_i\widetilde{y}_i)\right)^2,$$

$$\|\nabla_{\mathbf{W}_l}[L_S(\widetilde{\mathbf{W}})]\|_F \le -\frac{C'LM^{1/2}}{n}\sum_{i=1}^n \ell'(y_i\widetilde{y}_i),$$

for all $l \in [L]$, where $\widetilde{y}_i = f_{\widetilde{\mathbf{W}}}(\mathbf{x}_i)$.

Then we provide the following lemma that characterizes the training loss decreasing after one-step gradient descent.

Lemma 5.3 (Sufficient descent) *Let $\mathbf{W}_1^{(0)}, \ldots, \mathbf{W}_L^{(0)}$ be generated via Gaussian random initialization. Let $\mathbf{W}^{(k)} = \{\mathbf{W}_l^{(k)}\}_{l=1,\ldots,L}$ be the k-th iterate in the gradient descent and $\tau = O(L^{-11}\log^{-3/2}(M))$. If $\mathbf{W}^{(k)}, \mathbf{W}^{(k+1)} \in \mathcal{B}(\mathbf{W}^{(0)}, \tau)$, then there exist constants C' and C'' such that with probability at least $1 - \exp\big(-O(m\phi/n)\big)$ the following holds,*

$$\begin{aligned} L_S(\mathbf{W}^{(k+1)}) - L_S(\mathbf{W}^{(k)}) \le &-\big(\eta - C'ML^3\eta^2\big)\|\nabla L_S(\mathbf{W}^{(k)})\|_F^2 \\ &- \frac{C''L^{8/3}\tau^{1/3}\sqrt{M\log(M)}\cdot\eta\|\nabla L_S(\mathbf{W}^{(k)})\|_F}{n}\sum_{i=1}^n \ell'(y_i\widehat{y}_i^{(k)}) \end{aligned}$$

The second term on the R.H.S. of the result in Lemma 5.3 is due to the non-smoothness of ReLU activation, which can be characterized by counting how many nodes would change their activation patterns during the training process. Clearly, in order to guarantee that the gradient descent can bring sufficient descent in each step, we require the radius τ to be sufficiently small. In the following, we are going to complete the proof of Theorem 4.4 based on Lemmas 5.1–5.3.

Proof of Theorem 4.4 We first prove that GD is able to achieve ϵ training loss under the condition that all iterates are staying inside the perturbation region $\mathcal{B}(\mathbf{W}^{(0)}, \tau)$. Note that by Lemma 5.2, we know that there exists a constant c_0 such that

$$\|\nabla L_S(\mathbf{W}^{(k)})\|_F^2 \geq \|\nabla_{\mathbf{W}_L}[L_S(\mathbf{W}^{(k)})]\|_F^2 \geq \frac{c_0 m\phi}{n^5}\left(\sum_{i=1}^n \ell'(y_i\widehat{y}_i^{(k)})\right)^2.$$

We set the radius τ and the step size η as follows,

$$\tau = \left(\frac{c_0^{1/2} m^{1/2}\phi^{1/2}}{4C'' L^{8/3} n^{3/2}\sqrt{M\log(M)}}\right)^3 = \widetilde{O}(n^{-9/2}L^{-8}\phi^{3/2}),$$
$$\eta = \frac{1}{4C' M L^3} = O(M^{-1}L^{-3}).$$

Then we have

$$\begin{aligned}
&L_S(\mathbf{W}^{(k+1)}) - L_S(\mathbf{W}^{(k)})\\
&\quad\leq -\frac{3\eta}{4}\|\nabla L_S(\mathbf{W}^{(k)})\|_F^2 - \frac{c_0\eta m^{1/2}\phi^{1/2}}{4n^{5/2}}\|\nabla L_S(\mathbf{W}^{(k)})\|_F\cdot\sum_{i=1}^n \ell'(y_i\widehat{y}_i^{(k)})\\
&\quad\leq -\frac{\eta}{2}\|\nabla L_S(\mathbf{W}^{(k)})\|_F^2\\
&\quad\leq -\eta\frac{c_0 m\phi}{2n^5}\left(\sum_{i=1}^n \ell'(y_i\widehat{y}_i^{(k)})\right)^2, \qquad (5.1)
\end{aligned}$$

where the first inequality is by Lemma 5.3 and the choices of η and τ, the second inequality follows from Lemma 5.2, and the last inequality is due to the gradient lower bound we derived above. Note that $\ell(x) = \log(1+\exp(-x))$, which satisfies $-\ell'(x) = 1/(1+\exp(x)) \geq \min\{\alpha_0, \alpha_1\ell(x)\}$ where $\alpha_0 = 1/2$ and $\alpha_1 = 1/(2\log(2))$. This implies that

$$-\sum_{i=1}^n \ell'(y_i\widehat{y}_i^{(k)}) \geq \min\left\{\alpha_0, \sum_{i=1}^n \alpha_1\ell(y_i\widehat{y}_i^{(k)})\right\} \geq \min\{\alpha_0, n\alpha_1 L_S(\mathbf{W}^{(k)})\}.$$

Note that $\min\{a, b\} \geq 1/(1/a+1/b)$, we have the following by plugging the above inequality into (5.1)

$$\begin{aligned}
L_S(\mathbf{W}^{(k+1)}) - L_S(\mathbf{W}^{(k)}) &\leq -\eta\min\left\{\frac{c_0 m\phi\alpha_0^2}{2n^5}, \frac{c_0 m\phi\alpha_1^2}{2n^3}L_S^2(\mathbf{W}^{(k)})\right\}\\
&\leq -\eta\left(\frac{2n^5}{c_0 m\phi\alpha_0^2} + \frac{2n^3}{c_0 m\phi\alpha_1^2 L_S^2(\mathbf{W}^{(k)})}\right)^{-1}.
\end{aligned}$$

Rearranging terms gives

$$\frac{2n^5}{c_0 m\phi\alpha_0^2}\big(L_S(\mathbf{W}^{(k+1)}) - L_S(\mathbf{W}^{(k)})\big) + \frac{2n^3\big(L_S(\mathbf{W}^{(k+1)}) - L_S(\mathbf{W}^{(k)})\big)}{c_0 m\phi\alpha_1^2 L_S^2(\mathbf{W}^{(k)})} \leq -\eta. \qquad (5.2)$$

Applying the inequality $(x-y)/y^2 \geq y^{-1} - x^{-1}$ and taking telescope sum over k give

$$k\eta \leq \frac{2n^5}{c_0 m\phi\alpha_0^2}\big(L_S(\mathbf{W}^{(0)}) - L_S(\mathbf{W}^{(k)})\big) + \frac{2n^3\big(L_S^{-1}(\mathbf{W}^{(k)}) - L_S^{-1}(\mathbf{W}^{(0)})\big)}{c_0 m\phi\alpha_1^2}$$

$$\leq \frac{2n^5}{c_0 m\phi\alpha_0^2} L_S(\mathbf{W}^{(0)}) + \frac{2n^3\big(L_S^{-1}(\mathbf{W}^{(k)}) - L_S^{-1}(\mathbf{W}^{(0)})\big)}{c_0 m\phi\alpha_1^2}. \tag{5.3}$$

Now we need to guarantee that after K gradient descent steps the loss function $L_S(\mathbf{W}^{(K)})$ is smaller than the target accuracy ϵ. By Lemma 5.1, we know that the training loss $L_S(\mathbf{W}^{(0)}) = \widetilde{O}(1)$. Therefore, by (5.3) and our choice of η, the maximum iteration number K satisfies

$$K = \widetilde{O}\big(n^5L^3/\phi + n^3L^3\epsilon^{-1}/\phi\big). \tag{5.4}$$

Then we are going to verify the condition that all iterates stay inside the perturbation region $\mathcal{B}(\mathbf{W}^{(0)}, \tau)$. We prove this by induction. Clearly, $\mathbf{W}^{(0)} \in \mathcal{B}(\mathbf{W}^{(0)}, \tau)$. Then we are going to prove $\mathbf{W}^{(k+1)} \in \mathcal{B}(\mathbf{W}^{(0)}, \tau)$ under the induction hypothesis that $\mathbf{W}^{(t)} \in \mathcal{B}(\mathbf{W}^{(0)}, \tau)$ holds for all $t \leq k$. According to (5.1), we have

$$L_S(\mathbf{W}^{(t+1)}) - L_S(\mathbf{W}^{(t)}) \leq -\frac{\eta}{2}\|\nabla L_S(\mathbf{W}^{(t)})\|_F^2, \tag{5.5}$$

for any $t < k$. Therefore, by triangle inequality, we have

$$\begin{aligned}
\|\mathbf{W}_l^{(k)} - \mathbf{W}_l^{(0)}\|_2 &\leq \eta \sum_{t=0}^{k-1} \big\|\nabla_{\mathbf{W}_l}[L_S(\mathbf{W}^{(t)})]\big\|_2 \\
&\leq \eta\sqrt{k\sum_{t=0}^{k-1}\big\|\nabla L_S(\mathbf{W}^{(t)})\big\|_F^2} \\
&\leq \sqrt{2k\eta\sum_{t=0}^{k-1}\big[L_S(\mathbf{W}^{(t)}) - L_S(\mathbf{W}^{(t+1)})\big]} \\
&\leq \sqrt{2k\eta L_S(\mathbf{W}^{(0)})}.
\end{aligned}$$

By Lemma 5.1, we know that $L_S(\mathbf{W}^{(0)}) = \widetilde{O}(1)$. Then applying our choices of η and K, we have

$$\|\mathbf{W}_l^{(k)} - \mathbf{W}_l^{(0)}\|_2 \leq \sqrt{2K\eta L_S(\mathbf{W}^{(0)})} = \widetilde{O}\big(n^{5/2}\phi^{-1/2}m^{-1/2} + n^{3/2}\epsilon^{-1/2}\phi^{-1/2}m^{-1/2}\big).$$

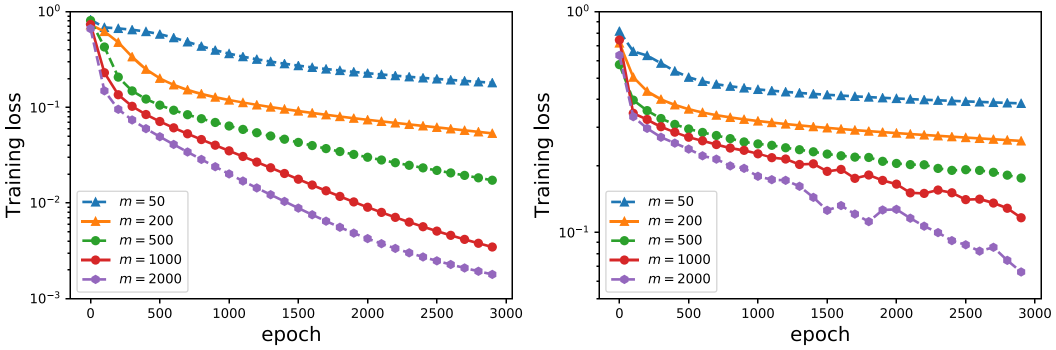
In addition, by Lemma 5.2 and our choice of η, we have

$$\begin{aligned}
\eta\|\nabla_{\mathbf{W}_l}[L_S(\mathbf{W}^{(k)})]\|_2 &\leq -\frac{\eta C'LM^{1/2}}{n}\sum_{i=1}^n \ell'\big(y_i \cdot f_{\mathbf{W}^{(k)}}(\mathbf{x}_i)\big) \\
&\leq \widetilde{O}(L^{-2}M^{-1/2}),
\end{aligned}$$

where the second inequality follows from the choice of η and the fact that $-1 \leq \ell'(\cdot) \leq 0$. Then by triangle inequality, we have

$$\begin{aligned}
\|\mathbf{W}_l^{(k+1)} - \mathbf{W}_l^{(0)}\|_2 &\leq \eta\|\nabla_{\mathbf{W}_l}[L_S(\mathbf{W}^{(k)})]\|_2 + \|\mathbf{W}_l^{(k)} - \mathbf{W}_l^{(0)}\|_2 \\
&= \widetilde{O}(n^{-9/2}L^{-8}\phi^{3/2}),
\end{aligned}$$

which is exactly in the same order of τ, where the last equality follows from the over-parameterization assumption $m = \widetilde{\Omega}\big(n^{14}L^{16}\phi^{-4} + n^{12}L^{16}\phi^{-4}\epsilon^{-1}\big)$. This verifies that $\mathbf{W}^{(k+1)} \in \mathcal{B}(\mathbf{W}^{(0)}, \tau)$ and completes the induction for k. Thus we can complete the proof. □



(a) *MNIST* **(b)** *CIFAR10*

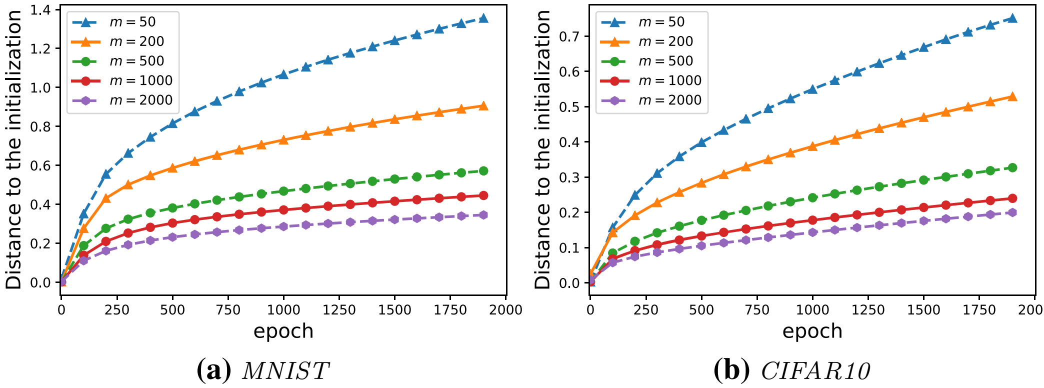
Fig. 1 The convergence of GD for training deep ReLU network with different network widths. **a** MNIST dataset. **b** CIFAR10 dataset

6 Experiments

In this section we carry out experiments on two real datasets (MNIST LeCun et al. 1998 and CIFAR10 Krizhevsky 2009) to support our theory. Since we mainly focus on binary classification, we extract a subset with digits 3 and 8 from the original MNIST dataset, which consists of 9, 943 training examples. In addition, we also extract two classes of images (“cat” and “ship”) from the original CIFAR10 dataset, which consists of 7, 931 training examples. Regarding the neural network architecture, we use a fully-connected deep ReLU network with $L = 15$ hidden layers, each layer has width m. The network architecture is consistent with the setting of our theory.

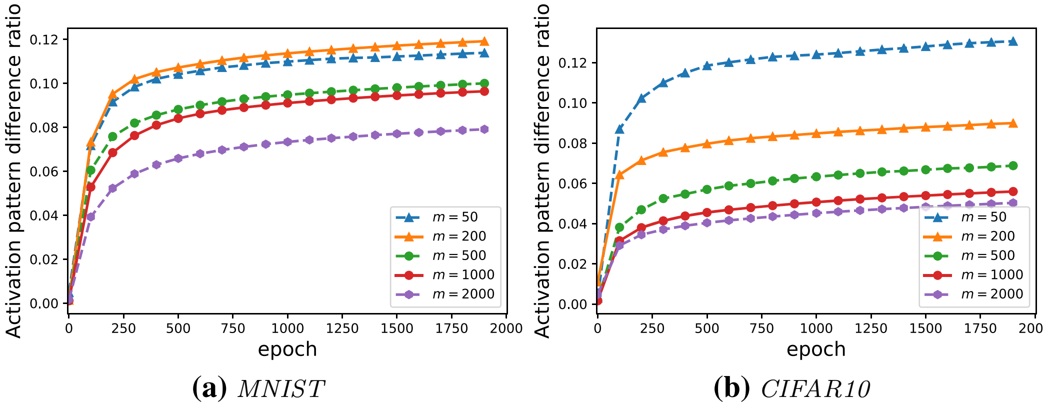
We first demonstrate that over-parameterization indeed helps optimization. We run GD for training deep ReLU networks with different network widths and plot the training loss in Fig. 1, where we apply cross-entropy loss on both MNIST and CIFAR10 datasets. In addition, the step sizes are set to be small enough and fixed for ReLU networks with different width. It can be observed that over-parameterization indeed speeds up the convergence of gradient descent, which is consistent with Lemmas 5.2 and 5.3, since the square of gradient norm scales with m, which further implies that wider network leads to larger function decrease if the step size is fixed. We also display the distance between the iterates of GD and the initialization in Fig. 2. It shows that when the network becomes wider, GD is more likely to converge to a point closer to the initialization. This suggests that the iterates of GD for training an over-parameterized deep ReLU network are harder to exceed the required perturbation region, thus can be guaranteed to converge to a global minimum. This corroborates our theory.

Finally, we monitor the activation pattern changes of all hidden neurons during the training process, and show the results in Fig. 3, where we use cross-entropy loss on both MNIST and CIFAR10 datasets. Specifically, in each iteration, we compare the activation status of all hidden nodes regarding all inputs with those at the initialization, and compute the number of nodes whose activation status differs from that at the initialization. From Fig. 3 it is clear that the activation pattern difference ratio dramatically decreases as the neural network becomes wider, which brings less non-smoothness during the training process. This implies that wider ReLU network can better guarantee sufficient function decrease after one-step gradient descent, which is consistent with our theory.



(a) *MNIST* **(b)** *CIFAR10*

Fig. 2 Distance between the iterates of GD and the initialization. **a** MNIST dataset. **b** CIFAR10 dataset



(a) *MNIST* **(b)** *CIFAR10*

Fig. 3 Activation pattern difference ratio between iterates of GD and the initialization. **a** MNIST dataset. **b** CIFAR10 dataset

7 Conclusions and future work

In this paper, we studied training deep neural networks by gradient descent. We proved that gradient descent can achieve global minima of the training loss for over-parameterized deep ReLU networks with random initialization, with milder assumption on the training data. Compared with the state-of-the-art results, our theoretical guarantees are sharper in terms of both over-parameterization condition and convergence rate. Our result can also be extended to stochastic gradient descent (SGD) and other loss functions (e.g., square hinge loss and smoothed hinge loss). Such extensions can be found in the longer version of this paper (Zou et al. 2018). In the future, we will further improve the over-parameterization condition such that it is closer to width of neural networks used in practice. Our proof technique can also be extended to other neural network architectures including convolutional neural networks (CNNs) (Krizhevsky et al. 2012), residual networks (ResNets) (He et al. 2016) and recurrent neural networks (RNNs) (Hochreiter and Schmidhuber 1997), and give sharper over-parameterization conditions than existing results for CNNs, ResNets (Du et al. 2018a; Allen-Zhu et al. 2018a) and RNNs (Allen-Zhu et al. 2018b). Moreover, it is also interesting to explore how our optimization guarantees of over-parameterized neural networks can be integrated with existing universal approximation ability results such as Hornik (1991), Telgarsky (2016), Lin and Jegelka (2018), Zhou (2019).

Acknowledgements We would like to thank the anonymous reviewers for their helpful comments. This research was sponsored in part by the National Science Foundation BIGDATA IIS-1855099, IIS-1903202 and IIS-1906169. QG is also partially supported by the Salesforce Deep Learning Research Grant. We also thank AWS for providing cloud computing credits associated with the NSF BIGDATA award. The views and conclusions contained in this paper are those of the authors.

A Proofs of lemmas in Sect. 5

In this section we provide the proof of all lemmas in Sect. 5.

A.1 Proof of Lemma 5.1

We first provide the following lemma that bounds the output of all hidden layer.

Lemma A.1 *With Gaussian random initialization, for any $\delta \in (0,1)$, if $m \geq \overline{C} L^2 \log(nL/\delta)$ for some large enough constant $\overline{C}$, then with probability at least $1-\delta$, the following holds for all $l \in [L]$,*

$$\big| \|\mathbf{x}_{l,i}\|_2 - 1 \big| \leq Cl \sqrt{\frac{\log(nL/\delta)}{m}},$$

where $m = \min\{m_1, \ldots, m_L\}$, and C is an absolute constant.

Proof of Lemma 5.1 Note that half of the entries of $\mathbf{v}$ are 1's and the other half of the entries are -1's. Therefore, without loss of generality, here we assume that $v_1 = \cdots = v_{m_L/2} = 1$ and $v_{m_L/2+1} = \cdots = v_{m_L} = -1$. Clearly, we have $\mathbb{E}(\widehat{y}_i) = 0$. Moreover, plugging in the value of $\mathbf{v}$ gives

$$\widehat{y}_i = \sum_{j=1}^{m_L/2} [\sigma(\mathbf{w}_{L,j}^\top \mathbf{x}_{L-1,i}) - \sigma(\mathbf{w}_{L,j+m_L/2}^\top \mathbf{x}_{L-1,i})].$$

Apparently, we have $\|\sigma(\mathbf{w}_{L,j}^\top \mathbf{x}_{L-1,i}) - \sigma(\mathbf{w}_{L,j+m_L/2}^\top \mathbf{x}_{L-1,i})\|_{\psi_2} \leq C_1 m_L^{-1/2}$ for some absolute constant C_1. Therefore by Hoeffding's inequality and Lemma A.1, with probability at least $1-\delta$, it holds that

$$|\widehat{y}_i| \leq C_2 \sqrt{\log(n/\delta)}$$

for all $i = 1, \ldots, n$. Then substituting the above bound into the formula of loss function $\ell(y_i \widehat{y}_i)$, we are able to complete the proof. □

A.2 Proof of Lemma 5.2

In order to prove Lemma 5.2, we require the following lemmas. We first establish the gradient lower bound at the initialization. Specifically, the following lemma gives a lower bound of gradient norm with respect to the weight matrix in the last hidden layer.

Lemma A.2 *There exist absolute constants $C, C', C'', C''' > 0$ such that, if $m \geq Cn^2\phi^{-1}\log(n)$, then with probability at least $1 - \exp(-C' m_L \phi/n)$, for any $\mathbf{a} = (a_1, \ldots, a_n)^\top \in \mathbb{R}_+^n$, there exist at least $C'' m_L \phi/n$ nodes in $\{1, \ldots, j, \ldots, m_L\}$ that satisfy*

$$\left\|\frac{1}{n}\sum_{i=1}^{n} a_i y_i \sigma'(\langle \mathbf{w}_{L,j}, \mathbf{x}_{L-1,i}\rangle)\mathbf{x}_{L-1,i}\right\|_2 \ge C'''\|\mathbf{a}\|_\infty/n$$

The following lemma characterizes the Lipschitz continuity of the gradients when the neural network parameters are staying inside the required perturbation region, which is essential to bound the norms of gradients.

Lemma A.3 [Lemmas B.1 and B.2 in Zou et al. (2018)] *Suppose that* $\mathbf{W}_1, \ldots, \mathbf{W}_L$ *are generated via Gaussian initialization. For* $\tau > 0$, *let* $\widetilde{\mathbf{W}}_1, \ldots, \widetilde{\mathbf{W}}_L$ *with* $\|\widetilde{\mathbf{W}}_l - \mathbf{W}_l\|_2 \le \tau$, $l = 1, \ldots, L$ *be the perturbed matrices. Let* $\widetilde{\mathbf{\Sigma}}_{l,i}$, $l = 1, \ldots, L$, $i = 1, \ldots, n$ *be diagonal matrices satisfying* $\|\widetilde{\mathbf{\Sigma}}_{l,i} - \mathbf{\Sigma}_{l,i}\|_0 \le s$ *and* $|(\widetilde{\mathbf{\Sigma}}_{l,i} - \mathbf{\Sigma}_{l,i})_{jj}|, |(\widetilde{\mathbf{\Sigma}}_{l,i})_{jj}| \le 1$ *for all* $l = 1, \ldots, L$, $i = 1, \ldots, n$ *and* $j = 1, \ldots, m_l$. *If* $\tau, \sqrt{s\log(M)/m} \le \kappa L^{-3/2}$ *for some small enough absolute constant* κ, *then*

$$\left\|\prod_{r=l_1}^{l_2}\widetilde{\mathbf{\Sigma}}_{r,i}\widetilde{\mathbf{W}}_r^\top\right\|_2 \le C\sqrt{L}, \quad \left\|\mathbf{v}^\top\prod_{r=l_1}^{L}\widetilde{\mathbf{\Sigma}}_{r,i}\widetilde{\mathbf{W}}_r^\top\right\|_2 \le C'\sqrt{M}, \quad \left\|\mathbf{v}^\top\prod_{r=l_1}^{L}\widetilde{\mathbf{\Sigma}}_{r,i}\widetilde{\mathbf{W}}_r^\top\mathbf{u}\right\|_2 \le C''\sqrt{s\log(M)}$$

for any $1 \le l_1 < l_2 \le L$ *and vector* $\mathbf{u}$ *with* $\|\mathbf{u}\|_2 = 1$ *and* $\|\mathbf{u}\|_0 \le s$, *where* C, C' *and* C'' *are absolute constants.*

We then provide the following lemma which characterizes the difference between activation patterns and outputs of all hidden layers generated by any two different neural networks.

Lemma A.4 [Lemma B.3 in Zou et al. (2018)] *Suppose that* $\mathbf{W}_1, \ldots, \mathbf{W}_L$ *are generated via Gaussian initialization. Let* $\widetilde{\mathbf{W}} = \{\widetilde{\mathbf{W}}_1, \ldots, \widetilde{\mathbf{W}}_L\}$, $\widehat{\mathbf{W}} = \{\widehat{\mathbf{W}}_1, \ldots, \widehat{\mathbf{W}}_L\}$ *be two collections of weight matrices satisfying* $\|\widetilde{\mathbf{W}}_l - \mathbf{W}_l\|_2, \|\widehat{\mathbf{W}}_l - \mathbf{W}_l\|_2 \le \tau$, $l = 1, \ldots, L$. *Let* $\mathbf{\Sigma}_{l,i}, \widetilde{\mathbf{\Sigma}}_{l,i}, \widehat{\mathbf{\Sigma}}_{l,i}$ *and* $\mathbf{x}_{l,i}, \widetilde{\mathbf{x}}_{l,i}, \widehat{\mathbf{x}}_{l,i}$ *be the binary matrices and hidden layer outputs at the* l*-th layer with parameter matrices* $\mathbf{W}, \widetilde{\mathbf{W}}, \widehat{\mathbf{W}}$ *respectively. If* $\tau \le \kappa' L^{-11}(\log(M))^{-3/2}$ *for some small enough absolute constant* $\kappa' > 0$, *then there exits constants* C *and* C' *such that*

$$\|\widehat{\mathbf{x}}_{l,i} - \widetilde{\mathbf{x}}_{l,i}\|_2 \le CL\cdot\sum_{r=1}^{l}\|\widehat{\mathbf{W}}_r - \widetilde{\mathbf{W}}_r\|_2, \qquad \|\widehat{\mathbf{\Sigma}}_{l,i} - \widetilde{\mathbf{\Sigma}}_{l,i}\|_0 \le C'L^{4/3}\tau^{2/3}m_l,$$

for all $l = 1, \ldots, L$ *and* $i = 1, \ldots, n$.

Now we ready to prove Lemma 5.2.

Proof of Lemma 5.2 We first prove the gradient upper bound. For the training example $(\mathbf{x}_i, y_i)$, let $\widetilde{y}_i = f_{\widetilde{\mathbf{W}}}(\mathbf{x}_i)$, the gradient $\nabla_{\mathbf{W}_l}\ell(y_i\widetilde{y}_i)$ can be written as follows,

$$\begin{aligned}\nabla_{\mathbf{W}_l}\ell(y_i\widetilde{y}_i) &= \ell'(y_i\widetilde{y}_i)y_i\nabla_{\mathbf{W}_l}[f_{\widetilde{\mathbf{W}}}(\mathbf{x}_i)]\\ &= \ell'(y_i\widetilde{y}_i)y_i\widetilde{\mathbf{x}}_{l-1,i}\mathbf{v}^\top\left(\prod_{r=l+1}^{L}\widetilde{\mathbf{\Sigma}}_{r,i}\widetilde{\mathbf{W}}_l^\top\right)\widetilde{\mathbf{\Sigma}}_{l,i}.\end{aligned}$$

Note that by Lemma A.3, there exists an absolute constant C_0 such that $\|\prod_{r=l_1}^{l_2}\widetilde{\mathbf{\Sigma}}_{r,i}\widetilde{\mathbf{W}}_r\|_2 \le C_0\sqrt{L}$. Hence, we have the following upper bound on $\left\|\nabla_{\mathbf{W}_l}\ell(y_i\widetilde{y}_i)\right\|_F$,

$$\begin{aligned}\left\|\nabla_{\mathbf{W}_l}\ell(y_i\widetilde{y}_i)\right\|_F = \left\|\nabla_{\mathbf{W}_l}\ell(y_i\widetilde{y}_i)\right\|_2 &\le -\ell'(y_i\widetilde{y}_i)\left\|\prod_{r=1}^{l-1}\widetilde{\mathbf{\Sigma}}_{r,i}\widetilde{\mathbf{W}}_r^\top\mathbf{x}_i\right\|_2\left\|\prod_{r=l+1}^{L}\widetilde{\mathbf{\Sigma}}_{r,i}\widetilde{\mathbf{W}}_l^\top\right\|_2\|\mathbf{v}\|_2\\ &\le -\ell'(y_i\widetilde{y}_i)C_0^2LM^{1/2},\end{aligned}$$

where the first equality holds due to the fact that $\nabla_{\mathbf{W}_l}\ell(y_i\widetilde{y}_i) = \ell'(y_i\widetilde{y}_i)y_i\widetilde{\mathbf{x}}_{l-1,i}\mathbf{v}^\top\big(\prod_{r=l+1}^{L}\widetilde{\boldsymbol{\Sigma}}_{r,i}\widetilde{\mathbf{W}}_l^\top\big)\widetilde{\boldsymbol{\Sigma}}_{l,i}$ is a rank-one matrix, and the last inequality follows from the fact that $\|\mathbf{v}\|_2 = m_L^{1/2} \le M^{1/2}$. Moreover, we have the following for $\nabla_{\mathbf{W}_l}[L_S(\widetilde{\mathbf{W}})]$:

$$\big\|\nabla_{\mathbf{W}_l}[L_S(\widetilde{\mathbf{W}})]\big\|_F = \left\|\frac{1}{n}\sum_{i=1}^{n}\nabla_{\mathbf{W}_l}\ell(y_i\widetilde{y}_i)\right\|_F \le \frac{1}{n}\sum_{i=1}^{n}\big\|\nabla_{\mathbf{W}_l}\ell(y_i\widetilde{y}_i)\big\|_F \le -\frac{C_0^2LM^{1/2}}{n}\sum_{i=1}^{n}\ell'(y_i\widetilde{y}_i),$$

which completes the proof of gradient upper bound.

Now we are going to prove the gradient lower bound. Given initialization $\mathbf{W}^{(0)}$ and any $\widetilde{\mathbf{W}} \in \mathcal{B}(\mathbf{W}^{(0)}, \tau)$, let $\widetilde{y}_i = f_{\widetilde{\mathbf{W}}}(\mathbf{x}_i)$, we define

$$\mathbf{g}_j = \frac{1}{n}\sum_{i=1}^{n}\ell'(y_i\widetilde{y}_i)y_i\mathbf{v}_j\sigma'(\langle\mathbf{w}_{L,j}^{(0)}, \mathbf{x}_{L-1,i}\rangle)\mathbf{x}_{L-1,i},$$

where $\mathbf{x}_{L,i}$ denotes the output of the last hidden layer with input $\mathbf{x}_i$ at the initialization. Then since $\mathbf{W}^{(0)}$ is generated via Gaussian random initialization, by Lemma A.2, we have the following holds for at least $C_2m_L\phi/n$ nodes,

$$\|\mathbf{g}_j\|_2 \ge C_1\max_i|\ell'(y_i\widetilde{y}_i)|/n$$

where $C_1, C_2 > 0$ are positive absolute constants. Moreover, we rewrite the gradient $\nabla_{\mathbf{W}_{L,j}}L_S(\widetilde{\mathbf{W}})$ as follows:

$$\nabla_{\mathbf{W}_{L,j}}L_S(\widetilde{\mathbf{W}}) = \frac{1}{n}\sum_{i=1}^{n}\ell'(y_i\widetilde{y}_i)y_i\mathbf{v}_j\sigma'(\langle\widetilde{\mathbf{w}}_{L,j}, \widetilde{\mathbf{x}}_{L-1,i}\rangle)\widetilde{\mathbf{x}}_{L-1,i},$$

where $\widetilde{\mathbf{x}}_{l,i}$ denotes the output of the l-th hidden layer with input $\mathbf{x}_i$ and weight matrices $\widetilde{\mathbf{W}}$. Let $b_{i,j} = \ell'(y_i\widetilde{y}_i)y_i\mathbf{v}_j$, we have

$$\begin{aligned}
&\|\mathbf{g}_j\|_2 - \|\nabla_{\mathbf{W}_{L,j}}L_S(\widetilde{\mathbf{W}})\|_2\\
&\le \left\|\frac{1}{n}\sum_{i=1}^{n}b_{i,j}\big(\sigma'(\langle\widetilde{\mathbf{w}}_{L,j}, \widetilde{\mathbf{x}}_{L-1,i}\rangle)\widetilde{\mathbf{x}}_{L-1,i} - \sigma'(\langle\mathbf{w}_{L,j}^{(0)}, \mathbf{x}_{L-1,i}\rangle)\mathbf{x}_{L-1,i}\big)\right\|_2\\
&\le \left\|\frac{1}{n}\sum_{i=1}^{n}b_{i,j}\Big[\big(\sigma'(\langle\widetilde{\mathbf{w}}_{L,j}, \widetilde{\mathbf{x}}_{L-1,i}\rangle) - \sigma'(\langle\mathbf{w}_{L,j}^{(0)}, \mathbf{x}_{L-1,i}\rangle\big)\mathbf{x}_{L-1,i} + \sigma'(\langle\widetilde{\mathbf{w}}_{L,j}, \widetilde{\mathbf{x}}_{L-1,i}\rangle)(\widetilde{\mathbf{x}}_{L-1,i} - \mathbf{x}_{L-1,i})\Big]\right\|_2.
\end{aligned}$$

According to Lemma A.4, the number of nodes satisfying $\sigma'(\langle\widetilde{\mathbf{w}}_{L,j}, \widetilde{\mathbf{x}}_{L-1,i}\rangle) - \sigma'(\langle\mathbf{w}_{L,j}^{(0)}, \mathbf{x}_{L-1,i}\rangle \ne 0$ for at least one i is at most $C_3nL^{4/3}\tau^{2/3}m_L$, where C_3 is an absolute constant. For the rest of the nodes in this layer, we have

$$\begin{aligned}
\|\mathbf{g}_j\|_2 - \|\nabla_{\mathbf{W}_{L,j}}L_S(\widetilde{\mathbf{W}})\|_2 &\le \left\|\frac{1}{n}\sum_{i=1}^{n}b_{i,j}\sigma'(\langle\widetilde{\mathbf{w}}_{L,j}, \widetilde{\mathbf{x}}_{L-1,i}\rangle)(\widetilde{\mathbf{x}}_{L-1,i} - \mathbf{x}_{L-1,i})\right\|_2\\
&\le \frac{1}{n}\sum_{i=1}^{n}C_4L^2\tau|b_{i,j}|\\
&\le C_4L^2\tau\max_i|\ell'(y_i\widetilde{y}_i)|,
\end{aligned}$$

where C_4 is an absolute constant, the first inequality holds since these nodes satisfy $\sigma'(\langle\widetilde{\mathbf{w}}_{L,j}, \widetilde{\mathbf{x}}_{L-1,i}\rangle) - \sigma'(\langle\mathbf{w}_{L,j}^{(0)}, \mathbf{x}_{L-1,i}\rangle) = 0$ for all i, the second inequality follows from Lemma A.4 and triangle inequality. Let

$$\tau \le \left(\frac{C_2\phi}{2C_3n^2L^{4/3}}\right)^{3/2} \wedge \frac{C_1}{2nL^2C_4} = O\big(\phi^{3/2}n^{-3}L^{-2}\big).$$

Note that we have at least $C_2m_L\phi/n$ nodes satisfying $\|\mathbf{g}_j\|_2 \ge C_1 \max_i |\ell'(y_i\widetilde{y}_i)|/n$, thus there are at least $C_2m_L\phi/n - C_3nL^{4/3}\tau^{2/3}m_L = C_2m_L\phi/(2n)$ nodes satisfying

$$\|\nabla_{\mathbf{W}_{L,j}}L_S(\widetilde{\mathbf{W}})\|_2 \ge C_1\max_i|\ell'(y_i\widetilde{y}_i)|/n - C_4L^2\tau\max_i|\ell'(y_i\widetilde{y}_i)|/n \ge \frac{C_1\max_i|\ell'(y_i\widetilde{y}_i)|}{2n}.$$

Therefore,

$$\begin{aligned}
\|\nabla_{\mathbf{W}_L}L_S(\widetilde{\mathbf{W}})\|_F^2 &= \sum_{j=1}^{m_L}\|\nabla_{\mathbf{W}_{L,j}}L_S(\widetilde{\mathbf{W}})\|_2^2\\
&\ge \frac{C_2\phi m_L}{2n}\left(\frac{C_1\max_i|\ell'(y_i\widehat{y}_i^{(k)})y_i\mathbf{v}_j|}{2n}\right)^2\\
&\ge \frac{C_2C_1^2\phi m_L}{8n^5}\left(\sum_{i=1}^{n}\ell'(y_i\widehat{y}_i^{(k)})\right)^2,
\end{aligned}$$

where the last inequality follows from the fact that $\ell'(\cdot) < 0$ and $|y_i\mathbf{v}_j| = 1$. Let $C = C_2C_1^2/8$, we complete the proof. □

A. 3 Proof of Lemma 5.3

Proof of Lemma 5.3 Note that $\ell(x)$ is 1/4-smooth, thus the following holds for any Δ and x,

$$\ell(x+\Delta) \le \ell(x) + \ell'(x)\Delta + \frac{1}{8}\Delta^2.$$

Then we have the following upper bound on $L_S(\mathbf{W}^{(k+1)}) - L_S(\mathbf{W}^{(k)})$,

$$\begin{aligned}
L_S(\mathbf{W}^{(k+1)}) - L_S(\mathbf{W}^{(k)}) &= \frac{1}{n}\sum_{i=1}^{n}\Big[\ell\big(y_i\widehat{y}_i^{(k+1)}\big) - \ell\big(y_i\widehat{y}_i^{(k)}\big)\Big]\\
&\le \frac{1}{n}\sum_{i=1}^{n}\Big[\ell'(y_i\widehat{y}_i^{(k)})\Delta_i^{(k)} + \frac{1}{8}(\Delta_i^{(k)})^2\Big], \qquad \text{(A.1)}
\end{aligned}$$

where $\Delta_i^{(k)} = y_i\big(\widehat{y}_i^{(k+1)} - \widehat{y}_i^{(k)}\big)$. Therefore, our next goal is to bound the quantity $\Delta_i^{(k)}$.

The upper bound of $|\Delta_i^{(k)}|$ can be derived straightforwardly. By Lemma A.4, we know that there exists a constant C_1 such that

$$\begin{aligned}
\|\mathbf{x}_{L,i}^{(k+1)} - \mathbf{x}_{L,i}^{(k)}\|_2 &\le C_1L\cdot\sum_{l=1}^{L}\|\mathbf{W}_l^{(k+1)} - \mathbf{W}_l^{(k)}\|_2\\
&= C_1L\eta\sum_{l=1}^{L}\big\|\nabla_{\mathbf{W}_l}[L_S(\mathbf{W}^{(k)})]\big\|_2\\
&\le C_1L^{1.5}\eta\|\nabla L_S(\mathbf{W}^{(k)})\|_F. \qquad \text{(A.2)}
\end{aligned}$$

Therefore, it follows that

$$|\Delta_i^{(k)}| = |y_i\mathbf{v}^\top(\mathbf{x}_{L,i}^{(k+1)} - \mathbf{x}_{L,i}^{(k)})| \le \|\mathbf{v}\|_2\|\mathbf{x}_{L,i}^{(k+1)} - \mathbf{x}_{L,i}^{(k)}\|_2 \le C_1L^{1.5}M^{1/2}\|\nabla L_S(\mathbf{W}^{(k)})\|_F,$$

where we use the fact that $\|\mathbf{v}\|_2 \le M^{1/2}$. In what follows we are going to prove the lower bound of $\Delta_i^{(k)}$. Note that $\Delta_i^{(k)} = y_i \mathbf{v}^\top\big(\mathbf{x}_{L,i}^{(k+1)} - \mathbf{x}_{L,i}^{(k)}\big)$, thus we mainly focus on bounding the term $\mathbf{x}_{L,i}^{(k+1)} - \mathbf{x}_{L,i}^{(k)}$. For $l = 1, \dots, L$, we define the diagonal matrix $\widetilde{\boldsymbol{\Sigma}}_{l,i}^{(k)}$ as

$$\big(\widetilde{\boldsymbol{\Sigma}}_{l,i}^{(k)}\big)_{jj} = \big(\boldsymbol{\Sigma}_{l,i}^{(k+1)} - \boldsymbol{\Sigma}_{l,i}^{(k)}\big)_{jj} \cdot \frac{\mathbf{w}_{l,j}^{(k)\top}\mathbf{x}_{l-1,i}^{(k)}}{\mathbf{w}_{l,j}^{(k+1)\top}\mathbf{x}_{l-1,i}^{(k+1)} - \mathbf{w}_{l,j}^{(k)\top}\mathbf{x}_{l-1,i}^{(k)}}.$$

Given the above definition of $\widetilde{\boldsymbol{\Sigma}}_{l,i}^{(k)}$, we have

$$\begin{aligned}
&\mathbf{x}_{L,i}^{(k+1)} - \mathbf{x}_{L,i}^{(k)}\\
&= \big(\boldsymbol{\Sigma}_{L,i}^{(k)} + \widetilde{\boldsymbol{\Sigma}}_{L,i}^{(k)}\big)\big(\mathbf{W}_L^{(k+1)\top}\mathbf{x}_{L-1,i}^{(k+1)} - \mathbf{W}_L^{(k)\top}\mathbf{x}_{L-1,i}^{(k)}\big)\\
&= \big(\boldsymbol{\Sigma}_{L,i}^{(k)} + \widetilde{\boldsymbol{\Sigma}}_{L,i}^{(k)}\big)\mathbf{W}_L^{(k+1)\top}\big(\mathbf{x}_{L-1,i}^{(k+1)} - \mathbf{x}_{L-1,i}^{(k)}\big) + \big(\boldsymbol{\Sigma}_{L,i}^{(k)} + \widetilde{\boldsymbol{\Sigma}}_{L,i}^{(k)}\big)\big(\mathbf{W}_L^{(k+1)\top} - \mathbf{W}_L^{(k)\top}\big)\mathbf{x}_{L-1,i}^{(k)}\\
&= \sum_{l=1}^{L}\left(\prod_{r=l+1}^{L}\big(\boldsymbol{\Sigma}_{r,i}^{(k)} + \widetilde{\boldsymbol{\Sigma}}_{r,i}^{(k)}\big)\mathbf{W}_r^{(k+1)\top}\right)\big(\boldsymbol{\Sigma}_{l,i}^{(k)} + \widetilde{\boldsymbol{\Sigma}}_{l,i}^{(k)}\big)\big(\mathbf{W}_l^{(k+1)\top} - \mathbf{W}_l^{(k)\top}\big)\mathbf{x}_{l-1,i}^{(k)}.
\end{aligned}$$

Then we define

$$\boldsymbol{D}_{l,i}^{(k)} = \left(\prod_{r=l+1}^{L}\boldsymbol{\Sigma}_{r,i}^{(k)}\mathbf{W}_r^{(k)\top}\right)\boldsymbol{\Sigma}_{l,i}^{(k)}, \quad \widetilde{\boldsymbol{D}}_{l,i}^{(k)} = \left(\prod_{r=l+1}^{L}\big(\boldsymbol{\Sigma}_{r,i}^{(k)} + \widetilde{\boldsymbol{\Sigma}}_{r,i}^{(k)}\big)\mathbf{W}_r^{(k+1)\top}\right)\big(\boldsymbol{\Sigma}_{l,i}^{(k)} + \widetilde{\boldsymbol{\Sigma}}_{l,i}^{(k)}\big).$$

Then by triangle inequality, we have

$$\big\|\mathbf{v}^\top\big(\boldsymbol{D}_{l,i}^{(k)} - \widetilde{\boldsymbol{D}}_{l,i}^{(k)}\big)\big\|_2 \le \big\|\mathbf{v}^\top\big(\boldsymbol{D}_{l,i}^{(k)} - \boldsymbol{D}_{l,i}^{(0)}\big)\big\|_2 + \big\|\mathbf{v}^\top\big(\boldsymbol{D}_{l,i}^{(0)} - \widetilde{\boldsymbol{D}}_{l,i}^{(k)}\big)\big\|_2.$$

Note that,

$$\begin{aligned}
&\big\|\mathbf{v}^\top\big(\boldsymbol{D}_{l,i}^{(k)} - \boldsymbol{D}_{l,i}^{(0)}\big)\big\|_2\\
&\le \sum_{r=l}^{L}\left\|\mathbf{v}^\top\left(\prod_{t=r+1}^{L}\boldsymbol{\Sigma}_{t,i}^{(k)}\mathbf{W}_t^{(k)\top}\right)\left(\boldsymbol{\Sigma}_{t,i}^{(k)}\mathbf{W}_t^{(k)\top} - \boldsymbol{\Sigma}_{t,i}^{(0)}\mathbf{W}_t^{(0)\top}\right)\left(\prod_{t=l+1}^{L}\boldsymbol{\Sigma}_{t,i}^{(0)}\mathbf{W}_t^{(0)\top}\right)\right\|_2\\
&\le \sum_{r=l}^{L}\left\|\mathbf{v}^\top\left(\prod_{t=r+1}^{L}\boldsymbol{\Sigma}_{t,i}^{(k)}\mathbf{W}_t^{(k)\top}\right)\left(\boldsymbol{\Sigma}_{t,i}^{(k)} - \boldsymbol{\Sigma}_{t,i}^{(0)}\right)\right\|_2\left\|\mathbf{W}_t^{(0)\top}\left(\prod_{t=l+1}^{L}\boldsymbol{\Sigma}_{t,i}^{(0)}\mathbf{W}_t^{(0)\top}\right)\right\|_2\\
&\quad + \sum_{r=l}^{L}\left\|\mathbf{v}^\top\left(\prod_{t=r+1}^{L}\boldsymbol{\Sigma}_{t,i}^{(k)}\mathbf{W}_t^{(k)\top}\right)\boldsymbol{\Sigma}_{t,i}^{(k)}\right\|_2\left\|\mathbf{W}_t^{(k)} - \mathbf{W}_t^{(0)\top}\right\|_2\left\|\prod_{t=l+1}^{L}\boldsymbol{\Sigma}_{t,i}^{(0)}\mathbf{W}_t^{(0)\top}\right\|_2.
\end{aligned}$$

Then by Lemma A.3, and use the fact that $\|\boldsymbol{\Sigma}_{t,i}^{(k)} - \boldsymbol{\Sigma}_{t,i}^{(0)}\|_0 \le O\big(L^{4/3}\tau^{2/3}M\big)$, we have

$$\big\|\mathbf{v}^\top\big(\boldsymbol{D}_{l,i}^{(k)} - \boldsymbol{D}_{l,i}^{(0)}\big)\big\|_2 \le C_2 L^{13/6}\tau^{1/3}\sqrt{M\log(M)} + C_3 L^{3/2}\sqrt{M}\tau,$$

where C_2 and C_3 are absolute constants and we use the fact that $\|\mathbf{v}\|_2 \le \sqrt{M}$. Then note that $\tau \le 1$, the second term on the R.H.S. of the above inequality is dominated by the first one. Then we have

$$\big\|\mathbf{v}^\top\big(\boldsymbol{D}_{l,i}^{(k)} - \boldsymbol{D}_{l,i}^{(0)}\big)\big\|_2 \le C_5 L^{13/6}\tau^{1/3}\sqrt{M\log(M)}, \tag{A.3}$$

where C_5 is an absolute constant. This inequality also holds for $\big\|\mathbf{v}^\top\big(\widetilde{\boldsymbol{D}}_{l,i}^{(k)} - \boldsymbol{D}_{l,i}^{(0)}\big)\big\|_2$. Therefore, we have

$$\begin{aligned}
\Delta_i^{(k)} &= y_i \mathbf{v}^\top\big(\mathbf{x}_{L,i}^{(k+1)} - \mathbf{x}_{L,i}^{(k)}\big)\\
&= y_i \mathbf{v}^\top \sum_{l=1}^{L} \widetilde{\boldsymbol{D}}_{l,i}^{(k)}\big(\mathbf{W}_l^{(k+1)} - \mathbf{W}_l^{(k)}\big)\mathbf{x}_{l-1,i}^{(k)}\\
&= \underbrace{-y_i \mathbf{v}^\top \sum_{l=1}^{L}\big(\widetilde{\boldsymbol{D}}_{l,i}^{(k)} - \boldsymbol{D}_{l,i}^{(k)}\big)\big(\nabla_{\mathbf{W}_l}[L_S(\mathbf{W}^{(k)})]\big)^\top \mathbf{x}_{l-1,i}^{(k)}}_{I_{1,i}^{(k)}}\\
&\quad \underbrace{-y_i \mathbf{v}^\top \sum_{l=1}^{L} \boldsymbol{D}_{l,i}^{(k)}\big(\nabla_{\mathbf{W}_l}[L_S(\mathbf{W}^{(k)})]\big)^\top \mathbf{x}_{l-1,i}^{(k)}}_{I_{2,i}^{(k)}}.
\end{aligned}$$

By (A.3) and triangle inequality, we know that

$$\begin{aligned}
|I_{1,i}^{(k)}| &\le 2C_5 L^{13/6}\tau^{1/3}\sqrt{M\log(M)}\eta \cdot \sum_{l=1}^{L}\big\|\nabla_{\mathbf{W}_l}[L_S(\mathbf{W}^{(k)})]\big\|_2\\
&\le 2C_5 L^{8/3}\tau^{1/3}\sqrt{M\log(M)}\eta \cdot \big\|\nabla L_S(\mathbf{W}^{(k)})\big\|_F.
\end{aligned}$$

Moreover, we have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\ell'(y_i\widehat{y}_i^{(k)})I_{2,i}^{(k)} &= -\frac{\eta}{n}\sum_{i=1}^{n}\ell'(y_i\widehat{y}_i^{(k)})y_i\mathbf{v}^\top\left(\prod_{r=l+1}^{L}\boldsymbol{\Sigma}_{r,i}^{(k)}\mathbf{W}_r^{(k)\top}\right)\boldsymbol{\Sigma}_{l,i}^{(k)}\big(\nabla_{\mathbf{W}_l}[L_S(\mathbf{W}^{(k)})]\big)^\top\mathbf{x}_{l-1,i}^{(k)}\\
&= -\frac{\eta}{n^2}\left\|\sum_{i=1}^{n}\ell'(y_i\widehat{y}_i^{(k)})y_i\mathbf{x}_{l-1,i}^{(k)}\mathbf{v}^\top\left(\prod_{r=l+1}^{L}\boldsymbol{\Sigma}_{r,i}^{(k)}\mathbf{W}_r^{(k)\top}\right)\boldsymbol{\Sigma}_{l,i}^{(k)}\right\|_F^2\\
&= -\eta\|\nabla_{\mathbf{W}_l}[L_S(\mathbf{W}^{(k)})]\|_F^2.
\end{aligned}$$

Therefore, putting everything together, we have

$$\begin{aligned}
&L_S(\mathbf{W}^{(k+1)}) - L_S(\mathbf{W}^{(k)})\\
&\le \frac{1}{n}\sum_{i=1}^{n}\Big[\ell'(y_i\widehat{y}_i^{(k)})\Delta_i^{(k)} + \frac{1}{8}(\Delta_i^{(k)})^2\Big]\\
&\le \frac{1}{n}\sum_{i=1}^{n}\ell'(y_i\widehat{y}_i^{(k)})\big(I_{1,i}^{(k)} + I_{2,i}^{(k)}\big) + C_3 M L^3\eta^2\|\nabla L_S(\mathbf{W}^{(k)})\|_F^2\\
&\le -\big(\eta - C_6 M L^3\eta^2\big)\|\nabla L_S(\mathbf{W}^{(k)})\|_F^2 - \frac{C_7 L^{8/3}\tau^{1/3}\sqrt{M\log(M)}\cdot\|\nabla L_S(\mathbf{W}^{(k)})\|_F}{n}\sum_{i=1}^{n}\ell'(y_i\widehat{y}_i^{(k)}),
\end{aligned}$$

where C_6 and C_7 are absolute constants. Thus we complete the proof. □

B Proofs of lemmas in Appendix A

B.1 Proof of Lemma A.1

Proof of Lemma A.1 In order to prove the desired results, it suffices to prove the inequality

$$\big|\|\mathbf{x}_{l,i}\|_2^2 - \|\mathbf{x}_{l-1,i}\|_2^2\big| \le C\|\mathbf{x}_{l-1,i}\|_2^2 \cdot \sqrt{\frac{\log(nL/\delta)}{m_l}}$$

for all $i = 1, \dots, n$ and $l = 1, \dots, L$, since this inequality implies that

$$\begin{aligned}\|\mathbf{x}_{l,i}\|_2 &\le \left[1 + C'\sqrt{\frac{\log(nL/\delta)}{m}}\right]^{1/2}\|\mathbf{x}_{l-1,i}\|_2 \le \cdots \le \left[1 + C'\sqrt{\frac{\log(nL/\delta)}{m}}\right]^{l/2}\|\mathbf{x}_i\|_2\\ &\le 1 + C'l\sqrt{\frac{\log(nL/\delta)}{m}},\end{aligned}$$

where C' is an absolute constant, and the last inequality follows by the fact that $(1+x)^{l/2} \le 1 + lx$ for $x \in (0, 1/(2L))$, which is applicable here because of the assumption $m \ge \overline{C}L^2\log(nL/\delta)$ for some large enough constant $\overline{C}$. Similarly, we can also proved that

$$\|\mathbf{x}_{l,i}\|_2 \ge 1 - C''l\sqrt{\frac{\log(nL/\delta)}{m}}$$

for some absolute constant C''. Combining the upper and lower bounds of $\|\mathbf{x}_{l,i}\|_2$ derived above gives the result of Lemma A.1.

For any fixed $i \in \{1, \dots, n\}$, $l \in \{1, \dots, L\}$ and $j \in \{1, \dots, m_l\}$, condition on $\mathbf{x}_{l-1,i}$ we have $\mathbf{w}_{l,j}^\top\mathbf{x}_{l-1,i} \sim N(\mathbf{0}, 2\|\mathbf{x}_{l-1,i}\|_2^2/m_l)$. Therefore,

$$\mathbb{E}[\sigma^2(\mathbf{w}_{l,j}^\top\mathbf{x}_{l-1,i})|\mathbf{x}_{l-1,i}] = \frac{1}{2}\mathbb{E}[(\mathbf{w}_{l,j}^\top\mathbf{x}_{l-1,i})^2|\mathbf{x}_{l-1,i}] = \frac{1}{m_l}\|\mathbf{x}_{l-1,i}\|_2^2.$$

Since $\|\mathbf{x}_{l,i}\|_2^2 = \sum_{j=1}^{m_l}\sigma^2(\mathbf{w}_{l,j}^\top\mathbf{x}_{l-1,i})$ and condition on $\mathbf{x}_{l-1}$, $\|\sigma^2(\mathbf{w}_{l,j}^\top\mathbf{x}_{l-1,i})\|_{\psi_1} \le C_1\|\mathbf{x}_{l-1,i}\|_2^2/m_l$ for some absolute constant C_1, by Bernstein inequality (see Proposition 5.16 in Vershynin 2010), for any $\xi \ge 0$ we have

$$\mathbb{P}\Big(\big|\|\boldsymbol{\Sigma}_{l,i}\mathbf{W}_l^\top\mathbf{x}_{l-1,i}\|_2^2 - \|\mathbf{x}_{l-1,i}\|_2^2\big| \ge \|\mathbf{x}_{l-1,i}\|_2^2\xi\,\Big|\,\mathbf{x}_{l-1,i}\Big) \le 2\exp(-C_2m_l\min\{\xi^2, \xi\}).$$

Taking union bound over l and i gives

$$\begin{aligned}\mathbb{P}\Big(\big|\|\mathbf{x}_{l,i}\|_2^2 - \|\mathbf{x}_{l-1,i}\|_2^2\big| \le \|\mathbf{x}_{l-1,i}\|_2^2\xi, i = 1, \dots, n, l = 1, \dots, L\Big)\\ \ge 1 - 2nL\exp(-C_2m_l\min\{\xi^2, \xi\}).\end{aligned}$$

The inequality above further implies that if $m_l \ge C_3^2\log(nL/\delta)$, then with probability at least $1-\delta$, we have

$$\big|\|\mathbf{x}_{l,i}\|_2^2 - \|\mathbf{x}_{l-1,i}\|_2^2\big| \le C_3\|\mathbf{x}_{l-1,i}\|_2^2 \cdot \sqrt{\frac{\log(nL/\delta)}{m_l}}$$

for any $i = 1, \dots, n$ and $l = 1, \dots, L$, where C_3 is an absolute constant. This completes the proof. □

B.2 Proof of Lemma A.2

In order to prove the gradient bounds, one key aspect is that the separation property for training data can be well preserved after passing through layers. The following lemma shows that the separation distance can be well preserved for all intermediate layers.

Lemma B.1 *Under the same conditions in Lemma A.2, with probability at least* $1 - \delta$,

$$\|\bar{\mathbf{x}}_{l,i} - \bar{\mathbf{x}}_{l,i'}\|_2 \geq \phi/2$$

for all $i, i' = 1, \ldots, n$ *with* $y_i \neq y_{i'}$, $l = 1, \ldots, L$.

Lemma B.2 *Let* $\mathbf{z}_1, \ldots, \mathbf{z}_n \in S^{d-1}$ *be n unit vectors and* $y_1, \ldots, y_n \in \{-1, 1\}$ *be the corresponding labels. Assume that for any* $i \neq j$ *such that* $y_i \neq y_j$, $\|\mathbf{z}_i - \mathbf{z}_j\|_2 \geq \widetilde{\phi}$ *and* $\mathbf{z}_i^\top \mathbf{z}_j \geq \widetilde{\mu}^2$ *for some* $\widetilde{\phi}, \widetilde{\mu} > 0$. *For any* $\mathbf{a} = (a_1, \ldots, a_n)^\top \in \mathbb{R}_+^n$, *let* $\mathbf{h}(\mathbf{w}) = \sum_{i=1}^n a_i y_i \sigma'(\langle \mathbf{w}, \mathbf{z}_i \rangle)\mathbf{z}_i$ *where* $\mathbf{w} \sim N(\mathbf{0}, \mathbf{I})$ *is a Gaussian random vector. If* $\widetilde{\phi} \leq \widetilde{\mu}/2$, *then there exist absolute constants* $C, C' > 0$ *such that*

$$\mathbb{P}\big[\|\mathbf{h}(\mathbf{w})\|_2 \geq C\|\mathbf{a}\|_\infty\big] \geq C'\widetilde{\phi}/n.$$

The following lemma is essential to show that deep ReLU network can provide significantly large gradient at the initialization.

Proof of Lemma A.2 For any given $j \in \{1, \ldots, m_L\}$ and $\widehat{\mathbf{a}}$ with $\|\widehat{\mathbf{a}}\|_\infty = 1$. By Lemma B.1, we know that for any $i \neq j$ and $y_i \neq y_j$, $\|\bar{\mathbf{x}}_{L-1,i} - \bar{\mathbf{x}}_{L-1,j}\|_2 \geq \widetilde{\phi}$, where $\bar{\mathbf{x}}_{L-1,i} = \mathbf{x}_{L-1,i}/\|\mathbf{x}_{L-1,i}\|_2$ and $\bar{\mathbf{x}}_{L-1,j} = \mathbf{x}_{L-1,j}/\|\mathbf{x}_{L-1,i}\|_2$. Then by Lemma B.2, we have

$$\mathbb{P}\left[\left\|\frac{1}{n}\sum_{i=1}^n \widehat{a}_i y_i \sigma'(\langle \mathbf{w}_{L,j}, \mathbf{x}_{L-1,i}\rangle)\mathbf{x}_{L-1,i}\right\|_2 \geq \frac{C_1}{n}\right] \geq \frac{C_2\phi}{n},$$

where $C_1, C_2 > 0$ are absolute constants. Let $S_{\infty,+}^{n-1} = \{\mathbf{a} \in \mathbb{R}_+^n : \|\mathbf{a}\|_\infty = 1\}$, and $\mathcal{N} = \mathcal{N}[S_{\infty,+}^{n-1}, C_1/(4n)]$ be a $C_1/(4n)$-net covering $S_{\infty,+}^{n-1}$ in ℓ_∞ norm. Then we have

$$|\mathcal{N}| \leq (4n/C_1)^n.$$

For $j = 1, \ldots, m_L$, define

$$Z_j = \mathbb{1}\left[\left\|\frac{1}{n}\sum_{i=1}^n \widehat{a}_i y_i \sigma'(\langle \mathbf{w}_{L,j}, \mathbf{x}_{L-1,i}\rangle)\mathbf{x}_{L-1,i}\right\|_2 \geq \frac{C_1}{n}\right].$$

Let $p_\phi = C_2\phi/n$. Then by Bernstein inequality and union bound, with probability at least $1 - \exp[-C_3 m_L p_\phi + n\log(4n/C_1)] \geq 1 - \exp(C_4 m_L \phi/n)$, we have

$$\frac{1}{m_L}\sum_{j=1}^{m_L} Z_j \geq p_\phi/2, \tag{B.1}$$

where C_3, C_4 are absolute constants. For any $\mathbf{a} \in S_{\infty,+}^{n-1}$, there exists $\widehat{\mathbf{a}} \in \mathcal{N}$ such that

$$\|\mathbf{a} - \widehat{\mathbf{a}}\|_\infty \leq C_1/(4n).$$

Therefore, we have

$$\begin{aligned}&\left|\left\|\frac{1}{n}\sum_{i=1}^{n}a_i y_i\sigma'(\langle \mathbf{w}_{L,j},\mathbf{x}_{L-1,i}\rangle)\mathbf{x}_{L-1,i}\right\|_2-\left\|\frac{1}{n}\sum_{i=1}^{n}\widehat{a}_i y_i\sigma'(\langle \mathbf{w}_{L,j},\mathbf{x}_{L-1,i}\rangle)\mathbf{x}_{L-1,i}\right\|_2\right|\\&\quad\le\left\|\frac{1}{n}\sum_{i=1}^{n}a_i y_i\sigma'(\langle \mathbf{w}_{L,j},\mathbf{x}_{L-1,i}\rangle)\mathbf{x}_{L-1,i}-\frac{1}{n}\sum_{i=1}^{n}\widehat{a}_i y_i\sigma'(\langle \mathbf{w}_{L,j},\mathbf{x}_{L-1,i}\rangle)\mathbf{x}_{L-1,i}\right\|_2\\&\quad\le\frac{2}{n}\sum_{i=1}^{n}|a_i-\widehat{a}_i|\le\frac{C_1}{2n}.\end{aligned}\tag{B.2}$$

By (B.1) and (B.2), it is clear that with probability at least $1-\exp(C_4 m_L\phi/n)$, for any $\mathbf{a}\in S_{\infty,+}^{n-1}$, there exist at least $m_L p_\phi/2$ nodes on layer L that satisfy

$$\left\|\frac{1}{n}\sum_{i=1}^{n}a_i y_i\sigma'(\langle \mathbf{w}_{L,j},\mathbf{x}_{L-1,i}\rangle)\mathbf{x}_{L-1,i}\right\|_2\ge\frac{C_1}{2n}.$$

This completes the proof. □

C Proofs of lemmas in Appendix B

C.1 Proof of Lemma B.1

The following lemma is necessary for proving Lemma B.1.

Lemma C.1 [Lemma A.3 in Zou et al. (2018)] *For $\theta>0$, let Z_1, Z_2 be two jointly Gaussian random variables with $\mathbb{E}(Z_1)=\mathbb{E}(Z_2)=0$, $\mathbb{E}(Z_1^2)=\mathbb{E}(Z_2^2)=1$ and $\mathbb{E}(Z_1Z_2)\le 1-\theta^2/2$. If $\theta\le\kappa$ for some small enough absolute constant κ, then*

$$\mathbb{E}[\sigma(Z_1)\sigma(Z_2)]\le\frac{1}{2}-\frac{1}{4}\theta^2+C\theta^3,$$

where C is an absolute constant.

Proof of Lemma B.1 We first consider any fixed $l\ge 1$. Suppose that $\|\overline{\mathbf{x}}_{l-1,i}-\overline{\mathbf{x}}_{l-1,i'}\|_2\ge[1-(2L)^{-1}\log(2)]^{l-1}\phi$. If we can show that under this condition, with high probability

$$\|\overline{\mathbf{x}}_{l,i}-\overline{\mathbf{x}}_{l,i'}\|_2\ge[1-(2L)^{-1}\log(2)]^{l}\phi,$$

then the result of the lemma follows by union bound and induction. Denote

$$\phi_{l-1}=[1-(2L)^{-1}\log(2)]^{l-1}\phi.$$

Then by assumption we have $\|\overline{\mathbf{x}}_{l-1,i}-\overline{\mathbf{x}}_{l-1,i'}\|_2^2\ge\phi_{l-1}^2$. Therefore $\overline{\mathbf{x}}_{l-1,i}^\top\overline{\mathbf{x}}_{l-1,i'}\le 1-\phi_{l-1}^2/2$. It follows by direct calculation that

$$\begin{aligned}\mathbb{E}\big(\|\mathbf{x}_{l,i}-\mathbf{x}_{l,i'}\|_2^2\big|\mathbf{x}_{l-1,i},\mathbf{x}_{l-1,i'}\big)&=\mathbb{E}\big(\|\mathbf{x}_{l,i}\|_2^2+\|\mathbf{x}_{l,i'}\|_2^2\big|\mathbf{x}_{l-1,i},\mathbf{x}_{l-1,i'}\big)-2\mathbb{E}\big(\mathbf{x}_{l,i}^\top\mathbf{x}_{l,i'}\big|\mathbf{x}_{l-1,i},\mathbf{x}_{l-1,i'}\big)\\&=(\|\mathbf{x}_{l-1,i}\|_2^2+\|\mathbf{x}_{l-1,i'}\|_2^2)-2\mathbb{E}\big(\mathbf{x}_{l,i}^\top\mathbf{x}_{l,i'}\big|\mathbf{x}_{l-1,i},\mathbf{x}_{l-1,i'}\big).\end{aligned}$$

By Lemma C.1 and the assumption that $\phi_{l-1} \leq \phi \leq \kappa$, we have

$$\begin{aligned}
&\mathbb{E}\big(\mathbf{x}_{l,i}^{\top}\mathbf{x}_{l,i'}\big|\mathbf{x}_{l-1,i},\mathbf{x}_{l-1,i'}\big)\\
&\quad= \mathbb{E}\left[\sum_{j=1}^{m_l}\sigma(\mathbf{w}_{l,j}^{\top}\mathbf{x}_{l-1,i})\sigma(\mathbf{w}_{l,j}^{\top}\mathbf{x}_{l-1,i'})\middle|\mathbf{x}_{l-1,i},\mathbf{x}_{l-1,i'}\right]\\
&\quad= \|\mathbf{x}_{l-1,i}\|_2\|\mathbf{x}_{l-1,i'}\|_2\cdot\mathbb{E}\left[\sum_{j=1}^{m_l}\sigma(\mathbf{w}_{l,j}^{\top}\overline{\mathbf{x}}_{l-1,i})\sigma(\mathbf{w}_{l,j}^{\top}\overline{\mathbf{x}}_{l-1,i'})\middle|\mathbf{x}_{l-1,i},\mathbf{x}_{l-1,i'}\right]\\
&\quad\leq \frac{2}{m}\|\mathbf{x}_{l-1,i}\|_2\|\mathbf{x}_{l-1,i'}\|_2\cdot m\cdot\left(\frac{1}{2}-\frac{1}{4}\phi_{l-1}^2+C\phi_{l-1}^3\right)\\
&\quad= \|\mathbf{x}_{l-1,i}\|_2\|\mathbf{x}_{l-1,i'}\|_2\cdot\left(1-\frac{1}{2}\phi_{l-1}^2+2C\phi_{l-1}^3\right).
\end{aligned}$$

Therefore,

$$\mathbb{E}\big(\|\mathbf{x}_{l,i}-\mathbf{x}_{l,i'}\|_2^2\big|\mathbf{x}_{l-1,i},\mathbf{x}_{l-1,i'}\big) \geq (\|\mathbf{x}_{l-1,i}\|_2-\|\mathbf{x}_{l-1,i'}\|_2)^2+\|\mathbf{x}_{l-1,i}\|_2\|\mathbf{x}_{l-1,i'}\|_2(\phi_{l-1}^2-4C\phi_{l-1}^3). \tag{C.1}$$

Condition on $\mathbf{x}_{l-1,i}$ and $\mathbf{x}_{l-1,i'}$, by Lemma 5.14 in Vershynin (2010) we have

$$\begin{aligned}
\big\|[\sigma(\mathbf{w}_{l,j}^{\top}\mathbf{x}_{l-1,i})-\sigma(\mathbf{w}_{l,j}^{\top}\mathbf{x}_{l-1,i'})]^2\big\|_{\psi_1} &\leq 2\big[\big\|[\sigma(\mathbf{w}_{l,j}^{\top}\mathbf{x}_{l-1,i})\big\|_{\psi_2}+\big\|\sigma(\mathbf{w}_{l,j}^{\top}\mathbf{x}_{l-1,i'})\big\|_{\psi_2}\big]^2\\
&\leq C_1(\|\mathbf{x}_{l-1,i}\|_2+\|\mathbf{x}_{l-1,i'}\|_2)^2/m_l,
\end{aligned}$$

where C_1 is an absolute constant. Therefore if $m_l \geq C_2^2\log(4n^2L/\delta)$, by Bernstein inequality and union bound, with probability at least $1-\delta/(4n^2L)$ we have

$$\big|\|\mathbf{x}_{l,i}-\mathbf{x}_{l,i'}\|_2^2-\mathbb{E}\big(\|\mathbf{x}_{l,i}-\mathbf{x}_{l,i'}\|_2^2\big|\mathbf{x}_{l-1,i},\mathbf{x}_{l-1,i'}\big)\big| \leq C_2(\|\mathbf{x}_{l-1,i}\|_2+\|\mathbf{x}_{l-1,i'}\|_2)^2\cdot\sqrt{\frac{\log(8n^2L/\delta)}{m_l}},$$

where C_2 is an absolute constant. Therefore with probability at least $1-\delta/(4n^2L)$ we have

$$\begin{aligned}
\|\mathbf{x}_{l,i}-\mathbf{x}_{l,i'}\|_2^2 \geq{}& (\|\mathbf{x}_{l-1,i}\|_2-\|\mathbf{x}_{l-1,i'}\|_2)^2+\|\mathbf{x}_{l-1,i}\|_2\|\mathbf{x}_{l-1,i'}\|_2(\phi_{l-1}^2-4C\phi_{l-1}^3)\\
&-C_2(\|\mathbf{x}_{l-1,i}\|_2+\|\mathbf{x}_{l-1,i'}\|_2)^2\cdot\sqrt{\frac{\log(8n^2L/\delta)}{m_l}}.
\end{aligned}$$

By union bound and Lemma A.1, if $m_r \geq C_3L^4\phi_l^{-4}\log(4n^2L/\delta)$, $r=1,\ldots,l$ for some large enough absolute constant C_3 and $\phi \leq \kappa L^{-1}$ for some small enough absolute constant κ, then with probability at least $1-\delta/(2n^2L)$ we have

$$\|\mathbf{x}_{l,i}-\mathbf{x}_{l,i'}\|_2^2 \geq [1-(4L)^{-1}\log(2)]\phi_{l-1}^2 \geq [1-(4L)^{-1}\log(2)]^2\phi_{l-1}^2.$$

Moreover, by Lemma A.1, with probability at least $1-\delta/(2n^2L)$ we have

$$\begin{aligned}
\big|\|\overline{\mathbf{x}}_{l,i}-\overline{\mathbf{x}}_{l,i'}\|_2-\|\mathbf{x}_{l,i}-\mathbf{x}_{l,i'}\|_2\big| &\leq \|\overline{\mathbf{x}}_{l,i}-\mathbf{x}_{l,i}\|_2+\|\overline{\mathbf{x}}_{l,i'}-\mathbf{x}_{l,i'}\|_2\\
&= \big|1-\|\mathbf{x}_{l,i}\|_2\big|+\big|1-\|\mathbf{x}_{l,i'}\|_2\big|\\
&\leq (4L)^{-1}\log(2)\cdot\phi_{l-1}^2,
\end{aligned}$$

and therefore with probability at least $1-\delta/(n^2L)$, we have

$$\|\overline{\mathbf{x}}_{l,i}-\overline{\mathbf{x}}_{l,i'}\|_2 \geq [1-(2L)^{-1}\log(2)]\phi_{l-1} = [1-(2L)^{-1}\log(2)]^l\phi.$$

Applying union bound and induction over $l = 1, \dots, L$ completes the proof. □

C.2 Proof of Lemma B.2

***Proof of Lemma* B.2** Without loss of generality, assume that $a_1 = \|\mathbf{a}\|_\infty$. Since $\|\mathbf{z}_1\|_2 = 1$, we can construct an orthonormal matrix $\mathbf{Q} = [\mathbf{z}_1, \mathbf{Q}'] \in \mathbb{R}^{d \times d}$. Let $\mathbf{u} = \mathbf{Q}^\top \mathbf{w} \sim N(\mathbf{0}, \mathbf{I})$ be a standard Gaussian random vector. Then we have

$$\mathbf{w} = \mathbf{Q}\mathbf{u} = u_1 \mathbf{z}_1 + \mathbf{Q}'\mathbf{u}',$$

where $\mathbf{u}' := (u_2, \dots, u_d)^\top$ is independent of u_1. We define the following two events based on a parameter $\gamma \in (0, 1]$:

$$\mathcal{E}_1(\gamma) = \big\{|u_1| \le \gamma\big\},\ \mathcal{E}_2(\gamma) = \big\{|\langle \mathbf{Q}'\mathbf{u}', \mathbf{z}_i \rangle| \ge \gamma \text{ for all } \mathbf{z}_i \text{ such that } \|\mathbf{z}_i - \mathbf{z}_1\|_2 \ge \widetilde{\phi}\big\}.$$

Let $\mathcal{E}(\gamma) = \mathcal{E}_1(\gamma) \cap \mathcal{E}_2(\gamma)$. We first give lower bound for $\mathbb{P}(\mathcal{E}) = \mathbb{P}(\mathcal{E}_1)\mathbb{P}(\mathcal{E}_2)$. Since u_1 is a standard Gaussian random variable, we have

$$\mathbb{P}(\mathcal{E}_1) = \frac{1}{\sqrt{2\pi}} \int_{-\gamma}^{\gamma} \exp\bigg(-\frac{1}{2}x^2\bigg) \mathrm{d}x \ge \sqrt{\frac{2}{\pi e}}\gamma.$$

Moreover, by definition, for any $i = 1, \dots, n$ we have

$$\langle \mathbf{Q}'\mathbf{u}', \mathbf{z}_i \rangle \sim N\big[0, 1 - (\mathbf{z}_1^\top \mathbf{z}_i)^2\big].$$

Let $\mathcal{I} = \{i : \|\mathbf{z}_i - \mathbf{z}_1\|_2 \ge \widetilde{\phi}\}$. By the assumption that $\widetilde{\phi} \le \widetilde{\mu}/2$, for any $i \in \mathcal{I}$, we have

$$-1 + \widetilde{\phi}^2/2 \le -(1 - \widetilde{\mu}^2) + \widetilde{\mu}^2 \le \langle \mathbf{z}_i, \mathbf{z}_1 \rangle \le 1 - \widetilde{\phi}^2/2,$$

and if $\widetilde{\phi}^2 \le 2$, then

$$1 - (\mathbf{z}_1^\top \mathbf{z}_i)^2 \ge \widetilde{\phi}^2 - \widetilde{\phi}^4/4 \ge \widetilde{\phi}^2/2.$$

Therefore for any $i \in \mathcal{I}$,

$$\mathbb{P}[|\langle \mathbf{Q}'\mathbf{u}', \mathbf{z}_i \rangle| < \gamma] = \frac{1}{\sqrt{2\pi}} \int_{-[1-(\mathbf{z}_1^\top \mathbf{z}_i)^2]^{-1/2}\gamma}^{[1-(\mathbf{z}_1^\top \mathbf{z}_i)^2]^{-1/2}\gamma} \exp\bigg(-\frac{1}{2}x^2\bigg) \mathrm{d}x \le \sqrt{\frac{2}{\pi}} \frac{\gamma}{[1 - (\mathbf{z}_1^\top \mathbf{z}_i)^2]^{1/2}} \le \frac{2}{\sqrt{\pi}}\gamma\widetilde{\phi}^{-1}.$$

By union bound over $\mathcal{I}$, we have

$$\mathbb{P}(\mathcal{E}_2) = \mathbb{P}[|\langle \mathbf{Q}'\mathbf{u}', \mathbf{z}_i \rangle| \ge \gamma, i \in \mathcal{I}] \ge 1 - \frac{2}{\sqrt{\pi}} n\gamma\widetilde{\phi}^{-1}.$$

Therefore we have

$$\mathbb{P}(\mathcal{E}) \ge \sqrt{\frac{2}{\pi e}}\gamma \cdot \bigg(1 - \frac{2}{\sqrt{\pi}} n\gamma\widetilde{\phi}^{-1}\bigg).$$

Setting $\gamma = \sqrt{\pi}\widetilde{\phi}/(4n)$, we obtain $\mathbb{P}(\mathcal{E}) \ge \widetilde{\phi}/(\sqrt{32e}n)$. Now let $\mathcal{I}' = [n] \setminus (\mathcal{I} \cup \{1\})$. Then conditioning on event $\mathcal{E}$, we have

$$\begin{aligned}
\mathbf{h}(\mathbf{w}) &= \sum_{i=1}^n a_i y_i \sigma'(\langle \mathbf{w}, \mathbf{z}_i \rangle)\mathbf{z}_i \\
&= a_1 y_1 \sigma'(u_1)\mathbf{z}_1 + \sum_{i \in \mathcal{I}} a_i y_i \sigma'\big(u_1 \langle \mathbf{z}_1, \mathbf{z}_i \rangle + \langle \mathbf{Q}'\mathbf{u}', \mathbf{z}_i \rangle\big)\mathbf{z}_i
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{i\in\mathcal{I}'} a_i y_i \sigma'\big(u_1\langle \mathbf{z}_1, \mathbf{z}_i\rangle + \langle \mathbf{Q}'\mathbf{u}, \mathbf{z}_i\rangle\big)\mathbf{z}_i\\
&= a_1 y_1 \sigma'(u_1)\mathbf{z}_1 + \sum_{i\in\mathcal{I}} a_i y_i \sigma'\big(\langle \mathbf{Q}'\mathbf{u}', \mathbf{z}_i\rangle\big)\mathbf{z}_i + \sum_{i\in\mathcal{I}'} a_i y_i \sigma'\big(u_1\langle \mathbf{z}_1, \mathbf{z}_i\rangle + \langle \mathbf{Q}'\mathbf{u}', \mathbf{z}_i\rangle\big)\mathbf{z}_i,
\end{aligned} \tag{C.2}
$$

where the last equality follows from the fact that conditioning on event $\mathcal{E}$, for all $i \in \mathcal{I}$, it holds that $|\langle \mathbf{Q}'\mathbf{u}', \mathbf{z}_i\rangle| \ge |u_1| \ge |u_1\langle \mathbf{z}_1, \mathbf{z}_i\rangle|$. We then consider two cases: $u_1 > 0$ and $u_1 < 0$, which occur equally likely conditioning on $\mathcal{E}$. Therefore we have

$$
\mathbb{P}\bigg[\|\mathbf{h}(\mathbf{w})\|_2 \ge \inf_{u_1^{(1)}>0,u_1^{(2)}<0} \max\Big\{\big\|\mathbf{h}(u_1^{(1)}\mathbf{z}_1 + \mathbf{Q}'\mathbf{u}')\big\|_2, \big\|\mathbf{h}(u_1^{(2)}\mathbf{z}_1 + \mathbf{Q}'\mathbf{u}')\big\|_2\Big\}\bigg|\mathcal{E}\bigg] \ge 1/2.
$$

By the inequality $\max\{\|\mathbf{a}\|_2, \|\mathbf{b}\|_2\} \ge \|\mathbf{a}-\mathbf{b}\|_2/2$, we have

$$
\mathbb{P}\bigg[\|\mathbf{h}(\mathbf{w})\|_2 \ge \inf_{u_1^{(1)}>0,u_1^{(2)}<0} \big\|\mathbf{h}(u_1^{(1)}\mathbf{z}_1 + \mathbf{Q}'\mathbf{u}') - \mathbf{h}(u_1^{(2)}\mathbf{z}_1 + \mathbf{Q}'\mathbf{u}')\big\|_2/2\bigg|\mathcal{E}\bigg] \ge 1/2. \tag{C.3}
$$

For any $u_1^{(1)} > 0$ and $u_1^{(2)} < 0$, denote $\mathbf{w}_1 = u_1^{(1)}\mathbf{z}_1 + \mathbf{Q}'\mathbf{u}'$, $\mathbf{w}_2 = u_1^{(2)}\mathbf{z}_1 + \mathbf{Q}'\mathbf{u}'$. We now proceed to give lower bound for $\|\mathbf{h}(\mathbf{w}_1) - \mathbf{h}(\mathbf{w}_2)\|_2$. By (C.2), we have

$$
\mathbf{h}(\mathbf{w}_1) - \mathbf{h}(\mathbf{w}_2) = a_1 y_1 \mathbf{z}_1 + \sum_{i\in\mathcal{I}'} a_i' y_i \mathbf{z}_i, \tag{C.4}
$$

where

$$
a_i' = a_i\big[\sigma'\big(u_1^{(1)}\langle \mathbf{z}_1, \mathbf{z}_i\rangle + \langle \mathbf{Q}'\mathbf{u}', \mathbf{z}_i\rangle\big) - \sigma'\big(u_1^{(2)}\langle \mathbf{z}_1, \mathbf{z}_i\rangle + \langle \mathbf{Q}'\mathbf{u}', \mathbf{z}_i\rangle\big)\big].
$$

Note that for all $i \in \mathcal{I}'$, we have $y_i = y_1$ and $\langle \mathbf{z}_1, \mathbf{z}_i\rangle \ge 1 - \widetilde{\phi}^2/2 \ge 0$. Therefore, since $u_1^{(1)} > 0 > u_1^{(2)}$, we have

$$
\sigma'\big(u_1^{(1)}\langle \mathbf{z}_1, \mathbf{z}_i\rangle + \langle \mathbf{Q}'\mathbf{u}', \mathbf{z}_i\rangle\big) - \sigma'\big(u_1^{(2)}\langle \mathbf{z}_1, \mathbf{z}_i\rangle + \langle \mathbf{Q}'\mathbf{u}', \mathbf{z}_i\rangle\big) \ge 0.
$$

Therefore $a_i' \ge 0$ for all $i \in \mathcal{I}'$ and

$$
\mathbf{h}(\mathbf{w}_1) - \mathbf{h}(\mathbf{w}_2) = a_1 y_1 \mathbf{z}_1 + \sum_{i\in\mathcal{I}'} a_i' y_1 \mathbf{z}_i = y_1\bigg(a_1\mathbf{z}_1 + \sum_{i\in\mathcal{I}'} a_i'\mathbf{z}_i\bigg),
$$

We have shown that $\langle \mathbf{z}_i, \mathbf{z}_1\rangle \ge 0$ for all $i \in \mathcal{I}'$. Therefore we have

$$
\|\mathbf{h}(\mathbf{w}_1) - \mathbf{h}(\mathbf{w}_2)\|_2 \ge \bigg\|y_1\bigg(a_1\mathbf{z}_1 + \sum_{i\in\mathcal{I}'} a_i'\mathbf{z}_i\bigg)\bigg\|_2 \ge \bigg\langle a_1\mathbf{z}_1 + \sum_{i\in\mathcal{I}'} a_i'\mathbf{z}_i, \mathbf{z}_1\bigg\rangle \ge a_1.
$$

Since the inequality above holds for any $u_1^{(1)} > 0$ and $u_1^{(2)} < 0$, taking infimum gives

$$
\inf_{u_1^{(1)}>0,u_1^{(2)}<0} \|\mathbf{h}(\mathbf{w}_1) - \mathbf{h}(\mathbf{w}_2)\|_2 \ge a_1. \tag{C.5}
$$

Plugging (C.5) back to (C.3), we obtain

$$
\mathbb{P}\big[\|\mathbf{h}(\mathbf{w})\|_2 \ge a_1/2\big|\mathcal{E}\big] \ge 1/2,
$$

Since $a_1 = \|\mathbf{a}\|_\infty$ and $\mathbb{P}(\mathcal{E}) \geq \widetilde{\phi}/(\sqrt{32}en)$, we have

$$\mathbb{P}\big[\|\mathbf{h}(\mathbf{w})\|_2 \geq C\|\mathbf{a}\|_\infty\big] \geq C'\widetilde{\phi}/n,$$

where C and C' are absolute constants. This completes the proof. □

References

Allen-Zhu, Z., Li, Y., & Song, Z. (2018a). A convergence theory for deep learning via over-parameterization. arXiv preprint arXiv:1811.03962

Allen-Zhu, Z., Li, Y., & Song, Z. (2018b) On the convergence rate of training recurrent neural networks. arXiv preprint arXiv:1810.12065

Arora, S., Cohen, N., Golowich, N., & Hu, W. (2018a). A convergence analysis of gradient descent for deep linear neural networks. arXiv preprint arXiv:1810.02281

Arora, S., Cohen, N., & Hazan, E. (2018b). On the optimization of deep networks: Implicit acceleration by overparameterization. arXiv preprint arXiv:1802.06509

Bartlett, P., Helmbold, D., & Long, P. (2018). Gradient descent with identity initialization efficiently learns positive definite linear transformations. In *International conference on machine learning*, pp. 520–529.

Brutzkus, A., & Globerson, A. (2017). Globally optimal gradient descent for a convnet with gaussian inputs. arXiv preprint arXiv:1702.07966

Du, S. S., Lee, J. D., & Tian, Y. (2017). When is a convolutional filter easy to learn? arXiv preprint arXiv:1709.06129

Du, S. S., Lee, J. D., Li, H., Wang, L., & Zhai, X. (2018a). Gradient descent finds global minima of deep neural networks. arXiv preprint arXiv:1811.03804

Du, S. S., Zhai, X., Poczos, B., & Singh, A. (2018b). Gradient descent provably optimizes over-parameterized neural networks. arXiv preprint arXiv:1810.02054

Gunasekar, S., Lee, J., Soudry, D., & Srebro, N. (2018). Implicit bias of gradient descent on linear convolutional networks. arXiv preprint arXiv:1806.00468

Hanin, B. (2017). Universal function approximation by deep neural nets with bounded width and ReLU activations. arXiv preprint arXiv:1708.02691

Hanin, B., Sellke, M. (2017). Approximating continuous functions by ReLU nets of minimal width. arXiv preprint arXiv:1710.11278

Hardt, M., & Ma, T. (2016). Identity matters in deep learning. arXiv preprint arXiv:1611.04231

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp 1026–1034.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 770–778.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Ar, Mohamed, Jaitly, N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, *29*(6), 82–97.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, *4*(2), 251–257.

Kawaguchi, K. (2016). Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, pp 586–594.

Krizhevsky, A. (2009). *Learning multiple layers of features from tiny images*. Tech. rep., Citeseer.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp 1097–1105.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Li, Y., & Liang, Y. (2018). Learning overparameterized neural networks via stochastic gradient descent on structured data. arXiv preprint arXiv:1808.01204

Li, Y., & Yuan, Y. (2017). Convergence analysis of two-layer neural networks with ReLU activation. arXiv preprint arXiv:1705.09886

Liang, S., & Srikant, R. (2016). Why deep neural networks for function approximation? arXiv preprint arXiv:1610.04161

Lin, H., & Jegelka, S. (2018). Resnet with one-neuron hidden layers is a universal approximator. In *Advances in neural information processing systems*, pp. 6172–6181.

Lu, Z., Pu, H., Wang, F., Hu, Z., & Wang, L. (2017). The expressive power of neural networks: A view from the width. arXiv preprint arXiv:1709.02540

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., et al. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489.

Telgarsky, M. (2015). Representation benefits of deep feedforward networks. arXiv preprint arXiv:1509.08101

Telgarsky, M. (2016). Benefits of depth in neural networks. arXiv preprint arXiv:1602.04485

Tian, Y. (2017). An analytical formula of population gradient for two-layered ReLU network and its applications in convergence and critical point analysis. arXiv preprint arXiv:1703.00560

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. arXiv preprint arXiv:1011.3027

Yarotsky, D. (2017). Error bounds for approximations with deep ReLU networks. *Neural Networks*, *94*, 103–114.

Yarotsky, D. (2018). Optimal approximation of continuous functions by very deep ReLU networks. arXiv preprint arXiv:1802.03620

Zhang, X., Yu, Y., Wang, L., & Gu, Q. (2018). Learning one-hidden-layer ReLU networks via gradient descent. arXiv preprint arXiv:1806.07808

Zhou, D. X. (2019). Universality of deep convolutional neural networks. In *Applied and computational harmonic analysis*.

Zou, D., Cao, Y., Zhou. D., & Gu, Q. (2018). Stochastic gradient descent optimizes over-parameterized deep relu networks. arXiv preprint arXiv:1811.08888

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.