An Improved Analysis of Training Over-parameterized Deep Neural Networks

Difan Zou
Department of Computer Science
University of California, Los Angeles
Los Angeles, CA 90095
`knowzou@cs.ucla.edu`

Quanquan Gu
Department of Computer Science
University of California, Los Angeles
Los Angeles, CA 90095
`qgu@cs.ucla.edu`

Abstract

A recent line of research has shown that gradient-based algorithms with random initialization can converge to the global minima of the training loss for over-parameterized (i.e., sufficiently wide) deep neural networks. However, the condition on the width of the neural network to ensure the global convergence is very stringent, which is often a high-degree polynomial in the training sample size n (e.g., $O(n^{24})$). In this paper, we provide an improved analysis of the global convergence of (stochastic) gradient descent for training deep neural networks, which only requires a milder over-parameterization condition than previous work in terms of the training sample size and other problem-dependent parameters. The main technical contributions of our analysis include (a) a tighter gradient lower bound that leads to a faster convergence of the algorithm, and (b) a sharper characterization of the trajectory length of the algorithm. By specializing our result to two-layer (i.e., one-hidden-layer) neural networks, it also provides a milder over-parameterization condition than the best-known result in prior work.

1 Introduction

Recent study [20] has revealed that deep neural networks trained by gradient-based algorithms can fit training data with random labels and achieve zero training error. Since the loss landscape of training deep neural network is highly nonconvex or even nonsmooth, conventional optimization theory cannot explain why gradient descent (GD) and stochastic gradient descent (SGD) can find the global minimum of the loss function (i.e., achieving zero training error). To better understand the training of neural networks, there is a line of research [18, 5, 10, 16, 23, 8, 22, 12] studying two-layer (i.e., one-hidden-layer) neural networks, where it assumes there exists a teacher network (i.e., an underlying ground-truth network) generating the output given the input, and casts neural network learning as weight matrix recovery for the teacher network. However, these studies not only make strong assumptions on the training data (existence of ground-truth network with the same architecture as the learned network), but also need special initialization methods that are very different from the commonly used initialization method [13] in practice. Li and Liang [15], Du et al. [11] advanced this line of research by proving that under much milder assumptions on the training data, (stochastic) gradient descent can attain a global convergence for training over-parameterized (i.e.,sufficiently wide) two-layer ReLU network with widely used random initialization method [13]. More recently, Allen-Zhu et al. [2], Du et al. [9], Zou et al. [24] generalized the global convergence results from two-layer networks to deep neural networks. However, there is a huge gap between the theory and practice since all these work Li and Liang [15], Du et al. [11], Allen-Zhu et al. [2], Du et al. [9], Zou et al. [24] require unrealistic over-parameterization conditions on the width of neural networks, especially for deep networks. In specific, in order to establish the global convergence for training two-layer ReLU networks, Du et al. [11] requires the network width, i.e., number of hidden

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

nodes, to be at least $\Omega(n^6/\lambda_0^4)$, where n is the training sample size and λ_0 is the smallest eigenvalue of the so-called Gram matrix defined in Du et al. [11], which is essentially the neural tangent kernel [14, 7] on the training data. Under the same assumption on the training data, Wu et al. [19] improved the iteration complexity of GD in Du et al. [11] from $O\big(n^2\log(1/\epsilon)/\lambda_0^2\big)$ to $O\big(n\log(1/\epsilon)/\lambda_0\big)$ and Oymak and Soltanolkotabi [17] improved the over-parameterization condition to $\Omega(n\|\mathbf{X}\|_2^6/\lambda_0^4)$, where ϵ is the target error and $\mathbf{X}\in\mathbb{R}^{n\times d}$ is the input data matrix. For deep ReLU networks, the best known result was established in Allen-Zhu et al. [2], which requires the network width to be at least $\widetilde{\Omega}(kn^{24}L^{12}\phi^{-8})$[1] to ensure the global convergence of GD and SGD, where L is the number of hidden layers, ϕ is the minimum data separation distance and k is the output dimension.

This paper continues the line of research, and improves the over-parameterization condition and the global convergence rate of (stochastic) gradient descent for training deep neural networks. In specific, under the same setting as in Allen-Zhu et al. [2], we prove faster global convergence rates for both GD and SGD under a significantly milder condition on the neural network width. Furthermore, when specializing our result to two-layer ReLU networks, it also outperforms the best-known result proved in Oymak and Soltanolkotabi [17]. The improvement in our result is due to the following two innovative proof techniques: (a) a tighter gradient lower bound, which leads to a faster rate of convergence for GD/SGD; and (b) a sharper characterization of the trajectory length for GD/SGD until convergence.

We highlight our main contributions as follows:

- We show that, with Gaussian random initialization [13] on each layer, when the number of hidden nodes per layer is $\widetilde{\Omega}\big(kn^8L^{12}\phi^{-4}\big)$, GD can achieve ϵ training loss within $\widetilde{O}\big(n^2L^2\log(1/\epsilon)\phi^{-1}\big)$ iterations, where L is the number of hidden layers, ϕ is the minimum data separation distance, n is the number of training examples, and k is the output dimension. Compared with the state-of-the-art result [2], our over-parameterization condition is milder by a factor of $\widetilde{\Omega}(n^{16}\phi^{-4})$, and our iteration complexity is better by a factor of $\widetilde{O}(n^4\phi^{-1})$.
- We also prove a similar convergence result for SGD. We show that with Gaussian random initialization [13] on each layer, when the number of hidden nodes per layer is $\widetilde{\Omega}\big(kn^{17}L^{12}B^{-4}\phi^{-8}\big)$, SGD can achieve ϵ expected training loss within $\widetilde{O}\big(n^5\log(1/\epsilon)B^{-1}\phi^{-2}\big)$ iterations, where B is the minibatch size of SGD. Compared with the corresponding results in Allen-Zhu et al. [2], our results are strictly better by a factor of $\widetilde{\Omega}(n^7B^5)$ and $\widetilde{O}(n^2)$ respectively regarding over-parameterization condition and iteration complexity.
- When specializing our results of training deep ReLU networks with GD to two-layer ReLU networks, it also outperforms the corresponding results [11, 19, 17]. In addition, for training two-layer ReLU networks with SGD, we are able to show much better result than training deep ReLU networks with SGD.

For the ease of comparison, we summarize the best-known results [11, 2, 9, 19, 17] of training overparameterized neural networks with GD and compare with them in terms of over-parameterization condition and iteration complexity in Table 1. We will show in Section 3 that, under the assumption that all training data points have unit ℓ_2 norm, which is the common assumption made in all these work [11, 2, 9, 19, 17], $\lambda_0>0$ is equivalent to the fact that all training data are separated by some distance ϕ, and we have $\lambda_0=O(n^{-2}\phi)$ [17]. Substituting $\lambda_0=\Omega(n^{-2}\phi)$ into Table 1, it is evident that our result outperforms all the other results under the same assumptions.

Notation For scalars, vectors and matrices, we use lower case, lower case bold face, and upper case bold face letters to denote them respectively. For a positive integer, we denote by $[k]$ the set $\{1,\dots,k\}$. For a vector $\mathbf{x}=(x_1,\dots,x_d)^\top$ and a positive integer p, we denote by $\|\mathbf{x}\|_p=\big(\sum_{i=1}^d|x_i|^p\big)^{1/p}$ the ℓ_p norm of $\mathbf{x}$. In addition, we denote by $\|\mathbf{x}\|_\infty=\max_{i=1,\dots,d}|x_i|$ the ℓ_∞ norm of $\mathbf{x}$, and $\|\mathbf{x}\|_0=|\{x_i:x_i\neq 0, i=1,\dots,d\}|$ the ℓ_0 norm of $\mathbf{x}$. For a matrix $\mathbf{A}\in\mathbb{R}^{m\times n}$, we denote by $\|\mathbf{A}\|_F$ the Frobenius norm of $\mathbf{A}$, $\|\mathbf{A}\|_2$ the spectral norm (maximum singular value), $\lambda_{\min}(\mathbf{A})$ the smallest singular value, $\|\mathbf{A}\|_0$ the number of nonzero entries, and $\|\mathbf{A}\|_{2,\infty}$ the maximum ℓ_2 norm over all row vectors, i.e., $\|\mathbf{A}\|_{2,\infty}=\max_{i=1,\dots,m}\|\mathbf{A}_{i*}\|_2$. For a collection of matrices $\mathbf{W}=\{\mathbf{W}_1,\dots,\mathbf{W}_L\}$, we denote $\|\mathbf{W}\|_F=\sqrt{\sum_{l=1}^L\|\mathbf{W}_l\|_F^2}$, $\|\mathbf{W}\|_2=\max_{l\in[L]}\|\mathbf{W}_l\|_2$ and

---

[1] Here $\widetilde{\Omega}(\cdot)$ hides constants and the logarithmic dependencies on problem dependent parameters except ϵ.

Table 1: Over-parameterization conditions and iteration complexities of GD for training overparamterized neural networks. $\mathbf{K}^{(L)}$ is the Gram matrix for L-hidden-layer neural network [9]. Note that the dimension of the output is $k = 1$ in Du et al. [11, 9], Wu et al. [19], Oymak and Soltanolkotabi [17].

	Over-para. condition	Iteration complexity	Deep?	ReLU?
Du et al. [11]	$\Omega\left(\frac{n^6}{\lambda_0^4}\right)$	$O\left(\frac{n^2\log(1/\epsilon)}{\lambda_0^2}\right)$	no	yes
Wu et al. [19]	$\Omega\left(\frac{n^6}{\lambda_0^4}\right)$	$O\left(\frac{n\log(1/\epsilon)}{\lambda_0}\right)$	no	yes
Oymak and Soltanolkotabi [17]	$\Omega\left(\frac{n\|\mathbf{X}\|_2^6}{\lambda_0^4}\right)$	$O\left(\frac{\|\mathbf{X}\|_2^2\log(1/\epsilon)}{\lambda_0}\right)$	no	yes
Du et al. [9]	$\Omega\left(\frac{2^{O(L)}\cdot n^4}{\lambda_{\text{min}}^4(\mathbf{K}^{(L)})}\right)$	$O\left(\frac{2^{O(L)}\cdot n^2\log(1/\epsilon)}{\lambda_{\text{min}}^2(\mathbf{K}^{(L)})}\right)$	yes	no
Allen-Zhu et al. [2]	$\widetilde{\Omega}\left(\frac{kn^{24}L^{12}}{\phi^8}\right)$	$O\left(\frac{n^6L^2\log(1/\epsilon)}{\phi^2}\right)$	yes	yes
This paper	$\widetilde{\Omega}\left(\frac{kn^{8}L^{12}}{\phi^4}\right)$	$O\left(\frac{n^2L^2\log(1/\epsilon)}{\phi}\right)$	yes	yes

$\|\mathbf{W}\|_{2,\infty} = \max_{l\in[L]}\|\mathbf{W}_l\|_{2,\infty}$. Given two collections of matrices $\widetilde{\mathbf{W}} = \{\widetilde{\mathbf{W}}_1, \ldots, \widetilde{\mathbf{W}}_L\}$ and $\widehat{\mathbf{W}} = \{\widehat{\mathbf{W}}_1, \ldots, \widehat{\mathbf{W}}_L\}$, we define their inner product as $\langle\widetilde{\mathbf{W}}, \widehat{\mathbf{W}}\rangle = \sum_{l=1}^L\langle\widetilde{\mathbf{W}}_l, \widehat{\mathbf{W}}_l\rangle$. 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 use $\widetilde{O}(\cdot)$ and $\widetilde{\Omega}(\cdot)$ to hide logarithmic factors.

2 Problem setup and algorithms

In this section, we introduce the problem setup and the training algorithms.

Following Allen-Zhu et al. [2], we consider the training of an L-hidden layer fully connected neural network, which takes $\mathbf{x} \in \mathbb{R}^d$ as input, and outputs $\mathbf{y} \in \mathbb{R}^k$. In specific, the neural network is a vector-valued function $\mathbf{f}_{\mathbf{W}}: \mathbb{R}^d \to \mathbb{R}^k$, which is defined as

$$\mathbf{f}_{\mathbf{W}}(\mathbf{x}) = \mathbf{V}\sigma(\mathbf{W}_L\sigma(\mathbf{W}_{L-1}\cdots\sigma(\mathbf{W}_1\mathbf{x})\cdots)),$$

where $\mathbf{W}_1 \in \mathbb{R}^{m\times d}$, $\mathbf{W}_2, \ldots, \mathbf{W}_L \in \mathbb{R}^{m\times m}$ denote the weight matrices for the hidden layers, and $\mathbf{V} \in \mathbb{R}^{k\times m}$ denotes the weight matrix in the output layer, $\sigma(x) = \max\{0, x\}$ is the entry-wise ReLU activation function. In addition, we denote by $\sigma'(x) = \mathbb{1}(x)$ the derivative of ReLU activation function and $\mathbf{w}_{l,j}$ the weight vector of the j-th node in the l-th layer.

Given a training set $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1,\ldots,n}$ where $\mathbf{x}_i \in \mathbb{R}^d$ and $\mathbf{y}_i \in \mathbb{R}^k$, the empirical loss function for training the neural network is defined as

$$L(\mathbf{W}) := \frac{1}{n}\sum_{i=1}^n \ell(\widehat{\mathbf{y}}_i, \mathbf{y}_i), \tag{2.1}$$

where $\ell(\cdot,\cdot)$ is the loss function, and $\widehat{\mathbf{y}}_i = \mathbf{f}_{\mathbf{W}}(\mathbf{x}_i)$. In this paper, for the ease of exposition, we follow Allen-Zhu et al. [2], Du et al. [11, 9], Oymak and Soltanolkotabi [17] and consider square loss as follows

$$\ell(\widehat{\mathbf{y}}_i, \mathbf{y}_i) = \frac{1}{2}\|\mathbf{y}_i - \widehat{\mathbf{y}}_i\|_2^2,$$

where $\widehat{\mathbf{y}}_i = \mathbf{f}_{\mathbf{W}}(\mathbf{x}_i) \in \mathbb{R}^k$ denotes the output of the neural network given input $\mathbf{x}_i$. It is worth noting that our result can be easily extended to other loss functions such as cross entropy loss [24] as well.

We will study both gradient descent and stochastic gradient descent as training algorithms, which are displayed in Algorithm 1. For gradient descent, we update the weight matrix $\mathbf{W}_l^{(t)}$ using full partial gradient $\nabla_{\mathbf{W}_l}L(\mathbf{W}^{(t)})$. For stochastic gradient descent, we update the weight matrix $\mathbf{W}_l^{(t)}$ using stochastic partial gradient $1/B\sum_{s\in\mathcal{B}^{(t)}}\nabla_{\mathbf{W}_l}\ell\big(\mathbf{f}_{\mathbf{W}^{(t)}}(\mathbf{x}_s), \mathbf{y}_s\big)$, where $\mathcal{B}^{(t)}$ with $|\mathcal{B}^{(t)}| = B$ denotes the minibatch of training examples at the t-th iteration. Both algorithms are initialized in the same

Algorithm 1 (Stochastic) Gradient descent with Gaussian random initialization

```
input: Training data $\{\mathbf{x}_i, \mathbf{y}_i\}_{i\in[n]}$, step size $\eta$, total number of iterations $T$, minibatch size $B$.
initialization: For all $l \in [L]$, each row of weight matrix $\mathbf{W}_l^{(0)}$ is independently generated from $\mathcal{N}(0, 2/m\mathbf{I})$, each row of $\mathbf{V}$ is independently generated from $\mathcal{N}(0, \mathbf{I}/k)$
Gradient Descent
for $t = 0, \ldots, T$ do
    $\mathbf{W}_l^{(t+1)} = \mathbf{W}_l^{(t)} - \eta\nabla_{\mathbf{W}_l} L(\mathbf{W}^{(t)})$ for all $l \in [L]$
end for
output: $\{\mathbf{W}_l^{(T)}\}_{l\in[L]}$
Stochastic Gradient Descent
for $t = 0, \ldots, T$ do
    Uniformly sample a minibatch of training data $\mathcal{B}^{(t)} \in [n]$
    $\mathbf{W}_l^{(t+1)} = \mathbf{W}_l^{(t)} - \frac{\eta}{B}\sum_{s\in\mathcal{B}^{(t)}} \nabla_{\mathbf{W}_l}\ell\big(\mathbf{f}_{\mathbf{W}^{(t)}}(\mathbf{x}_s), \mathbf{y}_s\big)$ for all $l \in [L]$
end for
output: $\{\mathbf{W}_l^{(T)}\}_{l\in[L]}$
```

way as Allen-Zhu et al. [2], which is essentially the initialization method [13] widely used in practice. In the remaining of this paper, we denote by

$$\nabla L(\mathbf{W}^{(t)}) = \{\nabla_{\mathbf{W}_l} L(\mathbf{W}^{(t)})\}_{l\in[L]} \quad \text{and} \quad \nabla\ell\big(\mathbf{f}_{\mathbf{W}^{(t)}}(\mathbf{x}_i), \mathbf{y}_i\big) = \{\nabla_{\mathbf{W}_l}\ell\big(\mathbf{f}_{\mathbf{W}^{(t)}}(\mathbf{x}_i), \mathbf{y}_i\big)\}_{l\in[L]}$$

the collections of all partial gradients of $L(\mathbf{W}^{(t)})$ and $\ell\big(\mathbf{f}_{\mathbf{W}^{(t)}}(\mathbf{x}_i), \mathbf{y}_i\big)$.

3 Main theory

In this section, we present our main theoretical results. We make the following assumptions on the training data.

Assumption 3.1. For any $\mathbf{x}_i$, it holds that $\|\mathbf{x}_i\|_2 = 1$ and $(\mathbf{x}_i)_d = \mu$, where μ is an positive constant.

The same assumption has been made in all previous work along this line [9, 2, 24, 17]. Note that requiring the norm of all training examples to be 1 is not essential, and this assumption can be relaxed to be $\|\mathbf{x}_i\|_2$ is lower and upper bounded by some constants.

Assumption 3.2. For any two different training data points $\mathbf{x}_i$ and $\mathbf{x}_j$, there exists a positive constant $\phi > 0$ such that $\|\mathbf{x}_i - \mathbf{x}_j\|_2 \geq \phi$.

This assumption has also been made in Allen-Zhu et al. [3, 2], which is essential to guarantee zero training error for deep neural networks. It is a quite mild assumption for the regression problem as studied in this paper. Note that Du et al. [9] made a different assumption on training data, which requires the Gram matrix $\mathbf{K}^{(L)}$ (See their paper for details) defined on the L-hidden-layer networks is positive definite. However, their assumption is not easy to verify for neural networks with more than two layers.

Based on Assumptions 3.1 and 3.2, we are able to establish the global convergence rates of GD and SGD for training deep ReLU networks. We start with the result of GD for L-hidden-layer networks.

3.1 Training L-hidden-layer ReLU networks with GD

The global convergence of GD for training deep neural networks is stated in the following theorem.

Theorem 3.3. Under Assumptions 3.1 and 3.2, and suppose the number of hidden nodes per layer satisfies

$$m = \Omega\big(kn^8L^{12}\log^3(m)/\phi^4\big). \tag{3.1}$$

Then if set the step size $\eta = O\big(k/(L^2m)\big)$, with probability at least $1 - O(n^{-1})$, gradient descent is able to find a point that achieves ϵ training loss within

$$T = O\big(n^2L^2\log(1/\epsilon)/\phi\big)$$

iterations.

Remark 3.4. The state-of-the-art results for training deep ReLU network are provided by Allen-Zhu et al. [2], where the authors showed that GD can achieve ϵ-training loss within $O\big(n^6L^2\log(1/\epsilon)/\phi^2\big)$ iterations if the neural network width satisfies $m = \widetilde{\Omega}\big(kn^{24}L^{12}/\phi^8\big)$. As a clear comparison, our result on the iteration complexity is better than theirs by a factor of $O(n^4/\phi)$, and our over-parameterization condition is milder than theirs by a factor of $\widetilde{\Omega}(n^{16}/\phi^4)$. Du et al. [9] also proved the global convergence of GD for training deep neural network with smooth activation functions. As shown in Table 1, the over-parameterization condition and iteration complexity in Du et al. [9] have an exponential dependency on L, which is much worse than the polynomial dependency on L as in Allen-Zhu et al. [2] and our result.

We now specialize our results in Theorem 3.3 to two-layer networks by removing the dependency on the number of hidden layers, i.e., L. We state this result in the following corollary.

Corollary 3.5. Under the same assumptions made in Theorem 3.3. For training two-layer ReLU networks, if set the number of hidden nodes $m = \Omega\big(kn^8\log^3(m)/\phi^4\big)$ and step size $\eta = O(k/m)$, then with probability at least $1 - O(n^{-1})$, GD is able to find a point that achieves ϵ-training loss within $T = O\big(n^2\log(1/\epsilon)/\phi\big)$ iterations.

For training two-layer ReLU networks, Du et al. [11] made a different assumption on the training data to establish the global convergence of GD. Specifically, Du et al. [11] defined a Gram matrix, which is also known as neural tangent kernel [14], based on the training data $\{\mathbf{x}_i\}_{i=1,\ldots,n}$ and assumed that the smallest eigenvalue of such Gram matrix is strictly positive. In fact, for two-layer neural networks, their assumption is equivalent to Assumption 3.2, as shown in the following proposition.

Proposition 3.6. Under Assumption 3.1, define the Gram matrix $\mathbf{H} \in \mathbb{R}^{n\times n}$ as follows

$$\mathbf{H}_{ij} = \mathbb{E}_{\mathbf{w}\sim\mathcal{N}(0,\mathbf{I})}[\mathbf{x}_i^\top\mathbf{x}_j\sigma'(\mathbf{w}^\top\mathbf{x}_i)\sigma'(\mathbf{w}^\top\mathbf{x}_j)],$$

then the assumption $\lambda_0 = \lambda_{\min}(\mathbf{H}) > 0$ is equivalent to Assumption 3.2. In addition, there exists a sufficiently small constant C such that $\lambda_0 \ge C\phi n^{-2}$.

Remark 3.7. According to Proposition 3.6, we can make a direct comparison between our convergence results for two-layer ReLU networks in Corollary 3.5 with those in Du et al. [11], Oymak and Soltanolkotabi [17]. In specific, as shown in Table 1, the iteration complexity and over-parameterization condition proved in Du et al. [11] can be translated to $O(n^6\log(1/\epsilon)/\phi^2)$ and $\Omega(n^{14}/\phi^4)$ respectively under Assumption 3.2. Oymak and Soltanolkotabi [17] improved the result in Du et al. [11] and the improved iteration complexity and over-parameterization condition can be translated to $O\big(n^2\|\mathbf{X}\|_2^2\log(1/\epsilon)/\phi\big)$ [2] and $\Omega\big(n^9\|\mathbf{X}\|_2^6/\phi^4\big)$ respectively, where $\mathbf{X} = [\mathbf{x}_1,\ldots,\mathbf{x}_n]^\top \in \mathbb{R}^{d\times n}$ is the input data matrix. Our iteration complexity for two-layer ReLU networks is better than that in Oymak and Soltanolkotabi [17] by a factor of $O(\|\mathbf{X}\|_2^2)$ [3], and the over-parameterization condition is also strictly milder than the that in Oymak and Soltanolkotabi [17] by a factor of $O(n\|\mathbf{X}\|_2^6)$.

3.2 Extension to training L-hidden-layer ReLU networks with SGD

Then we extend the convergence results of GD to SGD in the following theorem.

Theorem 3.8. Under Assumptions 3.1 and 3.2, and suppose the number of hidden nodes per layer satisfies

$$m = \Omega\big(kn^{17}L^{12}\log^3(m)/(B^4\phi^8)\big). \tag{3.2}$$

Then if set the step size as $\eta = O\big(kB\phi/(n^3m\log(m))\big)$, with probability at least $1 - O(n^{-1})$, SGD is able to achieve ϵ expected training loss within

$$T = O\big(n^5L^2\log(m)\log^2(1/\epsilon)/(B\phi^2)\big)$$

iterations.

---

[2] It is worth noting that $\|\mathbf{X}\|_2^2 = O(1)$ if $d \lesssim n$, $\|\mathbf{X}\|_2^2 = O(n/d)$ if $\mathbf{X}$ is randomly generated, and $\|\mathbf{X}\|_2^2 = O(n)$ in the worst case.

[3] Here we set $k = 1$ in order to match the problem setting in Du et al. [11], Oymak and Soltanolkotabi [17].

Remark 3.9. We first compare our result with the state-of-the-art proved in Allen-Zhu et al. [2], where they showed that SGD can find a point with ϵ-training loss within $\widetilde{O}\big(n^7L^2\log(1/\epsilon)/(B\phi^2)\big)$ iterations if $m = \widetilde{\Omega}\big(n^{24}L^{12}Bk/\phi^8\big)$. In stark contrast, our result on the over-parameterization condition is strictly better than it by a factor of $\widetilde{\Omega}(n^7B^5)$, and our result on the iteration complexity is also faster by a factor of $O(n^2)$.

Moreover, we also characterize the convergence rate and over-parameterization condition of SGD for training two-layer networks. Unlike the gradient descent, which has the same convergence rate and over-parameterization condition for training both deep and two-layer networks in terms of training data size n, we find that the over-parameterization condition of SGD can be further improved for training two-layer neural networks. We state this improved result in the following theorem.

Theorem 3.10. Under the same assumptions made in Theorem 3.8. For two-layer ReLU networks, if set the number of hidden nodes and step size as

$$m = \Omega\big(k^{5/2}n^{11}\log^3(m)/(\phi^5B)\big), \quad \eta = O\big(kB\phi/(n^3m\log(m))\big),$$

then with probability at least $1-O(n^{-1})$, stochastic gradient descent is able to achieve ϵ training loss within $T = O\big(n^5\log(m)\log(1/\epsilon)/(B\phi^2)\big)$ iterations.

Remark 3.11. From Theorem 3.8, we can also obtain the convergence results of SGD for two-layer ReLU networks by choosing $L = 1$. However, the resulting over-parameterization condition is $m = \Omega\big(kn^{17}\log^3(m)B^{-4}\phi^{-8}\big)$, which is much worse than that in Theorem 3.10. This is because for two-layer networks, the training loss enjoys nicer local properties around the initialization, which can be leveraged to improve the convergence of SGD. Due to space limit, we defer more details to Appendix A.

4 Proof sketch of the main theory

In this section, we provide the proof sketch for Theorems 3.3, and highlight our technical contributions and innovative proof techniques.

4.1 Overview of the technical contributions

The improvements in our result are mainly attributed to the following two aspects: (1) a tighter gradient lower bound leading to faster convergence; and (2) a sharper characterization of the trajectory length of the algorithm.

We first define the following perturbation region based on the initialization,

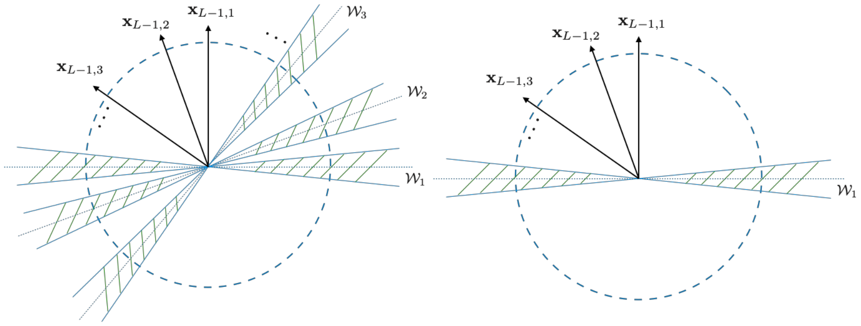
$$\mathcal{B}(\mathbf{W}^{(0)},\tau) = \{\mathbf{W} : \|\mathbf{W}_l - \mathbf{W}_l^{(0)}\|_2 \le \tau \text{ for all } l \in [L]\},$$

where $\tau > 0$ is the preset perturbation radius for each weight matrix $\mathbf{W}_l$.

Tighter gradient lower bound. By the definition of $\nabla L(\mathbf{W})$, we have $\|\nabla L(\mathbf{W})\|_F^2 = \sum_{l=1}^L \|\nabla_{\mathbf{W}_l} L(\mathbf{W})\|_F^2 \ge \|\nabla_{\mathbf{W}_L} L(\mathbf{W})\|_F^2$. Therefore, we can focus on the partial gradient of $L(\mathbf{W})$ with respect to the weight matrix at the last hidden layer. Note that we further have $\|\nabla_{\mathbf{W}_L} L(\mathbf{W})\|_F^2 = \sum_{j=1}^m \|\nabla_{\mathbf{w}_{L,j}} L(\mathbf{W})\|_2^2$, where

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

and $\mathbf{x}_{L-1,i}$ denotes the output of the $(L-1)$-th hidden layer with input $\mathbf{x}_i$. In order to prove the gradient lower bound, for each $\mathbf{x}_{L-1,i}$, we introduce a region namely "gradient region", denoted by $\mathcal{W}_j$, which is almost orthogonal to $\mathbf{x}_{L-1,i}$. Then we prove two major properties of these n regions $\{\mathcal{W}_1, \ldots, \mathcal{W}_n\}$: (1) $\mathcal{W}_i \cap \mathcal{W}_j = \emptyset$ if $i \ne j$, and (2) if $\mathbf{w}_{L,j} \in \mathcal{W}_i$ for any i, with probability at least $1/2$, $\|\nabla_{\mathbf{w}_{L,j}} L(\mathbf{W})\|_2$ is sufficiently large. We visualize these "gradient regions" in Figure 1(a). Since $\{\mathbf{w}_{L,j}\}_{j\in[m]}$ are randomly generated at the initialization, in order to get a larger bound of $\|\nabla_{\mathbf{W}_L} L(\mathbf{W})\|_F^2$, we hope the size of these "gradient regions" to be as large as possible. We take the union of the "gradient regions" for all training data, i.e., $\cup_{i=1}^n \mathcal{W}_i$, which is shown in Figure 1(a). As a



(a) "gradient region" for $\{\mathbf{x}_{L-1,i}\}_{i\in[n]}$ (b) "gradient region" for $\mathbf{x}_{L-1,1}$

Figure 1: (a): "gradient region" for all training data (b): "gradient region" for one training example.

comparison, Allen-Zhu et al. [2], Zou et al. [24] only leveraged the "gradient region" for one training data point to establish the gradient lower bound, which is shown in Figure 1(b). Roughly speaking, the size of "gradient regions" utilized in our proof is n times larger than those used in Allen-Zhu et al. [2], Zou et al. [24], which consequently leads to an $O(n)$ improvement on the gradient lower bound. The improved gradient lower bound is formally stated in the following lemma.

Lemma 4.1 (Gradient lower bound)**.** Let $\tau = O\big(\phi^{3/2} n^{-3} L^{-6} \log^{-3/2}(m)\big)$, then for all $\mathbf{W} \in \mathcal{B}(\mathbf{W}^{(0)}, \tau)$, with probability at least $1 - \exp\big(O(m\phi/(dn))\big)$, it holds that

$$\|\nabla L(\mathbf{W})\|_F^2 \geq O\big(m\phi L(\mathbf{W})/(kn^2)\big).$$

Sharper characterization of the trajectory length. The improved analysis of the trajectory length is motivated by the following observation: at the t-th iteration, the decrease of the training loss after one-step gradient descent is proportional to the gradient norm, i.e., $L(\mathbf{W}^{(t)}) - L(\mathbf{W}^{(t+1)}) \propto \|\nabla L(\mathbf{W}^{(t)})\|_F^2$. In addition, the gradient norm $\|\nabla L(\mathbf{W}^{(t)})\|_F$ determines the trajectory length in the t-th iteration. Putting them together, we can obtain

$$\|\mathbf{W}_l^{(t+1)} - \mathbf{W}_l^{(t)}\|_2 = \eta \|\nabla_{\mathbf{W}_l} L(\mathbf{W}^{(t)})\|_2 \leq \sqrt{Ckn^2/(m\phi)} \cdot \Big(\sqrt{L(\mathbf{W}^{(t)})} - \sqrt{L(\mathbf{W}^{(t+1)})}\Big), \tag{4.1}$$

where C is an absolute constant. (4.1) enables the use of telescope sum, which yields $\|\mathbf{W}_l^{(t)} - \mathbf{W}_l^{(0)}\|_2 \leq \sqrt{Ckn^2 L(\mathbf{W}^{(0)})/m\phi}$. In stark contrast, Allen-Zhu et al. [2] bounds the trajectory length as

$$\|\mathbf{W}_l^{(t+1)} - \mathbf{W}_l^{(t)}\|_2 = \eta \|\nabla_{\mathbf{W}_l} L(\mathbf{W}^{(t)})\|_2 \leq \eta\sqrt{C' m L(\mathbf{W}^{(t)})/k},$$

and further prove that $\|\mathbf{W}_l^{(t)} - \mathbf{W}_l^{(0)}\|_2 \leq \sqrt{C' k n^6 L^2(\mathbf{W}^{(0)})/(m\phi^2)}$ by taking summation over t, where C' is an absolute constant. Our sharp characterization of the trajectory length is formally summarized in the following lemma.

Lemma 4.2. Assuming all iterates are staying inside the region $\mathcal{B}(\mathbf{W}^{(0)}, \tau)$ with $\tau = O\big(\phi^{3/2} n^{-3} L^{-6} \log^{-3/2}(m)\big)$, if set the step size $\eta = O\big(k/(L^2 m)\big)$, with probability least $1 - O(n^{-1})$, the following holds for all $t \geq 0$ and $l \in [L]$,

$$\|\mathbf{W}_l^{(t)} - \mathbf{W}_l^{(0)}\|_2 \leq O\big(\sqrt{kn^2 \log(n)/(m\phi)}\big).$$

4.2 Proof of Theorem 3.3

Our proof road map can be organized in three steps: (i) prove that the training loss enjoys good curvature properties within the perturbation region $\mathcal{B}(\mathbf{W}^{(0)}, \tau)$; (ii) show that gradient descent is able to converge to global minima based on such good curvature properties; and (iii) ensure all iterates stay inside the perturbation region until convergence.

Step (i) Training loss properties. We first show some key properties of the training loss within $\mathcal{B}(\mathbf{W}^{(0)}, \tau)$, which are essential to establish the convergence guarantees of gradient descent.

Lemma 4.3. If $m \geq O(L\log(nL))$, with probability at least $1 - O(n^{-1})$ it holds that $L(\mathbf{W}^{(0)}) \leq \widetilde{O}(1)$.

Lemma 4.3 suggests that the training loss $L(\mathbf{W})$ at the initial point does not depend on the number of hidden nodes per layer, i.e., m.

Moreover, the training loss $L(\mathbf{W})$ is nonsmooth due to the non-differentiable ReLU activation function. Generally speaking, smoothness is essential to achieve linear rate of convergence for gradient-based algorithms. Fortunately, Allen-Zhu et al. [2] showed that the training loss satisfies locally semi-smoothness property, which is summarized in the following lemma.

Lemma 4.4 (Semi-smoothness [2])**.** Let

$$\tau \in \big[\Omega\big(k^{3/2}/(m^{3/2}L^{3/2}\log^{3/2}(m))\big), O\big(1/(L^{4.5}\log^{3/2}(m))\big)\big].$$

Then for any two collections $\widehat{\mathbf{W}} = \{\widehat{\mathbf{W}}_l\}_{l\in[L]}$ and $\widetilde{\mathbf{W}} = \{\widetilde{\mathbf{W}}_l\}_{l\in[L]}$ satisfying $\widehat{\mathbf{W}}, \widetilde{\mathbf{W}} \in \mathcal{B}(\mathbf{W}^{(0)}, \tau)$, with probability at least $1 - \exp(-\Omega(-m\tau^{3/2}L))$, there exist two constants C' and C'' such that

$$\begin{aligned} L(\widetilde{\mathbf{W}}) &\leq L(\widehat{\mathbf{W}}) + \langle \nabla L(\widehat{\mathbf{W}}), \widetilde{\mathbf{W}} - \widehat{\mathbf{W}}\rangle \\ &\quad + C'\sqrt{L(\widehat{\mathbf{W}})} \cdot \frac{\tau^{1/3}L^2\sqrt{m\log(m)}}{\sqrt{k}} \cdot \|\widetilde{\mathbf{W}} - \widehat{\mathbf{W}}\|_2 + \frac{C''L^2m}{k}\|\widetilde{\mathbf{W}} - \widehat{\mathbf{W}}\|_2^2. \end{aligned} \tag{4.2}$$

Lemma 4.4 is a rescaled version of Theorem 4 in Allen-Zhu et al. [2], since the training loss $L(\mathbf{W})$ in (2.1) is divided by the training sample size n, as opposed to the training loss in Allen-Zhu et al. [2]. This lemma suggests that if the perturbation region is small, i.e., $\tau \ll 1$, the non-smooth term (third term on the R.H.S. of (4.2)) is small and dominated by the gradient term (the second term on the the R.H.S. of (4.2)). Therefore, the training loss behaves like a smooth function in the perturbation region and the linear rate of convergence can be proved.

Step (ii) Convergence rate of GD. Now we are going to establish the convergence rate for gradient descent under the assumption that all iterates stay inside the region $\mathcal{B}(\mathbf{W}^{(0)}, \tau)$, where τ will be specified later.

Lemma 4.5. Assume all iterates stay inside the region $\mathcal{B}(\mathbf{W}^{(0)}, \tau)$, where $\tau = O\big(\phi^{3/2}n^{-3}L^{-6}\log^{-3/2}(m)\big)$. Then under Assumptions 3.1 and 3.2, if set the step size $\eta = O\big(k/(L^2m)\big)$, with probability least $1 - \exp\big(-O(m\tau^{3/2}L)\big)$, it holds that

$$L(\mathbf{W}^{(t)}) \leq \bigg(1 - O\bigg(\frac{m\phi\eta}{kn^2}\bigg)\bigg)^t L(\mathbf{W}^{(0)}).$$

Lemma 4.5 suggests that gradient descent is able to decrease the training loss to zero at a linear rate.

Step (iii) Verifying all iterates of GD stay inside the perturbation region. Then we are going to ensure that all iterates of GD are staying inside the required region $\mathcal{B}(\mathbf{W}^{(0)}, \tau)$. Note that we have proved the distance $\|\mathbf{W}_l^{(t)} - \mathbf{W}_l^{(0)}\|_2$ in Lemma 4.2. Therefore, it suffices to verify that such distance is smaller than the preset value τ. Thus, we can complete the proof of Theorem 3.3 by verifying the conditions based on our choice of m. Note that we have set the required number of m in (3.1), plugging (3.1) into the result of Lemma 4.2, we have with probability at least $1 - O(n^{-1})$, the following holds for all $t \leq T$ and $l \in [L]$

$$\|\mathbf{W}_l^{(t)} - \mathbf{W}_l^{(0)}\|_2 \leq O\big(\phi^{3/2}n^{-3}L^{-6}\log^{-3/2}(m)\big),$$

which is exactly in the same order of τ in Lemma 4.5. Therefore, our choice of m guarantees that all iterates are inside the required perturbation region. In addition, by Lemma 4.5, in order to achieve ϵ accuracy, we require

$$T\eta = O\big(kn^2\log\left(1/\epsilon\right)m^{-1}\phi^{-1}\big). \tag{4.3}$$

Then substituting our choice of step size $\eta = O\big(k/(L^2m)\big)$ into (4.3) and applying Lemma 4.3, we can get the desired result for T.

4.3 Optimizing both top and hidden layers

Here we would like to briefly discuss the extension to the case where the top layer is also optimized. The proof sketch is as follows: similar to our current proof, we can also define a small perturbation region around the initialization, but the new definition involves a constraint on the top layer weights. Specifically, such new perturbation region can be defined as follows,

$$\mathcal{B}(\mathbf{W}^{(0)}, \tau) = \{\mathbf{W} : \|\mathbf{W}_l - \mathbf{W}_l^{(0)}\|_2 \le \tau \text{ for all } l \in [L],\ \|\mathbf{V} - \mathbf{V}^{(0)}\|_2 \le \tau'\}.$$

Then, it can be proved that the neural network also enjoys good properties inside such region. Similar to the proof in this paper, based on these good properties, we can prove that until convergence the neural network weights, including the top layer weights, would not escape from such region. Note that optimizing more parameter can lead to larger gradient, thus we can prove a larger gradient lower bound during the training process which can potential speed up the convergence of optimization algorithm (e.g., GD, SGD).

5 Conclusions and future work

In this paper, we studied the global convergence of (stochastic) gradient descent for training over-parameterized ReLU networks, and improved the state-of-the-art results. Our proof technique can be also applied to prove similar results for other loss functions such as cross-entropy loss and other neural network architectures such as convolutional neural networks (CNN) [2, 11] and ResNet [2, 11, 21]. One important future work is to investigate whether the over-parameterization condition and the convergence rate can be further improved. It is promising that if we can further improve the characterization of "gradient region", as it may provide a tighter gradient lower bound and consequently sharpen the over-parameterization condition. Another interesting future direction is to explore the use of our proof technique to improve the generalization analysis of overparameterized neural networks trained by gradient-based algorithms [1, 6, 4].

Acknowledgement

We thank the anonymous reviewers and area chair for their helpful comments. This research was sponsored in part by the National Science Foundation CAREER Award IIS-1906169, BIGDATA IIS-1855099, and Salesforce Deep Learning Research Award. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

[1] ALLEN-ZHU, Z., LI, Y. and LIANG, Y. (2018). Learning and generalization in overparameterized neural networks, going beyond two layers. *arXiv preprint arXiv:1811.04918* .

[2] ALLEN-ZHU, Z., LI, Y. and SONG, Z. (2018). A convergence theory for deep learning via over-parameterization. *arXiv preprint arXiv:1811.03962* .

[3] ALLEN-ZHU, Z., LI, Y. and SONG, Z. (2018). On the convergence rate of training recurrent neural networks. *arXiv preprint arXiv:1810.12065* .

[4] ARORA, S., DU, S. S., HU, W., LI, Z. and WANG, R. (2019). Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *arXiv preprint arXiv:1901.08584* .

[5] BRUTZKUS, A. and GLOBERSON, A. (2017). Globally optimal gradient descent for a convnet with gaussian inputs. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org.

[6] CAO, Y. and GU, Q. (2019). A generalization theory of gradient descent for learning over-parameterized deep relu networks. *arXiv preprint arXiv:1902.01384* .

[7] CHIZAT, L. and BACH, F. (2018). A note on lazy training in supervised differentiable programming. *arXiv preprint arXiv:1812.07956* .

[8] Du, S. S. and Lee, J. D. (2018). On the power of over-parametrization in neural networks with quadratic activation. *arXiv preprint arXiv:1803.01206* .

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

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

[11] Du, S. S., Zhai, X., Poczos, B. and Singh, A. (2018). Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054* .

[12] Gao, W., Makkuva, A., Oh, S. and Viswanath, P. (2019). Learning one-hidden-layer neural networks under general input distributions. In *The 22nd International Conference on Artificial Intelligence and Statistics*.

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

[14] Jacot, A., Gabriel, F. and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*.

[15] Li, Y. and Liang, Y. (2018). Learning overparameterized neural networks via stochastic gradient descent on structured data. *arXiv preprint arXiv:1808.01204* .

[16] Li, Y. and Yuan, Y. (2017). Convergence analysis of two-layer neural networks with ReLU activation. *arXiv preprint arXiv:1705.09886* .

[17] Oymak, S. and Soltanolkotabi, M. (2019). Towards moderate overparameterization: global convergence guarantees for training shallow neural networks. *arXiv preprint arXiv:1902.04674* .

[18] 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* .

[19] Wu, X., Du, S. S. and Ward, R. (2019). Global convergence of adaptive gradient methods for an over-parameterized neural network. *arXiv preprint arXiv:1902.07111* .

[20] Zhang, C., Bengio, S., Hardt, M., Recht, B. and Vinyals, O. (2016). Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530* .

[21] Zhang, H., Yu, D., Chen, W. and Liu, T.-Y. (2019). Training over-parameterized deep resnet is almost as easy as training a two-layer network. *arXiv preprint arXiv:1903.07120* .

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

[23] Zhong, K., Song, Z., Jain, P., Bartlett, P. L. and Dhillon, I. S. (2017). Recovery guarantees for one-hidden-layer neural networks. *arXiv preprint arXiv:1706.03175* .

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

A Proof of the Main Theory

A.1 Proof of Proposition 3.6

We prove this proposition by two steps: (1) we prove that if there is no duplicate training data, it must hold that $\lambda_{\min}(\mathbf{H}) > 0$; (2) we prove that if there exists at least one duplicate training data, we have $\lambda_{\min}(\mathbf{H}) = 0$.

The first step can be done by applying Theorem 3 in Du et al. [11], where the author showed that if for any $i \neq j$, $\mathbf{x}_i \nparallel \mathbf{x}_j$, then it holds that $\lambda_{\min}(\mathbf{H}) > 0$. Since under Assumption 3.1, we have $\|\mathbf{x}_i\|_2 = \|\mathbf{x}_j\|_2$. Then it can be shown that $\mathbf{x}_i \neq \mathbf{x}_j$ for all $i \neq j$ is an sufficient condition to $\lambda_{\min}(\mathbf{H})$.

Then we conduct the second step. Clearly, if we have two training data with $\mathbf{x}_i = \mathbf{x}_j$, it can be shown that $\mathbf{H}_{ik} = \mathbf{H}_{jk}$ for all $k = 1, \ldots, n$. This immediately implies that there exist two identical rows in $\mathbf{H}$, which further suggests that $\lambda_{\min}(\mathbf{H}) = 0$.

The last argument can be directly proved by Lemma I.1 in Oymak and Soltanolkotabi [17], where the authors showed that $\lambda_0 = \lambda_{\min}(\mathbf{H}) \geq \phi/(100n^2)$.

By combining the above discussions, we are able to complete the proof.

A.2 Proof of Theorem 3.8

Now we sketch the proof of Theorem 3.8. Following the same idea of proving Theorem 3.3, we split the whole proof into three steps.

Step (i) Initialization and perturbation region characterization. Unlike the proof for GD, in addition to the crucial gradient lower bound specified in Lemma 4.1, we also require the gradient upper bound, which is stated in the following lemma.

Lemma A.1 (Gradient upper bounds [2]). Let $\tau = O\big(\phi^{3/2} n^{-3} L^{-6} \log^{-3/2}(m)\big)$, then for all $\mathbf{W} \in \mathcal{B}(\mathbf{W}^{(0)}, \tau)$, with probability at least $1 - \exp\big(O(m\phi/(dn))\big)$, the following holds for all $l \in [L]$

$$\|\nabla_{\mathbf{W}_l} L(\mathbf{W})\|_F^2 \leq O\bigg(\frac{mL(\mathbf{W})}{k}\bigg), \quad \|\nabla_{\mathbf{W}_l} \ell(\mathbf{f}_{\mathbf{W}}(\mathbf{x}_i), \mathbf{y}_i)\|_F^2 \leq O\bigg(\frac{m\ell(\mathbf{f}_{\mathbf{W}}(\mathbf{x}_i), \mathbf{y}_i)}{k}\bigg).$$

In later analysis, we show that the gradient upper bound will be exploited to bound the distance between iterates of SGD and its initialization. Besides, note that Lemmas 4.3 and 4.4 hold for both GD and SGD, we do not state them again in this part.

Step (ii) Convergence rate of SGD. Analogous to the proof for GD, the following lemma shows that SGD is able to converge to the global minima at a linear rate.

Lemma A.2. Assume all iterates stay inside the region $\mathcal{B}(\mathbf{W}^{(0)}, \tau)$, where $\tau = O\big(\phi^3 B^{3/2} n^{-6} L^{-6} \log^{-3/2}(m)\big)$. Then under Assumptions 3.1 and 3.2, if set the step size $\eta = O\big(B\phi/(L^2 m n^2)\big)$, with probability least $1 - \exp\big(-O(m\tau^{3/2}L)\big)$, it holds that

$$\mathbb{E}[L(\mathbf{W}^{(t)})] \leq \bigg(1 - O\bigg(\frac{m\phi\eta}{kn^2}\bigg)\bigg)^t L(\mathbf{W}^{(0)}).$$

Step (iii) Verifying all iterates of SGD stay inside the perturbation region. Similar to the proof for GD, the following lemma characterizes the distance from each iterate to the initial point for SGD.

Lemma A.3. Under the same assumptions made in Lemma A.2, if set the step size $\eta = O\big(kB\phi/(n^3 L^2 m \log(m))\big)$, suppose $m \geq O(T \cdot n)$, with probability at least $1 - O(n^{-1})$, the following holds for all $t \leq T$ and $l \in [L]$,

$$\|\mathbf{W}_l^{(t)} - \mathbf{W}_l^{(0)}\|_2 \leq O\big(k^{1/2} n^{5/2} B^{-1/2} m^{-1/2} \phi^{-1}\big).$$

Proof of Theorem 3.8. Compared with Lemma 4.2, the trajectory length of SGD is much larger than that of GD. In addition, we require a much smaller step size to guarantee that the iterates do not go too far away from the initial point. This makes over-parameterization condition of SGD worse than that of GD.

We complete the proof of Theorem 3.8 by verifying our choice of m in (3.2). By substituting (3.2) into Lemma A.3, we have with probability at least $1 - O(n^{-1})$, the following holds for all $t \le T$ and $l \in [L]$

$$\|\mathbf{W}_l^{(t)} - \mathbf{W}_l^{(0)}\|_2 = O\big(\phi^{3/2}B^{3/2}n^{-6}L^{-6}\log^{-3/2}(m)\big),$$

which is exactly in the same order of τ in Lemma A.2. Then by Lemma A.2, we know that in order to achieve ϵ expected training loss, it suffices to set

$$T\eta = O\big(kn^2m^{-1}\phi^{-1}\log(1/\epsilon)\big).$$

Then applying our choice of step size, i.e., $\eta = O\big(kB\phi/(n^3L^2m\log(m))\big)$, we can get the desired result for T. This completes the proof. $\square$

A.3 Proof of Theorem 3.10

Before proving Theorem 3.10, we first deliver the following two lemmas. The first lemma states the upper bound of stochastic gradient in $\|\cdot\|_{2,\infty}$ norm.

Lemma A.4. With probability at least $1 - O(m^{-1})$, it holds that

$$\|\nabla\ell(\mathbf{f}_{\mathbf{W}}(\mathbf{x}_i), \mathbf{y}_i)\|_{2,\infty}^2 \le O\big(\ell(\mathbf{f}_{\mathbf{W}}(\mathbf{x}_i), \mathbf{y}_i)\cdot\log(m)\big)$$

for all $\mathbf{W} \in \mathbb{R}^{m\times d}$ and $i \in [n]$.

The following lemma gives a different version of semi-smoothness for two-layer ReLU network.

Lemma A.5 (Semi-smoothness for two-layer ReLU network)**.** For any two collections $\widehat{\mathbf{W}} = \{\widehat{\mathbf{W}}_l\}_{l\in[L]}$ and $\widetilde{\mathbf{W}} = \{\widetilde{\mathbf{W}}_l\}_{l\in[L]}$ satisfying $\widehat{\mathbf{W}}, \widetilde{\mathbf{W}} \in \mathcal{B}(\mathbf{W}^{(0)}, \tau)$, with probability at least $1 - \exp(-O(-m\tau^{2/3}))$, there exist two constants C' and C'' such that

$$\begin{aligned} L(\widetilde{\mathbf{W}}) &\le L(\widehat{\mathbf{W}}) + \langle\nabla L(\widehat{\mathbf{W}}), \widetilde{\mathbf{W}} - \widehat{\mathbf{W}}\rangle \\ &\quad + C'\sqrt{L(\widehat{\mathbf{W}})}\cdot\frac{\tau^{2/3}m\sqrt{\log(m)}}{\sqrt{k}}\cdot\|\widetilde{\mathbf{W}} - \widehat{\mathbf{W}}\|_{2,\infty} + \frac{C''m}{k}\|\widetilde{\mathbf{W}} - \widehat{\mathbf{W}}\|_2^2. \end{aligned}$$

It is worth noting that Lemma 4.4 can also imply a $\|\cdot\|_{2,\infty}$ norm based semi-smoothness result by applying the inequality $\|\widetilde{\mathbf{W}} - \widehat{\mathbf{W}}\|_2 \le \|\widetilde{\mathbf{W}} - \widehat{\mathbf{W}}\|_F \le \sqrt{m}\|\widetilde{\mathbf{W}} - \widehat{\mathbf{W}}\|_{2,\infty}$. However, this operation will maintain the dependency on τ, i.e., $\tau^{1/3}$, which is worse than that in Lemma A.5 (e.g. $\tau^{2/3}$) since typically we have $\tau \ll 1$. Therefore, Lemma A.5 is crucial to establish a better convergence guarantee for SGD in training two-layer ReLU network.

Proof of Theorem 3.10. To simplify the proof, we use the following short-hand notation to define mini-batch stochastic gradient at the t-th iteration

$$\mathbf{G}^{(t)} = \frac{1}{|\mathcal{B}^{(t)}|}\sum_{s\in\mathcal{B}^{(t)}}\nabla\ell\big(\mathbf{f}_{\mathbf{W}^{(t)}}(\mathbf{x}_s), \mathbf{y}_s\big),$$

where $\mathcal{B}^{(t)}$ is the minibatch of data indices with $|\mathcal{B}^{(t)}| = B$. Then we bound its variance as follows,

$$\begin{aligned} \mathbb{E}[\|\mathbf{G}^{(t)} - \nabla L(\mathbf{W}^{(t)})\|_F^2] &\le \frac{1}{B}\mathbb{E}_s[\|\nabla\ell\big(\mathbf{f}_{\mathbf{W}^{(t)}}(\mathbf{x}_s), \mathbf{y}_s\big) - \nabla L(\mathbf{W}^{(t)})\|_F^2] \\ &\le \frac{2}{B}\big[\mathbb{E}_s[\|\nabla\ell\big(\mathbf{f}_{\mathbf{W}^{(t)}}(\mathbf{x}_s), \mathbf{y}_s\big)\|_F^2] + \|\nabla L(\mathbf{W}^{(t)})\|_F^2\big] \\ &\le \frac{4CL(\mathbf{W}^{(t)})}{Bk}, \end{aligned} \tag{A.1}$$

where C is an absolute constant, the expectation is taken over the random choice of training data and the second inequality follows from Young's inequality and the last inequality is by Lemma A.1. Moreover, we can further bound the expectation $\mathbb{E}[\|\mathbf{G}^{(t)}\|_2^2]$ as follows,

$$\mathbb{E}[\|\mathbf{G}^{(t)}\|_2^2] \le 2\mathbb{E}[\|\mathbf{G}^{(t)} - \nabla L(\mathbf{W}^{(t)})\|_F^2] + 2\|\nabla L(\mathbf{W}^{(t)})\|_F^2 \le \frac{8CmL(\mathbf{W}^{(t)})}{Bk} + 2\|\nabla L(\mathbf{W}^{(t)})\|_F^2. \tag{A.2}$$

By Lemma A.5, we have the following for one-step stochastic gradient descent

$$\begin{aligned} L(\mathbf{W}^{(t+1)}) &\le L(\mathbf{W}^{(t)}) - \eta\langle \nabla L(\mathbf{W}^{(t)}), \mathbf{G}^{(t)}\rangle \\ &\quad + C'\eta\sqrt{L(\mathbf{W}^{(t)})} \cdot \frac{\tau^{2/3}m\sqrt{\log(m)}}{\sqrt{k}} \cdot \|\mathbf{G}^{(t)}\|_{2,\infty} + \frac{C''m\eta^2}{k} \cdot \|\mathbf{G}^{(t)}\|_2^2. \end{aligned}$$

Taking expectation conditioned on $\mathbf{W}^{(t)}$, we obtain

$$\begin{aligned} \mathbb{E}[L(\mathbf{W}^{(t+1)})|\mathbf{W}^{(t)}] &\le L(\mathbf{W}^{(t)}) - \eta\langle \nabla L(\mathbf{W}^{(t)}), \nabla L(\mathbf{W}^{(t)})\rangle \\ &\quad + C'\eta\sqrt{L(\mathbf{W}^{(t)})} \cdot \frac{\tau^{2/3}m\sqrt{\log(m)}}{\sqrt{k}} \cdot \mathbb{E}[\|\mathbf{G}^{(t)}\|_{2,\infty}|\mathbf{W}^{(t)}] \\ &\quad + \frac{C''m\eta^2}{k} \cdot \mathbb{E}[\|\mathbf{G}^{(t)}\|_2^2|\mathbf{W}^{(t)}]. \end{aligned} \tag{A.3}$$

By Lemma A.4, with probability at least $1 - O(m^{-1})$ we have the following upper bound on the quantity $\mathbb{E}[\|\mathbf{G}^{(t)}\|_{2,\infty}|\mathbf{W}^{(t)}]$ for all $t = 1, \dots, T$,

$$\mathbb{E}[\|\mathbf{G}^{(t)}\|_{2,\infty}|\mathbf{W}^{(t)}] \le \mathbb{E}[\|\nabla\ell(\mathbf{f}_{\mathbf{W}^{(t)}}(\mathbf{x}_i), \mathbf{y}_i)\|_{2,\infty}|\mathbf{W}^{(t)}] \le O\big(\sqrt{L(\mathbf{W}^{(t)})\log(m)}\big).$$

Then based on Lemma 4.1, plugging (A.2) and the above inequality into (A.3), and set

$$\eta = O\bigg(\frac{k}{mn^2}\bigg) \quad \text{and} \quad \tau = O\bigg(\frac{\phi^3}{n^3k^{3/4}\log^{3/2}(m)}\bigg).$$

Then with proper adjustment of constants we can obtain

$$\mathbb{E}[L(\mathbf{W}^{(t+1)})|\mathbf{W}^{(t)}] \le L(\mathbf{W}^{(t)}) - \frac{\eta}{2}\|\nabla L(\mathbf{W}^{(t)})\|_F^2 \le \bigg(1 - \frac{m\phi\eta}{2kn^2}\bigg)L(\mathbf{W}^{(t)}),$$

where the last inequality follows from Lemma 4.1. Then taking expectation on $\mathbf{W}^{(t)}$, we have with probability $1 - O(m^{-1})$,

$$\mathbb{E}[L(\mathbf{W}^{(t+1)})] \le \bigg(1 - \frac{m\phi\eta}{2kn^2}\bigg)\mathbb{E}[L(\mathbf{W}^{(t)})] \le \bigg(1 - \frac{m\phi\eta}{2kn^2}\bigg)^{t+1}\mathbb{E}[L(\mathbf{W}^{(0)})] \tag{A.4}$$

holds for all $t > 0$. Then by Lemma A.3, we know that if set $\eta = O\big(kB\phi/(n^3m\log(m))\big)$, with probability at least $1 - O(n^{-1})$, it holds that

$$\|\mathbf{W}_l^{(t)} - \mathbf{W}_l^{(0)}\|_2 \le O\bigg(\frac{k^{1/2}n^{5/2}}{B^{1/2}m^{1/2}\phi}\bigg),$$

for all $t \le T$. Then by our choice of m, it is easy to verify that with probability at least $1 - O(n^{-1}) - O(m^{-1}) = 1 - O(n^{-1})$,

$$\|\mathbf{W}_l^{(t)} - \mathbf{W}_l^{(0)}\|_2 \le O\bigg(\frac{k^{1/2}n^{5/2}}{B^{1/2}\phi} \cdot \frac{\phi^4B^{1/2}}{k^{5/4}n^{11/2}\log^{3/2}(m)}\bigg) = \tau.$$

Moreover, note that in Lemma A.3 we set the step size as $\eta = O\big(kB\phi/(n^3m\log(m))\big)$ and (A.4) suggests that we need

$$T\eta = O\bigg(\frac{kn^2}{m\phi}\bigg)$$

to achieve ϵ expected training loss. Therefore we can derive the number of iteration as

$$T = O\bigg(\frac{n^5\log(m)\log(1/\epsilon)}{B\phi^2}\bigg).$$

This completes the proof. □

B Proof of Lemmas in Section 4 and Appendix A

B.1 Proof of Lemma 4.1

We first provide the following useful lemmas before starting the proof of Lemma 4.1.

The following lemma states that with high probability the norm of the output of each hidden layer is bounded by constants.

Lemma B.1 ([24])**.** At the initialization, if $m \ge O(L\log(nL))$, with probability at least $1 - \exp(-O(m/L))$, it holds that $1/2 \le \|\mathbf{x}_{l,i}\|_2 \le 2$ and $\big\|\mathbf{x}_{l,i}/\|\mathbf{x}_{l,i}\|_2 - \mathbf{x}_{l,j}/\|\mathbf{x}_{l,j}\|_2\big\|_2 \ge \phi/2$ for all $i, j \in [n]$ and $l \in [L]$, where $\mathbf{x}_{l,i}$ denotes the output of the l-th hidden layer given the input $\mathbf{x}_i$ and initial weight matrices $\mathbf{W}^{(0)}$.

The following lemma states

Lemma B.2. Assume $m \ge \widetilde{O}(n^2k^2\phi^{-1})$, then there exist an absolute constant $C > 0$ such that with probability at least $1 - \exp\big(-O(m\phi/(kn))\big)$, it holds that

$$\sum_{j=1}^m \bigg\| \frac{1}{n}\sum_{i=1}^n \langle \mathbf{u}_i, \mathbf{v}_j\rangle \sigma'\big(\langle \mathbf{w}_{L,j}^{(0)}, \mathbf{x}_{L-1,i}\rangle\big)\mathbf{x}_{L-1,i}\bigg\|_2^2 \ge \frac{C\phi m \sum_{i=1}^n \|\mathbf{u}_i\|_2^2}{kn^3}.$$

If we set $\mathbf{u}_i = \mathbf{f}_{\mathbf{W}^{(0)}}(\mathbf{x}_i) - \mathbf{y}_i$, Lemma B.2 corresponds to the gradient lower bound at the initialization. Then the next step is to prove the bounds for all $\mathbf{W}$ in the required perturbation region. Before proceeding to our final proof, we present the following lemma that provides useful results regarding the neural network within the perturbation region.

Lemma B.3 ([2])**.** Consider a collection of weight matrices $\widetilde{\mathbf{W}} = \{\widetilde{\mathbf{W}}_l\}_{l=1,\ldots,L}$ such that $\widetilde{\mathbf{W}} \in \mathcal{B}(\mathbf{W}^{(0)}, \tau)$, with probability at least $1 - \exp(-O(m\tau^{2/3}L))$, there exists constants C', C'' and C''' such that

- $\|\widetilde{\boldsymbol{\Sigma}}_{L,i} - \boldsymbol{\Sigma}_{L,i}\|_0 \le C'\tau^{2/3}L$
- $\|\mathbf{V}(\widetilde{\boldsymbol{\Sigma}}_{L,i} - \boldsymbol{\Sigma}_{L,i})\|_2 \le C''\tau^{1/3}L^2\sqrt{m\log(m)}/\sqrt{k}$
- $\|\widetilde{\mathbf{x}}_{L-1,i} - \mathbf{x}_{L-1,i}\|_2 \le C'''\tau L^{5/2}\sqrt{\log(m)}$,

for all $i = 1, \ldots, n$, where $\mathbf{x}_{L-1,i}$ and $\widetilde{\mathbf{x}}_{L-1,i}$ denote the outputs of the $L-1$-th layer of the neural network with weight matrices $\mathbf{W}^{(0)}$ and $\widetilde{\mathbf{W}}$, and $\boldsymbol{\Sigma}_{L,i}$ and $\widetilde{\boldsymbol{\Sigma}}_{L,i}$ are diagonal matrices with $(\boldsymbol{\Sigma}_{L,i})_{jj} = \sigma'(\langle \mathbf{w}_{L,j}^{(0)}, \mathbf{x}_{L-1}\rangle)$ and $(\widetilde{\boldsymbol{\Sigma}}_{L,i})_{jj} = \sigma'(\langle \widetilde{\mathbf{w}}_{L,j}, \widetilde{\mathbf{x}}_{L-1}\rangle)$ respectively.

Now we are ready to prove the lower and upper bounds of the Frobenious norm of the gradient.

Proof of Lemma 4.1. Consider any $\widetilde{\mathbf{W}} \in \mathcal{B}(\mathbf{W}^{(0)}, \tau)$, the gradient $\nabla_{\mathbf{W}_L} L(\widetilde{\mathbf{W}})$ takes form

$$\nabla_{\mathbf{W}_L} L(\widetilde{\mathbf{W}}) = \frac{1}{n}\sum_{i=1}^n \Big((\mathbf{f}_{\widetilde{\mathbf{W}}}(\mathbf{x}_i) - \mathbf{y}_i)^\top \mathbf{V}\widetilde{\boldsymbol{\Sigma}}_{L,i}\Big)^\top \widetilde{\mathbf{x}}_{L-1,i}^\top,$$

where $\widetilde{\boldsymbol{\Sigma}}_{L,i}$ is a diagonal matrix with $(\widetilde{\boldsymbol{\Sigma}}_{L,i})_{jj} = \sigma'(\widetilde{\mathbf{w}}_{L-1,j}, \widetilde{\mathbf{x}}_{L-1,i})$ and $\widetilde{\mathbf{x}}_{l-1,i}$ denotes the output of the l-th hidden layer with input $\mathbf{x}_i$ and model weight matrices $\widetilde{\mathbf{W}}$. Let $\mathbf{v}_j^\top$ denote the j-th row of matrix $\mathbf{V}$, and define

$$\widetilde{\mathbf{G}} = \frac{1}{n}\sum_{i=1}^n \Big((\mathbf{f}_{\widetilde{\mathbf{W}}}(\mathbf{x}_i) - \mathbf{y}_i)^\top \mathbf{V}\boldsymbol{\Sigma}_{L,i}\Big)^\top \mathbf{x}_{L-1,i}^\top,$$

where $\boldsymbol{\Sigma}_{L,i}$ is a diagonal matrix with $(\boldsymbol{\Sigma}_{L,i})_{jj} = \sigma'(\mathbf{w}_{L-1,j}^{(0)}, \mathbf{x}_{L-1,i})$. Then by Lemma B.2, we have with probability at least $1 - \exp\big(-O(m\phi/(kn))\big)$, the following holds for any $\widetilde{\mathbf{W}}$,

$$\begin{aligned}\|\widetilde{\mathbf{G}}\|_F^2 &= \frac{1}{n^2}\sum_{j=1}^{m}\bigg\|\sum_{i=1}^{n}\langle \mathbf{f}_{\widetilde{\mathbf{W}}}(\mathbf{x}_i) - \mathbf{y}_i, \mathbf{v}_j\rangle \sigma'(\langle \mathbf{w}_{L,j}, \mathbf{x}_{L-1,i})\mathbf{x}_{L-1,i}\bigg\|_2^2 \\ &\ge \frac{C_0\phi m\sum_{i=1}^n\|\mathbf{f}_{\widetilde{\mathbf{W}}}(\mathbf{x}_i) - \mathbf{y}_i\|_2^2}{kn^3},\end{aligned}$$

where C_0 is an absolute constant. Then we have

$$\begin{aligned}&\big\|\widetilde{\mathbf{G}} - \nabla_{\mathbf{W}_L}L(\widetilde{\mathbf{W}})\big\|_F \\ &= \frac{1}{n}\bigg\|\sum_{i=1}^{n}\Big((\mathbf{f}_{\widetilde{\mathbf{W}}}(\mathbf{x}_i) - \mathbf{y}_i)^\top\mathbf{V}\boldsymbol{\Sigma}_{L,i}\Big)^\top\mathbf{x}_{L-1,i}^\top - \sum_{i=1}^{n}\Big((\mathbf{f}_{\widetilde{\mathbf{W}}}(\mathbf{x}_i) - \mathbf{y}_i)^\top\mathbf{V}\widetilde{\boldsymbol{\Sigma}}_{L,i}\Big)^\top\widetilde{\mathbf{x}}_{L-1,i}^\top\bigg\|_F \\ &\le \frac{1}{n}\bigg[\bigg\|\sum_{i=1}^{n}\Big((\mathbf{f}_{\widetilde{\mathbf{W}}}(\mathbf{x}_i) - \mathbf{y}_i)^\top\mathbf{V}(\boldsymbol{\Sigma}_{L,i} - \widetilde{\boldsymbol{\Sigma}}_{L,i})\Big)^\top\mathbf{x}_{L-1,i}^\top\bigg\|_F \\ &\quad + \bigg\|\sum_{i=1}^{n}\Big((\mathbf{f}_{\widetilde{\mathbf{W}}}(\mathbf{x}_i) - \mathbf{y}_i)^\top\mathbf{V}\widetilde{\boldsymbol{\Sigma}}_{L,i}\Big)^\top(\mathbf{x}_{L-1,i} - \widetilde{\mathbf{x}}_{L-1,i})^\top\bigg\|_F\bigg].\end{aligned}$$

By Lemmas B.1 and B.3, we have

$$\begin{aligned}&\bigg\|\sum_{i=1}^{n}\Big((\mathbf{f}_{\widetilde{\mathbf{W}}}(\mathbf{x}_i) - \mathbf{y}_i)^\top\mathbf{V}(\boldsymbol{\Sigma}_{L,i} - \widetilde{\boldsymbol{\Sigma}}_{L,i})\Big)^\top\mathbf{x}_{L-1,i}^\top\bigg\|_F \\ &\le \sum_{i=1}^{n}\big\|\mathbf{f}_{\widetilde{\mathbf{W}}}(\mathbf{x}_i) - \mathbf{y}_i\big\|_2\big\|\mathbf{V}(\boldsymbol{\Sigma}_{L,i} - \widetilde{\boldsymbol{\Sigma}}_{L,i})\big\|_2\|\mathbf{x}_{L-1,i}\|_2 \\ &\le \frac{C_1\tau^{1/3}L^2\sqrt{m\log(m)}}{\sqrt{k}}\cdot\sum_{i=1}^{n}\big\|\mathbf{f}_{\widetilde{\mathbf{W}}(\mathbf{x}_i)} - \mathbf{y}_i\big\|_2,\end{aligned}$$

where the second inequality follows from Lemma B.3 and C_1 is an absolute constant. In addition, we also have

$$\begin{aligned}&\bigg\|\sum_{i=1}^{n}\Big((\mathbf{f}_{\widetilde{\mathbf{W}}}(\mathbf{x}_i) - \mathbf{y}_i)^\top\mathbf{V}\widetilde{\boldsymbol{\Sigma}}_{L,i}\Big)^\top(\mathbf{x}_{L-1,i} - \widetilde{\mathbf{x}}_{L-1,i})^\top\bigg\|_F \\ &\le \sum_{i=1}^{n}\|\mathbf{f}_{\widetilde{\mathbf{W}}}(\mathbf{x}_i) - \mathbf{y}_i\|_2\|\mathbf{V}\|_2\|\mathbf{x}_{L-1,i} - \widetilde{\mathbf{x}}_{L-1,i}\|_2 \\ &\le \frac{C_2\tau L^{5/2}\sqrt{m\log(m)}}{\sqrt{k}}\cdot\sum_{i=1}^{n}\|\mathbf{f}_{\widetilde{\mathbf{W}}}(\mathbf{x}_i) - \mathbf{y}_i\|_2,\end{aligned}$$

where the second inequality follows from Lemma B.3 and C_2 is an absolute constant. Combining the above bounds we have

$$\begin{aligned}\big\|\widetilde{\mathbf{G}} - \nabla_{\mathbf{W}_L}L(\widetilde{\mathbf{W}})\big\|_F &\le \frac{\sum_{i=1}^n\|\mathbf{f}_{\widetilde{\mathbf{W}}}(\mathbf{x}_i) - \mathbf{y}_i\|_2}{n}\cdot\bigg(\frac{C_1\tau^{1/3}L^2\sqrt{m\log(m)}}{\sqrt{k}} + \frac{C_2\tau L^{5/2}\sqrt{m\log(m)}}{\sqrt{k}}\bigg) \\ &\le \frac{\sum_{i=1}^n\|\mathbf{f}_{\widetilde{\mathbf{W}}}(\mathbf{x}_i) - \mathbf{y}_i\|_2}{n}\cdot\frac{C_3\tau^{1/3}L^2\sqrt{m\log(m)}}{\sqrt{k}},\end{aligned}$$

where the second inequality follows from the fact that $\tau \le O(L^{-4/3})$. Then by triangle inequality, we have the following lower bound of $\|\nabla_{\mathbf{W}_L}L(\widetilde{\mathbf{W}})\|_F$

$$\begin{aligned}\|\nabla_{\mathbf{W}_L}L(\widetilde{\mathbf{W}})\|_F &\ge \|\widetilde{\mathbf{G}}\|_F - \|\widetilde{\mathbf{G}} - \nabla_{\mathbf{W}_L}L(\widetilde{\mathbf{W}})\|_F \\ &\ge \frac{C_0\phi^{1/2}m^{1/2}\sqrt{n\sum_{i=1}^n\|\mathbf{f}_{\widetilde{\mathbf{W}}}(\mathbf{x}_i) - \mathbf{y}_i\|_2^2}}{\sqrt{k}n^2} \\ &\quad - \frac{\sum_{i=1}^n\|\mathbf{f}_{\widetilde{\mathbf{W}}}(\mathbf{x}_i) - \mathbf{y}_i\|_2}{n}\cdot\frac{C_3\tau^{1/3}L^2\sqrt{m\log(m)}}{\sqrt{k}}.\end{aligned}$$

By Jensen's inequality we know that $n\sum_{i=1}^n\|\mathbf{f}_{\widetilde{\mathbf{W}}}(\mathbf{x}_i)-\mathbf{y}_i\|_2^2\ge\big(\sum_{i=1}^n\|\mathbf{f}_{\widetilde{\mathbf{W}}}(\mathbf{x}_i)-\mathbf{y}_i\|_2\big)^2$. Then we set

$$\tau=\frac{C_3\phi^{3/2}}{2C_0n^3L^6\log^{3/2}(m)}=O\bigg(\frac{\phi^{3/2}}{n^3L^6\log^{3/2}(m)}\bigg),$$

and obtain

$$\|\nabla_{\mathbf{W}_L}L(\widetilde{\mathbf{W}})\|_F\ge\frac{C_0\phi^{1/2}m^{1/2}\sqrt{n\sum_{i=1}^n\|\mathbf{f}_{\widetilde{\mathbf{W}}}(\mathbf{x}_i)-\mathbf{y}_i\|_2^2}}{2\sqrt{k}n^2}.$$

Then plugging the fact that $1/(2n)\sum_{i=1}^n\|\mathbf{f}_{\widetilde{\mathbf{W}}}(\mathbf{x}_i)-\mathbf{y}_i\|_2^2=L(\widetilde{\mathbf{W}})$, we are able to complete the proof. □

B.2 Proof of Lemma 4.2

Proof of Lemma 4.2. In order to prove Lemma 4.2, we first establish the function decrease of gradient descent. Note that we assume that all iterate are staying inside the region $\mathcal{B}\big(\mathbf{W}^{(0)},\tau\big)$, then by Lemma 4.4, with probability at least $1-\exp(-O(m\tau^{2/3}L))$, we have the following after one-step gradient descent

$$\begin{aligned}L\big(\mathbf{W}^{(t+1)}\big)&\le L\big(\mathbf{W}^{(t)}\big)-\eta\|\nabla L\big(\mathbf{W}^{(t)}\big)\|_F^2\\&\quad+C'\eta\sqrt{L\big(\mathbf{W}^{(t)}\big)}\cdot\frac{\tau^{1/3}L^2\sqrt{m\log(m)}}{\sqrt{k}}\cdot\|\nabla L\big(\mathbf{W}^{(t)}\big)\|_2+\frac{C''L^2m\eta^2}{k}\|\nabla L\big(\mathbf{W}^{(t)}\big)\|_2^2.\end{aligned}\tag{B.1}$$

We first choose the step size

$$\eta=\frac{k}{4C''L^2m}=O\bigg(\frac{k}{L^2m}\bigg),$$

then (B.1) yields

$$\begin{aligned}L\big(\mathbf{W}^{(t+1)}\big)&\le L\big(\mathbf{W}^{(t)}\big)-\frac{3\eta}{4}\|\nabla L\big(\mathbf{W}^{(t)}\big)\|_F^2+C'\eta\sqrt{L\big(\mathbf{W}^{(t)}\big)}\cdot\frac{\tau^{1/3}L^2\sqrt{m\log(m)}}{\sqrt{k}}\cdot\|\nabla L\big(\mathbf{W}^{(t)}\big)\|_2\\&\le L\big(\mathbf{W}^{(t)}\big)-\eta\|\nabla L\big(\mathbf{W}^{(t)}\big)\|_F\bigg(\frac{3\|\nabla L\big(\mathbf{W}^{(t)}\big)\|_F}{4}-C'\sqrt{L\big(\mathbf{W}^{(t)}\big)}\cdot\frac{\tau^{1/3}L^2\sqrt{m\log(m)}}{\sqrt{k}}\bigg),\end{aligned}$$

where we use the fact that $\|\nabla L\big(\mathbf{W}^{(t)}\big)\|_2\le\|\nabla L\big(\mathbf{W}^{(t)}\big)\|_F$. Then by Lemma 4.1, we know that with probability at least $1-\exp\big(-O(m\phi/(kn))\big)$

$$\|\nabla L(\mathbf{W}^{(t)})\|_F^2\ge\|\nabla_{\mathbf{W}_L}L(\mathbf{W}^{(t)})\|_F^2\ge\frac{Cm\phi}{kn^2}L\big(\mathbf{W}^{(t)}\big),\tag{B.2}$$

where C is an absolute constant. Thus, we can choose the radius τ as

$$\tau=\frac{C^{3/2}\phi^{3/2}}{64n^3C'^3L^6\log^{3/2}(m)}=O\bigg(\frac{\phi^{3/2}}{n^3L^6\log^{3/2}(m)}\bigg),\tag{B.3}$$

and thus the following holds with probability at least $1-\exp(-O(m\tau^{2/3}L))-\exp\big(-O(m\phi/(kn))\big)=1-\exp(-O(m\tau^{2/3}L))$,

$$L\big(\mathbf{W}^{(t+1)}\big)\le L\big(\mathbf{W}^{(t)}\big)-\frac{\eta}{2}\|\nabla L\big(\mathbf{W}^{(t)}\big)\|_F^2,\tag{B.4}$$

where the second inequality follows from (B.2). By triangle inequality, we have

$$\|\mathbf{W}_l^{(t)}-\mathbf{W}_l^{(0)}\|_2\le\sum_{s=0}^{t-1}\eta\|\nabla_{\mathbf{W}_l}L\big(\mathbf{W}^{(s)}\big)\|_2\le\sum_{s=0}^{t-1}\eta\|\nabla L\big(\mathbf{W}^{(s)}\big)\|_F.\tag{B.5}$$

Moreover, we have

$$\begin{aligned}\sqrt{L(\mathbf{W}^{(s)})}-\sqrt{L(\mathbf{W}^{(s+1)})}&=\frac{L(\mathbf{W}^{(s)})-L(\mathbf{W}^{(s+1)})}{\sqrt{L(\mathbf{W}^{(s)})}+\sqrt{L(\mathbf{W}^{(s+1)})}}\\&\ge\frac{\eta\|\nabla L(\mathbf{W}^{(s)})\|_F^2}{4\sqrt{L(\mathbf{W}^{(s)})}}\\&\ge\sqrt{\frac{Cm\phi}{kn^2}}\cdot\frac{\eta\|\nabla L(\mathbf{W}^{(s)})\|_F}{4},\end{aligned}$$

where the second inequality is by (B.4) and the fact that $L(\mathbf{W}^{(s+1)})\le L(\mathbf{W}^{(s)})$, and the last inequality follows from (B.2). Plugging the above result into (B.5), we have with probability at least $1-\exp(-O(m\tau^{2/3}L))$,

$$\begin{aligned}\|\mathbf{W}_l^{(t)}-\mathbf{W}_l^{(0)}\|_2&\le\sum_{s=0}^{t-1}\eta\|\nabla L(\mathbf{W}^{(s)})\|_F\\&\le 4\sqrt{\frac{kn^2}{Cm\phi}}\sum_{s=0}^{t-1}\Big[\sqrt{L(\mathbf{W}^{(s)})}-\sqrt{L(\mathbf{W}^{(s+1)})}\Big]\\&\le 4\sqrt{\frac{kn^2}{Cm\phi}}\cdot\sqrt{L(\mathbf{W}^{(0)})}.\end{aligned}\tag{B.6}$$

Note that (B.6) holds for all l and t. Then apply Lemma 4.3, we are able to complete the proof.

□

B.3 Proof of Lemma 4.3

Proof of Lemma 4.3. Note that the output of the neural network can be formulated as

$$f_{\mathbf{W}^{(0)}}(\mathbf{x}_i)=\mathbf{V}\mathbf{x}_{L,i},$$

where $\mathbf{x}_{L,i}$ denotes the output of the last hidden layer with input $\mathbf{x}_i$. Note that each entry in $\mathbf{V}$ is i.i.d. generated from Gaussian distribution $\mathcal{N}(0,1/k)$. Thus, we know that with probability at least $1-\delta$, it holds that $\|\mathbf{V}\mathbf{x}_{L,i}\|_2\le\sqrt{\log(1/\delta)}\cdot\|\mathbf{x}_{L,i}\|_2$. Then by Lemma B.1 and union bound, we have $\|\mathbf{V}\mathbf{x}_{L,i}\|_2\le 2\sqrt{\log(1/\delta)}$ for all $i\in[n]$ with probability at least $1-\exp(-O(m/L))-n\delta$. Then we set $\delta=O(n^{-2})$ and use the fact that $m\ge O(L\log(nL))$, we have

$$f_{\mathbf{W}^{(0)}}(\mathbf{x}_i)=\|\mathbf{V}\mathbf{x}_{L,i}\|_2^2\le O(\log(n))$$

for all $i\in[n]$ with probability at least $1-O(n^{-1})$. Then by our definition of training loss, it follows that

$$\begin{aligned}L(\mathbf{W}^{(0)})&=\frac{1}{2n}\sum_{i=1}^n\|\mathbf{f}_{\mathbf{W}^{(0)}}(\mathbf{x}_i)-\mathbf{y}_i\|_2^2\\&\le\frac{1}{n}\sum_{i=1}^n[\|\mathbf{f}_{\mathbf{W}^{(0)}}(\mathbf{x}_i)\|_2^2+\|\mathbf{y}_i\|_2^2]\\&\le O(\log(n))\end{aligned}$$

with probability at least $1-O(n^{-1})$, where the first inequality is by Young's inequality and we assume that $\|\mathbf{y}_i\|_2=O(1)$ for all $i\in[n]$ in the second inequality. This completes the proof. □

B.4 Proof of Lemma 4.5

Proof of Lemma 4.5. By (B.4), we have

$$\begin{aligned} L\big(\mathbf{W}^{(t+1)}\big) &\le L\big(\mathbf{W}^{(t)}\big) - \frac{\eta}{2}\|\nabla L\big(\mathbf{W}^{(t)}\big)\|_F^2 \\ &\le \bigg(1 - \frac{Cm\phi\eta}{2kn^2}\bigg) L\big(\mathbf{W}^{(t)}\big) \\ &\le \bigg(1 - \frac{Cm\phi\eta}{2kn^2}\bigg)^{t+1} L\big(\mathbf{W}^{(0)}\big), \end{aligned} \tag{B.7}$$

where the second inequality follows from (B.2). This completes the proof. ☐

B.5 Proof of Lemma A.2

Proof of Lemma A.2. Let $\mathbf{G}^{(t)}$ denote the stochastic gradient leveraged in the t-th iteration, where the corresponding minibatch is defined as $\mathcal{B}^{(t)}$. By Lemma 4.4, we have the following inequality regarding one-step stochastic gradient descent

$$\begin{aligned} L(\mathbf{W}^{(t+1)}) &\le L(\mathbf{W}^{(t)}) - \eta\langle \nabla L(\mathbf{W}^{(t)}), \mathbf{G}^{(t)}\rangle \\ &\quad + C'\eta\sqrt{L\big(\mathbf{W}^{(t)}\big)} \cdot \frac{\tau^{1/3}L^2\sqrt{m\log(m)}}{\sqrt{k}} \cdot \|\mathbf{G}^{(t)}\|_2 + \frac{C''L^2m\eta^2}{k} \cdot \|\mathbf{G}^{(t)}\|_2^2. \end{aligned}$$

Then conditioned on $\mathbf{W}^{(t)}$, taking expectation over $\mathbf{G}^{(t)}$ on both sides gives

$$\begin{aligned} &\mathbb{E}\big[L(\mathbf{W}^{(t+1)})\big|\mathbf{W}^{(t)}\big] \\ &\le L(\mathbf{W}^{(t)}) - \eta\|\nabla L(\mathbf{W}^{(t)})\|_F^2 + C'\eta\sqrt{L\big(\mathbf{W}^{(t)}\big)} \cdot \frac{\tau^{1/3}L^2\sqrt{m\log(m)}}{\sqrt{k}} \cdot \mathbb{E}\big[\|\mathbf{G}^{(t)}\|_2\big|\mathbf{W}^{(t)}\big] \\ &\quad + \frac{C''L^2m\eta^2}{k} \cdot \mathbb{E}\big[\|\mathbf{G}^{(t)}\|_2^2\big|\mathbf{W}^{(t)}\big]. \end{aligned} \tag{B.8}$$

Note that given $\mathbf{W}^{(t)}$, the expectations on $\|\mathbf{G}^{(t)}\|_2$ and $\|\mathbf{G}^{(t)}\|_2^2$ are only taken over the random minibatch $\mathcal{B}^{(t)}$. Similar to (A.1) and (A.2), we have

$$\begin{aligned} \mathbb{E}_{\mathcal{B}^{(t)}}[\|\mathbf{G}^{(t)} - \nabla L(\mathbf{W}^{(t)})\|_2^2] &\le \frac{1}{B}\mathbb{E}_s[\|\nabla \ell\big(\mathbf{f}_{\mathbf{W}^{(t)}}(\mathbf{x}_s), \mathbf{y}_s\big) - \nabla L(\mathbf{W}^{(t)})\|_2^2] \\ &\le \frac{2}{B}\big[\mathbb{E}_s[\|\nabla \ell\big(\mathbf{f}_{\mathbf{W}^{(t)}}(\mathbf{x}_s), \mathbf{y}_s\big)\|_2^2] + \|\nabla L(\mathbf{W}^{(t)})\|_2^2\big]. \end{aligned}$$

Based on the definition of the ℓ_2-norm of a collection of matrices and Lemma A.1, we have

$$\mathbb{E}_s[\|\nabla \ell(f_{\mathbf{W}^{(t)}}(\mathbf{x}_s), y_s)\|_2^2] \le \mathbb{E}_s\big[\max_{l\in[L]}\|\nabla_{\mathbf{W}_l}\ell(f_{\mathbf{W}^{(t)}}(\mathbf{x}_s), y_s)\|_2^2\big] \le \frac{C_0L(\mathbf{W}^{(t)})}{k},$$

$$\|L(\mathbf{W}^{(t)})\|_2^2 \le \max_{l\in[L]}\|\nabla_{\mathbf{W}_l}L(\mathbf{W}^{(t)})\|_2^2 \le \frac{C_0L(\mathbf{W}^{(t)})}{k},$$

where C_0 is an absolute constant. Then we have

$$\begin{aligned} \mathbb{E}[\|\mathbf{G}^{(t)}\|_2|\mathbf{W}^{(t)}]^2 &\le \mathbb{E}[\|\mathbf{G}^{(t)}\|_2^2|\mathbf{W}^{(t)}] \\ &\le 2\mathbb{E}_{\mathcal{B}^{(t)}}[\|\mathbf{G}^{(t)} - \nabla L(\mathbf{W}^{(t)})\|_2^2] + 2\|\nabla L(\mathbf{W}^{(t)})\|_2^2 \\ &\le \frac{8C_0mL(\mathbf{W}^{(t)})}{Bk} + 2\|\nabla L(\mathbf{W}^{(t)})\|_F^2. \end{aligned}$$

By (B.2), we know that there is a constant C such that $\|\nabla L(\mathbf{W}^{(t)})\|_F^2 \ge Cm\phi L(\mathbf{W}^{(t)})/(kn^2)$. Then we set the step size η and radius τ as follows

$$\begin{aligned} \eta &= \frac{Cd}{64C_0C''L^2mn^2} = O\bigg(\frac{1}{L^2mn^2}\bigg) \\ \tau &= \frac{C^3\phi^{3/2}B^3}{64^2n^6C_0^3C'^3L^6\log^{3/2}(m)} = O\bigg(\frac{\phi^{3/2}B^3}{n^6L^6\log^{3/2}(m)}\bigg) \end{aligned} \tag{B.9}$$

Then (B.8) yields that

$$\begin{aligned}
\mathbb{E}\big[L(\mathbf{W}^{(t+1)})\big|\mathbf{W}^{(t)}\big] &\le L\big(\mathbf{W}^{(t)}\big) - \eta\|\nabla L\big(\mathbf{W}^{(t)}\big)\|_F^2 + \frac{C''L^2m\eta^2}{k}\bigg(\frac{8C_0n^2}{C\phi B}+2\bigg)\cdot\|\nabla L(\mathbf{W}^{(t)})\|_F^2 \\
&\quad + C'\eta\sqrt{L\big(\mathbf{W}^{(t)}\big)}\cdot\frac{\tau^{1/3}L^2\sqrt{m\log(m)}}{\sqrt{k}}\cdot\sqrt{\frac{8C_0n^2}{C\phi B}+2}\cdot\|\nabla L(\mathbf{W}^{(t)})\|_F \\
&\le L(\mathbf{W}^{(t)}) - \frac{\eta}{2}\|\nabla L(\mathbf{W}^{(t)})\|_F^2. \qquad\qquad \text{(B.10)}
\end{aligned}$$

Then applying (B.2) again and taking expectation over $\mathbf{W}^{(t)}$ on both sides of (B.10), we obtain

$$\mathbb{E}\big[L(\mathbf{W}^{(t+1)})\big] \le \bigg(1-\frac{Cm\phi\eta}{2kn^2}\bigg)\mathbb{E}[L(\mathbf{W}^{(t)})] \le \bigg(1-\frac{Cm\phi\eta}{2kn^2}\bigg)^{t+1}L(\mathbf{W}^{(0)}).$$

This completes the proof.

□

B.6 Proof of Lemma A.3

Proof of Lemma A.3. We prove this by standard martingale inequality. By Lemma 4.4, we have

$$\begin{aligned}
L(\mathbf{W}^{(t+1)}) &\le L(\mathbf{W}^{(t)}) - \eta\langle\nabla L(\mathbf{W}^{(t)}), \mathbf{G}^{(t)}\rangle \\
&\quad + \eta C'\sqrt{L(\mathbf{W}^{(t)})}\cdot\frac{\tau^{1/3}L^2\sqrt{m\log(m)}}{\sqrt{k}}\cdot\|\mathbf{G}^{(t)}\|_2 + \frac{C''L^2m\eta^2}{k}\|\mathbf{G}^{(t)}\|_2^2.
\end{aligned}$$

Then by our choice of τ in (B.9), we have

$$\frac{\tau^{1/3}L^2\sqrt{m\log(m)}}{\sqrt{k}} = O\bigg(\frac{B\sqrt{m\phi}}{n^2\sqrt{k}}\bigg).$$

Note that by Lemma 4.1, we have

$$C'\sqrt{L(\mathbf{W}^{(t)})}\cdot\frac{\tau^{1/3}L^2\sqrt{m\log(m)}}{\sqrt{k}} = O\big(B\sqrt{m\phi L(\mathbf{W})/(kn^2)}/n\big) \le \|\nabla L(\mathbf{W}^{(t)})\|_F.$$

Therefore, by inequality $\langle\mathbf{A},\mathbf{B}\rangle \le \|\mathbf{A}\|_F\|\mathbf{B}\|_F$, we have

$$L(\mathbf{W}^{(t+1)}) \le L(\mathbf{W}^{(t)}) + 2\eta\|\nabla L(\mathbf{W}^{(t)})\|_F\cdot\|\mathbf{G}^{(t)}\|_F + \frac{C''L^2m\eta^2}{k}\|\mathbf{G}^{(t)}\|_2^2. \qquad \text{(B.11)}$$

By Lemma A.1, we know that there exists an absolute constant C such that

$$\|\nabla L(\mathbf{W}^{(t)})\|_F^2 \le \frac{CLmL(\mathbf{W}^{(t)})}{k} \text{ and } \|\mathbf{G}^{(t)}\|_F^2 \le \frac{CLmnL(\mathbf{W}^{(t)})}{Bk},$$

where B denotes the minibatch size and we use the fact that $\sum_{i\in\mathcal{B}^{(t)}}\ell(f_{\mathbf{W}^{(t)}}(\mathbf{x}_i),\mathbf{y}_i) \le nL(\mathbf{W}^{(t)})$. Then note that $\eta \le O\big(B^{1/2}k/(mL^2n^{1/2})\big)$, we have the following according to (B.11)

$$L(\mathbf{W}^{(t+1)}) \le \bigg(1+\frac{C'Lmn^{1/2}\eta}{B^{1/2}k}\bigg)L(\mathbf{W}^{(t)}),$$

where C' is an absolute constant. Taking logarithm on both sides further leads to

$$\log\big(L(\mathbf{W}^{(t+1)})\big) \le \log\big(L(\mathbf{W}^{(t)})\big) + \frac{C'Lmn^{1/2}\eta}{B^{1/2}k},$$

where we use the inequality $\log(1+x) \le x$. By (B.2) and (B.10), we know that

$$\mathbb{E}[L(\mathbf{W}^{(t+1)})|\mathbf{W}^{(t)}] \le L(\mathbf{W}^{(t)}) - \frac{\eta}{2}\|\nabla L(\mathbf{W}^{(t)})\|_F^2 \le \bigg(1-\frac{C''m\phi\eta}{kn^2}\bigg)L(\mathbf{W}^{(t)}).$$

Then by Jensen's inequality and the inequality $\log(1+x) \le x$, we have

$$\mathbb{E}\big[\log\big(L(\mathbf{W}^{(t+1)})\big)|\mathbf{W}^{(t)}\big] \le \log\big(\mathbb{E}[L(\mathbf{W}^{(t+1)})|\mathbf{W}^{(t)}]\big) \le \log\big(L(\mathbf{W}^{(t)})\big) - \frac{C''m\phi\eta}{kn^2}.$$

Therefore we have $\{\log(L(\mathbf{W}^{(t)})) + C''m\phi t\eta/(kn^2)\}_{t=0,1\ldots,}$ is a super-martingale, and the martingale difference can be upper bounded by

$$\begin{aligned}\log(L(\mathbf{W}^{(t)})) + \frac{C''m\phi t\eta}{kn^2} - \log(L(\mathbf{W}^{(t-1)})) - \frac{C''m\phi(t-1)\eta}{kn^2} &\le \frac{C'Lmn^{1/2}\eta}{B^{1/2}k} + \frac{C''m\phi\eta}{kn^2}\\ &\le \frac{C'''Lmn^{1/2}\eta}{B^{1/2}k},\end{aligned}$$

where C''' is an absolute constant. By one-side Azuma's inequality for super-martingale, we know that for any t, with probability at least $1-\delta$, the following holds

$$\begin{aligned}\log\big(L(\mathbf{W}^{(t)})\big) &\le \log\big(L(\mathbf{W}^{(0)})\big) - \frac{tC''m\phi\eta}{kn^2} + \frac{C'''Lmn^{1/2}\eta}{B^{1/2}k}\sqrt{2t\log(1/\delta)}\\ &\le \log\big(L(\mathbf{W}^{(0)})\big) - \frac{tC''m\phi\eta}{2kn^2} + \frac{C'''^2L^2mn^3\log(1/\delta)\eta}{C''kB\phi},\end{aligned} \tag{B.12}$$

where the last inequality follows the fact that $-at + b\sqrt{t} \le b^2/(4a)$ in the last inequality. Then we chose $\delta = O(m^{-1})$ and

$$\eta = \frac{\log(2)C''kB\phi}{C'^2L^2mn^3\log(1/\delta)} = O\bigg(\frac{kB\phi}{L^2n^3m\log(m)}\bigg).$$

Plugging these into (B.12) gives

$$\log\big(L(\mathbf{W}^{(t)})\big) \le \log\big(2L(\mathbf{W}^{(0)})\big) - \frac{tC''m\phi\eta}{2kn^2},$$

which implies that

$$L(\mathbf{W}^{(t)}) \le 2L(\mathbf{W}^{(0)}) \cdot \exp\bigg(-\frac{tC''m\phi\eta}{2kn^2}\bigg). \tag{B.13}$$

By Lemma A.1 and the definition of $\mathbf{G}^{(t)}$, we have

$$\|\mathbf{G}^{(t)}\|_2 \le O\bigg(\frac{m^{1/2}n^{1/2}\sqrt{L(\mathbf{W}^{(t)})}}{B^{1/2}k^{1/2}}\bigg) \tag{B.14}$$

for all $t \le T$. Therefore, plugging (B.14) into (B.13) and taking union bound over all $t \le T$, and apply the result in Lemma 4.3, the following holds for all $t \le T$ with probability at least $1 - O(T\cdot m^{-1}) - O(n^{-1}) = 1 - O(n^{-1})$,

$$\|\mathbf{W}_l^{(t)} - \mathbf{W}_l^{(0)}\|_2 \le \sum_{s=0}^{t-1}\eta\|\mathbf{G}^{(t)}\|_2 \le O\bigg(\frac{m^{1/2}n^{1/2}}{B^{1/2}k^{1/2}}\bigg)\cdot\sum_{s=0}^{t-1}\eta\sqrt{L(\mathbf{W}^{(s)})} \le \widetilde{O}\bigg(\frac{k^{1/2}n^{5/2}}{B^{1/2}m^{1/2}\phi}\bigg),$$

where the first inequality is by triangle inequality, the second inequality follows from (B.14) and the last inequality is by (B.13) and Lemma 4.3. This completes the proof.

□

B.7 Proof of Lemma A.4

Proof of Lemma A.4. We first write the formula of $\nabla\ell\big(f_{\mathbf{W}}(\mathbf{x}_i), \mathbf{y}_i\big)$ as follows

$$\nabla\ell\big(f_{\mathbf{W}}(\mathbf{x}_i), \mathbf{y}_i\big) = \big[\big(f_{\mathbf{W}}(\mathbf{x}_i) - y_i\big)^\top \mathbf{V}\mathbf{\Sigma}_i\big]^\top \mathbf{x}_i^\top.$$

Since $\mathbf{\Sigma}_i$ is an diagonal matrix with $\big(\mathbf{\Sigma}_i\big)_{jj} = \sigma'(\langle\mathbf{w}_j, \mathbf{x}_i\rangle)$. Therefore, it holds that

$$\|\nabla\ell(f_{\mathbf{W}}(\mathbf{x}_i), \mathbf{y}_i)\|_{2,\infty} = \max_{j\in[m]}\langle f_{\mathbf{W}}(\mathbf{x}_i) - \mathbf{y}_i, \mathbf{v}_j\rangle \cdot \|\mathbf{x}_i\|_2 \le \max_{j\in[m]}\|f_{\mathbf{W}}(\mathbf{x}_i) - \mathbf{y}_i\|_2\|\mathbf{v}_j\|_2, \tag{B.15}$$

where $\mathbf{v}_j \in \mathbb{R}^k$ denotes the j-th column of $\mathbf{V}$ and we use the fact that $\|\mathbf{x}_i\|_2 = 1$. Note that $\mathbf{v}_j \sim \mathcal{N}(0, \mathbf{I}/k)$, we have

$$\mathbb{P}\big(\|\mathbf{v}_j\|_2^2 \ge O\big(\log(m)\big)\big) \le O(m^{-c}),$$

for any positive constant c. Setting $c = 2$ and applying union bound over $\mathbf{v}_1, \dots, \mathbf{v}_m$, we have with probability at least $1 - O(m^{-1})$,

$$\max_{j\in[m]} \|\mathbf{v}_j\|_2 \le O\big(\log^{1/2}(m)\big).$$

Plugging this into (B.15) and applying the fact that $\|f_{\mathbf{W}}(\mathbf{x}_i) - \mathbf{y}_i\|_2 = \sqrt{\ell(\mathbf{f}_{\mathbf{W}}(\mathbf{x}_i), \mathbf{y}_i)}$, we are able to complete the proof. $\square$

B.8 Proof of Lemma A.5

Recall that the output of two-layer ReLU network can be formulated as

$$\mathbf{f}_{\mathbf{W}}(\mathbf{x}_i) = \mathbf{V}\boldsymbol{\Sigma}_i\mathbf{W}\mathbf{x}_i,$$

where $\boldsymbol{\Sigma}_i$ is a diagonal matrix with only non-zero diagonal entry $(\boldsymbol{\Sigma}_i)_{jj} = \sigma'(\mathbf{w}_j^\top \mathbf{x}_i)$. Then based on the definition of $L(\mathbf{W})$, we have

$$\begin{aligned}
&L(\widetilde{\mathbf{W}}) - L(\widehat{\mathbf{W}}) \\
&= \frac{1}{2n}\sum_{i=1}^n \|\mathbf{V}\widetilde{\boldsymbol{\Sigma}}_i\widetilde{\mathbf{W}}\mathbf{x}_i - \mathbf{y}_i\|_2^2 - \frac{1}{2n}\sum_{i=1}^n \|\mathbf{V}\widehat{\boldsymbol{\Sigma}}_i\widehat{\mathbf{W}}\mathbf{x}_i - \mathbf{y}_i\|_2^2 \\
&= \underbrace{\frac{1}{2n}\sum_{i=1}^n \big\langle \mathbf{V}\widehat{\boldsymbol{\Sigma}}_i\widehat{\mathbf{W}}\mathbf{x}_i - \mathbf{y}_i, \mathbf{V}\widetilde{\boldsymbol{\Sigma}}_i\widetilde{\mathbf{W}}\mathbf{x}_i - \mathbf{V}\widehat{\boldsymbol{\Sigma}}_i\widehat{\mathbf{W}}\mathbf{x}_i\big\rangle}_{I_1} + \underbrace{\frac{1}{2n}\sum_{i=1}^n \big\|\mathbf{V}\widetilde{\boldsymbol{\Sigma}}_i\widetilde{\mathbf{W}}\mathbf{x}_i - \mathbf{V}\widehat{\boldsymbol{\Sigma}}_i\widehat{\mathbf{W}}\mathbf{x}_i\big\|_2^2}_{I_2}.
\end{aligned}$$

Then we tackle the two terms on the R.H.S. of the above equation separately. Regarding the first term, i.e., I_1, we have

$$\begin{aligned}
I_1 &= \frac{1}{2n}\sum_{i=1}^n \big\langle \mathbf{V}\widehat{\boldsymbol{\Sigma}}_i\widehat{\mathbf{W}}\mathbf{x}_i - \mathbf{y}_i, \mathbf{V}\widehat{\boldsymbol{\Sigma}}_i(\widetilde{\mathbf{W}} - \widehat{\mathbf{W}})\mathbf{x}_i\big\rangle \\
&\quad + \frac{1}{2n}\sum_{i=1}^n \big\langle \mathbf{V}\widehat{\boldsymbol{\Sigma}}_i\widehat{\mathbf{W}}\mathbf{x}_i - \mathbf{y}_i, \mathbf{V}(\widetilde{\boldsymbol{\Sigma}}_i - \widehat{\boldsymbol{\Sigma}}_i)\widetilde{\mathbf{W}}\mathbf{x}_i\big\rangle \\
&\le \langle \nabla L(\widehat{\mathbf{W}}), \widetilde{\mathbf{W}} - \widehat{\mathbf{W}}\rangle + \frac{1}{2n}\sum_{i=1}^n \sqrt{\ell(f_{\widehat{\mathbf{W}}}(\mathbf{x}_i), \mathbf{y}_i)} \cdot \|\mathbf{V}(\widetilde{\boldsymbol{\Sigma}}_i - \widehat{\boldsymbol{\Sigma}}_i)\widetilde{\mathbf{W}}\mathbf{x}_i\|_2.
\end{aligned}$$

Note that the non-zero entries in $\widetilde{\boldsymbol{\Sigma}}_i - \widehat{\boldsymbol{\Sigma}}_i$ represent the nodes, say j, satisfying $\mathrm{sign}(\widetilde{\mathbf{w}}_j^\top \mathbf{x}_i) \ne \mathrm{sign}(\widehat{\mathbf{w}}_j^\top \mathbf{x}_i)$, which implies $\big|\widetilde{\mathbf{w}}_j^\top \mathbf{x}_i\big| \le \big|(\widetilde{\mathbf{w}}_j - \widehat{\mathbf{w}}_j)^\top \mathbf{x}_i\big|$. Therefore, we have

$$\|\mathbf{V}(\widetilde{\boldsymbol{\Sigma}}_i - \widehat{\boldsymbol{\Sigma}}_i)\widetilde{\mathbf{W}}\mathbf{x}_i\|_2^2 \le \|\mathbf{V}(\widetilde{\boldsymbol{\Sigma}}_i - \widehat{\boldsymbol{\Sigma}}_i)(\widetilde{\mathbf{W}} - \widehat{\mathbf{W}})\mathbf{x}_i\|_2^2.$$

By Lemma B.3, we have $\|\widetilde{\boldsymbol{\Sigma}}_i - \widehat{\boldsymbol{\Sigma}}_i\|_0 \le \|\widetilde{\boldsymbol{\Sigma}}_i - \boldsymbol{\Sigma}_i\|_0 + \|\widehat{\boldsymbol{\Sigma}}_i - \boldsymbol{\Sigma}_i\|_0 = O(m\tau^{2/3})$. Then we define $\bar{\boldsymbol{\Sigma}}_i$ as

$$\big(\bar{\boldsymbol{\Sigma}}_i\big)_{jk} = \big|\big(\widetilde{\boldsymbol{\Sigma}}_i - \widehat{\boldsymbol{\Sigma}}_i\big)_{jk}\big| \quad \text{for all } j, k.$$

Then we have

$$\begin{aligned}
\|\mathbf{V}(\widetilde{\boldsymbol{\Sigma}}_i - \widehat{\boldsymbol{\Sigma}}_i)\widetilde{\mathbf{W}}\mathbf{x}_i\|_2 &\le \|\mathbf{V}(\widetilde{\boldsymbol{\Sigma}}_i - \widehat{\boldsymbol{\Sigma}}_i)\bar{\boldsymbol{\Sigma}}_i(\widetilde{\mathbf{W}} - \widehat{\mathbf{W}})\mathbf{x}_i\|_2 \\
&\le \|\mathbf{V}(\widetilde{\boldsymbol{\Sigma}}_i - \widehat{\boldsymbol{\Sigma}}_i)\|_2 \cdot \|\bar{\boldsymbol{\Sigma}}_i(\widetilde{\mathbf{W}} - \widehat{\mathbf{W}})\|_F \\
&= \|\mathbf{V}(\widetilde{\boldsymbol{\Sigma}}_i - \widehat{\boldsymbol{\Sigma}}_i)\|_2 \cdot \sqrt{\sum_{j=1}^m (\bar{\boldsymbol{\Sigma}}_i)_{jj}\|\widetilde{\mathbf{w}}_j - \widehat{\mathbf{w}}_j\|_2^2} \\
&\le \|\mathbf{V}(\widetilde{\boldsymbol{\Sigma}}_i - \widehat{\boldsymbol{\Sigma}}_i)\|_2 \cdot \|\bar{\boldsymbol{\Sigma}}_i\|_0^{1/2} \cdot \|\widetilde{\mathbf{W}} - \widehat{\mathbf{W}}\|_{2,\infty},
\end{aligned}$$

where $\widetilde{\mathbf{w}}_j$ and $\widehat{\mathbf{w}}_j$ denote the j-th columns of $\widetilde{\mathbf{W}}$ and $\widehat{\mathbf{W}}$ respectively. By Lemma B.3 and the fact that $\|\bar{\boldsymbol{\Sigma}}_i\|_0 = O(m\tau^{2/3})$, we have with probability $1 - O(m\tau^{2/3})$

$$\|\mathbf{V}(\widetilde{\boldsymbol{\Sigma}}_i - \widehat{\boldsymbol{\Sigma}}_i)\widetilde{\mathbf{W}}\mathbf{x}_i\|_2 \le O(m\sqrt{\log(m)}\tau^{2/3}k^{-1}) \cdot \|\widetilde{\mathbf{W}} - \widehat{\mathbf{W}}\|_{2,\infty}. \tag{B.16}$$

Therefore, we have

$$\begin{aligned} I_1 &\le \langle \nabla L(\widehat{\mathbf{W}}), \widetilde{\mathbf{W}} - \widehat{\mathbf{W}}\rangle + \frac{1}{2n}\sum_{i=1}^n \sqrt{\ell(f_{\widehat{\mathbf{W}}}(\mathbf{x}_i), \mathbf{y}_i)} \cdot \|\mathbf{V}(\widetilde{\boldsymbol{\Sigma}}_i - \widehat{\boldsymbol{\Sigma}}_i)\widetilde{\mathbf{W}}\mathbf{x}_i\|_2 \\ &\le \langle \nabla L(\widehat{\mathbf{W}}), \widetilde{\mathbf{W}} - \widehat{\mathbf{W}}\rangle + O(m\sqrt{\log(m)}\tau^{2/3}k^{-1/2}) \cdot \sqrt{L(\widehat{\mathbf{W}})} \cdot \|\widetilde{\mathbf{W}} - \widehat{\mathbf{W}}\|_{2,\infty}, \end{aligned}$$

where the last inequality follows from (B.16) and Young's inequality. In what follows we are going to tackle the term I_2. Note that for each i, we have

$$\begin{aligned} \|\mathbf{V}\widetilde{\boldsymbol{\Sigma}}_i\widetilde{\mathbf{W}}\mathbf{x}_i - \mathbf{V}\widehat{\boldsymbol{\Sigma}}_i\widehat{\mathbf{W}}\mathbf{x}_i\|_2 &= \|\mathbf{V}\widehat{\boldsymbol{\Sigma}}_i(\widetilde{\mathbf{W}} - \widehat{\mathbf{W}})\mathbf{x}_i\|_2 + \|\mathbf{V}(\widetilde{\boldsymbol{\Sigma}}_i - \widehat{\boldsymbol{\Sigma}}_i)\widetilde{\mathbf{W}}\mathbf{x}_i\|_2 \\ &\le \|\mathbf{V}\|_2\|\widetilde{\mathbf{W}} - \widehat{\mathbf{W}}\|_2 + \|\mathbf{V}(\widetilde{\boldsymbol{\Sigma}}_i - \widehat{\boldsymbol{\Sigma}}_i)\|_2 \cdot \|\widetilde{\mathbf{W}} - \widehat{\mathbf{W}}\|_2 \\ &= O(m^{1/2}/k^{1/2}) \cdot \|\widetilde{\mathbf{W}} - \widehat{\mathbf{W}}\|_2, \end{aligned}$$

where the last inequality holds due to the fact that $\|\mathbf{V}\|_2 = O(m^{1/2}/k^{1/2})$ with probability at least $1 - \exp(-O(m/k))$. This leads to $I_2 \le O(m/k) \cdot \|\widetilde{\mathbf{W}} - \widehat{\mathbf{W}}\|_2^2$. Now we can put everything together, and obtain

$$\begin{aligned} L(\widetilde{\mathbf{W}}) - L(\widehat{\mathbf{W}}) &= I_1 + I_2 \\ &\le \langle \nabla L(\widehat{\mathbf{W}}), \widetilde{\mathbf{W}} - \widehat{\mathbf{W}}\rangle + O(m\sqrt{\log(m)}\tau^{2/3}k^{-1/2}) \cdot \sqrt{L(\widehat{\mathbf{W}})} \cdot \|\widetilde{\mathbf{W}} - \widehat{\mathbf{W}}\|_{2,\infty} \\ &\quad + O(m/k) \cdot \|\widetilde{\mathbf{W}} - \widehat{\mathbf{W}}\|_2^2. \end{aligned}$$

Then applying union bound on the inequality for I_1 and I_2, we are able to complete the proof.

C Proof of Technical Lemmas in Appendix B

C.1 Proof of Lemma B.2

Let $\mathbf{z}_1, \ldots, \mathbf{z}_n \in \mathbb{R}^d$ be n vectors with $1/2 \le \min_i\{\|\mathbf{z}_i\|_2\} \le \max_i\{\|\mathbf{z}_i\|_2\} \le 2$. Let $\bar{\mathbf{z}}_i = \mathbf{z}_i/\|\mathbf{z}_i\|_2$ and assume $\min_{i,j}\|\bar{\mathbf{z}}_i - \bar{\mathbf{z}}_j\|_2, \min_{i,j}\|\bar{\mathbf{z}}_i + \bar{\mathbf{z}}_j\|_2 \ge \widetilde{\phi}$. For each $\mathbf{z}_i$, we construct an orthonormal matrix $\mathbf{Q}_i = [\bar{\mathbf{z}}_i, \mathbf{Q}_i'] \in \mathbb{R}^{d\times d}$. Then consider a random vector $\mathbf{w} \in \mathbb{R}^d$ following distribution $\mathcal{N}(0, \mathbf{I})$, it follows that $\mathbf{u}_i := \mathbf{Q}_i^\top\mathbf{w} \sim \mathcal{N}(0, \mathbf{I})$. Then we can decompose $\mathbf{w}$ as

$$\mathbf{w} = \mathbf{Q}_i\mathbf{u}_i = \mathbf{u}_i^{(1)}\bar{\mathbf{z}}_i + \mathbf{Q}_i'\mathbf{u}_i', \tag{C.1}$$

where $\mathbf{u}_i^{(1)}$ denotes the first coordinate of $\mathbf{u}_i$ and $\mathbf{u}_i' := [\mathbf{u}_i^{(2)}, \ldots, \mathbf{u}_i^{(d)}]^\top$. Then let $\gamma = \sqrt{\pi}\widetilde{\phi}/(8n)$, we define the following set of $\mathbf{w}$ based on $\mathbf{z}_i$,

$$\mathcal{W}_i = \big\{\mathbf{w} : |\mathbf{u}_i^{(1)}| \le \gamma, |\langle \mathbf{Q}_i'\mathbf{u}_i', \bar{\mathbf{z}}_j\rangle| \ge 2\gamma \text{ for all } \bar{\mathbf{z}}_j \text{ such that } j \ne i\big\}.$$

Regarding the class of sets $\{\mathcal{W}_1, \ldots, \mathcal{W}_n\}$, we have the following lemmas.

Lemma C.1. For any $\mathcal{W}_i$ and $\mathcal{W}_j$ with $i \ne j$, we have

$$\mathbb{P}(\mathbf{w} \in \mathcal{W}_i) \ge \frac{\widetilde{\phi}}{n\sqrt{128e}} \quad \text{and} \quad \mathcal{W}_i \cap \mathcal{W}_j = \emptyset.$$

Then we deliver the following two lemmas which are useful to establish the required lower bound.

Lemma C.2. For any $\mathbf{a} = (a_1, \ldots, a_n)^\top \in \mathbb{R}^n$, let $\mathbf{h}(\mathbf{w}) = \sum_{i=1}^n a_i\sigma'(\langle \mathbf{w}, \mathbf{z}_i\rangle)\mathbf{z}_i$ where $\mathbf{w} \sim N(\mathbf{0}, \mathbf{I})$ is a Gaussian random vector. Then it holds that

$$\mathbb{P}\big[\|\mathbf{h}(\mathbf{w})\|_2 \ge |a_i|/4 \big| \mathbf{w} \in \mathcal{W}_i\big] \ge 1/2.$$

Now we are able to prove Lemma B.2.

Proof of Lemma B.2. We first prove the result for any fixed $\mathbf{u}_1, \ldots, \mathbf{u}_n$. Then we define $a_i(\mathbf{v}_j) = \langle \mathbf{u}_i, \mathbf{v}_j \rangle$, $\mathbf{w}_j = \sqrt{m/2}\mathbf{w}_{L,j}^{(0)}$ and

$$\mathbf{h}(\mathbf{v}_j, \mathbf{w}_j) = \sum_{i=1}^{n} a_i(\mathbf{v}_j)\sigma'(\langle \mathbf{w}_j, \mathbf{x}_{L-1,i} \rangle)\mathbf{x}_{L-1,i}.$$

Then we define the event

$$\mathcal{E}_i = \big\{ j \in [m] : \mathbf{w}_j' \in \mathcal{W}_i, \|\mathbf{h}(\mathbf{v}_j, \mathbf{w}_j)\|_2 \geq |a_i(\mathbf{v}_j)|/4, |a_i(\mathbf{v}_j)| \geq \|\mathbf{u}_i\|_2/\sqrt{k} \big\}.$$

By Lemma B.1, we know that with high probability $1/2 \leq \|\mathbf{x}_{L-1,i}\|_2 \leq 2$ for all i and $\big\|\mathbf{x}_{L-1,i}/\|\mathbf{x}_{L-1,i}\|_2 - \mathbf{x}_{L-1,j}/\|\mathbf{x}_{L-1,j}\|_2\big\| \geq \phi/2$ and $\big\|\mathbf{x}_{L-1,i}/\|\mathbf{x}_{L-1,i}\|_2 + \mathbf{x}_{L-1,j}/\|\mathbf{x}_{L-1,j}\|_2\big\| \geq \phi/2$ for all $i \neq j$. Then by Lemmas C.1 and C.2 we know that $\mathcal{E}_i \cap \mathcal{E}_j = \emptyset$ if $i \neq j$ and

$$\begin{aligned} \mathbb{P}(j \in \mathcal{E}_i) &= \mathbb{P}\big[\|\mathbf{h}(\mathbf{v}_j, \mathbf{w}_j)\|_2 \geq |a_i(\mathbf{v}_j)|/4 | \mathbf{w}_j' \in \mathcal{W}_i\big] \cdot \mathbb{P}\big[\mathbf{w}_j' \in \mathcal{W}_i\big] \cdot \mathbb{P}\big[|a_i(\mathbf{v}_j)| \geq \|\mathbf{u}_i\|_2/\sqrt{k}\big] \\ &\geq \frac{\phi}{64\sqrt{2}en}, \end{aligned} \tag{C.2}$$

where the first equality holds because $\mathbf{w}_j$ and $\mathbf{v}_j$ are independent, and the second inequality follows from Lemmas C.1, C.2 and the fact that $\mathbb{P}(|a_i(\mathbf{v}_j)| \geq \|\mathbf{u}_i\|_2/\sqrt{k}) \geq 1/2$. Then we have

$$\begin{aligned} \|\nabla_{\mathbf{W}_L} L(\mathbf{W})\|_F^2 &= \frac{1}{n^2} \sum_{j=1}^{m} \|\mathbf{h}(\mathbf{v}_j, \mathbf{w}_j)\|_2^2 \\ &\geq \frac{1}{n^2} \sum_{j=1}^{m} \|\mathbf{h}(\mathbf{v}_j, \mathbf{w}_j)\|_2^2 \sum_{s=1}^{n} \mathbb{1}\left(j \in \mathcal{E}_s\right) \\ &\geq \frac{1}{n^2} \sum_{j=1}^{m} \sum_{s=1}^{n} \frac{\|\mathbf{u}_s\|_2^2}{16k} \mathbb{1}\left(j \in \mathcal{E}_s\right), \end{aligned}$$

where the second inequality holds due to the fact that

$$\begin{aligned} \|\mathbf{h}(\mathbf{v}_j, \mathbf{w}_j)\|_2^2 \mathbb{1}\left(j \in \mathcal{E}_s\right) &\geq \frac{a_s^2(\mathbf{v}_j)}{16} \mathbb{1}(|a_s(\mathbf{v}_j)| \geq \|\mathbf{u}_s\|_2/\sqrt{k}) \cdot \mathbb{1}(j \in \mathcal{E}_s) \\ &\geq \frac{\|\mathbf{u}_s\|_2^2}{16k} \mathbb{1}(j \in \mathcal{E}_s), \end{aligned}$$

where the first inequality follows from the definition of $\mathcal{E}_s$. Then we further define

$$Z_j = \sum_{s=1}^{n} \frac{\|\mathbf{u}_s\|_2^2}{16k} \mathbb{1}\left(j \in \mathcal{E}_s\right),$$

and provide the following results for $\mathbb{E}[Z(\mathbf{w}_j)]$ and $\mathrm{var}[Z(\mathbf{w}_j)]$

$$\mathbb{E}[Z_j] = \sum_{s=1}^{n} \frac{\|\mathbf{u}_s\|_2^2}{16k} \mathbb{P}\big(j \in \mathcal{E}_s\big), \qquad \mathrm{var}[Z(\mathbf{w})] = \sum_{s=1}^{n} \frac{\|\mathbf{u}_s\|_2^4}{256k^2} \mathbb{P}\big(j \in \mathcal{E}_s\big)\big[1 - \mathbb{P}\big(j \in \mathcal{E}_s\big)\big].$$

Then by Bernstein inequality, with probability at least $1 - \exp\big(-O\big(m\mathbb{E}[Z(\mathbf{w})]/\max_{i\in[n]} \|\mathbf{u}_i\|_2^2\big)\big)$, it holds that

$$\sum_{j=1}^{m} Z_j \geq \frac{m}{2}\mathbb{E}[Z_j] \geq \sum_{i=1}^{n} \frac{\|\mathbf{u}_i\|_2^2}{32k} \cdot \frac{m\phi}{64\sqrt{2}en} = \frac{C\phi m \sum_{i=1}^{n} \|\mathbf{u}_i\|_2^2}{kn},$$

where the second inequality follows from (C.2) and $C = 1/(2096\sqrt{2}e)$ is an absolute constant. Therefore, with probability at least $1 - \exp\big(-O(m\phi/(kn))\big)$ we have

$$\sum_{j=1}^{m} \left\| \frac{1}{n} \sum_{i=1}^{n} \langle \mathbf{u}_i, \mathbf{v}_j \rangle \sigma'\big(\langle \mathbf{w}_{L,j}, \mathbf{x}_{L-1,i} \rangle\big)\mathbf{x}_{L-1,i} \right\|_2^2 \geq \frac{1}{n^2} \sum_{j=1}^{m} Z(\mathbf{w}_j) \geq \frac{C\phi m \sum_{i=1}^{n} \|\mathbf{u}_i\|_2^2}{kn^3}.$$

Till now we have completed the proof for one particular vector collection $\{\mathbf{u}_i\}_{i=1,\ldots,n}$. Then we are going to prove that the above inequality holds for arbitrary $\{\mathbf{u}_i\}_{i=1,\ldots,n}$ with high probability. Taking ϵ-net over all possible vectors $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\} \in (\mathbb{R}^k)^n$ and applying union bound, the above inequality holds with probability at least $1 - \exp\big(-O(m\phi/(kn)) + \widetilde{O}(nk)\big)$. Since we have $m \geq \widetilde{O}\big(\phi^{-1}n^2k^2\big)$, the desired result holds for all choices of $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$.

□

D Proof of Auxiliary Lemmas in Appendix C

Proof of Lemma C.1. We first prove that any two sets $\mathcal{W}_i$ and $\mathcal{W}_j$ have not overlap region. Consider an vector $\mathbf{w} \in \mathcal{W}_i$ with the decomposition

$$\mathbf{w} = \mathbf{u}_i^{(1)} \bar{\mathbf{z}}_i + \mathbf{Q}_i' \mathbf{u}_i'.$$

Then based on the definition of $\mathcal{W}_i$ we have,

$$\langle \mathbf{w}, \bar{\mathbf{z}}_j \rangle = \langle \mathbf{u}_i^{(1)} \bar{\mathbf{z}}_i + \mathbf{Q}_i' \mathbf{u}_i', \bar{\mathbf{z}}_j \rangle = \mathbf{u}_i^{(1)} \langle \bar{\mathbf{z}}_i, \bar{\mathbf{z}}_j \rangle + \langle \mathbf{Q}_i' \mathbf{u}_i', \bar{\mathbf{z}}_j \rangle.$$

Since $\mathbf{w} \in \mathcal{W}_i$, we have $|\mathbf{u}_i^{(1)}| \le \gamma$ and $|\langle \mathbf{Q}' \mathbf{u}_i', \bar{\mathbf{z}}_j \rangle| \ge 2\gamma$. Therefore, note that $|\langle \bar{\mathbf{z}}_i, \bar{\mathbf{z}}_j \rangle| \le 1$, it holds that

$$|\langle \mathbf{w}, \bar{\mathbf{z}}_j \rangle| \ge \big| |\langle \mathbf{Q}_i' \mathbf{u}_i', \bar{\mathbf{z}}_j \rangle| - |\mathbf{u}_i^{(1)}| \big| > \gamma. \tag{D.1}$$

Note that set $\mathcal{W}_j$ requires $|\mathbf{u}_j^{(1)}| = \langle \mathbf{w}, \bar{\mathbf{z}}_j \rangle \le \gamma$, which conflicts with (D.1). This immediately implies that $\mathcal{W}_i \cap \mathcal{W}_j = \emptyset$.

Then we are going to compute the probability $\mathbb{P}(\mathbf{w} \in \mathcal{W}_i)$. Based on the parameter γ, we define the following two events

$$\mathcal{E}_1(\gamma) = \big\{ |\mathbf{u}_i^{(1)}| \le \gamma \big\},\ \mathcal{E}_2(\gamma) = \big\{ |\langle \mathbf{Q}_i' \mathbf{u}_i', \bar{\mathbf{z}}_j \rangle| \ge 2\gamma \text{ for all } \bar{\mathbf{z}}_j, j \ne i \big\}.$$

Evidently, we have $\mathbb{P}(\mathbf{w} \in \mathcal{W}_i) = \mathbb{P}(\mathcal{E}_1)\mathbb{P}(\mathcal{E}_2)$. Since $\mathbf{u}_i^{(1)}$ 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 $j = 1, \ldots, n$ we have

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

Note that for any $j \ne i$ we have $\|\bar{\mathbf{z}}_i - \bar{\mathbf{z}}_j\|_2 \ge \widetilde{\phi}$ and $\|\bar{\mathbf{z}}_i + \bar{\mathbf{z}}_j\|_2 \ge \widetilde{\phi}$, then it follows that

$$|\langle \bar{\mathbf{z}}_i, \bar{\mathbf{z}}_j \rangle| \le 1 - \widetilde{\phi}^2/2,$$

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

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

Therefore for any $j \ne i$,

$$\mathbb{P}[|\langle \mathbf{Q}_i' \mathbf{u}_i', \bar{\mathbf{z}}_j \rangle| < 2\gamma] = \frac{1}{\sqrt{2\pi}} \int_{-2[1-(\bar{\mathbf{z}}_i^\top \bar{\mathbf{z}}_j)^2]^{-1/2}\gamma}^{2[1-(\bar{\mathbf{z}}_i^\top \bar{\mathbf{z}}_j)^2]^{-1/2}\gamma} \exp\bigg(-\frac{1}{2} x^2 \bigg) \mathrm{d}x \le \sqrt{\frac{8}{\pi}} \frac{\gamma}{[1 - (\bar{\mathbf{z}}_i^\top \bar{\mathbf{z}}_j)^2]^{1/2}} \le \frac{4}{\sqrt{\pi}} \gamma \widetilde{\phi}^{-1}.$$

By union bound over $[n]$, we have

$$\mathbb{P}(\mathcal{E}_2) = \mathbb{P}[|\langle \mathbf{Q}_i' \mathbf{u}_i', \bar{\mathbf{z}}_j \rangle| \ge 2\gamma, j \in \mathcal{I}] \ge 1 - \frac{4}{\sqrt{\pi}} n \gamma \widetilde{\phi}^{-1}.$$

Therefore we have

$$\mathbb{P}(\mathbf{w} \in \mathcal{W}_i) \ge \sqrt{\frac{2}{\pi e}} \gamma \cdot \bigg(1 - \frac{4}{\sqrt{\pi}} n \gamma \widetilde{\phi}^{-1} \bigg).$$

Plugging $\gamma = \sqrt{\pi} \widetilde{\phi}/(8n)$, it holds that $\mathbb{P}(\mathcal{E}) \ge \widetilde{\phi}/(\sqrt{128e}n)$. This completes the proof. ☐

Proof of Lemma C.2. Recall the decomposition of $\mathbf{w}$ in (C.1),

$$\mathbf{w} = \mathbf{u}_i^{(1)} \bar{\mathbf{z}}_i + \mathbf{Q}_i' \mathbf{u}_i'.$$

Define the event $\mathcal{E}_i := \{\mathbf{w} \in \mathcal{W}_i\}$. Then conditioning on $\mathcal{E}_i$, we have

$$
\begin{aligned}
\mathbf{h}(\mathbf{w}) &= \sum_{i=1}^{n} a_i \sigma'(\langle \mathbf{w}, \mathbf{z}_i \rangle) \mathbf{z}_i \\
&= a_i \sigma'(\mathbf{u}_i^{(1)}) \mathbf{z}_i + \sum_{j \neq i} a_j \sigma'\big(\mathbf{u}_i^{(1)} \langle \bar{\mathbf{z}}_i, \bar{\mathbf{z}}_j \rangle + \langle \mathbf{Q}_i' \mathbf{u}_i', \bar{\mathbf{z}}_j \rangle\big) \mathbf{z}_j \\
&= a_i \sigma'(\mathbf{u}_i^{(1)}) \mathbf{z}_i + \sum_{j \neq i} a_j \sigma'\big(\langle \mathbf{Q}_i' \mathbf{u}_i', \mathbf{z}_j \rangle\big) \mathbf{z}_j
\end{aligned} \tag{D.2}
$$

where the last equality follows from the fact that conditioning on event $\mathcal{E}_i$, for all $j \neq i$, it holds that $|\langle \mathbf{Q}_i' \mathbf{u}_i', \bar{\mathbf{z}}_j \rangle| \ge 2\gamma > |\mathbf{u}_i^{(1)}| \ge |\mathbf{u}_i^{(1)} \langle \bar{\mathbf{z}}_i, \bar{\mathbf{z}}_j \rangle|$. We then consider two cases: $\mathbf{u}_i^{(1)} > 0$ and $\mathbf{u}_i^{(1)} < 0$, which occur equally likely conditioning on the event $\mathcal{E}_i$. Let $u_1 > 0$ and $u_2 < 0$ denote $\mathbf{u}_i^{(1)}$ in these two cases, we have

$$
\mathbb{P}\bigg[\|\mathbf{h}(\mathbf{w})\|_2 \ge \inf_{u_1 > 0, u_2 < 0} \max\big\{\big\|\mathbf{h}(u_1 \bar{\mathbf{z}}_i + \mathbf{Q}_i' \mathbf{u}_i')\big\|_2, \big\|\mathbf{h}(u_2 \bar{\mathbf{z}}_i + \mathbf{Q}_i' \mathbf{u}_i')\big\|_2\big\} \bigg| \mathcal{E}_i \bigg] \ge 1/2.
$$

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

$$
\mathbb{P}\bigg[\|\mathbf{h}(\mathbf{w})\|_2 \ge \inf_{u_1 > 0, u_1 < 0} \big\|\mathbf{h}(u_1 \bar{\mathbf{z}}_i + \mathbf{Q}_i' \mathbf{u}_i') - \mathbf{h}(u_2 \bar{\mathbf{z}}_i + \mathbf{Q}_i' \mathbf{u}_i')\big\|_2/2 \bigg| \mathcal{E}_i \bigg] \ge 1/2. \tag{D.3}
$$

For any $u_1 > 0$ and $u_2 < 0$, denote $\mathbf{w}_1 = u_1 \bar{\mathbf{z}}_i + \mathbf{Q}_i' \mathbf{u}_i'$, $\mathbf{w}_2 = u_2 \bar{\mathbf{z}}_i + \mathbf{Q}_i' \mathbf{u}_i'$. We now proceed to give lower bound for $\|\mathbf{h}(\mathbf{w}_1) - \mathbf{h}(\mathbf{w}_2)\|_2$. By (D.2), we have

$$
\|\mathbf{h}(\mathbf{w}_1) - \mathbf{h}(\mathbf{w}_2)\|_2 = \|a_i \mathbf{z}_i\|_2 \ge a_i/2, \tag{D.4}
$$

where we use the fact that $\|\mathbf{z}_i\|_2 \ge 1/2$. Plugging this back into (D.3), we have

$$
\mathbb{P}\big[\|\mathbf{h}(\mathbf{w})\|_2 \ge |a_i|/4 \big| \mathcal{E}_i\big] \ge 1/2.
$$

This completes the proof. $\square$