Generalization Error Bounds of Gradient Descent for Learning Over-parameterized Deep ReLU Networks

Yuan Cao[*] and Quanquan Gu[†]

Abstract

Empirical studies show that gradient-based methods can learn deep neural networks (DNNs) with very good generalization performance in the over-parameterization regime, where DNNs can easily fit a random labeling of the training data. While a line of recent work explains in theory that with over-parameterization and proper random initialization, gradient-based methods can find the global minima of the training loss for DNNs, it does not explain the good generalization performance of the gradient-based methods for learning over-parameterized DNNs. In this work, we take a step further, and prove that under certain assumption on the data distribution that is milder than linear separability, gradient descent (GD) with proper random initialization is able to train a sufficiently over-parameterized DNN to achieve arbitrarily small *expected* error (i.e., population error). This leads to an algorithmic-dependent generalization error bound for deep learning. To the best of our knowledge, this is the first result of its kind that can explain the good generalization performance of over-parameterized *deep* neural networks learned by gradient descent.

1 Introduction

Deep learning achieves great successes in almost all real-world applications ranging from image processing (Krizhevsky et al., 2012), speech recognition (Hinton et al., 2012) to Go games (Silver et al., 2016). Understanding and explaining the success of deep learning has thus become a central problem for theorists. One of the mysteries is that the neural networks used in practice are often heavily over-parameterized such that they can even fit random labels to the input data (Zhang et al., 2016), while they can still achieve very small generalization error (i.e., test error) when trained with real labels.

There are multiple recent attempts towards answering the above question and demystifying the success of deep learning. Soudry and Carmon (2016); Soudry and Hoffer (2017) explained why over-parametrization can remove bad local minima. Safran and Shamir (2016) showed that over-parametrization can improve the quality of the random initialization. Arora et al. (2018b) interpreted over-parametrization as a way of implicit acceleration during optimization. Haeffele and Vidal (2015); Nguyen and Hein (2017); Venturi et al. (2018) showed that for sufficiently over-parametrized networks, all local minima are global, but do not show how to find these minima via gradient descent. Li and Liang (2018); Du et al. (2018b) proved that with proper random

[*] Department of Computer Science, University of California, Los Angeles, CA 90095, USA; e-mail: yuancao@cs.ucla.edu

[†] Department of Computer Science, University of California, Los Angeles, CA 90095, USA; e-mail: qgu@cs.ucla.edu

initialization, (stochastic) gradient descent provably finds the global minimum for training over-parameterized one-hidden-layer ReLU networks. Du et al. (2018a) proved that gradient descent can converge to the global minima for over-parameterized deep nueral networks with smooth activation functions. Arora et al. (2018a) analyzed the convergence of GD to global optimum for training a deep linear neural network under a set of assumptions on the network width and initialization. Allen-Zhu et al. (2018c,b); Zou et al. (2018) proved the global convergence results of GD/SGD for deep neural networks with ReLU activation functions in the over-parameterization regime. However, in such an over-parametrized regime, the training loss function of deep neural networks may have potentially infinitely many global minima, but not all of them can generalize well. Hence, minimizing the training error is not sufficient to explain the good generalization performance of GD/SGD. There are only a few studies on the generalization theory for learning neural networks in the over-parameterization regime. Brutzkus et al. (2017) showed that SGD learns over-parameterized networks that provably generalize on linearly separable data. Li and Liang (2018) relaxed the linear separable data assumption and proved that SGD learns an over-parameterized network with a small generalization error when the data comes from mixtures of well-separated distributions. Neyshabur et al. (2018) proposed a novel complexity measure based on unit-wise capacities, and proved a tighter generalization bound for two-layer over-parameterized ReLU networks. However, this bound is independent of the specific training algorithms (e.g., GD or SGD). Allen-Zhu et al. (2018a) proved that under over-parameterization, SGD or its variants can learn some notable hypothesis classes, including two and three-layer neural networks with fewer parameters. Arora et al. (2019) provided a generalization bound of GD for two-layer ReLU networks based on a fine-grained analysis on how much the network parameters can move during GD. Nevertheless, all these results still do not explain the good generalization performance of gradient-based methods for learning over-parameterized *deep* neural networks.

In this paper, we aim to answer the following question:

Why gradient descent can learn an over-parameterized deep neural network with good generalization performance?

Without loss of generality, we focus on binary classification problems on the d-dimensional unit sphere S^{d-1}, and consider using gradient descent to solve the empirical risk minimization problem of learning deep fully connected neural networks with ReLU activation function and cross-entropy loss.

1.1 Our Main Results and Contributions

The following theorem gives an informal version of our main results.

Theorem 1.1 (Informal version of Theorem 4.4)**.** Under certain data distribution assumptions, for any $\epsilon > 0$, if the number of nodes per each hidden layer is set to $\widetilde{\Omega}(\max\{d, \epsilon^{-14}\})$ and the sample size $n = \widetilde{\Omega}(\epsilon^{-4})$, then with high probability, gradient descent with properly chosen step size and random initialization method learns a deep ReLU network and achieves a population classification error at most ϵ.

Theorem 1.1 holds for ReLU networks with arbitrary constant number of layers, as long as the data distribution satisfies certain separation conditions, which will be specified in Section 4.

Our contributions. Our main contributions can be summarized as follows:

- We show that, over-parameterized deep ReLU networks trained by gradient descent and random initialization provably generalize well. More specifically, we prove that, when solving the empirical loss minimization problem, with high probability, gradient descent starting from proper random initialization can find a network that gives arbitrarily small *population error*, as long as the number of hidden nodes per layer and the sample size are large enough. To the best of our knowledge, this is the first theoretical result in literature that reasonably explains the mysterious empirical observation that gradient-based methods can learn over-parameterized deep neural networks and achieve good generalization.

- Our result is based on a data distribution assumption that essentially requires that there exists a two-layer ReLU network with infinite hidden nodes and regularized weights that separates the two classes. This assumption is milder than linearly separable condition made in Brutzkus et al. (2017); Soudry et al. (2017).

- We provide a thorough landscape analysis for deep ReLU network that helps the study of both optimization and generalization. Compared with similar results given in earlier work (Allen-Zhu et al., 2018b; Zou et al., 2018) where the convergence guarantee heavily relies on a finite sample size n, our work removes this ill dependency on n and gives a series of results that hold uniformly over the input domain. These results characterize the geometry/landscape of population loss function near random initialization, which is of independent interest and may be used to revisit the optimization dynamics of (stochastic) gradient descent for training over-parameterized DNNs.

Comparison with most related work. Our result is most relevant to the recent line of work on the generalization for over-parameterized shallow (two- or three-layer) networks (Brutzkus et al., 2017; Li and Liang, 2018; Allen-Zhu et al., 2018a; Arora et al., 2019). While all these existing results only hold for two- or three-layer neural networks, our result holds for deep ReLU networks with arbitrarily many layers. Compared to standard uniform convergence based generalization error bounds (Neyshabur et al., 2015; Bartlett et al., 2017; Neyshabur et al., 2017; Golowich et al., 2017; Dziugaite and Roy, 2017; Arora et al., 2018c; Li et al., 2018a; Wei et al., 2018), the major advantage of our result is that, our generalization error bound is algorithm-dependent, almost independent of the width of the network, and provably achievable by gradient-based local search algorithms.

1.2 Notation

Throughout this paper, scalars, vectors and matrices are denoted by lower case, lower case bold face, and upper case bold face letters respectively. For a positive integer n, we denote $[n] = \{1, \ldots, n\}$. For a vector $\mathbf{x} = (x_1, \ldots, x_d)^\top$, we denote by $\|\mathbf{x}\|_p = \big(\sum_{i=1}^d |x_i|^p\big)^{1/p}$, $\|\mathbf{x}\|_\infty = \max_{i=1,\ldots,d} |x_i|$, and $\|\mathbf{x}\|_0 = |\{x_i : x_i \neq 0, i = 1, \ldots, d\}|$ the ℓ_p, ℓ_∞ and ℓ_0 norms of $\mathbf{x}$ respectively. We use $\mathrm{Diag}(\mathbf{x})$ to denote a square diagonal matrix with the entries of $\mathbf{x}$ on the main diagonal. For a matrix $\mathbf{A} = (A_{ij}) \in \mathbb{R}^{m\times n}$, we use $\|\mathbf{A}\|_2$ and $\|\mathbf{A}\|_F$ to denote the spectral norm (maximum singular value) and Frobenius norm of $\mathbf{A}$ respectively. We also denote by $\|\mathbf{A}\|_0$ the number of nonzero entries of $\mathbf{A}$. We denote by $S^{d-1} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 = 1\}$ the unit sphere in $\mathbb{R}^d$. For a function $f : \mathbb{R}^d \to \mathbb{R}$, we denote by $\|f(\cdot)\|_\infty = \inf\{C \geqslant 0 : |f(\mathbf{x})| \leqslant C$ for almost every $\mathbf{x}\}$ the essential supreme of f.

We use the following standard asymptotic notations. For two sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n = O(b_n)$ if $a_n \leqslant C_1 b_n$ for some absolute constant $C_1 > 0$, and $a_n = \Omega(b_n)$ if $a_n \geqslant C_2 b_n$ for some absolute constant $C_2 > 0$. In addition, we use $\widetilde{O}(\cdot)$ and $\widetilde{\Omega}(\cdot)$ to hide some logarithmic terms in Big-O and Big-Omega notations.

1.3 Organization of the Paper

In Section 2 we survey recent work most related to ours. In Section 3 we introduce basic definitions and some preliminary results. Our main theoretical results are presented in Section 4. We provide a proof sketch of our main theory in Section 5. In Section 6 we conclude the paper and briefly discuss potential future work. All remaining proofs are deferred to the appendix.

2 Additional Related Work

There is a huge body of literature towards building the foundations of deep learning, and we are not able to include every work in this paper. In this section, we briefly review and comment additional work that is most related to ours and was not discussed in Section 1.

Representation power of deep neural networks. A line of research has shown that deeper neural networks have higher expressive power (Telgarsky, 2015, 2016; Lu et al., 2017; Liang and Srikant, 2016; Yarotsky, 2017, 2018; Hanin, 2017; Hanin and Sellke, 2017) than shallow neural networks. This to certain extent explains the advantage of deep neural networks with over-parameterization. Lin and Jegelka (2018) proved that ResNet (He et al., 2016) with one hidden node per layer is a universal approximator to any Lebesgue integrable function.

Optimization landscape of neural networks. Many studies (Haeffele and Vidal, 2015; Kawaguchi, 2016; Freeman and Bruna, 2016; Hardt and Ma, 2016; Safran and Shamir, 2017; Xie et al., 2017; Nguyen and Hein, 2017; Soltanolkotabi et al., 2017; Zhou and Liang, 2017; Yun et al., 2017; Du and Lee, 2018; Venturi et al., 2018) investigated the optimization landscape of neural networks with different activation functions. However, these results only apply to one-hidden layer neural networks, or deep linear networks, or rely on some stringent assumptions on the data and/or activation functions. In fact, they do not hold for non-linear shallow neural networks (Yun et al., 2018a) or three-layer linear neural networks (Kawaguchi, 2016). Furthmore, Yun et al. (2018b) showed that small nonlinearities in activation functions create bad local minima in neural networks.

Implicit bias/regularization of GD and its variants. A bunch of papers (Gunasekar et al., 2017; Soudry et al., 2017; Ji and Telgarsky, 2018; Gunasekar et al., 2018a,b; Nacson et al., 2018; Li et al., 2018b) have studied implicit regularization/bias of GD, stochastic gradient descent (SGD) or mirror descent for matrix factorization, logistic regression, and deep linear networks. However, generalizing these results to deep non-linear neural networks turns out to be challenging and is still an open problem.

Connections between deep learning and kernel methods. Daniely (2017) uncovered the connection between deep neural networks with kernel methods and showed that SGD can learn a function that is comparable with the best function in the conjugate kernel space of the network. Jacot et al. (2018) showed that the evolution of a DNN during training can be described by a so-called neural tangent kernel, which makes it possible to study the training of DNNs in the functional space. Belkin et al. (2018); Liang and Rakhlin (2018) showed that good generalization performance of overfitted/interpolated classifiers is not only an intriguing feature for deep learning, but also for kernel methods.

Recovery guarantees for shallow neural networks. A series of work (Tian, 2017; Brutzkus and Globerson, 2017; Li and Yuan, 2017; Soltanolkotabi, 2017; Du et al., 2017a,b; Zhong et al., 2017;

Zhang et al., 2018) have attempted to study shallow one-hidden-layer neural networks with ground truth parameters, and proved recovery guarantees for gradient-based methods such as gradient descent (GD) and stochastic gradient descent (SGD). However, the assumption of the existence of ground truth parameters is not realistic and the analysis of the recovery guarantee can hardly be extended to deep neural networks. Moreover, many of these studies need strong assumptions on the input distribution such as Gaussian, sub-Gaussian or symmetric distributions.

Distributional view of over-parameterized networks. Mei et al. (2018); Chizat and Bach (2018); Sirignano and Spiliopoulos (2018); Rotskoff and Vanden-Eijnden (2018); Wei et al. (2018) took a distributional view of over-parametrized networks, used mean field analysis to show that the empirical distribution of the two-layer neural network parameters can be described as a Wasserstein gradient flow, and proved that Wasserstein gradient flow converges to global optimima under certain structural assumptions. However, their results are limited to two-layer infinitely wide neural networks. Very recently, Yang (2019) studied the scaling limit of wide multi-layer neural networks.

3 Preliminaries

In this paper, we study the binary classification problem with some unknown data distribution $\mathcal{D}$ over $\mathbb{R}^d \times \{+1,-1\}$. A data point $(\mathbf{x}, y)$ drawn from $\mathcal{D}$ consists of the input $\mathbf{x} \in \mathbb{R}^d$ and label $y \in \{+1,-1\}$. We denote by $\mathcal{D}_{\mathbf{x}}$ the marginal distribution of $\mathbf{x}$. Given an input $\mathbf{x}$, we consider predicting its corresponding label y using a deep neural network with the ReLU activation function $\sigma(z) := \max\{0, z\}$.

We consider L-hidden-layer neural networks with m_l hidden nodes on the l-th layer, $l = 1, \ldots, L$. We denote $m_0 = d$, and define

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

where $\sigma(\cdot)$ denotes the entry-wise ReLU activation function (with a slight abuse of notation), $\mathbf{W}_l = (\mathbf{w}_{l,1}, \ldots, \mathbf{w}_{l,m_l}) \in \mathbb{R}^{m_{l-1}\times m_l}$, $l = 1, \ldots, L$ are the weight matrices, and $\mathbf{v} \in (\mathbf{1}^\top, -\mathbf{1}^\top)^\top \in \{-1,+1\}^{m_L}$ is the fixed output layer weight vector with half 1 and half -1 entries. We denote by $\mathbf{W} = \{\mathbf{W}_l\}_{l=1}^L$ the collection of matrices $\mathbf{W}_1, \ldots, \mathbf{W}_L$.

Given n training examples $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ drawn independently from $\mathcal{D}$, the empirical loss minimization problem is defined as follows:

$$\min_{\mathbf{W}} L_S(\mathbf{W}) = \frac{1}{n}\sum_{i=1}^n \ell[y_i \cdot f_{\mathbf{W}}(\boldsymbol{x}_i)], \tag{3.1}$$

where $S = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ is the training sample set, and $\ell(z) = \log[1 + \exp(-z)]$ is the cross-entropy loss function.

3.1 Gradient Descent with Gaussian Initialization

Here we introduce the details of the algorithm we use to solve the empirical loss minimization problem (3.1).

Gaussian initialization. We say that the weight matrices $\mathbf{W}_1, \ldots, \mathbf{W}_L$ are generated via Gaussian initialization if for all $l = 1, \ldots, L$, each entries of $\mathbf{W}_l$ are generated independently from $N(0, 2/m_l)$.

Gradient descent. We study the generalization performance of deep ReLU networks trained by gradient descent. Let $\mathbf{W}_1^{(0)}, \ldots, \mathbf{W}_L^{(0)}$ be weight matrices generated via Gaussian initialization. We consider the following gradient descent update rule to solve the empirical loss minimization problem (3.1):

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

where $\eta > 0$ is the step size.

3.2 Matrix Product Representation for Deep ReLU Networks

Here we introduce the matrix product representation for deep ReLU networks, which plays an essential role in our analysis. Given parameter matrices $\mathbf{W}_1, \ldots, \mathbf{W}_L$ and an input $\mathbf{x}$, we set $\mathbf{x}_0 = \mathbf{x}$, and denote by $\mathbf{x}_l$ the output of the l-th layer of the ReLU network:

$$\mathbf{x}_l = \sigma(\mathbf{W}_l^\top \sigma(\mathbf{W}_{l-1}^\top \cdots \sigma(\mathbf{W}_1^\top \mathbf{x}) \cdots)),\ l = 1, \ldots, L.$$

We also define diagonal binary matrices as follows

$$\boldsymbol{\Sigma}_l(\mathbf{x}) = \mathrm{Diag}(\mathbb{1}\{\mathbf{w}_{l,1}^\top \mathbf{x}_{l-1} > 0\}, \ldots, \mathbb{1}\{\mathbf{w}_{l,m_l}^\top \mathbf{x}_{l-1} > 0\}),\ l = 1, \ldots, L.$$

Then we have the following representations for the neural network and its gradients:

$$f_{\mathbf{W}}(\mathbf{x}) = \mathbf{v}^\top \Bigg[\prod_{r=1}^{L} \boldsymbol{\Sigma}_r(\mathbf{x}) \mathbf{W}_r^\top \Bigg] \mathbf{x},\quad \nabla_{\mathbf{W}_l}[f_{\mathbf{W}}(\mathbf{x})] = \mathbf{x}_{l-1} \mathbf{v}^\top \Bigg[\prod_{r=l+1}^{L} \boldsymbol{\Sigma}_r(\mathbf{x}) \mathbf{W}_r^\top \Bigg] \boldsymbol{\Sigma}_l(\mathbf{x}),$$

where we use the following matrix product notation:

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

Since this paper studies the generalization performance, we frequently need to study the training examples $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ as well as a test example $(\mathbf{x}, y) \sim \mathcal{D}$. To distinguish the i-th example in the training sample and the l-th layer output of the test input $\mathbf{x}$, we implement the following notations:

- For $i = 1, \ldots, n$, $l = 1, \ldots, L$, we use $\boldsymbol{x}_i$ to denote the i-th training input, and $\boldsymbol{x}_{l,i}$ the output of the l-th layer with input $\boldsymbol{x}_i$.
- For $l = 1, \ldots, L$, we denote by $\mathbf{x}_l$ the output of the l-th layer with test input $\mathbf{x}$.

4 Main Theory

In this section we present our main result. We first introduce several assumptions.

Assumption 4.1. We assume that the input data are normalized: $\mathrm{supp}(\mathcal{D}_x) \subseteq S^{d-1}$.

Assumption 4.1 is widely made in most existing work on over-parameterized neural networks (Li and Liang, 2018; Allen-Zhu et al., 2018b; Du et al., 2017b, 2018b). This assumption can be relaxed to the case that $c_1 \leqslant \|\mathbf{x}\|_2 \leqslant c_2$ for all $\mathbf{x} \in \mathrm{supp}(\mathcal{D}_x)$, where $c_2 > c_1 > 0$ are absolute constants. Such relaxation will not affect our final generalization results.

Assumption 4.2. Denote by $p(\overline{\mathbf{u}})$ the density of standard Gaussian random vectors. Define

$$\mathcal{F} = \bigg\{ f(\mathbf{x}) = \int_{\mathbb{R}^d} c(\overline{\mathbf{u}})\sigma(\overline{\mathbf{u}}^\top \mathbf{x}) p(\overline{\mathbf{u}}) \mathrm{d}\overline{\mathbf{u}} : \|c(\cdot)\|_\infty \leqslant 1 \bigg\},$$

where $\sigma(\cdot)$ is the ReLU function. We assume that there exist an $f(\cdot) \in \mathcal{F}$ and a constant $\gamma > 0$ such that $y \cdot f(\mathbf{x}) \geqslant \gamma$ for all $(\mathbf{x}, y) \in \mathrm{supp}(\mathcal{D})$.

Assumption 4.2 essentially states that there exists a function in the function class $\mathcal{F}$ that gives a constant margin γ. $\mathcal{F}$ is a fairly large function class. In the definition, each value of $\overline{\mathbf{u}}$ can be considered as a node in an infinite-width single-hidden-layer ReLU network, and the corresponding product $c(\overline{\mathbf{u}})p(\overline{\mathbf{u}})$ can be considered as the second-layer weight. Therefore $\mathcal{F}$ consists of infinite-width single-hidden-layer ReLU networks whose second-layer weights decay faster than $p(\overline{\mathbf{u}})$. Assumption 4.2 is comparable with assumptions made in previous work:

- Assumption 4.2 covers the case when the data distribution is linearly separable. To show this, choosing $c(\overline{\mathbf{u}})$ to be a function of the form $[\overline{c}(\overline{\mathbf{u}}) - \overline{c}(-\overline{\mathbf{u}})]/2$ and $\|\overline{c}(\cdot)\|_\infty \leqslant 1$, we have

  $$f(\mathbf{x}) = \int_{\mathbb{R}^d} \overline{c}(\overline{\mathbf{u}})/2 \cdot [\sigma(\overline{\mathbf{u}}^\top \mathbf{x}) - \sigma(-\overline{\mathbf{u}}^\top \mathbf{x})] \cdot p(\overline{\mathbf{u}}) \mathrm{d}\overline{\mathbf{u}} = \bigg[\int_{\mathbb{R}^d} \overline{c}(\overline{\mathbf{u}})/2 \cdot \overline{\mathbf{u}} \cdot p(\overline{\mathbf{u}}) \mathrm{d}\overline{\mathbf{u}} \bigg]^\top \mathbf{x},$$

  which is a linear function of $\mathbf{x}$. So Assumption 4.2 is milder than linear separable assumption.

- $\mathcal{F}$ defined in Assumption 4.2 corresponds to the function class studied in Rahimi and Recht (2009) when the feature function is ReLU. In this sense, our work essentially studies the generalization performance of over-parameterized deep ReLU networks when the data can be classified by the Random Kitchen Sinks fitting procedure proposed by Rahimi and Recht (2009).

Assumption 4.3. Define $M = \max\{m_1, \ldots, m_L\}$, $m = \min\{m_1, \ldots, m_L\}$. We assume that $M/m = O(1)$.

We are now ready to present our main theoretical result.

Theorem 4.4. Suppose that $\mathbf{W}^{(0)} = \{\mathbf{W}_l^{(0)}\}_{l=1}^L$ is generated via Gaussian initialization. For any $\epsilon, \delta > 0$, there exist

$$m^*(\epsilon, d, L, \gamma, \delta) = \widetilde{O}(\mathrm{poly}(2^L, \gamma^{-1})) \cdot \max\{d, \epsilon^{-14}\} \cdot \log(1/\delta),$$
$$n^*(\epsilon, L, \gamma, \delta) = \widetilde{O}(\mathrm{poly}(2^L, \gamma^{-1})) \cdot \epsilon^{-4} \cdot \log(1/\delta)$$

such that, if $m \geqslant m^*(\epsilon, d, L, \gamma, \delta)$ and $n \geqslant n^*(\epsilon, L, \gamma, \delta)$, then with probability at least $1-\delta$, gradient descent initialized by $\mathbf{W}^{(0)}$ with step size $\eta = O(16^{-L} L^{-6} \gamma^4 m^{-1})$ finds a point $\mathbf{W}^{(k)}$ that satisfies

$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big[y \cdot f_{\mathbf{W}^{(k)}}(\mathbf{x}) > 0\big] \geqslant 1 - \epsilon$$

within $K = \widetilde{O}(\mathrm{poly}(2^L, \gamma^{-1})) \cdot \epsilon^{-2}$ iterations.

Here are a few remarks on Theorem 4.4.

Remark 4.5. In our setting the number of hidden layers L and the margin γ can both be considered as constants. However for the sake of completeness we still give the detailed dependency of n, m, η, K on L and γ. Our current result has an exponential dependency on L, which can potentially be further improved. We leave this as a future research direction. It is worth noting that such exponential dependency on L also appears in existing uniform convergence based generalization bounds (Neyshabur et al., 2015; Bartlett et al., 2017; Neyshabur et al., 2017; Golowich et al., 2017; Dziugaite and Roy, 2017; Arora et al., 2018c; Li et al., 2018a), since under our over-parameterization setting the spectral norms of weight matrices are constants greater than one.

Remark 4.6. Theorem 4.4 suggests an $n = \widetilde{\Omega}(\mathrm{poly}(2^L, \gamma^{-1})) \cdot \epsilon^{-4}$ sample complexity, which is almost (up to a logarithmic factor) independent of the minimum number of nodes per layer m. In terms of the condition $m = \widetilde{\Omega}(\mathrm{poly}(2^L, \gamma^{-1})) \cdot \max\{d, \epsilon^{-14}\} \cdot \log(1/\delta)$, we would like to point out that with a proof similar to the proof given in Zou et al. (2018), the dimension d can in fact easily be replaced by a $\log(n)$ term, since our optimization analysis focuses on the optimization of empirical loss. However, since the condition $m \geqslant d$ is commonly satisfied in practice, here we choose to present the version $m = \widetilde{\Omega}(\mathrm{poly}(2^L, \gamma^{-1})) \cdot \max\{d, \epsilon^{-14}\} \cdot \log(1/\delta)$ so that our detailed analysis can be directly extended to population loss. It is worth noting that out sample complexity result $n = \widetilde{\Omega}(\mathrm{poly}(2^L, \gamma^{-1})) \cdot \epsilon^{-4}$ is still dimension-free.

Remark 4.7. Theorem 4.4 holds for the over-parameterization setting in the sense that in order to provably achieve ϵ expected/population error, the minimum number of nodes per layer m should be chosen to be at least $O(\epsilon^{-14})$. This differs from the optimization results for over-parameterized deep ReLU networks given by Allen-Zhu et al. (2018b); Du et al. (2018b); Zou et al. (2018), where m is required to be of order $\widetilde{O}(\mathrm{poly}(n, \epsilon^{-1}))$. We remark that such polynomial dependency on the sample size n hinders generalization analysis, since it will lead to loose generalization error bounds. In stark contrast, our derived over-parameterization condition is independent of the sample size n and is the key to generalization analysis.

5 Proof Sketch of the Main Theory

In this section we sketch the proof of Theorem 4.4, and provide the insights of the analysis techniques. For the ease of exposition, we first introduce two auxiliary definitions.

Definition 5.1. For a collection of parameter matrices $\mathbf{W} = \{\mathbf{W}_l\}_{l=1}^L$, we call the set

$$\mathcal{W}(\mathbf{W}, \tau) := \big\{\widetilde{\mathbf{W}} = \{\widetilde{\mathbf{W}}_l\}_{l=1}^L : \|\widetilde{\mathbf{W}}_l - \mathbf{W}_l\|_F \leqslant \tau,\ l = 1, \ldots, L\big\}$$

the τ-neighborhood of $\mathbf{W}$.

The definition of $\mathcal{W}(\mathbf{W}, \tau)$ is motivated by the observation that in a small neighborhood of initialization, deep ReLU networks satisfy good scaling and landscape properties. It also provides a small parameter space and enables sharper bound based on Rademacher complexity for the generalization gap between empirical and expected/population errors.

Definition 5.2. For a collection of parameter matrices $\mathbf{W} = \{\mathbf{W}_l\}_{l=1}^L$, we define its empirical surrogate error $\mathcal{E}_S(\mathbf{W})$ and population surrogate error $\mathcal{E}_\mathcal{D}(\mathbf{W})$ as follows:

$$\mathcal{E}_S(\mathbf{W}) := -\frac{1}{n}\sum_{i=1}^n \ell'\big[y_i \cdot f_{\mathbf{W}}(\boldsymbol{x}_i)\big],\ \mathcal{E}_\mathcal{D}(\mathbf{W}) := \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\big\{-\ell'\big[y \cdot f_{\mathbf{W}}(\mathbf{x})\big]\big\}.$$

The intuition behind the definition of surrogate error is that, for cross-entropy loss we have $-\ell'(z) = 1/[1+\exp(-z)]$, which can be seen as a smooth version of the indicator function $\mathbb{1}\{z<0\}$, and therefore $-\ell'[y\cdot f_{\mathbf{W}}(\mathbf{x})]$ is related to the classification error of the classifier. Surrogate error plays a pivotal role in our generalization analysis: on the one hand, it is closely related to the derivative of the empirical loss function. On the other hand, by $-2\ell'(z) \geqslant \mathbb{1}\{z<0\}$, it also provides an upper bound on the classification error. It is worth noting that the surrogate error is comparable with the ramp loss studied in margin-based generalization error bounds (Neyshabur et al., 2015; Bartlett et al., 2017; Neyshabur et al., 2017; Golowich et al., 2017; Arora et al., 2018c; Li et al., 2018a) in the sense that it is Lipschitz continuous in $\mathbf{W}$, which ensures that $\mathcal{E}_S(\mathbf{W})$ concentrates around $\mathcal{E}_{\mathcal{D}}(\mathbf{W})$ uniformly over the parameter space $\mathcal{W}(\mathbf{W},\tau)$.

The proof of Theorem 4.4 consists of two parts:

- In Section 5.1, we study the scaling and landscape properties of Deep ReLU networks with parameter $\widetilde{\mathbf{W}} \in \mathcal{W}(\mathbf{W},\tau)$ for small enough τ, where $\mathbf{W}$ is the random parameter generated via Gaussian initialization.
- In Section 5.2, we focus on the gradient descent iterates $\mathbf{W}^{(0)},\ldots,\mathbf{W}^{(K)}$ where $\mathbf{W}^{(0)}$ is generated via Gaussian initialization, and establish a connection from gradient descent to empirical surrogate error, further to population surrogate error, and finally to population error to complete the proof.

5.1 Scaling and Landscape Analysis Around Initialization

Here we first study the properties of deep ReLU networks with parameter $\widetilde{\mathbf{W}} \in \mathcal{W}(\mathbf{W},\tau)$ for small enough τ, where $\mathbf{W}$ is generated via Gaussian initialization.

Scaling properties at random initialization. We first study the scaling properties of ReLU networks when all parameter matrices are generated via Gaussian initialization. Define

$$\Gamma_{l_1,l_2,s}(\mathbf{x}) = \sup_{\substack{\|\mathbf{a}\|_2=\|\mathbf{b}\|_2=1,\\ \|\mathbf{a}\|_0,\|\mathbf{b}\|_0\leqslant s}} \mathbf{b}^\top\mathbf{W}_{l_2}^\top\Bigg(\prod_{r=l_1}^{l_2-1}\mathbf{\Sigma}_r(\mathbf{x})\mathbf{W}_r^\top\Bigg)\mathbf{a},\quad 1\leqslant l_1<l_2\leqslant L,$$

$$\Gamma'_{l_1,l_2,s}(\mathbf{x}) = \sup_{\substack{\|\mathbf{a}\|_2=\|\mathbf{b}\|_2=1,\\ \|\mathbf{a}\|_0\leqslant s}} \mathbf{b}^\top\mathbf{W}_{l_2}^\top\Bigg(\prod_{r=l_1}^{l_2-1}\mathbf{\Sigma}_r(\mathbf{x})\mathbf{W}_r^\top\Bigg)\mathbf{a},\quad 1\leqslant l_1<l_2\leqslant L,$$

$$\Gamma''_{l_1,l_2,s}(\mathbf{x}) = \sup_{\substack{\|\mathbf{a}\|_2=\|\mathbf{b}\|_2=1,\\ \|\mathbf{b}\|_0\leqslant s}} \mathbf{b}^\top\mathbf{W}_{l_2}^\top\Bigg(\prod_{r=l_1}^{l_2-1}\mathbf{\Sigma}_r(\mathbf{x})\mathbf{W}_r^\top\Bigg)\mathbf{a},\quad 1\leqslant l_1<l_2\leqslant L,$$

$$\Gamma'''_{l,s}(\mathbf{x}) = \sup_{\|\mathbf{a}\|_2=1,\|\mathbf{a}\|_0\leqslant s} \mathbf{v}^\top\Bigg(\prod_{r=l}^{L}\mathbf{\Sigma}_r(\mathbf{x})\mathbf{W}_r^\top\Bigg)\mathbf{a},\quad 1\leqslant l\leqslant L,$$

$$\Lambda_{l_1,l_2}(\mathbf{x}) = \sup_{\|\mathbf{a}\|_2=\|\mathbf{b}\|_2=1} \mathbf{b}^\top\mathbf{W}_{l_2}^\top\Bigg(\prod_{r=l_1}^{l_2-1}\mathbf{\Sigma}_r(\mathbf{x})\mathbf{W}_r^\top\Bigg)\mathbf{a},\quad 1\leqslant l_1<l_2\leqslant L,$$

where in the definition of $\Gamma_{l_1,l_2,s}(\mathbf{x})$, $\Gamma'_{l_1,l_2,s}(\mathbf{x})$, $\Gamma''_{l_1,l_2,s}(\mathbf{x})$ and $\Gamma'''_{l,s}(\mathbf{x})$ we use a parameter s to measure the sparsity of vectors $\mathbf{a}$ and/or $\mathbf{b}$. Intuitively, these sparsity-based bounds can be combined with activation pattern analysis to give refined bounds on the network scaling. The following theorem summarizes the main scaling properties at Gaussian initialization.

Theorem 5.3. Suppose that $\mathbf{W}_1, \ldots, \mathbf{W}_L$ are generated via Gaussian initialization. There exists absolute constants $\overline{C}, \overline{C}' > 0$ such that for any $\delta > 0$, if

$$m \geqslant \overline{C} \max\{L^2 s, L^4 (m/s)^2 d, L^8 d [\log(m)]^2\} \cdot \log(m/\delta), \quad s \geqslant \overline{C} \max\{d \log(m), \log(L/\delta)\}$$

for some large enough absolute constant $\overline{C}$, then with probability at least $1-\delta$, the following results hold:

(i) $\|\mathbf{W}_l\|_2 \leqslant \overline{C}'$, $\big|\|\mathbf{x}_l\|_2 - 1\big| \leqslant \overline{C}' L \sqrt{d \log(m/\delta)/m}$ for all $\mathbf{x} \in S^{d-1}$ and $l \in [L]$

(ii) $\Gamma_{l_1,l_2,s}(\mathbf{x}) \leqslant \overline{C}' \sqrt{s \log(m)/m}$, $\Gamma'_{l_1,l_2,s}(\mathbf{x}), \Gamma''_{l_1,l_2,s}(\mathbf{x}) \leqslant \overline{C}'$, $\Gamma'''_{l,s}(\mathbf{x}) \leqslant \overline{C}' \sqrt{s \log(m)}$, $\Lambda_{l_1,l_2}(\mathbf{x}) \leqslant \overline{C}' L$ for all $\mathbf{x} \in S^{d-1}$, $l \in [L]$ and $1 \leqslant l_1 < l_2 \leqslant L$

(iii) $\|\mathbf{x}_l - \mathbf{x}'_l\|_2 \leqslant \overline{C}' L \|\mathbf{x} - \mathbf{x}'\|_2$ for all $\mathbf{x}, \mathbf{x}' \in S^{d-1}$ and $l \in [L]$, where $\mathbf{x}_l$, $\mathbf{x}'_l$ denote the output of the l-th layer of the network at initialization with inputs $\mathbf{x}$, $\mathbf{x}'$ respectively.

(iv) $|f_{\mathbf{W}}(\boldsymbol{x}_i)| \leqslant \overline{C}' \sqrt{\log(n/\delta)}$ for all $i \in [n]$.

It is worth noting that the condition of m given in Theorem 5.3 requires that $m \geqslant \overline{C} L^4 (m/s)^2 d$, which, for fixed s, is an upper bound of m. However, throughout our proof, whenever we apply Theorem 5.3 we always use s of order $s = m \cdot \mathrm{poly}(2^L, \gamma^{-1}, \log(m)) \cdot d \cdot \log(1/\delta)$. Therefore the condition on m in Theorem 5.3 is essentially $m = \Omega(\mathrm{poly}(2^L, \gamma^{-1}, \log(m))) \cdot d \cdot \log(1/\delta)$.

Uniform scaling analysis over $\mathcal{W}(\mathbf{W}, \tau)$. Based on the scaling analysis at initialization, we now show that for small enough τ, the key scaling properties obtained at initialized parameter $\mathbf{W}$ in fact hold around the τ-neighborhood of $\mathbf{W}$.

Theorem 5.4. Suppose that $\mathbf{W}_1, \ldots, \mathbf{W}_L$ are generated via Gaussian initialization. There exist absolute constants $\overline{C}, \overline{C}', \underline{C}$, such that for any $\delta > 0$, if

$$m \geqslant \overline{C} \max\{d L^2 \log(m/\delta), L^{-8/3} \tau^{-4/3} \log[m/(\tau\delta)]\}, \quad \tau \leqslant \underline{C} L^{-5} [\log(m)]^{-3/2},$$

then with probability at least $1-\delta$ the following results hold:

(i) $\|\widehat{\mathbf{x}}_l - \widetilde{\mathbf{x}}_l\|_2 \leqslant \overline{C}' L \cdot \sum_{r=1}^{l} \|\widehat{\mathbf{W}}_r - \widetilde{\mathbf{W}}_r\|_2$, $\|\widehat{\boldsymbol{\Sigma}}_l(\mathbf{x}) - \widetilde{\boldsymbol{\Sigma}}_l(\mathbf{x})\|_0 \leqslant \overline{C}' L^{4/3} \tau^{2/3} m_l$ for all $\widetilde{\mathbf{W}}, \widehat{\mathbf{W}} \in \mathcal{W}(\mathbf{W}, \tau)$, $\mathbf{x} \in S^{d-1}$ and $l \in [L]$;

(ii) $\big\| \prod_{r=l_1}^{l_2} \widetilde{\boldsymbol{\Sigma}}_r(\mathbf{x}) \widetilde{\mathbf{W}}_r^\top \big\|_2 \leqslant \overline{C}' L$ for all $\widetilde{\mathbf{W}} \in \mathcal{W}(\mathbf{W}, \tau)$, $\mathbf{x} \in S^{d-1}$ and $1 \leqslant l_1 < l_2 \leqslant L$;

(iii) $\mathbf{v}^\top \big(\prod_{r=l}^{L} \widetilde{\boldsymbol{\Sigma}}_r(\mathbf{x}) \widetilde{\mathbf{W}}_r^\top \big) \mathbf{a} \leqslant \overline{C}' L^{2/3} \tau^{1/3} \sqrt{m \log(m)}$ for all $\mathbf{a} \in \mathbb{R}^{m_{l-1}}$ satisfying $\|\mathbf{a}\|_2 = 1$, $\|\mathbf{a}\|_0 \leqslant \overline{C}' L^{4/3} \tau^{2/3} m_l$, $\widetilde{\mathbf{W}} \in \mathcal{W}(\mathbf{W}, \tau)$, $\mathbf{x} \in S^{d-1}$ and $1 \leqslant l \leqslant L$,

(iv) $\big| f_{\widehat{\mathbf{W}}}(\mathbf{x}) - F_{\widetilde{\mathbf{W}}, \widehat{\mathbf{W}}}(\mathbf{x}) \big| \leqslant \overline{C}' L^{8/3} \tau^{1/3} \sqrt{m \log(m)} \cdot \sum_{l=1}^{L} \big\| \widehat{\mathbf{W}}_l - \widetilde{\mathbf{W}}_l \big\|_2 + \overline{C}' L^3 \sqrt{m} \cdot \sum_{l=1}^{L} \|\widehat{\mathbf{W}}_l - \widetilde{\mathbf{W}}_l\|_2^2$ for all $\widetilde{\mathbf{W}}, \widehat{\mathbf{W}} \in \mathcal{W}(\mathbf{W}, \tau)$ and all $\mathbf{x} \in S^{d-1}$,

where $\widetilde{\mathbf{x}}_l$, $\widehat{\mathbf{x}}_l$ are the outputs of the l-th hidden layer of the ReLU network with input $\mathbf{x}$ and weight matrices $\widetilde{\mathbf{W}}$, $\widehat{\mathbf{W}}$ respectively, $\widetilde{\mathbf{x}}_0 = \widehat{\mathbf{x}}_0 = \mathbf{x}$,

$$\widetilde{\boldsymbol{\Sigma}}_l(\mathbf{x}) = \mathrm{Diag}\big(\mathbb{1}\{\widetilde{\mathbf{w}}_{l,1}^\top \widetilde{\mathbf{x}}_{l-1} > 0\}, \ldots, \mathbb{1}\{\widetilde{\mathbf{w}}_{l,m_l}^\top \widetilde{\mathbf{x}}_{l-1} > 0\} \big),$$
$$\widehat{\boldsymbol{\Sigma}}_l(\mathbf{x}) = \mathrm{Diag}\big(\mathbb{1}\{\widehat{\mathbf{w}}_{l,1}^\top \widehat{\mathbf{x}}_{l-1} > 0\}, \ldots, \mathbb{1}\{\widehat{\mathbf{w}}_{l,m_l}^\top \widehat{\mathbf{x}}_{l-1} > 0\} \big),$$

$l = 1, \ldots, L$, and

$$F_{\widetilde{\mathbf{W}},\widehat{\mathbf{W}}}(\mathbf{x}) := f_{\widetilde{\mathbf{W}}}(\mathbf{x}) + \sum_{l=1}^{L} \mathrm{Tr}\big[(\widehat{\mathbf{W}}_l - \widetilde{\mathbf{W}}_l)^\top \nabla_{\mathbf{W}_l} f_{\widetilde{\mathbf{W}}}(\mathbf{x})\big]$$

is the first order approximation of $f_{\widehat{\mathbf{W}}}(\mathbf{x})$ around $\widetilde{\mathbf{W}}$ in the parameter space.

The results of Theorems 5.3 and 5.4 are comparable with similar results given in Allen-Zhu et al. (2018b) and Zou et al. (2018). However, these previous results only give scaling properties over the training set $S = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$, while our results hold uniformly for all $\mathbf{x}$ in the input domain. As a consequence, our results not only lead to nice landscape properties of the empirical loss, but also play an important role in showing the concentration of empirical surrogate error around the population surrogate error.

Optimization landscape over $\mathcal{W}(\mathbf{W}, \tau)$. We are now ready to present the results on the landscape of the empirical loss function in the τ-neighborhood of initialization.

Theorem 5.5. There exist absolute constants $\overline{C}, \overline{C}', \underline{C}, \underline{C}'$ such that for any $\delta > 0$, if

$$m \geqslant \overline{C} \max\{dL^2 \log(m/\delta), L^{-8/3}\tau^{-4/3} \log[m/(\tau\delta)], 4^L \cdot L^4 \gamma^{-2} d \log[dL/(\gamma\delta)]\},$$
$$\tau \leqslant \underline{C} \min\{L^{-5}[\log(m)]^{-3/2}, 8^{-(L+1)} L^{-2}\gamma^3\},$$

then with probability at least $1 - \delta$,

(i) $\big\|\nabla_{\mathbf{W}_l} L_S(\widetilde{\mathbf{W}})\big\|_F \leqslant \overline{C}' L \cdot \sqrt{m} \cdot \mathcal{E}_S(\widetilde{\mathbf{W}})$ for all $\widetilde{\mathbf{W}} \in \mathcal{W}(\mathbf{W}, \tau)$ and $l = 1, \ldots, L$.

(ii) $\big\|\nabla_{\mathbf{W}_L} L_S(\widetilde{\mathbf{W}})\big\|_F \geqslant \underline{C}' 4^{-L} \cdot \gamma^2 \cdot \sqrt{m_L} \cdot \mathcal{E}_S(\widetilde{\mathbf{W}})$ for all $\widetilde{\mathbf{W}} \in \mathcal{W}(\mathbf{W}, \tau)$.

(iii) For all $\widetilde{\mathbf{W}}, \widehat{\mathbf{W}} \in \mathcal{W}(\mathbf{W}, \tau)$, it holds that

$$\begin{aligned} L_S(\widehat{\mathbf{W}}) - L_S(\widetilde{\mathbf{W}}) \leqslant C \sum_{l=1}^{L} L^{8/3} \tau^{1/3} \sqrt{m \log(m)} \cdot \big\|\widehat{\mathbf{W}}_l - \widetilde{\mathbf{W}}_l\big\|_2 \cdot \mathcal{E}_S(\widetilde{\mathbf{W}}) \\ + C \sum_{l=1}^{L} mL^3 \|\widehat{\mathbf{W}}_l - \widetilde{\mathbf{W}}_l\|_2^2 + \sum_{l=1}^{L} \mathrm{Tr}[(\widehat{\mathbf{W}}_l - \widetilde{\mathbf{W}}_l)^\top \nabla_{W_l} L_S(\widetilde{\mathbf{W}})]. \end{aligned}$$

Theorem 5.5 studies the landscape of the empirical loss function in three aspects: (i) gives upper bounds for the empirical gradients at each layer, (ii) gives a lower bound for the last layer derivative, and (iii) studies the smoothness property of $L_S(\cdot)$. All the results relate the empirical loss function to the empirical surrogate loss $\mathcal{E}_S(\widetilde{\mathbf{W}})$. It is worth noting that, compared with similar results for over-parameterized optimization problems in Allen-Zhu et al. (2018b); Zou et al. (2018), our results have more reasonable dependency on sample size n: if we formally let n go to infinity, the gradient lower bounds given in Allen-Zhu et al. (2018b) and Zou et al. (2018) both vanishes to zero; the objective semi-smoothness result given by Allen-Zhu et al. (2018b) also explodes and does not provide any meaningful result. In comparison, our results directly guarantee the same landscape properties of the population loss function as the empirical loss function.

5.2 Generalization Guarantee for Gradient Descent

We now study gradient descent starting at $\mathbf{W}^{(0)}$ generated via Gaussian initialization. Our analysis consists of the following two parts:

- We show that if the minimum number of nodes per layer m is large enough and the step size η is chosen properly, the iterates of gradient descent stay inside $\mathcal{W}(\mathbf{W}^{(0)}, \tau)$ for small enough τ. Moreover, gradient descent is able to find a point $\mathbf{W}^{(k)}$ such that the empirical surrogate error $\mathcal{E}_S(\mathbf{W}^{(k)})$ is small.
- We establish the relation between the empirical surrogate error $\mathcal{E}_S(\mathbf{W})$ and the population error $\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big[y \cdot f_{\mathbf{W}}(\mathbf{x}) < 0\big]$ uniformly for all $\mathbf{W} \in \mathcal{W}(\mathbf{W}^{(0)}, \tau)$, and show that when the sample size n is large enough, a small empirical surrogate error implies small expected/population error.

Convergence of gradient descent. We first give the convergence result of gradient descent based on the landscape analysis given by Theorems 5.4 and 5.5.

Theorem 5.6. Suppose that $\mathbf{W}_1^{(0)}, \ldots, \mathbf{W}_L^{(0)}$ are generated via Gaussian initialization. For any $\epsilon, \delta > 0$, there exists

$$m_1^* = \widetilde{O}(\mathrm{poly}(2^L, \gamma^{-1})) \cdot \max\{d, \epsilon^{-2}\} \cdot \log(1/\delta)$$

such that, for any $m \geqslant m_1^*$, with probability at least $1 - \delta/2$, gradient descent starting at $\mathbf{W}^{(0)}$ with step size $\eta = O(16^{-L} L^{-6} \gamma m^{-1})$ generates $K = \widetilde{O}(\mathrm{poly}(2^L, \gamma^{-1})) \cdot \epsilon^{-2}$ iterates $\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(K)}$ that satisfy:

(i) $\mathbf{W}^{(k)} \in \mathcal{W}(\mathbf{W}^{(0)}, \tau)$ for all $k = 0, \ldots, K$, where $\tau = O(16^L \cdot \gamma^{-4} \cdot \sqrt{\log(n/\delta)} \cdot \epsilon^{-1} m^{-1/2})$.

(ii) There exists $k \in \{0, \ldots, K-1\}$ such that $\mathcal{E}_S(\mathbf{W}^{(k)}) \leqslant \epsilon/4$.

Theorem 5.6 suggests that gradient descent is able to find a point $\mathbf{W}^{(k)}$ which gives small empirical surrogate error $\mathcal{E}_S(\mathbf{W}^{(k)})$ without escaping from $\mathcal{W}(\mathbf{W}^{(0)}, \tau)$. In the next step, we relate $\mathcal{E}_S(\mathbf{W}^{(k)})$ with the population error using a uniform convergence argument.

Population error bound and sample complexity. The following theorem gives an upper bound of the population error based on the empirical surrogate error $\mathcal{E}_S(\mathbf{W})$ and the sample size n.

Theorem 5.7. Suppose that $\mathbf{W}^{(0)} = \{\mathbf{W}_l^{(0)}\}_{l=1}^L$ is generated via Gaussian initialization and the results of Theorem 5.4 and Theorem 5.5 hold with $\mathbf{W} = \mathbf{W}^{(0)}$. Then there exists an absolute constant $\overline{C}$ and $m_2^* = \widetilde{O}(\mathrm{poly}(2^L, \gamma^{-1})) \cdot \epsilon^{-14}$ such that, for any $\epsilon, \delta > 0$, if $m \geqslant m_2^*$, then with probability at least $1 - \delta/2$,

$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big[y \cdot f_{\mathbf{W}}(\mathbf{x}) < 0\big] \leqslant 2 \cdot \mathcal{E}_S(\mathbf{W}) + \overline{C}\frac{16^L \cdot L^2 \cdot \gamma^{-4} \cdot \sqrt{\log(n/\delta)} \cdot \epsilon^{-1}}{\sqrt{n}} + \frac{\epsilon}{2}$$

for all $\mathbf{W} \in \mathcal{W}(\mathbf{W}^{(0)}, \tau)$, where $\tau = O(16^L \cdot \gamma^{-4} \cdot \sqrt{\log(n/\delta)} \cdot \epsilon^{-1} m^{-1/2})$.

By putting all the above pieces together, we are now able to prove Theorem 4.4.

Proof of Theorem 4.4. Let m_1^* and m_2^* be the necessary number of nodes per layer defined in Theorem 5.6 and Theorem 5.7 respectively, and set

$$\begin{aligned} m^* &= \max\{m_1^*, m_2^*\} = \widetilde{O}(\mathrm{poly}(2^L, \gamma^{-1})) \cdot \max\{d, \epsilon^{-14}\} \cdot \log(1/\delta), \\ n^* &= \widetilde{O}(\mathrm{poly}(2^L, \gamma^{-1})) \cdot \epsilon^{-4}. \end{aligned}$$

Then by Theorems 5.6 and 5.7, if $m \geqslant m^*$ and $n \geqslant n^*$, we can immediately see that gradient descent finds a point $\mathbf{W}^{(k)}$ satisfying

$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big[y \cdot f_{\mathbf{W}^{(k)}}(\mathbf{x}) < 0\big] \leqslant \epsilon.$$

This completes the proof. □

6 Conclusions and Future Work

In this paper, we provided a generalization guarantee of gradient descent for training deep ReLU networks under over-parameterization. Our results hold under a data distribution assumption that is milder than linear separability, and may be used to revisit the convergence analysis of training DNNs. Although we only analyzed the generalization performance of gradient descent and cross-entropy loss, our results can be extended to stochastic gradient descent and other loss functions such as those discussed in Zou et al. (2018). In addition, our results can be extended to other types of deep neural networks such as convolutional neural networks (CNNs) and ResNet. We leave it as future work. Another interesting direction is to derive generalization error bounds for deep learning based on the "small-ball" assumption proposed in Mendelson (2014).

A Proof of Theorem 5.3

In this section we provide the proof of Theorem 5.3. Let $\tau_0 = \nu \min\{L^{-4}[\log(m)]^{-3/2}, L^{-1}(s/m)^{3/2}\}$, where $\nu > 0$ is a small enough absolute constant, and $\mathcal{N}^* = \mathcal{N}(S^{d-1}, \tau_0)$ be a τ_0-net covering the unit sphere in the input space. By Lemma 5.2 in Vershynin (2010), we have

$$|\mathcal{N}^*| \leqslant (1 + 2/\tau_0)^d \leqslant (4/\tau_0)^d.$$

Throughout this section, we denote by $\widehat{\mathbf{x}}_l$ the output of the l-th layer of the network at initialization with input $\widehat{\mathbf{x}}$. Note that this is slightly different from the notation used in Theorem 5.4. Since in this section focuses on parameter matrices at initialization, this slight abuse of notation won't cause confusion.

Lemma A.1. Suppose that $\mathbf{W}_1, \ldots, \mathbf{W}_L$ are generated via Gaussian initialization. Then there exist absolute constants $C, C' > 0$ such that, for any $\delta > 0$, as long as $m \geqslant C\log(2/\delta)$, with probability at least $1-\delta$, $\|\mathbf{W}\|_2 \leqslant C'$.

Proof of Lemma A.1. The result directly follows by Corollary 5.35 in Vershynin (2010). □

Lemma A.2. If $C\max\{d\log(1/\tau_0), \log(L^2/\delta)\} \leqslant s \leqslant C'L^{-2}m[\log(m)]^{-1}$ for some large enough absolute constant C and small enough absolute constant C', then with probability at least $1-\delta$

we have

$$\Gamma_{l_1,l_2,s}(\widehat{\mathbf{x}}) \leqslant C''\sqrt{\frac{s\log(m)}{m}}$$

for all $\widehat{\mathbf{x}} \in \mathcal{N}^*$ and $1 \leqslant l_1 < l_2 \leqslant L$, where C'' is an absolute constant.

Proof of Lemma A.2. Denote $l_0 = l_1 - 1$, and

$$g_{l_1,l_2}(\mathbf{a},\mathbf{b},\mathbf{x}) = \mathbf{b}^\top \mathbf{W}_{l_2}^\top \Bigg(\prod_{r=l_1}^{l_2-1} \mathbf{\Sigma}_r(\mathbf{x})\mathbf{W}_r^\top\Bigg)\mathbf{a}.$$

Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be subspaces of $\mathbb{R}^{m_{l_0}}$ and $\mathbb{R}^{m_{l_2}}$ with sparsity s respectively. Then there are $\binom{m_{l_0}}{s}$ choices of $\mathcal{M}_1$ and $\binom{m_{l_2}}{s}$ choices of $\mathcal{M}_2$. let $\mathcal{N}_1 = \cup_{\mathcal{M}_1}\mathcal{N}[\mathcal{M}_1, 1/4]$, $\mathcal{N}_2 = \cup_{\mathcal{M}_1}\mathcal{N}[\mathcal{M}_2, 1/4]$ be the unions of 1/4-nets covering each choice of $\mathcal{M}_1$ and $\mathcal{M}_2$ respectively. Then by Lemma 5.2 in Vershynin (2010) and the assumption that s is larger than a sufficiently large absolute constant, we have

$$|\mathcal{N}_1| \leqslant \binom{m_{l_0}}{s} 9^s \leqslant m_{l_0}^s,\ |\mathcal{N}_2| \leqslant \binom{m_{l_2}}{s} 9^s \leqslant m_{l_2}^s.$$

For any $\widehat{\mathbf{x}} \in \mathcal{N}^*$, $\widehat{\mathbf{a}} \in \mathcal{N}_1$, $l_0 \in [L-1] \cup \{0\}$ and $l \in [L - l_0]$, denote $\widehat{\mathbf{a}}_0 = \widehat{\mathbf{a}}$ and $\widehat{\mathbf{a}}_l = \big[\prod_{r=l_1}^{l} \mathbf{\Sigma}_r(\widehat{\mathbf{x}})\mathbf{W}_r^\top\big]\widehat{\mathbf{a}}$. Then for any $\widehat{\mathbf{a}} \in \mathcal{N}_1$ and $\widehat{\mathbf{x}} \in \mathcal{N}^*$, by definition we have $\mathbf{w} \overset{d}{=} -\mathbf{w}$, and therefore

$$\begin{aligned}
\mathbb{E}[\|\widehat{\mathbf{a}}_l\|_2^2|\widehat{\mathbf{a}}_{l-1}] &= \sum_{j=1}^{m_l} \mathbb{E}\Big[\big(\mathbf{w}_{l,j}^\top\widehat{\mathbf{a}}_{l-1}\big)^2 \cdot \mathbb{1}\big\{\mathbf{w}_{l,j}^\top\widehat{\mathbf{x}}_{l-1} > 0\big\}\Big|\widehat{\mathbf{a}}_{l-1}\Big] \\
&= \frac{1}{2}\sum_{j=1}^{m_l} \mathbb{E}\Big[\big(\mathbf{w}_{l,j}^\top\widehat{\mathbf{a}}_{l-1}\big)^2 \cdot \mathbb{1}\big\{\mathbf{w}_{l,j}^\top\widehat{\mathbf{x}}_{l-1} > 0\big\} + \big(-\mathbf{w}_{l,j}^\top\widehat{\mathbf{a}}_{l-1}\big)^2 \cdot \mathbb{1}\big\{-\mathbf{w}_{l,j}^\top\widehat{\mathbf{x}}_{l-1} > 0\big\}\Big|\widehat{\mathbf{a}}_{l-1}\Big] \\
&= \frac{1}{2}\mathbb{E}[\|\mathbf{W}_l^\top\widehat{\mathbf{a}}_{l-1}\|_2^2|\widehat{\mathbf{a}}_{l-1}] \\
&= \|\widehat{\mathbf{a}}_{l-1}\|_2^2.
\end{aligned}$$

Moreover, condition on $\widehat{\mathbf{a}}_{l-1}$, $\|\sigma^2(\mathbf{w}_{l,j}^\top\widehat{\mathbf{a}}_{l-1})\|_{\psi_1} \leqslant C_1\|\widehat{\mathbf{a}}_{l-1}\|_2^2/m_l$ for some absolute constant C_1. By Bernstein inequality and union bound, with probability at least $1-\delta/2$, we have

$$\begin{aligned}
\big|\|\widehat{\mathbf{a}}_l\|_2^2 - \|\widehat{\mathbf{a}}_{l-1}\|_2^2\big| &\leqslant C_3\|\widehat{\mathbf{a}}_{l-1}\|_2^2 \cdot \sqrt{\frac{\log(2\cdot m_{l_0}^s \cdot (4/\tau_0)^d \cdot L^2/\delta)}{m_l}} \\
&\leqslant C_4\|\widehat{\mathbf{a}}_{l-1}\|_2^2 \cdot \sqrt{\frac{s\log(m_{l_0}) + d\log(1/\tau_0) + \log(L^2/\delta)}{m_l}}
\end{aligned}$$

for all $\widehat{\mathbf{a}} \in \mathcal{N}_1$, $l_0 \in [L] \cup \{0\}$, $l \in [L - l_0]$ and $\widehat{\mathbf{x}} \in \mathcal{N}^*$, where C_3, C_4 are large enough absolute constants. By the assumption that $C\max\{d\log(1/\tau_0), \log(L^2/\delta)\} \leqslant s \leqslant C'L^{-2}m[\log(m)]^{-1}$ for some large enough absolute constant C and small enough absolute constant C', we have

$$\big|\|\widehat{\mathbf{a}}_l\|_2^2 - \|\widehat{\mathbf{a}}_{l-1}\|_2^2\big| \leqslant \|\widehat{\mathbf{a}}_{l-1}\|_2^2 \cdot L^{-1}$$

for all $\widehat{\mathbf{a}} \in \mathcal{N}_1$, $l_0 \in [L] \cup \{0\}$, $l \in [L - l_0]$ and $\widehat{\mathbf{x}} \in \mathcal{N}^*$. Moreover, since $\|\widehat{\mathbf{a}}_0\|_2 = \|\widehat{\mathbf{a}}\|_2 = 1$, we have

$$\|\widehat{\mathbf{a}}_l\|_2^2 \leqslant \|\widehat{\mathbf{a}}_{l-1}\|_2^2 \cdot (1 + L^{-1}) \leqslant \|\widehat{\mathbf{a}}_0\|_2^2 \cdot (1 + L^{-1})^L \leqslant e$$

for all $\widehat{\mathbf{a}} \in \mathcal{N}_1$, $l_0 \in [L] \cup \{0\}$, $l \in [L - l_0]$ and $\widehat{\mathbf{x}} \in \mathcal{N}^*$.

Now we proceed to bound $g_{l_1,l_2}(\widehat{\mathbf{a}}, \widehat{\mathbf{b}}, \widehat{\mathbf{x}})$ based on the randomness of $\mathbf{W}_{l_2}$. Condition on $\widehat{\mathbf{a}}_{l-1}$ and the event that $\|\widehat{\mathbf{a}}_{l-1}\|_2 \leqslant e$, we have

$$\|g_{l_1,l_2}(\widehat{\mathbf{a}}, \widehat{\mathbf{b}}, \widehat{\mathbf{x}})\|_{\psi_2} \leqslant C_5 \|\widehat{\mathbf{a}}_{l-1}\|_2 \|\widehat{\mathbf{b}}\|_2 m^{-1/2} \leqslant C_6 m^{-1/2},$$

where C_5 and C_6 are absolute constants. By Hoeffding inequality, union bound and the assumption that $C \max\{d \log(1/\tau_0), \log(L^2/\delta)\} \leqslant s \leqslant C' L^{-2} m [\log(m)]^{-1}$ for some large enough absolute constant C and small enough absolute constant C', with probability at least $1 - \delta/2$, we have

$$g_{l_1,l_2}(\widehat{\mathbf{a}}, \widehat{\mathbf{b}}, \widehat{\mathbf{x}}) \leqslant C_7 \sqrt{\frac{\log(2e \cdot m_{l_0}^s \cdot m_{l_2}^s \cdot (4/\tau_0)^d \cdot L^2/\delta)}{m}} \leqslant C_8 \sqrt{\frac{s \log(m) + s}{m}} \leqslant C_9 \sqrt{\frac{s \log(m)}{m}}$$

for all $\widehat{\mathbf{a}} \in \mathcal{N}_1$, $\widehat{\mathbf{b}} \in \mathcal{N}_2$, $\widehat{\mathbf{x}} \in \mathcal{N}^*$ and $1 \leqslant l_1 < l_2 \leqslant L$, where $C_7, C_8, C_9 > 0$ are absolute constants. For any $\mathbf{a} \in S^{m_{l_0}-1}$ with $\|\mathbf{a}\|_0 \leqslant s$ and $\mathbf{b} \in S^{m_l - 1}$ with $\|\mathbf{b}\|_0 \leqslant s$, there exist $\widehat{\mathbf{a}} \in \mathcal{N}_1$ and $\widehat{\mathbf{b}} \in \mathcal{N}_2$ such that $\|\mathbf{a} - \widehat{\mathbf{a}}\|_2 \leqslant 1/4$, $\|\mathbf{b} - \widehat{\mathbf{b}}\|_2 \leqslant 1/4$. Therefore, we have

$$\begin{aligned} g_{l_1,l_2}(\mathbf{a}, \mathbf{b}, \widehat{\mathbf{x}}) &\leqslant |g_{l_1,l_2}(\widehat{\mathbf{a}}, \widehat{\mathbf{b}}, \widehat{\mathbf{x}})| + |g_{l_1,l_2}(\mathbf{a}, \mathbf{b}, \widehat{\mathbf{x}}) - g_{l_1,l_2}(\widehat{\mathbf{a}}, \widehat{\mathbf{b}}, \widehat{\mathbf{x}})| \\ &\leqslant C_9 \sqrt{\frac{s \log(m)}{m}} + |g_{l_1,l_2}(\mathbf{a}, \mathbf{b}, \widehat{\mathbf{x}}) - g_{l_1,l_2}(\widehat{\mathbf{a}}, \mathbf{b}, \widehat{\mathbf{x}})| + |g_{l_1,l_2}(\widehat{\mathbf{a}}, \mathbf{b}, \widehat{\mathbf{x}}) - g_{l_1,l_2}(\widehat{\mathbf{a}}, \widehat{\mathbf{b}}, \widehat{\mathbf{x}})|. \end{aligned}$$

Since $g_{l_1,l_2}(\mathbf{a}, \mathbf{b}, \mathbf{x})$ is a bilinear function in $\mathbf{a}$ and $\mathbf{b}$, we have

$$\begin{aligned} |g_{l_1,l_2}(\mathbf{a}, \mathbf{b}, \widehat{\mathbf{x}}) - g_{l_1,l_2}(\widehat{\mathbf{a}}, \mathbf{b}, \widehat{\mathbf{x}})| &\leqslant \|\mathbf{a} - \widehat{\mathbf{a}}\|_2 \cdot \left| g_{l_1,l_2}\left(\frac{\mathbf{a} - \widehat{\mathbf{a}}}{\|\mathbf{a} - \widehat{\mathbf{a}}\|_2}, \mathbf{b}, \widehat{\mathbf{x}} \right) \right| \\ &\leqslant \|\mathbf{a} - \widehat{\mathbf{a}}\|_2 \cdot \Gamma_{l_1,l_2,s}(\widehat{\mathbf{x}}) \\ &\leqslant \Gamma_{l_1,l_2,s}(\widehat{\mathbf{x}})/4. \end{aligned}$$

Similarly, we also have

$$|g_{l_1,l_2}(\widehat{\mathbf{a}}, \mathbf{b}, \widehat{\mathbf{x}}) - g_{l_1,l_2}(\widehat{\mathbf{a}}, \widehat{\mathbf{b}}, \widehat{\mathbf{x}})| \leqslant \Gamma_{l_1,l_2,s}(\widehat{\mathbf{x}})/4.$$

Therefore, we have

$$\Gamma_{l_1,l_2,s}(\widehat{\mathbf{x}}) \leqslant C_{10} \sqrt{\frac{s \log(m)}{m}},$$

for all $\widehat{\mathbf{x}} \in \mathcal{N}^*$ and $1 \leqslant l_1 < l_2 \leqslant L$, where $C_{10} > 0$ is an absolute constant. This completes the proof. □

Lemma A.3. For any $\delta > 0$, if $C \max\{d \log(1/\tau_0), \log(L^2/\delta)\} \leqslant s \leqslant C' L^{-2} m [\log(m)]^{-1}$ for some large enough absolute constant C and small enough absolute constant C', then with probability at least $1-\delta$,

$$\Gamma'_{l_1,l_2,s}(\widehat{\mathbf{x}}), \Gamma''_{l_1,l_2,s}(\widehat{\mathbf{x}}) \leqslant C''$$

for all $\widehat{\mathbf{x}} \in \mathcal{N}^*$ and $1 \leqslant l_1 < l_2 \leqslant L$, where $C'' > 0$ is an absolute constant.

Proof of Lemma A.3. Let $K_0 > 0$ be a large enough constant integer and $\overline{s} = \lceil m_{l_1-1}K_0^{-1}L^{-2}\rceil$. For $t = 1, \ldots, K_0L^2$, define

$$\mathcal{M}_t = \big\{\mathbf{c} \in S^{m_{l_1-1}-1}, \mathrm{supp}(\mathbf{c}) \subseteq \{(t-1)\overline{s}+1, \ldots, \min\{m_{l_1-1}, t\overline{s}\}\big\}$$

and $\mathcal{M} = \cup_{t=1}^{K_0L^2}\mathcal{M}_t$. By Lemma 5.2 in Vershynin (2010), it is easy to see that $|\mathcal{N}(\mathcal{M}, 1/4)| \leqslant K_0L^2 \cdot 9^{\overline{s}}$. Therefore with the same proof of Lemma A.2, it can be shown that as long as

$$C_1 \max\{d\log(1/\tau_0), \log(L^2/\delta)\} \leqslant s \leqslant C_2L^{-2}m[\log(m)]^{-1}$$

for some large enough absolute constant C_1 and small enough absolute constant C_2, with probability at least $1-\delta$,

$$\sup_{\substack{\mathbf{a}\in\mathcal{M}, \mathbf{b}\in S^{m_{l_2}-1},\\ \|\mathbf{b}\|_0 \leqslant s}} \mathbf{b}^\top\mathbf{W}_{l_2}^\top\left(\prod_{r=l_1}^{l_2-1}\mathbf{\Sigma}_r(\widehat{\mathbf{x}})\mathbf{W}_r^\top\right)\mathbf{a} \leqslant C_3\sqrt{\frac{s\log(m)+\overline{s}\log(L/\delta)}{m}} \leqslant C_4L^{-1}$$

for all $\widehat{\mathbf{x}} \in \mathcal{N}^*$ and $1 \leqslant l_1 < l_2 \leqslant L$, where C_3, C_4 are absolute constants. Now any $\mathbf{a} \in S^{m_{l_1-1}-1}$ can be written as a summation of K_0L^2 sparse vectors $\mathbf{a}^{(t)} \in \mathcal{M}_t, t = 1, \ldots, K_0L^2$ with sparsity level at most $\overline{s}$. Therefore for any $\mathbf{b} \in S^{m_{l_2}-1}$, we have

$$\begin{aligned}
\mathbf{b}^\top\mathbf{W}_{l_2}^\top\left(\prod_{r=l_1}^{l_2-1}\mathbf{\Sigma}_r(\widehat{\mathbf{x}})\mathbf{W}_r^\top\right)\mathbf{a} &= \sum_{t=1}^{K_0L^2}\mathbf{b}^\top\mathbf{W}_{l_2}^\top\left(\prod_{r=l_1}^{l_2-1}\mathbf{\Sigma}_r(\widehat{\mathbf{x}})\mathbf{W}_r^\top\right)\mathbf{a}^{(t)}\\
&\leqslant \sum_{t=1}^{K_0L^2} C_5L^{-1}\cdot\|\mathbf{a}^{(t)}\|_2\\
&\leqslant C_6L^{-1}\cdot L\cdot\left[\sum_{t=1}^{K_0L^2}\|\mathbf{a}^{(t)}\|_2^2\right]^{1/2}\\
&= C_6,
\end{aligned}$$

where $C_5, C_6 > 0$ are absolute constants. This implies that $\Gamma''_{l_1,l_2,s}(\widehat{\mathbf{x}}) \leqslant C_6$ for all $\widehat{\mathbf{x}} \in \mathcal{N}^*$ and $1 \leqslant l_1 < l_2 \leqslant L$. Similarly, it can be shown that $\Gamma'_{l_1,l_2,s}(\widehat{\mathbf{x}}) \leqslant C_7$ for all $\widehat{\mathbf{x}} \in \mathcal{N}^*$ and $1 \leqslant l_1 < l_2 \leqslant L$, where C_7 is an absolute constant. This completes the proof. □

Lemma A.4. For any $\delta > 0$, if $m \geqslant CL^2\max\{d\log(1/\tau_0), \log(L^2/\delta)\}$ for some large enough absolute constant C, then with probability at least $1-\delta$,

$$\Lambda_{l_1,l_2}(\widehat{\mathbf{x}}) \leqslant C''L$$

for all $\widehat{\mathbf{x}} \in \mathcal{N}^*$ and $1 \leqslant l_1 < l_2 \leqslant L$, where $C'' > 0$ is an absolute constant.

Proof of Lemma A.4. Let $K_1, K_2 > 0$ be large enough constant integers and $\overline{s}_1 = \lceil m_{l_1-1}K_1^{-1}L^{-2}\rceil$,

$\overline{s}_2 = \lceil m_{l_2} K_2^{-1} L^{-2} \rceil$. For $t_1 = 1, \ldots, K_1L^2$ and $t_2 = 1, \ldots, K_2L^2$, define

$$\mathcal{M}_{1,t_1} = \big\{\mathbf{c} \in S^{m_{l_1-1}-1}, \mathrm{supp}(\mathbf{c}) \subseteq \{(t_1-1)\overline{s}_1 + 1, \ldots, \max\{m_{l_1-1}, t_1\overline{s}_1\}\big\},$$
$$\mathcal{M}_{2,t_2} = \big\{\mathbf{c} \in S^{m_{l_2}-1}, \mathrm{supp}(\mathbf{c}) \subseteq \{(t_2-1)\overline{s}_2 + 1, \ldots, \max\{m_{l_2}, t_2\overline{s}_2\}\big\},$$

and $\mathcal{M}_1 = \cup_{t_1=1}^{K_1L^2} \mathcal{M}_{1,t_1}$, $\mathcal{M}_2 = \cup_{t_2=1}^{K_2L^2} \mathcal{M}_{2,t_2}$. By Lemma 5.2 in Vershynin (2010), it is easy to see that $|\mathcal{N}(\mathcal{M}_1, 1/4)| \leqslant K_1L^2 \cdot 9^{\overline{s}_1}$, $|\mathcal{N}(\mathcal{M}_2, 1/4)| \leqslant K_2L^2 \cdot 9^{\overline{s}_2}$. Therefore with the same proof of Lemma A.2, it can be shown that with probability at least $1-\delta$,

$$\sup_{\mathbf{a}\in\mathcal{M}_1, \mathbf{b}\in\mathcal{M}_2} \mathbf{b}^\top \mathbf{W}_{l_2}^\top \Bigg(\prod_{r=l_1}^{l_2-1} \mathbf{\Sigma}_r(\hat{\mathbf{x}}) \mathbf{W}_r^\top\Bigg) \mathbf{a} \leqslant C_1 L^{-1}$$

for all $\hat{\mathbf{x}} \in \mathcal{N}^*$ and $1 \leqslant l_1 < l_2 \leqslant L$, where C_1 is an absolute constant. Similar to the proof of Lemma A.3, for any $\mathbf{a} \in S^{m_{l_1-1}-1}$ and $\mathbf{b} \in S^{m_{l_2}-1}$, we can rewrite them as $\mathbf{a} = \sum_{t_1=1}^{K_1L^2} \mathbf{a}^{(t_1)}$, $\mathbf{b} = \sum_{t_2=1}^{K_2L^2} \mathbf{b}^{(t_2)}$, where $\mathbf{a}^{(t_1)} \in \mathcal{M}_{1,t_1}, t_1 = 1, \ldots, K_1L^2$, $\mathbf{b}^{(t_2)} \in \mathcal{M}_{2,t_2}, t_2 = 1, \ldots, K_2L^2$. Therefore we have

$$\begin{aligned}
\mathbf{b}^\top \mathbf{W}_{l_2}^\top \Bigg(\prod_{r=l_1}^{l_2-1} \mathbf{\Sigma}_r(\hat{\mathbf{x}}) \mathbf{W}_r^\top\Bigg) \mathbf{a} &= \sum_{t_1=1}^{K_1L^2} \sum_{t_2=1}^{K_2L^2} \mathbf{b}^{(t_2)\top} \mathbf{W}_{l_2}^\top \Bigg(\prod_{r=l_1}^{l_2-1} \mathbf{\Sigma}_r(\hat{\mathbf{x}}) \mathbf{W}_r^\top\Bigg) \mathbf{a}^{(t_1)} \\
&\leqslant \sum_{t_1=1}^{K_1L^2} \sum_{t_2=1}^{K_2L^2} C_2 L^{-1} \cdot \|\mathbf{a}^{(t_1)}\|_2 \|\mathbf{b}^{(t_2)}\|_2 \\
&\leqslant C_3 L^{-1} \cdot L \cdot \Bigg[\sum_{t_1=1}^{K_1L^2} \|\mathbf{a}^{(t_1)}\|_2^2\Bigg]^{1/2} \cdot L \cdot \Bigg[\sum_{t_2=1}^{K_2L^2} \|\mathbf{a}^{(t_2)}\|_2^2\Bigg]^{1/2} \\
&= C_3 L,
\end{aligned}$$

where $C_2, C_3 > 0$ are absolute constants. This completes the proof. □

Lemma A.5. If $C \max\{d \log(1/\tau_0), \log(L^2/\delta)\} \leqslant s \leqslant C' L^{-2} m [\log(m)]^{-1}$ for some large enough absolute constant C and small enough absolute constant C', then with probability at least $1-\delta$, we have

$$\Gamma'''_{l,s}(\hat{\mathbf{x}}) \leqslant C'' \sqrt{s \log(m)}$$

for all $\hat{\mathbf{x}} \in \mathcal{N}^*$ and $l = 1, \ldots, L$, where C'' is an absolute constant.

Proof of Lemma A.5. Denote $l_0 = l - 1$, and

$$g_l(\mathbf{a}, \mathbf{x}) = \mathbf{v}^\top \Bigg(\prod_{r=l}^{L} \mathbf{\Sigma}_r(\mathbf{x}) \mathbf{W}_r^\top\Bigg) \mathbf{a}.$$

Let $\mathcal{M}$ be a fixed subspace of $\mathbb{R}^{m_{l_0}}$ with sparsity s. Then there are $\binom{m_{l_0}}{s}$ choices of $\mathcal{M}$. let $\mathcal{N} = \cup_{\mathcal{M}} \mathcal{N}[\mathcal{M}, 1/2]$ be the union of 1/2-nets covering each choice of $\mathcal{M}$. Then by Lemma 5.2 in Vershynin (2010) and the assumption that s is larger than a sufficiently large absolute constant, we have

$$|\mathcal{N}| \leqslant \binom{m_{l_0}}{s} 5^s \leqslant m_{l_0}^s.$$

For any $\widehat{\mathbf{x}} \in \mathcal{N}^*$, $\widehat{\mathbf{a}} \in \mathcal{N}$ and $l \in [L]$, by Lemma A.3, with probability at least $1 - \delta/2$ we have

$$\left\| \left(\prod_{r=l}^{L-1} \mathbf{\Sigma}_r(\widehat{\mathbf{x}}) \mathbf{W}_r^\top \right) \widehat{\mathbf{a}} \right\|_2 \leqslant C_1$$

for some constant C_1. Condition on $\mathbf{W}_1, \ldots, \mathbf{W}_{L-1}$, by union bound and Hoeffding inequality, with probability at least $1 - \delta/2$,

$$g_l(\widehat{\mathbf{a}}, \widehat{\mathbf{x}}) \leqslant C_2 \cdot \sqrt{m_L} \cdot \sqrt{\frac{\log[2e \cdot m_{l_0}^s \cdot (4/\tau_0)^d \cdot L/\delta]}{m_L}} \leqslant C_3 \sqrt{s \log(m)}$$

for all $\widehat{\mathbf{a}} \in \mathcal{N}$, $\widehat{\mathbf{x}} \in \mathcal{N}^*$ and $1 \leqslant l \leqslant L$, where $C_2, C_3 > 0$ are absolute constants. For any $\mathbf{a} \in S^{m_{l_0}-1}$ with $\|\mathbf{a}\|_0 \leqslant s$, there exists $\widehat{\mathbf{a}} \in \mathcal{N}$ such that $\|\mathbf{a} - \widehat{\mathbf{a}}\|_2 \leqslant 1/2$. Therefore, we have

$$g_l(\mathbf{a}, \widehat{\mathbf{x}}) \leqslant |g_l(\widehat{\mathbf{a}}, \widehat{\mathbf{x}})| + |g_l(\mathbf{a}, \widehat{\mathbf{x}}) - g_l(\widehat{\mathbf{a}}, \widehat{\mathbf{x}})| \leqslant C_3 \sqrt{s \log(m)} + |g_l(\mathbf{a}, \widehat{\mathbf{x}}) - g_l(\widehat{\mathbf{a}}, \widehat{\mathbf{x}})|.$$

Since $g_l(\mathbf{a}, \mathbf{x})$ is linear in $\mathbf{a}$, we have

$$|g_l(\mathbf{a}, \widehat{\mathbf{x}}) - g_l(\widehat{\mathbf{a}}, \widehat{\mathbf{x}})| \leqslant \|\mathbf{a} - \widehat{\mathbf{a}}\|_2 \left| g_l\left(\frac{\mathbf{a} - \widehat{\mathbf{a}}}{\|\mathbf{a} - \widehat{\mathbf{a}}\|_2}, \widehat{\mathbf{x}} \right) \right| \leqslant \|\mathbf{a} - \widehat{\mathbf{a}}\|_2 \cdot \Gamma'''_{l,s}(\widehat{\mathbf{x}}) \leqslant \Gamma'''_{l,s}(\widehat{\mathbf{x}})/2.$$

Therefore, we have

$$\Gamma'''_{l,s}(\widehat{\mathbf{x}}) \leqslant C_4 \sqrt{s \log(m)},$$

for all $\widehat{\mathbf{x}} \in \mathcal{N}^*$ and $l \in [L]$, where C_4 is an absolute constant. This completes the proof. □

Lemma A.6. Suppose that $\mathbf{W}_1, \ldots, \mathbf{W}_L$ are generated via Gaussian initialization. For any $\delta > 0$, if $m \geqslant CL^2 d \log[m/(\tau_0 \delta)]$ for some large enough absolute constant C, then with probability at least $1 - \delta$,

$$\big| \|\widehat{\mathbf{x}}_l\|_2 - 1 \big| \leqslant C' L \sqrt{\frac{d \log[m/(\tau_0 \delta)]}{m}}$$

for all $\widehat{\mathbf{x}} \in \mathcal{N}^*$ and $l \in [L]$, where C' is an absolute constant.

Proof of Lemma A.6. We can in fact show a stronger result:

$$\left| \left\| \left[\prod_{r=1}^{l} \mathbf{\Sigma}_r(\widehat{\mathbf{x}}) \mathbf{W}_r^\top \right] \mathbf{a} \right\|_2 - 1 \right| \leqslant CL \sqrt{\frac{d \log(m/\tau_0)}{m}}$$

for all $\mathbf{a} \in S^{d-1}$, $\widehat{\mathbf{x}} \in \mathcal{N}^*$ and $l \in [L]$. To show this result, let $\mathcal{N} = \mathcal{N}(s^{d-1}, \sqrt{d/m})$ be a $\sqrt{d/m}$-net covering S^{d-1}. Then by Lemma 5.2 in Vershynin (2010), we have

$$|\mathcal{N}| \leqslant \big(1 + 2\sqrt{m/d}\big)^s \leqslant (m/d)^s.$$

For any $\widehat{\mathbf{x}} \in \mathcal{N}^*$, $\widehat{\mathbf{a}} \in \mathcal{N}$ and $l \in [L]$, denote $\widehat{\mathbf{a}}_0 = \widehat{\mathbf{a}}$ and $\widehat{\mathbf{a}}_l = \big[\prod_{r=1}^l \boldsymbol{\Sigma}_r(\widehat{\mathbf{x}})\mathbf{W}_r^\top\big]\widehat{\mathbf{a}}$. Then for any $\widehat{\mathbf{a}} \in \mathcal{N}$ and $\widehat{\mathbf{x}} \in \mathcal{N}^*$, Similar to the proof of Lemma A.2, we have $\mathbf{w} \overset{d}{=} -\mathbf{w}$, and

$$\begin{aligned}\mathbb{E}[\|\widehat{\mathbf{a}}_l\|_2^2|\widehat{\mathbf{a}}_{l-1}] &= \frac{1}{2}\sum_{j=1}^{m_l}\mathbb{E}\Big[\big(\mathbf{w}_{l,j}^\top\widehat{\mathbf{a}}_{l-1}\big)^2 \cdot \mathbb{1}\,\{\mathbf{w}_{l,j}^\top\widehat{\mathbf{x}}_{l-1} > 0\} + \big(-\mathbf{w}_{l,j}^\top\widehat{\mathbf{a}}_{l-1}\big)^2 \cdot \mathbb{1}\,\{-\mathbf{w}_{l,j}^\top\widehat{\mathbf{x}}_{l-1} > 0\}\Big|\widehat{\mathbf{a}}_{l-1}\Big]\\ &= \|\widehat{\mathbf{a}}_{l-1}\|_2^2.\end{aligned}$$

Moreover, condition on $\widehat{\mathbf{a}}_{l-1}$, we have $\|\sigma^2(\mathbf{w}_{l,j}^\top\widehat{\mathbf{a}}_{l-1})\|_{\psi_1} \leqslant C_1\|\widehat{\mathbf{a}}_{l-1}\|_2^2/m_l$ for some absolute constant C_1. By Bernstein inequality and union bound, with probability at least $1-\delta/2$, we have

$$\begin{aligned}\big|\|\widehat{\mathbf{a}}_l\|_2^2 - \|\widehat{\mathbf{a}}_{l-1}\|_2^2\big| &\leqslant C_3\|\widehat{\mathbf{a}}_{l-1}\|_2^2 \cdot \sqrt{\frac{\log[2\cdot(m/d)^d\cdot(4/\tau_0)^d\cdot L/\delta]}{m}}\\ &\leqslant C_4\|\widehat{\mathbf{a}}_{l-1}\|_2^2 \cdot \sqrt{\frac{d\log[m/(\tau_0\delta)]}{m_l}}\end{aligned}$$

for all $\widehat{\mathbf{a}} \in \mathcal{N}$, $\widehat{\mathbf{x}} \in \mathcal{N}^*$ and $l \in [L]$, where C_3, C_4 are large enough absolute constants. By the assumption that $m \geqslant CL^2 d\log(m/\tau_0)$ for some large enough absolute constant C, we have

$$C_4\sqrt{\frac{d\log[m/(\tau_0\delta)]}{m_l}} \leqslant (2L)^{-1},$$

and therefore

$$\|\widehat{\mathbf{a}}_l\|_2^2 \leqslant \|\widehat{\mathbf{a}}_0\|_2^2 \cdot \left\{1 + C_4\sqrt{\frac{d\log[m/(\tau_0\delta)]}{m_l}}\right\}^l \leqslant 1 + C_5 L\sqrt{\frac{d\log[m/(\tau_0\delta)]}{m_l}},$$

for all $\widehat{\mathbf{a}} \in \mathcal{N}$, $\widehat{\mathbf{x}} \in \mathcal{N}^*$ and $l \in [L]$, where C_5 is an absolute constant. Similarly, we also have

$$\|\widehat{\mathbf{a}}_l\|_2^2 \geqslant 1 - C_6 L\sqrt{\frac{d\log[m/(\tau_0\delta)]}{m_l}}$$

for all $\widehat{\mathbf{a}} \in \mathcal{N}$, $\widehat{\mathbf{x}} \in \mathcal{N}^*$ and $l \in [L]$, where C_6 is an absolute constant. Therefore,

$$\left|\left\|\left[\prod_{r=1}^l \boldsymbol{\Sigma}_r(\widehat{\mathbf{x}})\mathbf{W}_r^\top\right]\widehat{\mathbf{a}}\right\|_2 - 1\right| \leqslant C_7 L\sqrt{\frac{d\log[m/(\tau_0\delta)]}{m}}$$

for all $\widehat{\mathbf{a}} \in \mathcal{N}$, $\widehat{\mathbf{x}} \in \mathcal{N}^*$ and $l \in [L]$, where C_7 is an absolute constant. Now for any $\mathbf{a} \in S^{d-1}$, there exists $\widehat{\mathbf{a}} \in \mathcal{N}$ such that $\|\mathbf{a} - \widehat{\mathbf{a}}\|_2 \leqslant \sqrt{d/m}$. By triangle inequality and Lemma A.4 we have

$$\left|\left\|\left[\prod_{r=1}^l \boldsymbol{\Sigma}_r(\widehat{\mathbf{x}})\mathbf{W}_r^\top\right]\mathbf{a}\right\|_2 - \left\|\left[\prod_{r=1}^l \boldsymbol{\Sigma}_r(\widehat{\mathbf{x}})\mathbf{W}_r^\top\right]\widehat{\mathbf{a}}\right\|_2\right| \leqslant \left\|\left[\prod_{r=1}^l \boldsymbol{\Sigma}_r(\widehat{\mathbf{x}})\mathbf{W}_r^\top\right](\mathbf{a} - \widehat{\mathbf{a}})\right\|_2 \leqslant C_8 L\sqrt{d/m},$$

where C_8 is an absolute constant. Therefore we have

$$\left|\left\|\left[\prod_{r=1}^{l}\mathbf{\Sigma}_r(\widehat{\mathbf{x}})\mathbf{W}_r^\top\right]\mathbf{a}\right\|_2 - 1\right| \leqslant C_9 L\sqrt{\frac{d\log[m/(\tau_0\delta)]}{m}}$$

for some absolute constant C_9. This completes the proof. □

Lemma A.7. Suppose that $\mathbf{W}_1,\ldots,\mathbf{W}_L$ are generated via Gaussian initialization, and the results of Lemmas A.2, A.3 and A.4 hold. Let $\widetilde{\mathbf{\Sigma}}_l(\widehat{\mathbf{x}})$, $\widehat{\mathbf{x}} \in \mathcal{N}^*$, $l = 1,\ldots,L$ be diagonal matrices, and define

$$\widetilde{\Gamma}_{l_1,l_2,s}(\widehat{\mathbf{x}}) = \sup_{\substack{\|\mathbf{a}\|_2=\|\mathbf{b}\|_2=1,\\ \|\mathbf{a}\|_0,\|\mathbf{b}\|_0\leqslant s}} \mathbf{b}^\top\mathbf{W}_{l_2}^\top\left(\prod_{r=l_1}^{l_2-1}\widetilde{\mathbf{\Sigma}}_r(\widehat{\mathbf{x}})\mathbf{W}_r^\top\right)\mathbf{a},$$

$$\widetilde{\Gamma}'_{l_1,l_2,s}(\widehat{\mathbf{x}}) = \sup_{\substack{\|\mathbf{a}\|_2=\|\mathbf{b}\|_2=1,\\ \|\mathbf{a}\|_0\leqslant s}} \mathbf{b}^\top\mathbf{W}_{l_2}^\top\left(\prod_{r=l_1}^{l_2-1}\widetilde{\mathbf{\Sigma}}_r(\widehat{\mathbf{x}})\mathbf{W}_r^\top\right)\mathbf{a},$$

$$\widetilde{\Gamma}''_{l_1,l_2,s}(\widehat{\mathbf{x}}) = \sup_{\substack{\|\mathbf{a}\|_2=\|\mathbf{b}\|_2=1,\\ \|\mathbf{b}\|_0\leqslant s}} \mathbf{b}^\top\mathbf{W}_{l_2}^\top\left(\prod_{r=l_1}^{l_2-1}\widetilde{\mathbf{\Sigma}}_r(\widehat{\mathbf{x}})\mathbf{W}_r^\top\right)\mathbf{a},$$

$$\widetilde{\Gamma}'''_{l,s}(\widehat{\mathbf{x}}) = \sup_{\|\mathbf{a}\|_2=1,\|\mathbf{a}\|_0\leqslant s} \mathbf{v}^\top\left(\prod_{r=l}^{L}\widetilde{\mathbf{\Sigma}}_r(\widehat{\mathbf{x}})\mathbf{W}_r^\top\right)\mathbf{a},$$

$$\widetilde{\Lambda}_{l_1,l_2}(\widehat{\mathbf{x}}) = \sup_{\|\mathbf{a}\|_2=\|\mathbf{b}\|_2=1} \mathbf{b}^\top\mathbf{W}_{l_2}^\top\left(\prod_{r=l_1}^{l_2-1}\widetilde{\mathbf{\Sigma}}_r(\widehat{\mathbf{x}})\mathbf{W}_r^\top\right)\mathbf{a}.$$

If $\|\widetilde{\mathbf{\Sigma}}_l(\widehat{\mathbf{x}}) - \mathbf{\Sigma}_l(\widehat{\mathbf{x}})\|_0 \leqslant s$ and $|[\widetilde{\mathbf{\Sigma}}_l(\widehat{\mathbf{x}}) - \mathbf{\Sigma}_l(\widehat{\mathbf{x}})]_{jj}| \leqslant C$ for all $l = 1,\ldots,L$ and $j = 1,\ldots,m_l$, where C is an absolute constant that can be arbitrarily large and $s\log(m)/m \leqslant \kappa L^{-2}$ for some small enough absolute constant κ, then there exists an absolute constant C' such that

$$\widetilde{\Gamma}_{l_1,l_2,s}(\widehat{\mathbf{x}}) \leqslant C'\sqrt{s\log(m)/m}, \qquad \widetilde{\Gamma}'_{l_1,l_2,s}(\widehat{\mathbf{x}}), \widetilde{\Gamma}''_{l_1,l_2,s}(\widehat{\mathbf{x}}) \leqslant C',$$
$$\widetilde{\Gamma}'''_{l,s}(\widehat{\mathbf{x}}) \leqslant C'\sqrt{s\log(m)}, \qquad \widetilde{\Lambda}_{l_1,l_2}(\widehat{\mathbf{x}}) \leqslant C'L$$

for all $\widehat{\mathbf{x}} \in \mathcal{N}^*$, $l \in [L]$ and $1 \leqslant l_1 < l_2 \leqslant L$.

Proof of Lemma A.7. Here we give the detailed proof of the bound $\widetilde{\Lambda}_{l_1,l_2}(\widehat{\mathbf{x}}) \leqslant C'L$. The proof of the other results are almost identical. Note that for $r = l_1,\ldots,l_2$ we have

$$\widetilde{\mathbf{\Sigma}}_r(\widehat{\mathbf{x}})\mathbf{W}_r^\top = \mathbf{\Sigma}_r(\widehat{\mathbf{x}})\mathbf{W}_r^\top + [\widetilde{\mathbf{\Sigma}}_r(\widehat{\mathbf{x}}) - \mathbf{\Sigma}_r(\widehat{\mathbf{x}})]\mathbf{W}_r^\top.$$

Therefore, let $\mathcal{A}_r = \{\mathbf{\Sigma}_r(\widehat{\mathbf{x}})\mathbf{W}_r^\top, [\widetilde{\mathbf{\Sigma}}_r(\widehat{\mathbf{x}}) - \mathbf{\Sigma}_r(\widehat{\mathbf{x}})]\mathbf{W}_r^\top\}$, $r = l_1,\ldots,l_2$, then we have

$$\prod_{r=l_1}^{l_2}\widetilde{\mathbf{\Sigma}}_r(\widehat{\mathbf{x}})\mathbf{W}_r^\top = \sum_{\mathbf{A}_{l_1}\in\mathcal{A}_{l_1},\ldots,\mathbf{A}_{l_2}\in\mathcal{A}_{l_2}}\left(\prod_{r=l_1}^{l_2}\mathbf{A}_r\right).$$

For the rest of the proof, we denote by $|\mathbf{\Sigma}|$ the diagonal matrix with absolute values of elements of

$\boldsymbol{\Sigma}$ on the corresponding entries. For each sequence $\mathbf{A}_{l_1}, \ldots, \mathbf{A}_{l_2}$, denote

$$\widehat{\boldsymbol{\Sigma}}_r(\widehat{\mathbf{x}}) = \begin{cases} |\widetilde{\boldsymbol{\Sigma}}_{r-1}(\widehat{\mathbf{x}}) - \boldsymbol{\Sigma}_{r-1}(\widehat{\mathbf{x}})|, & \text{if } r \geqslant 2 \text{ and } \mathbf{A}_{r-1} = [\widetilde{\boldsymbol{\Sigma}}_{r-1}(\widehat{\mathbf{x}}) - \boldsymbol{\Sigma}_{r-1}(\widehat{\mathbf{x}})]\mathbf{W}_{r-1}^\top, \\ \mathbf{I}, & \text{otherwise.} \end{cases}$$

Then we have

$$\prod_{r=l_1}^{l_2} \mathbf{A}_r = \prod_{r=l_1}^{l_2} \mathbf{A}_r \widehat{\boldsymbol{\Sigma}}_r(\widehat{\mathbf{x}}).$$

We consider the following cases:

- When $\mathbf{A}_r = \boldsymbol{\Sigma}_r(\widehat{\mathbf{x}})\mathbf{W}_r^\top$ for all $r = l_1, \ldots, l_2$, the bound of $\|\prod_{r=l_1}^{l_2} \mathbf{A}_r\|_2$ is given by Lemma A.4.
- If there exists only one $r \in \{l_1, \ldots, l_2\}$ such that $\mathbf{A}_r = [\widetilde{\boldsymbol{\Sigma}}_r(\widehat{\mathbf{x}}) - \boldsymbol{\Sigma}_r(\widehat{\mathbf{x}})]\mathbf{W}_r^\top$, then $\prod_{r=l_1}^{l_2} \mathbf{A}_r$ has the form

$$\left\{\left[\prod_{r'=r+1}^{l_2} \boldsymbol{\Sigma}_{r'}(\widehat{\mathbf{x}})\mathbf{W}_{r'}^\top\right] |\widetilde{\boldsymbol{\Sigma}}_r(\widehat{\mathbf{x}}) - \boldsymbol{\Sigma}_r(\widehat{\mathbf{x}})|\right\}\left\{[\widetilde{\boldsymbol{\Sigma}}_r(\widehat{\mathbf{x}}) - \boldsymbol{\Sigma}_r(\widehat{\mathbf{x}})]\mathbf{W}_r^\top \left[\prod_{r'=l_1}^{r-1} \boldsymbol{\Sigma}_{r'}(\widehat{\mathbf{x}})\mathbf{W}_{r'}^\top\right]\right\}.$$

  Therefore by Lemma A.3, $\|\prod_{r=l_1}^{l_2} \mathbf{A}_r\|_2$ is bounded by an absolute constant.
- For all the other terms in the expansion, we can always split it into the product of:
  - An identity matrix $\mathbf{I}$, or a matrix of the form

$$\left[\prod_{r=r_1+1}^{l_2} \boldsymbol{\Sigma}_r(\widehat{\mathbf{x}})\mathbf{W}_r^\top\right] |\widetilde{\boldsymbol{\Sigma}}_{r_1}(\widehat{\mathbf{x}}) - \boldsymbol{\Sigma}_{r_1}(\widehat{\mathbf{x}})|.$$

    By Lemma A.3, in both cases, the spectral norm of this matrix is bounded by an absolute constant.
  - A product of matrices of the form

$$[\widetilde{\boldsymbol{\Sigma}}_{r_2}(\widehat{\mathbf{x}}) - \boldsymbol{\Sigma}_{r_2}(\widehat{\mathbf{x}})]\mathbf{W}_{r_2}^\top \left[\prod_{r=r_1+1}^{r_2-1} \boldsymbol{\Sigma}_r(\widehat{\mathbf{x}})\mathbf{W}_r^\top\right] |\widetilde{\boldsymbol{\Sigma}}_{r_2}(\widehat{\mathbf{x}}) - \boldsymbol{\Sigma}_{r_2}(\widehat{\mathbf{x}})|,$$

    whose spectral norms are of order at most $O[\sqrt{s\log(m)/m}]$ according to Lemma A.2.
  - An identity matrix $\mathbf{I}$, or a matrix of the form

$$[\widetilde{\boldsymbol{\Sigma}}_{r_2}(\widehat{\mathbf{x}}) - \boldsymbol{\Sigma}_{r_2}(\widehat{\mathbf{x}})]\mathbf{W}_{r_2}^\top \left[\prod_{r=l_1}^{r_2-1} \boldsymbol{\Sigma}_r(\widehat{\mathbf{x}})\mathbf{W}_r^\top\right].$$

    By Lemma A.3, in both cases, the spectral norm of this matrix is bounded by an absolute constant.

By the discussion above, we have

$$\begin{aligned}\bigg\|\prod_{r=l_1}^{l_2}\widetilde{\mathbf{\Sigma}}_r(\widehat{\mathbf{x}})\mathbf{W}_r^\top\bigg\|_2 &\leqslant C_1L + L\cdot C_2 + \sum_{p=2}^{l_2-l_1+1}\binom{l_2-l_1+1}{p}\cdot C_3\cdot\big[C_4\sqrt{s\log(m)/m}\big]^{p-1}\\ &\leqslant C_5L + C_3\cdot\sum_{p=2}^{l_2-l_1+1}\binom{l_2-l_1+1}{p}L^{-(p-1)}\\ &\leqslant C_5L + C_3L\cdot\sum_{p=0}^{l_2-l_1+1}\binom{l_2-l_1+1}{p}L^{-p}\\ &\leqslant C_5L + C_3L\cdot(1+L^{-1})^L\\ &\leqslant (C_5+C_3e)L,\end{aligned}$$

where $C_1,\ldots,C_5$ are absolute constants. This completes the proof of $\widetilde{\Lambda}_{l_1,l_2}(\widehat{\mathbf{x}})\leqslant C'L$. For the other bounds, it is easy to see that an additional sparse vector introduces an additional term $\sqrt{s\log(m)/m}$ in the final bound, and therefore the results can be obtained with the same proof. □

Lemma A.8. For $\beta>0$, $\mathbf{x}\in S^{d-1}$ and $l\in[L]$, define

$$\mathcal{S}_l(\mathbf{x},\beta)=\{j\in[m_l]:|\langle\mathbf{w}_{l,j},\mathbf{x}_{l-1}\rangle|\leqslant\beta\}.$$

For any $\delta>0$, if $m\geqslant C\max\{L^2d\log[m/(\tau_0\delta)],\beta^{-1}\sqrt{d\log(1/\tau_0)},\beta^{-1}\sqrt{\log(L/\delta)}\}$ for some large enough constant $C>0$, then with probability at least $1-\delta$, $|\mathcal{S}_l(\widehat{\mathbf{x}},\beta)|\leqslant 2m_l^{3/2}\beta$ for all $\widehat{\mathbf{x}}\in\mathcal{N}^*$ and $l\in[L]$.

Proof of Lemma A.8. Let $Z_{l,j}(\widehat{\mathbf{x}})=\mathbb{1}\{|\langle\mathbf{w}_{l,j},\widehat{\mathbf{x}}_{l-1}\rangle|\leqslant\beta\}$, $j=1,\ldots,m_l$. Then we have $\sum_{j=1}^{m_l}Z_{l,j}(\widehat{\mathbf{x}})=|\mathcal{S}_l(\widehat{\mathbf{x}},\beta)|$. By Lemma A.6, with probability at least $1-\delta/2$, $\|\widehat{\mathbf{x}}_l\|_2\geqslant 1/2$ for all $\widehat{\mathbf{x}}\in\mathcal{N}^*$ and $l\in[L]$. Condition on $\widehat{\mathbf{x}}_{l-1}$, for any $j=1,\ldots,m_l$, we have

$$\mathbb{E}(Z_{l,j}(\widehat{\mathbf{x}})|\widehat{\mathbf{x}}_{l-1})=\frac{1}{\sqrt{2\pi}}\cdot\frac{\sqrt{m_l}}{\sqrt{2}\|\widehat{\mathbf{x}}_{l-1}\|_2}\int_{-\beta}^{\beta}e^{-x^2m_l/(4\|\widehat{\mathbf{x}}_{l-1}\|_2)}\mathrm{d}x\leqslant 2\pi^{-1/2}\beta m_l^{1/2}.$$

Therefore by Bernstein inequality and union bound, with probability at least $1-\delta/2$, we have

$$\frac{1}{m_l}\sum_{j=1}^{m_l}Z_{l,j}(\widehat{\mathbf{x}})\leqslant 2\pi^{-1/2}\beta m_l^{1/2}+C_1\sqrt{\frac{\log[8\cdot(4/\tau_0)^dL/\delta]}{m_l}}\leqslant 2\beta m_l^{1/2}$$

for all $\widehat{\mathbf{x}}\in\mathcal{N}^*$ and $l\in[L]$, where C_1 is an absolute constant, and the last inequality follows by the assumption that

$$m\geqslant C\beta^{-1}\max\{\sqrt{d\log(1/\tau_0)},\sqrt{\log(L/\delta)}\}\geqslant\sqrt{2}\cdot C\beta^{-1}\sqrt{d\log(1/\tau_0)+\log(L/\delta)}$$

for some large enough absolute constant C. Therefore with probability at least $1-\delta$, we have

$$|\mathcal{S}_l(\widehat{\mathbf{x}},\beta)|\leqslant 2\beta m_l^{3/2}$$

for all $\widehat{\mathbf{x}}\in\mathcal{N}^*$ and $l\in[L]$. This completes the proof. □

Lemma A.9. Suppose that $\mathbf{W}_1, \ldots, \mathbf{W}_L$ are generated via Gaussian initialization. For any $\delta > 0$, If $m \geqslant C\tau_0^{-4/3} L^2 d \log[m/(\tau_0\delta)]$, $\tau_0 \leqslant \nu L^{-4}[\log(m)]^{-3/2}$ for some large enough absolute constant C and small enough absolute constant ν, then with probability at least $1-\delta$, we have

$$\begin{aligned}
&\|\mathbf{x}_l - \mathbf{x}'_l\|_2 \leqslant C_0 L \|\mathbf{x} - \mathbf{x}'\|_2,\\
&\|\mathbf{\Sigma}_l(\mathbf{x}) - \mathbf{\Sigma}_l(\mathbf{x}')\|_0 \leqslant 12 C_0^{2/3} C_0'^{2/3} L^{2/3} \tau_0^{2/3} m_l
\end{aligned}$$

for all $\mathbf{x}, \mathbf{x}' \in S^{d-1}$ with $\|\mathbf{x}-\mathbf{x}'\|_2 \leqslant \tau_0$ and $l = 1, \ldots, L$, where C_0, C_0' are the absolute constants introduced in the bounds of $\big\|\prod_{r=l_1}^{l_2} \widetilde{\mathbf{\Sigma}}_r(\widehat{\mathbf{x}})\mathbf{W}_r^\top\big\|_2$ and $\|\mathbf{W}_l\|_2$ by Lemma A.7 and Lemma A.1 respectively.

Proof of Lemma A.9. By Lemma A.8 and Lemma A.7, with probability at least $1-\delta/2$, we have

$$|\mathcal{S}_l(\widehat{\mathbf{x}}, \beta)| \leqslant 2\beta m_l^{3/2} \tag{A.1}$$

for all $\widehat{\mathbf{x}} \in \mathcal{N}^*$ and $l \in [L]$, and

$$\left\| \prod_{r=l_1}^{l_2} \widetilde{\mathbf{\Sigma}}_r \mathbf{W}_r^\top \right\|_2 \leqslant C_0 L \tag{A.2}$$

for all $1 \leqslant l_1 < l_2 \leqslant L$, $\widehat{\mathbf{x}} \in \mathcal{N}^*$, and diagonal matrices $\widetilde{\mathbf{\Sigma}}_{l_1}(\widehat{\mathbf{x}}), \ldots, \widetilde{\mathbf{\Sigma}}_{l_2}(\widehat{\mathbf{x}})$ satisfying $\|\widetilde{\mathbf{\Sigma}}_l(\widehat{\mathbf{x}}) - \mathbf{\Sigma}_l(\widehat{\mathbf{x}})\|_0 \leqslant \kappa^2 L^{-2} m/\log(m)$ and $|[\widetilde{\mathbf{\Sigma}}_l(\widehat{\mathbf{x}}) - \mathbf{\Sigma}_l(\widehat{\mathbf{x}})]_{jj}| \leqslant 3$, where κ is an absolute constant.

Now for any $\mathbf{x}, \mathbf{x}' \in S^{d-1}$ satisfying $\|\mathbf{x}-\mathbf{x}'\|_2 \leqslant \tau_0$, there exists $\widehat{\mathbf{x}} \in \mathcal{N}^*$ such that $\|\mathbf{x}-\widehat{\mathbf{x}}\|_2 \leqslant \tau_0$, and $\|\mathbf{x}'-\widehat{\mathbf{x}}\|_2 \leqslant 2\tau_0$. In the following we prove the following stronger results for $\mathbf{x}, \mathbf{x}'$ that covers the original lemma results:

$$\begin{aligned}
&\|\mathbf{x}_l - \mathbf{x}'_l\|_2 \leqslant C_0 L \cdot \|\mathbf{x}-\mathbf{x}'\|_2,\ \|\mathbf{x}_l - \widehat{\mathbf{x}}_l\|_2, \|\mathbf{x}'_l - \widehat{\mathbf{x}}_l\|_2 \leqslant 2C_0 L\tau_0,\\
&\|\mathbf{\Sigma}_l(\mathbf{x}) - \mathbf{\Sigma}_l(\mathbf{x}')\|_0, \|\mathbf{\Sigma}_l(\mathbf{x}) - \mathbf{\Sigma}_l(\widehat{\mathbf{x}})\|_0, \|\mathbf{\Sigma}_l(\mathbf{x}') - \mathbf{\Sigma}_l(\widehat{\mathbf{x}})\|_0 \leqslant 12 C_0^{2/3} C_0'^{2/3} L^{2/3} \tau_0^{2/3} m_l.
\end{aligned}$$

We prove the results above by induction in l. Suppose that for $r = 1, \ldots, l-1$ it holds that

$$\begin{aligned}
&\|\mathbf{x}_r - \mathbf{x}'_r\|_2 \leqslant C_0 L \cdot \|\mathbf{x}-\mathbf{x}'\|_2,\ \|\mathbf{x}_r - \widehat{\mathbf{x}}_r\|_2, \|\mathbf{x}'_r - \widehat{\mathbf{x}}_r\|_2 \leqslant 2C_0 L\tau_0,\\
&\|\mathbf{\Sigma}_r(\mathbf{x}) - \mathbf{\Sigma}_r(\mathbf{x}')\|_0, \|\mathbf{\Sigma}_r(\mathbf{x}) - \mathbf{\Sigma}_r(\widehat{\mathbf{x}})\|_0, \|\mathbf{\Sigma}_r(\mathbf{x}') - \mathbf{\Sigma}_r(\widehat{\mathbf{x}})\|_0 \leqslant 12 C_0^{2/3} C_0'^{2/3} L^{2/3} \tau_0^{2/3} m_r.
\end{aligned}$$

We first prove the bounds for the diagonal matrices on the l-th layer. By definition, we have

$$\|\mathbf{\Sigma}_l(\mathbf{x}) - \mathbf{\Sigma}_l(\widehat{\mathbf{x}})\|_0 = |\{j \in [m_l] : (\mathbf{w}_{l,j}^\top \mathbf{x}_{l-1}) \cdot (\mathbf{w}_{l,j}^\top \widehat{\mathbf{x}}_{l-1}) < 0\}| = s_l^{(1)}(\beta) + s_l^{(2)}(\beta),$$

where

$$\begin{aligned}
s_l^{(1)}(\beta) &= |\{j \in \mathcal{S}_l(\widehat{\mathbf{x}}, \beta) : (\mathbf{w}_{l,j}^\top \mathbf{x}_{l-1}) \cdot (\mathbf{w}_{l,j}^\top \widehat{\mathbf{x}}_{l-1}) < 0\}|,\\
s_l^{(2)}(\beta) &= |\{j \in \mathcal{S}_l^c(\widehat{\mathbf{x}}, \beta) : (\mathbf{w}_{l,j}^\top \mathbf{x}_{l-1}) \cdot (\mathbf{w}_{l,j}^\top \widehat{\mathbf{x}}_{l-1}) < 0\}|.
\end{aligned}$$

For $s_l^{(1)}(\beta)$, by (A.1) we have

$$s_l^{(1)}(\beta) \leqslant |\mathcal{S}_l(\widehat{\mathbf{x}}, \beta)| \leqslant 2\beta m_l^{3/2}.$$

For $s_l^{(2)}(\beta)$, by definition, $j \in \{j \in \mathcal{S}_l^c(\widehat{\mathbf{x}}, \beta) : (\mathbf{w}_{l,j}^\top \mathbf{x}_{l-1}) \cdot (\mathbf{w}_{l,j}^\top \widehat{\mathbf{x}}_{l-1}) < 0\}$ implies that

$$\big|\langle \mathbf{w}_{l,j}, \mathbf{x}_{l-1}\rangle - \langle \mathbf{w}_{l,j}, \widehat{\mathbf{x}}_{l-1}\rangle\big| \geqslant \beta.$$

Therefore, by Lemma A.1 and the induction assumptions, we have

$$s_l^{(2)}\beta^2 \leqslant \|\mathbf{W}_l^\top \mathbf{x}_{l-1} - \mathbf{W}_l^\top \widehat{\mathbf{x}}_{l-1}\|_2^2 \leqslant C_0'^2 \|\mathbf{x}_{l-1} - \widehat{\mathbf{x}}_{l-1}\|_2^2 \leqslant 4C_0^2 C_0'^2 L^2 \tau_0^2.$$

Combining the bounds of $s_l^{(1)}(\beta)$ and $s_l^{(2)}(\beta)$, we obtain

$$\|\mathbf{\Sigma}_l(\mathbf{x}) - \mathbf{\Sigma}_l(\widehat{\mathbf{x}})\|_0 \leqslant s_l^{(1)}(\beta) + s_l^{(2)}(\beta) \leqslant 2m_l^{3/2}\beta + \frac{4C_0^2 C_0'^2 L^2 \tau_0^2}{\beta^2}.$$

Setting $\beta = C_0^{2/3} C_0'^{2/3} L^{2/3} \tau_0^{2/3} m_l^{-1/2}$, we obtain

$$\|\mathbf{\Sigma}_l(\mathbf{x}) - \mathbf{\Sigma}_l(\widehat{\mathbf{x}})\|_0 \leqslant 6C_0^{2/3} C_0'^{2/3} L^{2/3} \tau_0^{2/3} m_l.$$

By the exact same proof, we have

$$\|\mathbf{\Sigma}_l(\mathbf{x}') - \mathbf{\Sigma}_l(\widehat{\mathbf{x}})\|_0 \leqslant 6C_0^{2/3} C_0'^{2/3} L^{2/3} \tau_0^{2/3} m_l,$$

and therefore

$$\|\mathbf{\Sigma}_l(\mathbf{x}) - \mathbf{\Sigma}_l(\mathbf{x}')\|_0 \leqslant 12C_0^{2/3} C_0'^{2/3} L^{2/3} \tau_0^{2/3} m_l.$$

This completes the proof of the $\|\cdot\|_0$-bounds for diagonal matrices on the l-th layer.

Now, based on the bounds obtained above and the induction assumptions on $\|\mathbf{\Sigma}_r(\mathbf{x}) - \mathbf{\Sigma}_r(\mathbf{x}')\|_0$, $r = 1, \ldots, l-1$, we derive the bound for $\|\mathbf{x}_l - \mathbf{x}_l'\|_2$. By the assumption that $\tau_0 \leqslant \nu L^{-4}[\log(m)]^{-3/2}$ for some small enough absolute constant $\nu > 0$, we have

$$\|\mathbf{\Sigma}_r(\mathbf{x}) - \mathbf{\Sigma}_r(\mathbf{x}')\|_0 \leqslant (\kappa/2) \cdot L^{-2} m/\log(m),\ r = 1, \ldots, l.$$

Define diagonal matrices $\check{\mathbf{\Sigma}}_r$, $r = 1, \ldots, l$ as follows:

$$(\check{\mathbf{\Sigma}}_r)_{jj} := [\mathbf{\Sigma}_r(\mathbf{x}) - \mathbf{\Sigma}_r(\mathbf{x}')]_{jj} \cdot \frac{\mathbf{w}_{r,j}^\top \mathbf{x}_{r-1}'}{\mathbf{w}_{r,j}^\top \mathbf{x}_{r-1} - \mathbf{w}_{r,j}^\top \mathbf{x}_{r-1}'},\ j = 1, \ldots, m_l.$$

Then we have

$$\begin{aligned}
\mathbf{x}_l - \mathbf{x}_l' &= [\mathbf{\Sigma}_l(\mathbf{x}) + \check{\mathbf{\Sigma}}_l](\mathbf{W}_l^\top \mathbf{x}_{l-1} - \mathbf{W}_l^\top \mathbf{x}_{l-1}') \\
&= [\mathbf{\Sigma}_l(\mathbf{x}) + \check{\mathbf{\Sigma}}_l]\mathbf{W}_l^\top [\mathbf{\Sigma}_{l-1}(\mathbf{x}) + \check{\mathbf{\Sigma}}_{l-1}]\mathbf{W}_{l-1}^\top (\mathbf{x}_{l-2} - \mathbf{x}_{l-2}') \\
&= \cdots \\
&= \bigg\{\prod_{r=1}^{l} [\mathbf{\Sigma}_l(\mathbf{x}) + \check{\mathbf{\Sigma}}_l]\mathbf{W}_l^\top\bigg\}(\mathbf{x} - \mathbf{x}').
\end{aligned}$$

Note that $|(\breve{\boldsymbol{\Sigma}}_r)_{jj}| \leqslant 1$ and

$$\begin{aligned} \|\boldsymbol{\Sigma}_r(\mathbf{x}) + \breve{\boldsymbol{\Sigma}}_r - \boldsymbol{\Sigma}_r(\widehat{\mathbf{x}})\|_0 &\leqslant \|\boldsymbol{\Sigma}_r(\mathbf{x}) - \boldsymbol{\Sigma}_r(\widehat{\mathbf{x}})\|_0 + \|\breve{\boldsymbol{\Sigma}}_r\|_0 \\ &= \|\boldsymbol{\Sigma}_r(\mathbf{x}) - \boldsymbol{\Sigma}_r(\widehat{\mathbf{x}})\|_0 + \|\boldsymbol{\Sigma}_r(\mathbf{x}) - \boldsymbol{\Sigma}_r(\mathbf{x}')\|_0 \\ &\leqslant \kappa L^{-2} m/\log(m) \end{aligned}$$

for all $r = 1, \ldots, l$. Therefore by (A.2), we have

$$\|\mathbf{x}_l - \mathbf{x}_l'\|_2 \leqslant C_0 L \|\mathbf{x}_l - \mathbf{x}_l'\|_2.$$

With the exact same proof, it can be shown that

$$\|\mathbf{x}_l - \widehat{\mathbf{x}}_l\|_2 \leqslant C_0 L \|\mathbf{x}_l - \widehat{\mathbf{x}}_l\|_2 \leqslant 2C_0 L\tau_0,\ \|\mathbf{x}_l' - \widehat{\mathbf{x}}_l\|_2 \leqslant C_0 L \|\mathbf{x}_l' - \widehat{\mathbf{x}}_l\|_2 \leqslant 2C_0 L\tau_0.$$

This completes the proof. □

Lemma A.10. Under the same assumptions as Lemma A.9, for any $\delta > 0$, if

$$m \geqslant CL^8 d \log(m/\delta) \cdot [\log(m)]^2$$

for some large enough absolute constant C, then

$$\|\mathbf{x}_l - \mathbf{x}_l'\|_2 \leqslant C' L \|\mathbf{x} - \mathbf{x}'\|_2$$

for all $\mathbf{x}, \mathbf{x}' \in S^{d-1}$, where $C' > 0$ is an absolute constant.

Proof of Lemma A.10. Let ν be the absolute constant in the assumption on τ_0 in Lemma A.9, and set $\tau_0 = \nu L^{-4}[\log(m)]^{-3/2}$. For any $\mathbf{x}, \mathbf{x}' \in S^{d-1}$, we consider the geodesic path on S^{d-1} connecting $\mathbf{x}$ and $\mathbf{x}'$. Obviously, the path is part of the 2-dimensional unit circle in the subspace $\mathrm{span}\{\mathbf{x}, \mathbf{x}'\}$. Then there exist a large enough integer $N = N(\tau_0)$ and points $\mathbf{x}_1, \ldots, \mathbf{x}_N$ on this geodesic path such that $\|\mathbf{x}_i - \mathbf{x}_{i+1}\|_2 \leqslant \tau_0$, $i = 1, \ldots, N-1$, $\|\mathbf{x} - \mathbf{x}_1\|_2 \leqslant \tau_0$, $\|\mathbf{x}_N - \mathbf{x}'\|_2 \leqslant \tau_0$. Denote $\mathbf{x}_0 = \mathbf{x}$, $\mathbf{x}_{N+1} = \mathbf{x}'$. Then by definition we have

$$\sum_{i=0}^{N} \|\mathbf{x}_i - \mathbf{x}_{i+1}\|_2 \leqslant \pi \cdot \|\mathbf{x} - \mathbf{x}'\|_2.$$

Moreover, by Lemma A.9, we have

$$\|\mathbf{x}_{l,i} - \mathbf{x}_{l,i+1}\|_2 \leqslant C_1 L \|\mathbf{x}_i - \mathbf{x}_{i+1}\|_2,\ i = 0, \ldots, N,$$

where C_1 is an absolute constant. Therefore by triangle inequality we have

$$\|\mathbf{x}_l - \mathbf{x}_l'\|_2 \leqslant C_1 L \cdot \sum_{i=0}^{N} \|\mathbf{x}_i - \mathbf{x}_{i+1}\|_2 \leqslant C_1 \pi L \cdot \|\mathbf{x} - \mathbf{x}'\|_2.$$

This completes the proof. □

Proof of Theorem 5.3. We can prove the results given in Theorem 5.3. The Lipschitz continuity of mappings $\mathbf{x} \to \mathbf{x}_l$ has been proved in Lemma A.10. For the bounds of $\Gamma_{l_1,l_2,s}(\mathbf{x})$, $\Gamma'_{l_1,l_2,s}(\mathbf{x})$,

$\Gamma''_{l_1,l_2,s}(\mathbf{x})$ and $\Lambda_{l_1,l_2}(\mathbf{x})$, we know that as long as $\tau_0 \leqslant \nu \min\{L^{-4}[\log(m)]^{-3/2}, L^{-1}(s/m)^{3/2}\}$ for some small enough absolute constant ν, with probability at least $1-\delta/2$ the results of Lemma A.9 and Lemma A.7 hold. For any $\mathbf{x} \in S^{d-1}$, there exists $\widehat{\mathbf{x}} \in \mathcal{N}^*$ such that $\|\mathbf{x} - \widehat{\mathbf{x}}\|_2 \leqslant \tau_0$. By Lemma A.9 we have

$$\|\mathbf{\Sigma}_l(\mathbf{x}) - \mathbf{\Sigma}_l(\mathbf{x}')\|_0 \leqslant C_2 L^{2/3} \tau_0^{2/3} m_l \leqslant s.$$

for all $\widehat{\mathbf{x}} \in \mathcal{N}^*$ and $1 \leqslant l_1 < l_2 \leqslant L$, where C_1 is an absolute constant. Therefore by Lemma A.7, we have

$$\Gamma_{l_1,l_2,s}(\mathbf{x}) = \sup_{\substack{\|\mathbf{a}\|_2=\|\mathbf{b}\|_2=1,\\ \|\mathbf{a}\|_0,\|\mathbf{b}\|_0\leqslant s}} \mathbf{b}^\top \mathbf{W}_{l_2}^\top \Bigg(\prod_{r=l_1}^{l_2-1} \mathbf{\Sigma}_r(\mathbf{x}) \mathbf{W}_r^\top \Bigg) \mathbf{a} \leqslant C_2 \sqrt{s \log(m)/m}$$

for all $\mathbf{x} \in S^{d-1}$ and $1 \leqslant l_1 < l_2 \leqslant L$, where C_2 is an absolute constant. Similarly, combining the results of Lemma A.9 and Lemma A.7 also gives

$$\Gamma'_{l_1,l_2,s}(\mathbf{x}), \Gamma''_{l_1,l_2,s}(\mathbf{x}) \leqslant C_3,\ \Gamma'''_{l_1,l_2,s}(\mathbf{x}) \leqslant C_4 \sqrt{s \log(m)},\ \Lambda_{l_1,l_2}(\mathbf{x}) \leqslant C_5 L$$

for all $\mathbf{x} \in S^{d-1}$ and $1 \leqslant l_1 < l_2 \leqslant L$, where C_3, C_4, C_5 are absolute constants.

We now prove the bounds of $\|\mathbf{x}_l\|_2$. Let $\tau'_0 = \sqrt{d/m}$. Let $\mathcal{N}' = \mathcal{N}(S^{d-1}, \tau'_0)$. Then we have $|\mathcal{N}'| \leqslant (4/\tau'_0)^d$. With the same proof of Lemma A.6, as long as $m \geqslant CL^2 d \log[m/(\tau'_0\delta)]$ for some large enough absolute constant C, with probability at least $1-\delta/2$, we have

$$\big|\|\check{\mathbf{x}}_l\|_2 - 1\big| \leqslant C_6 L \sqrt{\frac{d \log[m/(\tau'_0\delta)]}{m}} \leqslant C_7 L \sqrt{\frac{d \log(m/\delta)}{m}}$$

for all $\check{\mathbf{x}} \in \mathcal{N}'$, where C_6, C_7 are absolute constants. For any $\mathbf{x} \in S^{d-1}$, there exists $\check{\mathbf{x}} \in \mathcal{N}'$ such that $\|\mathbf{x} - \check{\mathbf{x}}\|_2 \leqslant \tau'_0$. Then by Lemma A.10, as long as $m \geqslant CL^8 d \log(m/\delta) \cdot [\log(m)]^2$ for some large enough constant C, $\|\mathbf{x}_l - \check{\mathbf{x}}_l\|_2 \leqslant C_8 L \tau'_0$, where C_8 is an absolute constant. Therefore we have

$$\big|\|\mathbf{x}_l\|_2 - 1\big| \leqslant C_9 L \sqrt{\frac{d \log(m/\delta)}{m}}$$

for all $\mathbf{x} \in S^{d-1}$, where C_9 is an absolute constant.

Finally for the bound of $f_{\mathbf{W}}(\boldsymbol{x}_i)$ for $i = 1, \ldots, n$. We have

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

Therefore condition on $\boldsymbol{x}_{L-1,i}$, by the bound $\|\boldsymbol{x}_{L-1,i}\|_2 \leqslant 2$ we have

$$\|\sigma(\mathbf{w}_{L,j}^\top \boldsymbol{x}_{L-1,i}) - \sigma(-\mathbf{w}_{L,j+m_L/2}^\top \boldsymbol{x}_{L-1,i})\|_{\psi_2} \leqslant C_{10}$$

for some absolute constant C_{10}. Therefore by Hoeffding inequality and union bound,

$$|f_{\mathbf{W}}(\boldsymbol{x}_i)| \leqslant C_{11} \sqrt{\log(n/\delta)}$$

for all $i \in [n]$, where C_{11} is an absolute constant. This completes the proof. $\square$

B Proof of Theorem 5.4

Lemma B.1. For any $\delta > 0$, if $m \geqslant C \max\{dL^2 \log(m/\delta), \beta^{-1} d\sqrt{\log[L/(\beta\delta)]}\}$ for some large enough constant $C > 0$, then with probability at least $1 - \delta$, the sets

$$\mathcal{S}_l(\mathbf{x}, \beta) = \{j \in [m_l] : |\langle \mathbf{w}_{l,j}, \mathbf{x}_{l-1}\rangle| \leqslant \beta\},\ \mathbf{x} \in S^{d-1}, l = 1, \ldots, L$$

have cardinality bounds $|\mathcal{S}_l(\mathbf{x}, \beta)| \leqslant 4m_l^{3/2}\beta$ for all $\mathbf{x} \in S^{d-1}$ and $l \in [L]$.

Proof of Lemma B.1. By Theorem 5.3, with probability at least $1 - \delta/2$, we have $\|\mathbf{w}_{l,j}\|_2 \leqslant C_1$, $\mathbf{x}_{l-1} \geqslant 1/2$ for all $\mathbf{x} \in S^{d-1}$, $l \in [L]$ and $j \in [m_l]$, where C_1 is an absolute constant. Let $\mathcal{N} = \mathcal{N}[S^{d-1}, \beta/C_1]$ be a β/C_1-net covering S^{d-1}, then by Lemma 5.2 in Vershynin (2010), we have

$$|\mathcal{N}| \leqslant (4C_1/\beta)^d.$$

For any fixed $l \in \{1, \ldots, L\}$ and $\widehat{\mathbf{x}} \in \mathcal{N}$, let $Z_{l,j}(\widehat{\mathbf{x}}) = \mathbb{1}\{|\langle \mathbf{w}_{l,j}, \widehat{\mathbf{x}}_{l-1}\rangle| \leqslant 2\beta\}$, $j = 1, \ldots, m_l$. Then we have $\sum_{j=1}^{m_l} Z_{l,j}(\widehat{\mathbf{x}}) = |\mathcal{S}_l(\widehat{\mathbf{x}}, 2\beta)|$. Condition on $\mathbf{x}_{l-1}$, for any $j = 1, \ldots, m_l$, we have

$$\mathbb{E}(Z_{l,j}(\widehat{\mathbf{x}})|\widehat{\mathbf{x}}_{l-1}) \leqslant \frac{1}{\sqrt{2\pi}} \cdot \frac{\sqrt{m_l}}{\sqrt{2}\|\widehat{\mathbf{x}}_{l-1}\|_2} \int_{-2\beta}^{2\beta} e^{-x^2 m_l/(4\|\widehat{\mathbf{x}}_{l-1}\|_2)} \mathrm{d}x \leqslant 4\pi^{-1/2}\beta m_l^{1/2}.$$

Therefore by Bernstein inequality and union bound, with probability at least $1 - \delta/2$, we have

$$\frac{1}{m_l} \sum_{j=1}^{m_l} Z_{l,j}(\widehat{\mathbf{x}}) \leqslant 4\pi^{-1/2}\beta m_l^{1/2} + C_4 \sqrt{\frac{\log[8 \cdot (4C_1/\beta)^d L/\delta]}{m_l}} \leqslant 4\beta m_l^{1/2}$$

for all $\widehat{\mathbf{x}} \in \mathcal{N}$ and $l \in [L]$, where the last inequality follows by the assumption that

$$m \geqslant C\beta^{-1}\sqrt{d \log[L/(\beta\delta)]}$$

for some large enough absolute constant C. Therefore with probability at least $1 - \delta$, we have

$$|\mathcal{S}_l(\widehat{\mathbf{x}}, 2\beta)| \leqslant 4\beta m_l^{3/2}$$

for all $\widehat{\mathbf{x}} \in \mathcal{N}$ and $l \in [L]$. Now for any $\mathbf{x} \in S^{d-1}$, there exists $\widehat{\mathbf{x}} \in \mathcal{N}$ such that $\|\mathbf{x} - \widehat{\mathbf{x}}\|_2 \leqslant \beta/C_1$. By definition and Lemma A.10, we have

$$\mathcal{S}_l(\mathbf{x}, \beta) \subseteq \mathcal{S}_l(\widehat{\mathbf{x}}, 2\beta),$$

therefore

$$|\mathcal{S}_l(\mathbf{x}, \beta)| \leqslant |\mathcal{S}_l(\widehat{\mathbf{x}}, 2\beta)| \leqslant 4\beta m_l^{3/2}.$$

This completes the proof. $\square$

Lemma B.2. Suppose that $\mathbf{W}_1, \ldots, \mathbf{W}_L$ are generated via Gaussian initialization, and the results of Theorem 5.3 all hold. If diagonal matrices $\widetilde{\mathbf{\Sigma}}_l(\mathbf{x})$, $l = 1, \ldots, L$ satisfy $\|\widetilde{\mathbf{\Sigma}}_l(\mathbf{x}) - \mathbf{\Sigma}_l(\mathbf{x})\|_0 \leqslant s$ and

$|[\widetilde{\mathbf{\Sigma}}_l(\mathbf{x}) - \mathbf{\Sigma}_l(\mathbf{x})]_{jj}| \leqslant C$ for all $\mathbf{x} \in S^{d-1}$, $l = 1, \ldots, L$ and $j = 1, \ldots, m_l$, where C is an absolute constant that can be arbitrarily large and $s\log(m)/m \leqslant \kappa L^{-2}$ for some small enough absolute constant κ, then

$$\left\| \prod_{r=l_1}^{l_2} \widetilde{\mathbf{\Sigma}}_r(\mathbf{x})\mathbf{W}_r^\top \right\|_2 \leqslant C'L$$

for all $1 \leqslant l_1 < l_2 \leqslant L$, where C' is an absolute constant.

Proof of Lemma B.2. The proof is almost identical to the proof of Lemma A.7, except now that Theorem 5.3 gives the bound $\Lambda_{l_1,l_2}(\mathbf{x}) \leqslant C_1 L$ uniformly for all $\mathbf{x} \in S^{d-1}$, where C_1 is an absolute constant, the original result for $\widehat{\mathbf{x}} \in \mathcal{N}^*$ given in Lemma A.7 can be naturally extended to all $\mathbf{x} \in S^{d-1}$. We therefore omit the detailed proof. □

Lemma B.3. Suppose that $\mathbf{W}_1, \ldots, \mathbf{W}_L$ are generated via Gaussian initialization, and the results of Theorem 5.3 all hold. For $\tau > 0$, let $\widetilde{\mathbf{W}}_1, \ldots, \widetilde{\mathbf{W}}_L$ be perturbed matrices satisfying $\|\widetilde{\mathbf{W}}_l - \mathbf{W}_l\|_2 \leqslant \tau$, $l = 1, \ldots, L$. Moreover, let $\widetilde{\mathbf{\Sigma}}_l(\mathbf{x})$, $\mathbf{x} \in S^{d-1}$, $l = 1, \ldots, L$ be diagonal matrices satisfying $\|\widetilde{\mathbf{\Sigma}}_l(\mathbf{x}) - \mathbf{\Sigma}_l(\mathbf{x})\|_0 \leqslant s$ and $|[\widetilde{\mathbf{\Sigma}}_l(\mathbf{x}) - \mathbf{\Sigma}_l(\mathbf{x})]_{jj}| \leqslant C$ for all $\mathbf{x} \in S^{d-1}$, $l = 1, \ldots, L$ and $j = 1, \ldots, m_l$, where C is an absolute constant that can be arbitrarily large. If $\tau, s\log(m)/m \leqslant \kappa L^{-2}$ for some small enough absolute constant κ, then

$$\left\| \prod_{r=l_1}^{l_2} \widetilde{\mathbf{\Sigma}}_r(\mathbf{x})\widetilde{\mathbf{W}}_r^\top \right\|_2 \leqslant C'L$$

for any $1 \leqslant l_1 < l_2 \leqslant L$, where C' is an absolute constant.

Proof of Lemma B.3. The proof is similar to the proof of Lemma A.7 and Lemma B.2. Note that we have

$$\widetilde{\mathbf{\Sigma}}_r(\mathbf{x})\widetilde{\mathbf{W}}_r^\top = \widetilde{\mathbf{\Sigma}}_r(\mathbf{x})\mathbf{W}_r^\top + \widetilde{\mathbf{\Sigma}}_r(\mathbf{x})(\widetilde{\mathbf{W}}_r - \mathbf{W}_r)^\top.$$

Therefore, let $\mathcal{A}_r(\mathbf{x}) = \{\widetilde{\mathbf{\Sigma}}_r(\mathbf{x})\mathbf{W}_r^\top, \widetilde{\mathbf{\Sigma}}_r(\mathbf{x})(\widetilde{\mathbf{W}}_r - \mathbf{W}_r)^\top\}$, $r = l_1, \ldots, l_2$, then we have

$$\prod_{r=l_1}^{l_2} \widetilde{\mathbf{\Sigma}}_r(\mathbf{x})\widetilde{\mathbf{W}}_r^\top = \sum_{\mathbf{A}_{l_1}(\mathbf{x}) \in \mathcal{A}_{l_1}(\mathbf{x}), \ldots, \mathbf{A}_{l_2}(\mathbf{x}) \in \mathcal{A}_{l_2}(\mathbf{x})} \left[\prod_{r=l_1}^{l_2} \mathbf{A}_r(\mathbf{x}) \right].$$

For any fixed sequence $\mathbf{A}_{l_1}(\mathbf{x}) \in \mathcal{A}_{l_1}(\mathbf{x}), \ldots, \mathbf{A}_{l_2}(\mathbf{x}) \in \mathcal{A}_{l_2}(\mathbf{x})$, define

$$p = |\{r : \mathbf{A}_r(\mathbf{x}) = \widetilde{\mathbf{\Sigma}}_r(\mathbf{x})(\widetilde{\mathbf{W}}_r - \mathbf{W}_r)^\top\}|.$$

Then it is clear that $\prod_{r=l_1}^{l_2} \mathbf{A}_r(\mathbf{x})$ can be written as a product of p matrices of the form $\widetilde{\mathbf{\Sigma}}_r(\mathbf{x})(\widetilde{\mathbf{W}}_r - \mathbf{W}_r)^\top$ and at most $p+1$ matrices of the form $\prod_{r=r_1}^{r_2} \widetilde{\mathbf{\Sigma}}_r(\mathbf{x})\mathbf{W}_r^\top$. Therefore by (ii) in Theorem 5.3 and Lemma B.2, there exists an absolute constant C_1 such that

$$\prod_{r=l_1}^{l_2} \mathbf{A}_r(\mathbf{x}) \leqslant (C_1 L)^{p+1} \cdot (C\tau)^p.$$

Therefore, we have

$$\begin{aligned}\bigg\|\prod_{r=l_1}^{l_2}\widetilde{\boldsymbol{\Sigma}}_r(\mathbf{x})\widetilde{\mathbf{W}}_r^\top\bigg\|_2 &\leqslant \sum_{p=0}^{l_2-l_1+1}\binom{l_2-l_1+1}{p}(C_1L)^{p+1}\cdot(C\tau)^p\\ &\leqslant C_1L\cdot\sum_{p=0}^{l_2-l_1+1}\binom{l_2-l_1+1}{p}(C_2L\tau)^p\\ &\leqslant C_1L\cdot\sum_{p=0}^{l_2-l_1+1}\binom{l_2-l_1+1}{p}(L^{-1})^p\\ &\leqslant C_1L\cdot(1+L^{-1})^L\\ &\leqslant C_1eL,\end{aligned}$$

where C_2 is an absolute constant. This completes the proof. □

Lemma B.4. Suppose that $\mathbf{W}_1,\ldots,\mathbf{W}_L$ are generated via Gaussian initialization, and the results of Theorem 5.3 all hold. For $\tau>0$, let $\widetilde{\mathbf{W}}_1,\ldots,\widetilde{\mathbf{W}}_L$ be perturbed matrices satisfying $\|\widetilde{\mathbf{W}}_l-\mathbf{W}_l\|_2\leqslant\tau$, $l=1,\ldots,L$. Moreover, let $\widetilde{\boldsymbol{\Sigma}}_l(\mathbf{x})$, $\mathbf{x}\in S^{d-1}$, $l=1,\ldots,L$ be diagonal matrices satisfying $\|\widetilde{\boldsymbol{\Sigma}}_l(\mathbf{x})-\boldsymbol{\Sigma}_l(\mathbf{x})\|_0\leqslant s$ and $|[\widetilde{\boldsymbol{\Sigma}}_l(\mathbf{x})-\boldsymbol{\Sigma}_l(\mathbf{x})]_{jj}|\leqslant C$ for all $\mathbf{x}\in S^{d-1}$, $l=1,\ldots,L$ and $j=1,\ldots,m_l$, where C is an absolute constant that can be arbitrarily large. If $\tau,\sqrt{s\log(m)/m}\leqslant\kappa L^{-3}$ for some small enough absolute constant κ, then

$$\mathbf{v}^\top\Bigg(\prod_{r=l}^{L}\widetilde{\boldsymbol{\Sigma}}_r(\mathbf{x})\widetilde{\mathbf{W}}_r^\top\Bigg)\mathbf{a}\leqslant C'[L^2\tau\sqrt{m}+\sqrt{s\log(m)}].$$

for all $\mathbf{a}\in\mathbb{R}^{m_{l-1}}$ satisfying $\|\mathbf{a}\|_2=1$, $\|\mathbf{a}\|_0\leqslant s$, all $\mathbf{x}\in S^{d-1}$ and all $l\in[L]$, where C' is an absolute constant.

Proof of Lemma B.4. The proof is similar to the proof of Lemma B.3. For $r=l,\ldots,L$, let

$$\mathcal{A}_r(\mathbf{x})=\{\boldsymbol{\Sigma}_r(\mathbf{x})\mathbf{W}_r^\top,[\widetilde{\boldsymbol{\Sigma}}_r(\mathbf{x})-\boldsymbol{\Sigma}_r(\mathbf{x})]\mathbf{W}_r^\top,\widetilde{\boldsymbol{\Sigma}}_r(\mathbf{x})(\widetilde{\mathbf{W}}_r-\mathbf{W}_r)^\top\},$$

then we have

$$\mathbf{v}^\top\Bigg[\prod_{r=l}^{L}\widetilde{\boldsymbol{\Sigma}}_r(\mathbf{x})\widetilde{\mathbf{W}}_r^\top\Bigg]\mathbf{a}=m_L^{1/2}\cdot\sum_{\mathbf{A}_l(\mathbf{x})\in\mathcal{A}_l(\mathbf{x}),\ldots,\mathbf{A}_L(\mathbf{x})\in\mathcal{A}_L(\mathbf{x})}m_L^{-1/2}\mathbf{v}^\top\Bigg[\prod_{r=l}^{L}\mathbf{A}_r(\mathbf{x})\Bigg]\mathbf{a}.$$

We denote by $|\boldsymbol{\Sigma}|$ the diagonal matrix with absolute values of elements of $\boldsymbol{\Sigma}$ on the corresponding entries. For each sequence $\mathbf{A}_{l,i},\ldots,\mathbf{A}_{L,i}$, denote

$$\widehat{\boldsymbol{\Sigma}}_r(\mathbf{x})=\begin{cases}|\widetilde{\boldsymbol{\Sigma}}_{r-1}(\mathbf{x})-\boldsymbol{\Sigma}_{r-1}(\mathbf{x})|, & \text{if } r\geqslant l+1 \text{ and } \mathbf{A}_{r-1}(\mathbf{x})=[\widetilde{\boldsymbol{\Sigma}}_{r-1}(\mathbf{x})-\boldsymbol{\Sigma}_{r-1}(\mathbf{x})]\mathbf{W}_{r-1}^\top,\\ \mathbf{I}, & \text{otherwise.}\end{cases}$$

Then we have

$$\prod_{r=l}^{L}\mathbf{A}_r(\mathbf{x})=\prod_{r=l}^{L}\mathbf{A}_r(\mathbf{x})\widehat{\boldsymbol{\Sigma}}_r(\mathbf{x}).$$

When $\mathbf{A}_r(\mathbf{x}) = \mathbf{\Sigma}_r(\mathbf{x})\mathbf{W}_r^\top$ for all $r = l, \ldots, L$, then the bound of $m_L^{-1/2}\mathbf{v}^\top\big[\prod_{r=l}^L \mathbf{A}_r(\mathbf{x})\big]\mathbf{a}$ is given by (ii) in Theorem 5.3. For all the other terms in the expansion, we consider sequences of the form $\mathbf{B}_{r_2}(\mathbf{x})\big[\prod_{r=r_1+1}^{r_2-1} \mathbf{\Sigma}_r(\mathbf{x})\mathbf{W}_r^\top\big]\mathbf{B}_{r_1}(\mathbf{x})$, where

$$\begin{aligned}
&\mathbf{B}_{r_2}(\mathbf{x}) \in \{[\widetilde{\mathbf{\Sigma}}_{r_2}(\mathbf{x}) - \mathbf{\Sigma}_{r_2}(\mathbf{x})]\mathbf{W}_{r_2}^\top, \widetilde{\mathbf{\Sigma}}_{r_2}(\mathbf{x})(\widetilde{\mathbf{W}}_{r_2} - \mathbf{W}_{r_2})^\top, m_L^{-1/2}\mathbf{v}^\top\},\\
&\mathbf{B}_{r_1}(\mathbf{x}) \in \{|\widetilde{\mathbf{\Sigma}}_{r_1}(\mathbf{x}) - \mathbf{\Sigma}_{r_1}(\mathbf{x})|, \widetilde{\mathbf{\Sigma}}_{r_1}(\mathbf{x})(\widetilde{\mathbf{W}}_{r_1} - \mathbf{W}_{r_1})^\top, \mathbf{a}\}.
\end{aligned}$$

By (ii) in Theorem 5.3, there exists an absolute constant C_1 such that different choices of $\mathbf{B}_{r_1}(\mathbf{x})$ and $\mathbf{B}_{r_2}(\mathbf{x})$ give the following bounds of $\|\mathbf{B}_{r_2}(\mathbf{x})(\prod_{r=r_1+1}^{r_2-1} \mathbf{\Sigma}_r(\mathbf{x})\mathbf{W}_r^\top)\mathbf{B}_{r_1}(\mathbf{x})\|_2$:

1. If $\mathbf{B}_{r_1}(\mathbf{x}) \in \big\{|\widetilde{\mathbf{\Sigma}}_{r_1}(\mathbf{x}) - \mathbf{\Sigma}_{r_1}(\mathbf{x})|, \mathbf{a}\big\}$, $\mathbf{B}_{r_2}(\mathbf{x}) \in \big\{[\widetilde{\mathbf{\Sigma}}_{r_2}(\mathbf{x}) - \mathbf{\Sigma}_{r_2}(\mathbf{x})]\mathbf{W}_{r_2}^\top, m_L^{-1/2}\mathbf{v}^\top\big\}$, then
$$\left\|\mathbf{B}_{r_2}(\mathbf{x})\left[\prod_{r=r_1+1}^{r_2-1} \mathbf{\Sigma}_r(\mathbf{x})\mathbf{W}_r^\top\right]\mathbf{B}_{r_1}(\mathbf{x})\right\|_2 \leqslant C_1\sqrt{s\log(M)/m}.$$

2. If $\mathbf{B}_{r_1}(\mathbf{x}) = \widetilde{\mathbf{\Sigma}}_{r_1}(\mathbf{x})(\widetilde{\mathbf{W}}_{r_1} - \mathbf{W}_{r_1})^\top$, $\mathbf{B}_{r_2}(\mathbf{x}) = \widetilde{\mathbf{\Sigma}}_{r_2}(\mathbf{x})(\widetilde{\mathbf{W}}_{r_2} - \mathbf{W}_{r_2})^\top$, then
$$\left\|\mathbf{B}_{r_2}(\mathbf{x})\left[\prod_{r=r_1+1}^{r_2-1} \mathbf{\Sigma}_r(\mathbf{x})\mathbf{W}_r^\top\right]\mathbf{B}_{r_1}(\mathbf{x})\right\|_2 \leqslant C_1L\tau^2.$$

3. Otherwise,
$$\left\|\mathbf{B}_{r_2}(\mathbf{x})\left[\prod_{r=r_1+1}^{r_2-1} \mathbf{\Sigma}_r(\mathbf{x})\mathbf{W}_r^\top\right]\mathbf{B}_{r_1}(\mathbf{x})\right\|_2 \leqslant C_1\tau.$$

Let

$$p_1 = \big|\{r : \mathbf{A}_r(\mathbf{x}) = \widetilde{\mathbf{\Sigma}}_r(\mathbf{x})(\widetilde{\mathbf{W}}_r - \mathbf{W}_r)^\top\}\big|,\ p_2 = \big|\{r : \mathbf{A}_r(\mathbf{x}) = (\widetilde{\mathbf{\Sigma}}_r(\mathbf{x}) - \mathbf{\Sigma}_r(\mathbf{x}))\mathbf{W}_r^\top\}\big|.$$

Then it is easy to see that the bound of $\big\|\prod_{r=l}^L \mathbf{A}_r(\mathbf{x})\big\|_2$ contains a term $(C_1L\tau)^{p_1}$. Moreover, if $p_2 > p_1 - 1$, then the matrix products discussed in the first case above introduce another term in the bound, which is $(C_1\sqrt{s\log(M)/m})^{p_2-p_1+1}$. Therefore we have

$$m_L^{-1/2}\mathbf{v}^\top\left[\prod_{r=l}^L \widetilde{\mathbf{\Sigma}}_r(\mathbf{x})\widetilde{\mathbf{W}}_r^\top\right]\mathbf{a} \leqslant I_{1,i} + I_{2,i}, \tag{B.1}$$

where

$$\begin{aligned}
I_{1,i} &= \sum_{p_1=1}^{L-l+1}\sum_{p_2=0}^{L-l+1-p_1}\binom{L-l+1}{p_1}\binom{L-l+1-p_1}{p_2}(C_1\tau)^{p_1}(C_1L\sqrt{s\log(M)/m})^{\max\{p_2+1-p_1,0\}},\\
I_{2,i} &= \sum_{p_2=0}^{L-l+1}\binom{L-l+1}{p_2}(C_1\sqrt{s\log(M)/m})^{p_2+1}.
\end{aligned}$$

For $I_{1,i}$, we have

$$\begin{aligned}
I_{1,i} &\leqslant C_1 L\tau \cdot \sum_{p_1=1}^{L-l+1}\sum_{p_2=0}^{L-l+1-p_1}\binom{L-l+1}{p_1}\binom{L-l+1-p_1}{p_2}(C_1L\tau)^{p_1-1}(C_1\sqrt{s\log(M)/m})^{\max\{p_2+1-p_1,0\}}\\
&\leqslant C_1 L\tau \cdot \sum_{p_1=1}^{L-l+1}\sum_{p_2=0}^{L-l+1-p_1}\binom{L-l+1}{p_1}\binom{L-l+1-p_1}{p_2}L^{-2(p_1-1)}L^{-2\max\{p_2+1-p_1,0\}}\\
&\leqslant C_1 L\tau \sum_{p_1=1}^{L-l+1}\sum_{p_2=0}^{L-l+1-p_1}\binom{L-l+1}{p_1}\binom{L-l+1-p_1}{p_2}L^{-2\max\{p_2,p_1-1\}}\\
&\leqslant C_1 L^2\tau \sum_{p_1=1}^{L-l+1}\sum_{p_2=0}^{L-l+1-p_1}\binom{L-l+1}{p_1}\binom{L-l+1-p_1}{p_2}L^{-(p_1+p_2)}\cdot 1^{L-l+1-p_1-p_2}\\
&\leqslant C_1L^2\tau\cdot(1+2/L)^L\\
&\leqslant C_1e^2L^2\tau.
\end{aligned}$$

For $I_{2,i}$, we have

$$I_{2,i} \leqslant C_1\sqrt{s\log(M)/m}\cdot\sum_{p_2=1}^{L-l+1}\binom{L-l+1}{p_2}L^{-2p_2} \leqslant C_1e\sqrt{s\log(M)/m}.$$

Plugging the bounds of $I_{1,i}$ and $I_{2,i}$ into (B.1) completes the proof. □

Lemma B.5. Suppose that $\mathbf{W}_1,\dots,\mathbf{W}_L$ are generated via Gaussian initialization, and the results of Theorem 5.3 all hold. Let $\widetilde{\mathbf{W}} = (\widetilde{\mathbf{W}}_1,\dots,\widetilde{\mathbf{W}}_L)$, $\widehat{\mathbf{W}} = (\widehat{\mathbf{W}}_1,\dots,\widehat{\mathbf{W}}_L)$ be two collections of weight matrices satisfying $\|\widetilde{\mathbf{W}}_l - \mathbf{W}_l\|_2, \|\widehat{\mathbf{W}}_l - \mathbf{W}_l\|_2 \leqslant \tau$, $l = 1,\dots,L$. For $\mathbf{x}\in S^{d-1}$, let $\mathbf{\Sigma}_l(\mathbf{x}), \widetilde{\mathbf{\Sigma}}_l(\mathbf{x}), \widehat{\mathbf{\Sigma}}_l(\mathbf{x})$ and $\mathbf{x}_l, \widetilde{\mathbf{x}}_l, \widehat{\mathbf{x}}_l$ be the binary matrices and hidden layer outputs at the l-th layer with parameter matrices $\mathbf{W}, \widetilde{\mathbf{W}}, \widehat{\mathbf{W}}$ respectively. Let C_0, C_0' be the absolute constants in the bounds of $\|\prod_{r=l_1}^{l_2}\widetilde{\mathbf{\Sigma}}_r(\mathbf{x})\widetilde{\mathbf{W}}_r^\top\|_2$, $1\leqslant l_1 < l_2 \leqslant L$ and $\|\mathbf{W}_l\|_2$, $l=1,\dots,L$ given in Lemma B.3 and (i) in Theorem 5.3 respectively. For $\delta > 0$, if $m \geqslant C\max\{dL^2\log(mL/\delta), L^{-8/3}\tau^{-4/3}\log[m/(\tau\delta)]\}$ for some large enough constant C and $\tau \leqslant \nu L^{-5}[\log(m)]^{-3/2}$ for some small enough absolute constant $\nu > 0$, then with probability at least $1-\delta$ it holds that

- $\|\widehat{\mathbf{x}}_l - \widetilde{\mathbf{x}}_l\|_2 \leqslant C'L\cdot\sum_{r=1}^{l}\|\widehat{\mathbf{W}}_r - \widetilde{\mathbf{W}}_r\|_2$,
- $\|\widehat{\mathbf{\Sigma}}_l(\mathbf{x}) - \widetilde{\mathbf{\Sigma}}_l(\mathbf{x})\|_0 \leqslant C''L^{4/3}\tau^{2/3}m_l$,

for all $\mathbf{x}\in S^{d-1}$ and $l\in[L]$, where $C' = 2\max\{C_0, 1\}$ and $C'' = 16C'^{2/3}C_0'^{2/3}$.

Proof of Lemma B.5. For all $\mathbf{x}\in S^{d-1}$ and $l\in[L]$, we prove the following stronger results:

$$\begin{aligned}
&\|\widehat{\mathbf{x}}_l - \widetilde{\mathbf{x}}_l\|_2 \leqslant C'L\cdot\sum_{r=1}^{l}\|\widehat{\mathbf{W}}_r - \widetilde{\mathbf{W}}_r\|_2,\ \|\widetilde{\mathbf{x}}_l - \mathbf{x}_l\|_2, \|\widehat{\mathbf{x}}_l - \mathbf{x}_l\|_2 \leqslant C'L^2\tau,\\
&\|\widetilde{\mathbf{\Sigma}}_l(\mathbf{x}) - \mathbf{\Sigma}_l(\mathbf{x})\|_0, \|\widehat{\mathbf{\Sigma}}_l(\mathbf{x}) - \mathbf{\Sigma}_l(\mathbf{x})\|_0, \|\widehat{\mathbf{\Sigma}}_l(\mathbf{x}) - \widetilde{\mathbf{\Sigma}}_l(\mathbf{x})\|_0 \leqslant C''L^{4/3}\tau^{2/3}m_l.
\end{aligned}$$

We prove the results above by induction in l. Suppose that for $r = 1, \dots, l-1$ it holds that

$$\|\widehat{\mathbf{x}}_r - \widetilde{\mathbf{x}}_r\|_2 \leqslant C'L \cdot \sum_{r'=1}^{r} \|\widehat{\mathbf{W}}_{r'} - \widetilde{\mathbf{W}}_{r'}\|_2,\ \|\widetilde{\mathbf{x}}_r - \mathbf{x}_r\|_2, \|\widehat{\mathbf{x}}_r - \mathbf{x}_r\|_2 \leqslant C'L^2\tau,$$
$$\|\widetilde{\mathbf{\Sigma}}_r(\mathbf{x}) - \mathbf{\Sigma}_r(\mathbf{x})\|_0, \|\widehat{\mathbf{\Sigma}}_r(\mathbf{x}) - \mathbf{\Sigma}_r(\mathbf{x})\|_0, \|\widehat{\mathbf{\Sigma}}_r(\mathbf{x}) - \widetilde{\mathbf{\Sigma}}_r(\mathbf{x})\|_0 \leqslant C''L^{4/3}\tau^{2/3}m_r.$$

We first prove the bounds for the diagonal matrices on the l-th layer. We remind the reader the definition

$$\mathcal{S}_l(\mathbf{x}, \beta) = \{j \in [m_l] : |\langle \mathbf{w}_{l,j}, \mathbf{x}_{l-1}\rangle| \leqslant \beta\},\ \mathbf{x} \in S^{d-1}, l = 1, \dots, L.$$

given in Lemma A.8 and Lemma B.1. Then we have

$$\|\widetilde{\mathbf{\Sigma}}_l(\mathbf{x}) - \mathbf{\Sigma}_l(\mathbf{x})\|_0 = |\{j \in [m_l] : (\widetilde{\mathbf{w}}_{l,j}^\top \widetilde{\mathbf{x}}_{l-1}) \cdot (\mathbf{w}_{l,j}^\top \mathbf{x}_{l-1}) < 0\}| = s_l^{(1)}(\beta) + s_l^{(2)}(\beta),$$

where

$$s_l^{(1)}(\beta) = |\{j \in \mathcal{S}_l(\mathbf{x}, \beta) : (\widetilde{\mathbf{w}}_{l,j}^\top \widetilde{\mathbf{x}}_{l-1}) \cdot (\mathbf{w}_{l,j}^\top \mathbf{x}_{l-1}) < 0\}|,$$
$$s_l^{(2)}(\beta) = |\{j \in \mathcal{S}_l^c(\mathbf{x}, \beta) : (\widetilde{\mathbf{w}}_{l,j}^\top \widetilde{\mathbf{x}}_{l-1}) \cdot (\mathbf{w}_{l,j}^\top \mathbf{x}_{l-1}) < 0\}|.$$

For $s_l^{(1)}(\beta)$, by Lemma B.1 we have

$$s_l^{(1)}(\beta) \leqslant |\mathcal{S}_l(\mathbf{x}, \beta)| \leqslant 4\beta m_l^{3/2}.$$

For $s_l^{(2)}(\beta)$, by definition, $j \in \{j \in \mathcal{S}_l^c(\mathbf{x}, \beta) : (\widetilde{\mathbf{w}}_{l,j}^\top \widetilde{\mathbf{x}}_{l-1}) \cdot (\mathbf{w}_{l,j}^\top \mathbf{x}_{l-1}) < 0\}$ implies that

$$\big|\langle \widetilde{\mathbf{w}}_{l,j}, \widetilde{\mathbf{x}}_{l-1}\rangle - \langle \mathbf{w}_{l,j}, \mathbf{x}_{l-1}\rangle\big| \geqslant \beta.$$

Therefore, by Lemma A.1 and the induction assumptions, we have

$$\begin{aligned}
s_l^{(2)}\beta^2 &\leqslant \|\widetilde{\mathbf{W}}_l^\top \widetilde{\mathbf{x}}_{l-1} - \mathbf{W}_l^\top \mathbf{x}_{l-1}\|_2^2 \\
&= \big\|(\widetilde{\mathbf{W}}_l^\top - \mathbf{W}_l^\top)\mathbf{x}_{l-1} + \widetilde{\mathbf{W}}_l^\top(\widetilde{\mathbf{x}}_{l-1} - \mathbf{x}_{l-1})\big\|_2^2 \\
&\leqslant \big[\|\widetilde{\mathbf{W}}_l^\top - \mathbf{W}_l^\top\|_2\|\mathbf{x}_{l-1}\|_2 + \|\widetilde{\mathbf{W}}_l^\top\|_2\|\widetilde{\mathbf{x}}_{l-1} - \mathbf{x}_{l-1}\|_2\big]^2 \\
&\leqslant (2\tau + C_0' \cdot CL^2\tau)^2 \\
&\leqslant 4C'^2C_0'^2L^4\tau^2.
\end{aligned}$$

Combining the bounds of $s_l^{(1)}(\beta)$ and $s_l^{(2)}(\beta)$ gives

$$\|\widetilde{\mathbf{\Sigma}}_l(\mathbf{x}) - \mathbf{\Sigma}_l(\mathbf{x})\|_0 \leqslant s_l^{(1)}(\beta) + s_l^{(2)}(\beta) \leqslant 4m_l^{3/2}\beta + \frac{4C'^2C_0'^2L^4\tau^2}{\beta^2}.$$

Setting $\beta = C'^{2/3}C_0'^{2/3}L^{4/3}\tau^{2/3}m_l^{-1/2}$, we obtain

$$\|\widetilde{\mathbf{\Sigma}}_l(\mathbf{x}) - \mathbf{\Sigma}_l(\mathbf{x})\|_0 \leqslant 8C'^{2/3}C_0'^{2/3}L^{4/3}\tau^{2/3}m_l.$$

By the exact same proof, we have

$$\|\widehat{\boldsymbol{\Sigma}}_l(\mathbf{x}) - \boldsymbol{\Sigma}_l(\mathbf{x})\|_0 \leqslant 8C'^{2/3}C_0'^{2/3}L^{4/3}\tau^{2/3}m_l,$$

and therefore

$$\|\widehat{\boldsymbol{\Sigma}}_l(\mathbf{x}) - \widetilde{\boldsymbol{\Sigma}}_l(\mathbf{x})\|_0 \leqslant 16C'^{2/3}C_0'^{2/3}L^{4/3}\tau^{2/3}m_l.$$

This completes the proof of the $\|\cdot\|_0$-bounds for diagonal matrices on the l-th layer.

Now, based on the bounds obtained above and the induction assumptions on $\|\widetilde{\boldsymbol{\Sigma}}_r(\mathbf{x}) - \widehat{\boldsymbol{\Sigma}}_r(\mathbf{x})\|_0$, $r = 1, \ldots, l-1$, we derive the bound for $\|\widetilde{\mathbf{x}}_l - \widehat{\mathbf{x}}_l\|_2$. By the assumption that $\tau \leqslant \nu L^{-5}[\log(m)]^{-3/2}$ for some small enough absolute constant $\nu > 0$, we have

$$\|\widetilde{\boldsymbol{\Sigma}}_r(\mathbf{x}) - \widehat{\boldsymbol{\Sigma}}_r(\mathbf{x})\|_0 \leqslant (\kappa/2) \cdot L^{-2}m/\log(m),\ \ r = 1, \ldots, l.$$

Define diagonal matrices $\check{\boldsymbol{\Sigma}}_r$, $r = 1, \ldots, l$ as follows:

$$[\check{\boldsymbol{\Sigma}}_r(\mathbf{x})]_{jj} := [\widehat{\boldsymbol{\Sigma}}_r(\mathbf{x}) - \widetilde{\boldsymbol{\Sigma}}_r(\mathbf{x})]_{jj} \cdot \frac{\widetilde{\mathbf{w}}_{r,j}^\top \widetilde{\mathbf{x}}_{r-1}}{\widehat{\mathbf{w}}_{r,j}^\top \widehat{\mathbf{x}}_{r-1} - \widetilde{\mathbf{w}}_{r,j}^\top \widetilde{\mathbf{x}}_{r-1}},\ \ j = 1, \ldots, m_l.$$

Then we have

$$\begin{aligned}
\widehat{\mathbf{x}}_l - \widetilde{\mathbf{x}}_l &= [\widehat{\boldsymbol{\Sigma}}_l(\mathbf{x}) + \check{\boldsymbol{\Sigma}}_l(\mathbf{x})](\widehat{\mathbf{W}}_l^\top \widehat{\mathbf{x}}_{l-1} - \widetilde{\mathbf{W}}_l^\top \widetilde{\mathbf{x}}_{l-1}) \\
&= [\widehat{\boldsymbol{\Sigma}}_l(\mathbf{x}) + \check{\boldsymbol{\Sigma}}_l(\mathbf{x})]\widehat{\mathbf{W}}_l^\top (\widehat{\mathbf{x}}_{l-1} - \widetilde{\mathbf{x}}_{l-1}) + [\widehat{\boldsymbol{\Sigma}}_l(\mathbf{x}) + \check{\boldsymbol{\Sigma}}_l(\mathbf{x})](\widehat{\mathbf{W}}_l - \widetilde{\mathbf{W}}_l)^\top \widetilde{\mathbf{x}}_{l-1} \\
&= \cdots \\
&= \sum_{r=1}^{l} \Bigg\{ \prod_{t=r+1}^{l} [\widehat{\boldsymbol{\Sigma}}_t(\mathbf{x}) + \check{\boldsymbol{\Sigma}}_t(\mathbf{x})]\widehat{\mathbf{W}}_t^\top \Bigg\} [\widehat{\boldsymbol{\Sigma}}_r(\mathbf{x}) + \check{\boldsymbol{\Sigma}}_r(\mathbf{x})](\widehat{\mathbf{W}}_r - \widetilde{\mathbf{W}}_r)^\top \widetilde{\mathbf{x}}_{r-1}.
\end{aligned}$$

Now by Lemma B.3, the fact that $|[\check{\boldsymbol{\Sigma}}_r(\mathbf{x})]_{jj}|, |[\widehat{\boldsymbol{\Sigma}}_r(\mathbf{x}) + \check{\boldsymbol{\Sigma}}_r(\mathbf{x})]_{jj}| \leqslant 1$ and

$$\|\widehat{\boldsymbol{\Sigma}}_r(\mathbf{x}) + \check{\boldsymbol{\Sigma}}_r(\mathbf{x}) - \boldsymbol{\Sigma}_r(\mathbf{x})\|_0 \leqslant \|\widehat{\boldsymbol{\Sigma}}_r(\mathbf{x}) - \boldsymbol{\Sigma}_r(\mathbf{x})\|_0 + \|\widehat{\boldsymbol{\Sigma}}_r(\mathbf{x}) - \widetilde{\boldsymbol{\Sigma}}_r(\mathbf{x})\|_0 \leqslant \kappa L^{-2}m/\log(m)$$

for all $r = 1, \ldots, l$, we have

$$\|\widehat{\mathbf{x}}_l - \widetilde{\mathbf{x}}_l\|_2 \leqslant 2C_0 L \cdot \sum_{r=1}^{l} \|\widehat{\mathbf{W}}_r - \widetilde{\mathbf{W}}_r\|_2.$$

With the exact same proof, it can be shown that

$$\begin{aligned}
\|\widehat{\mathbf{x}}_l - \mathbf{x}_l\|_2 &\leqslant 2C_0 L \cdot \sum_{r=1}^{l} \|\widehat{\mathbf{W}}_r - \mathbf{W}_r\|_2 \leqslant 2C_0 L^2 \tau, \\
\|\widetilde{\mathbf{x}}_l - \mathbf{x}_l\|_2 &\leqslant 2C_0 L \cdot \sum_{r=1}^{l} \|\widetilde{\mathbf{W}}_r - \mathbf{W}_r\|_2 \leqslant 2C_0 L^2 \tau.
\end{aligned}$$

This completes the proof. □

Lemma B.6. Suppose that $\mathbf{W}_1, \ldots, \mathbf{W}_L$ are generated via Gaussian initialization, and the results of Theorem 5.3 and Lemma B.5 all hold. If $\tau \leqslant \nu L^{-5}[\log(m)]^{-3/2}$ for some small enough absolute constant $\nu > 0$, then

$$f_{\widehat{\mathbf{W}}}(\mathbf{x}) - f_{\widetilde{\mathbf{W}}}(\mathbf{x}) \leqslant C \sum_{l=1}^{L} L^{8/3} \tau^{1/3} \sqrt{m \log(m)} \cdot \big\|\widehat{\mathbf{W}}_l - \widetilde{\mathbf{W}}_l\big\|_2 + C \sum_{l=1}^{L} L^3 \cdot \sqrt{m} \cdot \|\widehat{\mathbf{W}}_l - \widetilde{\mathbf{W}}_l\|_2^2 + \sum_{l=1}^{L} \mathrm{Tr}[(\widehat{\mathbf{W}}_l - \widetilde{\mathbf{W}}_l)^\top \nabla_{W_l} f_{\widetilde{\mathbf{W}}}(\mathbf{x})]$$

for all $\widetilde{\mathbf{W}}, \widehat{\mathbf{W}} \in \mathcal{W}(\mathbf{W}, \tau)$ and all $\mathbf{x} \in S^{d-1}$, where C is an absolute constant.

Proof of Lemma B.6. Denote by $\widehat{y}$, $\widetilde{y}$ the outputs of the network with input $\mathbf{x}$ and parameter matrices $\widehat{\mathbf{W}}$, $\widetilde{\mathbf{W}}$ respectively. Then we have

$$\begin{aligned} f_{\widehat{\mathbf{W}}}(\mathbf{x}) - f_{\widetilde{\mathbf{W}}}(\mathbf{x}) &= \mathbf{v}^\top \Bigg[\prod_{l=1}^{L} \widehat{\boldsymbol{\Sigma}}_l(\mathbf{x}) \widehat{\mathbf{W}}_l^\top\Bigg]\mathbf{x} - \mathbf{v}^\top \Bigg[\prod_{l=1}^{L} \widetilde{\boldsymbol{\Sigma}}_l(\mathbf{x}) \widetilde{\mathbf{W}}_l^\top\Bigg]\mathbf{x} \\ &= \sum_{l=1}^{L} \Bigg[\prod_{r=l+1}^{L} \widetilde{\boldsymbol{\Sigma}}_r(\mathbf{x}) \widetilde{\mathbf{W}}_r^\top\Bigg][\widehat{\boldsymbol{\Sigma}}_l(\mathbf{x}) \widehat{\mathbf{W}}_l^\top - \widetilde{\boldsymbol{\Sigma}}_l(\mathbf{x}) \widetilde{\mathbf{W}}_l^\top]\widehat{\mathbf{x}}_{l-1} \\ &= I_1 + I_2 + I_3, \end{aligned}$$

where

$$\begin{aligned} I_1 &= \sum_{l=1}^{L} \mathbf{v}^\top \Bigg[\prod_{r=l+1}^{L} \widetilde{\boldsymbol{\Sigma}}_r(\mathbf{x}) \widetilde{\mathbf{W}}_r^\top\Bigg][\widehat{\boldsymbol{\Sigma}}_l(\mathbf{x}) - \widetilde{\boldsymbol{\Sigma}}_l(\mathbf{x})]\widehat{\mathbf{W}}_l^\top \widehat{\mathbf{x}}_{l-1}, \\ I_2 &= \sum_{l=1}^{L} \mathbf{v}^\top \Bigg[\prod_{r=l+1}^{L} \widetilde{\boldsymbol{\Sigma}}_r(\mathbf{x}) \widetilde{\mathbf{W}}_r^\top\Bigg]\widetilde{\boldsymbol{\Sigma}}_l(\mathbf{x})\big(\widehat{\mathbf{W}}_l - \widetilde{\mathbf{W}}_l\big)^\top (\widehat{\mathbf{x}}_{l-1} - \widetilde{\mathbf{x}}_{l-1}), \\ I_3 &= \sum_{l=1}^{L} \mathbf{v}^\top \Bigg[\prod_{r=l+1}^{L} \widetilde{\boldsymbol{\Sigma}}_r(\mathbf{x}) \widetilde{\mathbf{W}}_r^\top\Bigg]\widetilde{\boldsymbol{\Sigma}}_l(\mathbf{x})\big(\widehat{\mathbf{W}}_l - \widetilde{\mathbf{W}}_l\big)^\top \widetilde{\mathbf{x}}_{l-1}. \end{aligned}$$

For I_1, note that for any $l = 1, \ldots, L$, by Lemma B.5 we have

$$\begin{aligned} \big\|[\widehat{\boldsymbol{\Sigma}}_l(\mathbf{x}) - \widetilde{\boldsymbol{\Sigma}}_l(\mathbf{x})]\widehat{\mathbf{W}}_l^\top \widehat{\mathbf{x}}_{l-1}\big\|_2 &\leqslant \big\|\widehat{\mathbf{W}}_l^\top \widehat{\mathbf{x}}_{l-1} - \widetilde{\mathbf{W}}_l^\top \widetilde{\mathbf{x}}_{l-1}\big\|_2 \\ &\leqslant \big\|(\widehat{\mathbf{W}}_l^\top - \widetilde{\mathbf{W}}_l^\top)\widehat{\mathbf{x}}_{l-1}\big\|_2 + \big\|\widetilde{\mathbf{W}}_l^\top (\widehat{\mathbf{x}}_{l-1} - \widetilde{\mathbf{x}}_{l-1})\big\|_2 \\ &\leqslant C_1 \big\|\widehat{\mathbf{W}}_l - \widetilde{\mathbf{W}}_l\big\|_2 + C_1 L \sum_{l=1}^{L} \big\|\widehat{\mathbf{W}}_l - \widetilde{\mathbf{W}}_l\big\|_2 \\ &\leqslant C_2 L \cdot \sum_{l=1}^{L} \big\|\widehat{\mathbf{W}}_l - \widetilde{\mathbf{W}}_l\big\|_2, \end{aligned}$$

where the first inequality follows by checking the signs of the diagonal entries of $\widehat{\boldsymbol{\Sigma}}_l(\mathbf{x}) - \widetilde{\boldsymbol{\Sigma}}_l(\mathbf{x})$, and C_1, C_2 are absolute constants. Therefore by Lemma B.5 we have

$$
\begin{aligned}
|I_1| &\leqslant \sum_{l=1}^{L} \Bigg\| \mathbf{v}^\top \Bigg[\prod_{r=l+1}^{L} \widetilde{\boldsymbol{\Sigma}}_r(\mathbf{x}) \widetilde{\mathbf{W}}_r^\top \Bigg] \cdot \big| \widehat{\boldsymbol{\Sigma}}_l(\mathbf{x}) - \widetilde{\boldsymbol{\Sigma}}_l(\mathbf{x}) \big| \Bigg\|_2 \cdot \big\| \big[\widehat{\boldsymbol{\Sigma}}_l(\mathbf{x}) - \widetilde{\boldsymbol{\Sigma}}_l(\mathbf{x}) \big] \widehat{\mathbf{W}}_l^\top \widehat{\mathbf{x}}_{l-1} \big\|_2 \\
&\leqslant C_3 L \cdot L^{2/3} \tau^{1/3} \sqrt{m \log(m)} \cdot L \cdot \sum_{l=1}^{L} \big\| \widehat{\mathbf{W}}_l - \widetilde{\mathbf{W}}_l \big\|_2 \\
&= C_3 L^{8/3} \tau^{1/3} \sqrt{m \log(m)} \cdot \sum_{l=1}^{L} \big\| \widehat{\mathbf{W}}_l - \widetilde{\mathbf{W}}_l \big\|_2,
\end{aligned}
$$

where C_3 is an absolute constant.

For I_2, by Lemma B.5 we have

$$
|I_2| \leqslant C_4 \sqrt{m} \cdot L \cdot \sum_{l=1}^{L} \| \widehat{\mathbf{W}}_l - \widetilde{\mathbf{W}}_l \|_2 \cdot L \cdot \sum_{r=1}^{l} \| \widehat{\mathbf{W}}_r - \widetilde{\mathbf{W}}_r \|_2 \leqslant C_4 L^3 \cdot \sqrt{m} \cdot \sum_{l=1}^{L} \| \widehat{\mathbf{W}}_l - \widetilde{\mathbf{W}}_l \|_2^2,
$$

where C_4 is an absolute constant.

For I_3, we have

$$
\begin{aligned}
I_3 &= \sum_{l=1}^{L} \mathbf{v}^\top \Bigg[\prod_{r=l+1}^{L} \widetilde{\boldsymbol{\Sigma}}_r(\mathbf{x}) \widetilde{\mathbf{W}}_r^\top \Bigg] \widetilde{\boldsymbol{\Sigma}}_l(\mathbf{x}) \big(\widehat{\mathbf{W}}_l - \widetilde{\mathbf{W}}_l \big)^\top \widetilde{\mathbf{x}}_{l-1} \\
&= \sum_{l=1}^{L} \mathrm{Tr} \Bigg\{ \mathbf{v}^\top \Bigg[\prod_{r=l+1}^{L} \widetilde{\boldsymbol{\Sigma}}_r(\mathbf{x}) \widetilde{\mathbf{W}}_r^\top \Bigg] \widetilde{\boldsymbol{\Sigma}}_l(\mathbf{x}) \big(\widehat{\mathbf{W}}_l - \widetilde{\mathbf{W}}_l \big)^\top \widetilde{\mathbf{x}}_{l-1} \Bigg\} \\
&= \sum_{l=1}^{L} \mathrm{Tr} \Bigg\{ \big(\widehat{\mathbf{W}}_l - \widetilde{\mathbf{W}}_l \big)^\top \widetilde{\mathbf{x}}_{l-1} \mathbf{v}^\top \Bigg[\prod_{r=l+1}^{L} \widetilde{\boldsymbol{\Sigma}}_r(\mathbf{x}) \widetilde{\mathbf{W}}_r^\top \Bigg] \widetilde{\boldsymbol{\Sigma}}_l(\mathbf{x}) \Bigg\} \\
&= \sum_{l=1}^{L} \mathrm{Tr} \big[(\widehat{\mathbf{W}}_l - \widetilde{\mathbf{W}}_l)^\top \nabla_{\mathbf{W}_l} f_{\widetilde{\mathbf{W}}}(\mathbf{x}) \big].
\end{aligned}
$$

Combining the bounds of I_1, I_2 and I_3 completes the proof. □

Proof of Theorem 5.4. Most of the results given in Theorem 5.4 have already been proved by Lemma B.2, B.3, B.4, B.5 and B.6. Note that by B.5, for any $\widetilde{\mathbf{W}}$ with $\|\widetilde{\mathbf{W}}_l - \mathbf{W}_l\|_2 \leqslant \tau$, $l = 1, \ldots, L$, we have $\|\widetilde{\boldsymbol{\Sigma}}_l(\mathbf{x}) - \boldsymbol{\Sigma}_l(\mathbf{x})\|_0 \leqslant C_1 L^{4/3} \tau^{2/3} m_l$ for all $\mathbf{x} \in S^{d-1}$ and $l \in [L]$, where C_1 is an absolute constant. Therefore in Lemma B.4, setting $s = C_2 L^{4/3} \tau^{2/3} m$ for some large enough absolute constant C_2 gives $\mathbf{v}^\top \big(\prod_{r=l}^{L} \widetilde{\boldsymbol{\Sigma}}_r(\mathbf{x}) \widetilde{\mathbf{W}}_r^\top \big) \mathbf{a} \leqslant C_3 L^{2/3} \tau^{1/3} \sqrt{m \log(m)}$ for all $\mathbf{a} \in \mathbb{R}^{m_{l-1}}$ satisfying $\|\mathbf{a}\|_2 = 1$, $\|\mathbf{a}\|_0 \leqslant \overline{C} L^{4/3} \tau^{2/3} m_l$ and any $1 \leqslant l \leqslant L$, where C_3 is an absolute constant. This completes the proof. □

C Proof of Theorem 5.5

In this section we give the proof of Theorem 5.5. We remind the reader the following notations:

- For $i = 1, \ldots, n$, $l = 1, \ldots, L$, we use $\boldsymbol{x}_i$ to denote the i-th sample input, and $\boldsymbol{x}_{l,i}$ the output of the l-th layer with Gaussian initialization parameter $\mathbf{W}$ and input $\boldsymbol{x}_i$.

- For $l = 1, \ldots, L$, we denote by $\mathbf{x}_l$ the output of the l-th layer with Gaussian initialization parameter $\mathbf{W}$ and input $\mathbf{x}$.

Lemma C.1. Suppose that $\mathbf{W}_1, \ldots, \mathbf{W}_L$ are generated via Gaussian initialization. For any $\delta > 0$, if $m \geqslant C\gamma^{-2} \cdot [d \log(1/\gamma) + \log(1/\delta)]$ for some large enough absolute constant C, then with probability at least $1 - \delta$, there exists $\boldsymbol{\alpha}_1 \in S^{m_1-1}$ such that

$$y \cdot \langle \boldsymbol{\alpha}_1, \mathbf{x}_1 \rangle \geqslant \gamma/2$$

for all $(\mathbf{x}, y) \in \operatorname{supp}(\mathcal{D})$.

Proof of Lemma C.1. By Assumption 4.2, there exists $c(\overline{\mathbf{u}}) \in \mathcal{C}$ such that

$$f(\mathbf{x}) = \int_{\mathbb{R}^d} c(\overline{\mathbf{u}}) \sigma(\overline{\mathbf{u}}^\top \mathbf{x}) p(\overline{\mathbf{u}}) \mathrm{d}\overline{\mathbf{u}}$$

satisfies $y \cdot f(\mathbf{x}) \geqslant \gamma$ for all $(\mathbf{x}, y) \in \operatorname{supp}(\mathcal{D})$. Let

$$\widetilde{\boldsymbol{\alpha}} = (\sqrt{1/m_1} c(\sqrt{m_1/2}\mathbf{w}_1), \ldots, \sqrt{1/m_1} c(\sqrt{m_1/2}\mathbf{w}_{m_1}))^\top.$$

Then we have

$$\|\widetilde{\boldsymbol{\alpha}}\|_2^2 = \frac{1}{m_1} \sum_{j=1}^{m_1} c^2\bigg(\sqrt{\frac{m_1}{2}}\mathbf{w}_j\bigg).$$

Therefore

$$\mathbb{E}(\|\widetilde{\boldsymbol{\alpha}}\|_2^2) = \int_{\mathbb{R}^d} c^2(\overline{\mathbf{u}}) p(\overline{\mathbf{u}}) \mathrm{d}\overline{\mathbf{u}} \leqslant \int_{\mathbb{R}^d} p(\overline{\mathbf{u}}) \mathrm{d}\overline{\mathbf{u}} = 1.$$

Since $\sup_{\overline{\mathbf{u}} \in \mathbb{R}^d} |c(\overline{\mathbf{u}})| \leqslant 1$, by Hoeffding inequality, with probability at least $1 - \delta/4$ we have

$$\|\widetilde{\boldsymbol{\alpha}}\|_2^2 \leqslant 1 + C_1 \sqrt{\log(4e/\delta)/m_1} \leqslant 3/2,$$

where C_1 is an absolute constant. By (i) in Theorem 5.3, with probability at least $1 - \delta/4$, $\|\mathbf{W}_1\|_2 \leqslant C_2$ for some absolute constant C_2. Moreover, by the definition of $\mathcal{F}$ and Lipschitz continuity of the ReLU function, clearly there exists an absolute constant C_3 such that

$$|f(\mathbf{x}) - f(\mathbf{x}')| \leqslant C_3 \|\mathbf{x} - \mathbf{x}'\|_2$$

for all $\mathbf{x}, \mathbf{x}' \in S^{d-1}$. Let $R = \gamma/(16 \cdot \max\{2C_2, C_3\})$, and $\mathcal{N} = \mathcal{N}(S^{d-1}, R)$ be an R-net covering S^{d-1}. Then by Lemma 5.2 in Vershynin (2010) we have

$$|\mathcal{N}| \leqslant (1 + 2/R)^d \leqslant (C_4/\gamma)^d$$

for some absolute constant C_4. For any $\widehat{\mathbf{x}} \in \mathcal{N}$, we have

$$\begin{aligned}\widetilde{\boldsymbol{\alpha}}^\top \widehat{\mathbf{x}}_1 &= \sum_{j=1}^{m_1} \sqrt{\frac{1}{m_1}} c\bigg(\sqrt{\frac{m_1}{2}}\mathbf{w}_j\bigg) \cdot \sqrt{\frac{2}{m_1}} \sigma\bigg(\sqrt{\frac{m_1}{2}}\mathbf{w}_j^\top \widehat{\mathbf{x}}\bigg) \\ &= \frac{\sqrt{2}}{m_1} \sum_{j=1}^{m_1} c\bigg(\sqrt{\frac{m_1}{2}}\mathbf{w}_j\bigg) \cdot \sigma\bigg(\sqrt{\frac{m_1}{2}}\mathbf{w}_j^\top \widehat{\mathbf{x}}\bigg).\end{aligned}$$

Therefore we have $\mathbb{E}(\widetilde{\boldsymbol{\alpha}}^\top \widehat{\mathbf{x}}_1) = \sqrt{2} f(\widehat{\mathbf{x}})$. Moreover, since $\sup_{\overline{\mathbf{u}} \in \mathbb{R}^d} |c(\overline{\mathbf{u}})| \leqslant 1$, we have

$$\bigg\| c\bigg(\sqrt{\frac{m_1}{2}}\mathbf{w}_j\bigg) \cdot \sigma\bigg(\sqrt{\frac{m_1}{2}}\mathbf{w}_j^\top \widehat{\mathbf{x}}\bigg)\bigg\|_{\psi_2} \leqslant C_5$$

for some absolute constant C_5. Therefore by Hoeffding inequality and union bound, with probability at least $1 - \delta/2$ we have

$$|\widetilde{\boldsymbol{\alpha}}^\top \widehat{\mathbf{x}}_1 - \sqrt{2} f(\widehat{\mathbf{x}})| \leqslant C_6 \sqrt{\frac{\log(2e \cdot C_4^d \cdot \gamma^{-d}/\delta)}{m_1}} \leqslant C_7 \sqrt{\frac{d \log(1/\gamma) + \log(1/\delta)}{m_1}} \leqslant \gamma/8,$$

where C_6, C_7 are absolute constants. For any $(\mathbf{x}, y) \in \operatorname{supp}(\mathcal{D})$, there exists $\widehat{\mathbf{x}} \in \mathcal{N}$ such that $\|\mathbf{x} - \widehat{\mathbf{x}}\|_2 \leqslant R = \gamma/(16 \cdot \max\{2C_2, C_3\})$. Therefore we have

$$\begin{aligned}&|\widetilde{\boldsymbol{\alpha}}^\top \mathbf{x}_1 - \widetilde{\boldsymbol{\alpha}}^\top \widehat{\mathbf{x}}_1| \leqslant (3/2)\|\mathbf{x}_1 - \widehat{\mathbf{x}}_1\|_2 \leqslant (3/2) \cdot \|\mathbf{W}_1\|_2 \cdot \|\mathbf{x} - \widehat{\mathbf{x}}\|_2 < \gamma/16, \\ &|f(\mathbf{x}) - f(\widehat{\mathbf{x}})| \leqslant C_3 \|\mathbf{x} - \widehat{\mathbf{x}}\|_2 \leqslant \gamma/16,\end{aligned}$$

and

$$y \cdot \widetilde{\boldsymbol{\alpha}}^\top \mathbf{x}_1 \geqslant (3/4) \cdot \gamma.$$

Set $\boldsymbol{\alpha}_1 = \widetilde{\boldsymbol{\alpha}}/\|\widetilde{\boldsymbol{\alpha}}\|_2$. Then by the bound $\|\widetilde{\boldsymbol{\alpha}}\|_2 \leqslant 3/2$, we have

$$y \cdot \boldsymbol{\alpha}_1^\top \mathbf{x}_1 \geqslant \sqrt{2} \cdot (3/4) \cdot \gamma/(3/2) > \gamma/2.$$

This completes the proof. □

Lemma C.2. Suppose that $\mathbf{W}_1, \ldots, \mathbf{W}_L$ are generated via Gaussian initialization. For any $\delta > 0$, if $m \geqslant C \cdot 4^L \cdot L^4 \gamma^{-2} d \log[L/(\gamma\delta)]$ for some large enough absolute constant C, then with probability at least $1 - \delta$, there exist $\boldsymbol{\alpha}_1 \in S^{m_1 - 1}, \ldots, \boldsymbol{\alpha}_L \in S^{m_L - 1}$ such that

$$y \cdot \langle \boldsymbol{\alpha}_l, \mathbf{x}_l \rangle \geqslant 2^{-(l+1)}\gamma$$

for all $(\mathbf{x}, y) \in \operatorname{supp}(\mathcal{D})$ and $l \in [L]$.

Proof of Lemma C.2. By Lemma C.1, with probability at least $1 - \delta/4$, there exists $\boldsymbol{\alpha}_1 \in S^{m_1 - 1}$ such that $y \cdot \langle \boldsymbol{\alpha}_1, \mathbf{x}_1 \rangle \geqslant \gamma/2$ for all $(\mathbf{x}, y) \in \operatorname{supp}(\mathcal{D})$. Define $\widetilde{\boldsymbol{\alpha}}_1 = \boldsymbol{\alpha}_1$, and $\widetilde{\boldsymbol{\alpha}}_l = \mathbf{W}_l \widetilde{\boldsymbol{\alpha}}_{l-1}$, $l = 2, \ldots, L$.

For any $l = 2, \ldots, L$, by definition, we have

$$\|\widetilde{\boldsymbol{\alpha}}_l\|_2^2 = \sum_{j=1}^{m_l} (\mathbf{w}_{l,j}^\top \widetilde{\boldsymbol{\alpha}}_{l-1})^2.$$

Therefore we have $\mathbb{E}(\|\widetilde{\boldsymbol{\alpha}}_l\|_2^2 | \widetilde{\boldsymbol{\alpha}}_{l-1}) = 2\|\widetilde{\boldsymbol{\alpha}}_{l-1}\|_2^2$. Since $\|(\mathbf{w}_{l,j}^\top \widetilde{\boldsymbol{\alpha}}_{l-1})^2\|_{\psi_1} = O(\|\widetilde{\boldsymbol{\alpha}}_{l-1}\|_2^2)$, by Bernstein inequality and union bound, with probability at least $1 - \delta/4$,

$$\big| \|\widetilde{\boldsymbol{\alpha}}_l\|_2^2 - 2\|\widetilde{\boldsymbol{\alpha}}_{l-1}\|_2^2 \big| \leqslant C_1 \|\widetilde{\boldsymbol{\alpha}}_{l-1}\|_2^2 \cdot \sqrt{\frac{\log(8/\delta)}{m_l}} \leqslant 2\|\widetilde{\boldsymbol{\alpha}}_{l-1}\|_2^2$$

for all $l \in [L]$. Therefore since $\|\widetilde{\boldsymbol{\alpha}}_1\|_2 = 1$, we have $\|\widetilde{\boldsymbol{\alpha}}_l\|_2 \leqslant 2^{l-1}$ for all $l = 2, \ldots, L$. Moreover, for any $l = 2, \ldots, L$, by definition, we have $\mathbf{w} \overset{d}{=} -\mathbf{w}$, and therefore

$$\begin{aligned}
\mathbb{E}[\langle \widetilde{\boldsymbol{\alpha}}_l, \mathbf{x}_l \rangle | \widetilde{\boldsymbol{\alpha}}_{l-1}] &= \sum_{j=1}^{m_l} \mathbb{E}\Big[\big(\mathbf{w}_{l,j}^\top \widetilde{\boldsymbol{\alpha}}_{l-1}\big) \cdot \sigma\big(\mathbf{w}_{l,j}^\top \mathbf{x}_{l-1}\big) \Big| \widetilde{\boldsymbol{\alpha}}_{l-1} \Big] \\
&= \frac{1}{2} \sum_{j=1}^{m_l} \mathbb{E}\Big[\big(\mathbf{w}_{l,j}^\top \widetilde{\boldsymbol{\alpha}}_{l-1}\big) \cdot \sigma\big(\mathbf{w}_{l,j}^\top \mathbf{x}_{l-1}\big) + \big(-\mathbf{w}_{l,j}^\top \widetilde{\boldsymbol{\alpha}}_{l-1}\big) \cdot \sigma\big(-\mathbf{w}_{l,j}^\top \mathbf{x}_{l-1}\big) \Big| \widetilde{\boldsymbol{\alpha}}_{l-1} \Big] \\
&= \frac{1}{2} \sum_{j=1}^{m_l} \mathbb{E}\Big[\big(\mathbf{w}_{l,j}^\top \widetilde{\boldsymbol{\alpha}}_{l-1}\big) \cdot \big(\mathbf{w}_{l,j}^\top \mathbf{x}_{l-1}\big) \Big| \widetilde{\boldsymbol{\alpha}}_{l-1} \Big] \\
&= \langle \widetilde{\boldsymbol{\alpha}}_{l-1}, \mathbf{x}_{l-1} \rangle.
\end{aligned}$$

for all $\mathbf{x} \in S^{d-1}$. By Theorem 5.3, with probability at least $1 - \delta/2$, we have

$$\|\mathbf{W}_l\|_2 \leqslant C_2, \quad \|\mathbf{x}_l\|_2 \leqslant C_3, \quad \|\mathbf{x}_l - \mathbf{x}_l'\|_2 \leqslant C_4 L \|\mathbf{x} - \mathbf{x}'\|_2$$

for all $\mathbf{x}, \mathbf{x}' \in S^{d-1}$ and $l \in [L]$. Let $R = \gamma/(4 \cdot 2^L C_4 L)$ and $\mathcal{N} = \mathcal{N}(S^{d-1}, R)$. Then by Lemma 5.2 in Vershynin (2010) we have $|\mathcal{N}| \leqslant (C_5 L/\gamma)^{Ld}$ for some absolute constant C_5. Since

$$\begin{aligned}
(m_l/2) \cdot \big\| \mathbf{w}_{l,j}^\top \widetilde{\boldsymbol{\alpha}}_{l-1} \cdot \sigma(\mathbf{w}_{l,j}^\top \widehat{\mathbf{x}}_{l-1}) \big\|_{\psi_1} &\leqslant C_6 \big\| \langle \sqrt{m_l/2} \mathbf{w}_{l,j}, \widetilde{\boldsymbol{\alpha}}_{l-1} \rangle \big\|_{\psi_2} \big\| \langle \sqrt{m_l/2} \mathbf{w}_{l,j}, \widehat{\mathbf{x}}_{l-1} \rangle \big\|_{\psi_2} \\
&\leqslant C_7 \|\widetilde{\boldsymbol{\alpha}}_{l-1}\|_2,
\end{aligned}$$

where C_5, C_6 are absolute constants, by Bernstein inequality and union bound we have

$$\big| \langle \widetilde{\boldsymbol{\alpha}}_l, \widehat{\mathbf{x}}_l \rangle - \langle \widetilde{\boldsymbol{\alpha}}_{l-1}, \widehat{\mathbf{x}}_{l-1} \rangle \big| \leqslant C_8 \|\widetilde{\boldsymbol{\alpha}}_{l-1}\|_2 \cdot \sqrt{\frac{\log[2L \cdot (C_5 L/\gamma)^{Ld}/\delta]}{m_l}} \leqslant \gamma/(8L)$$

for all $\widehat{\mathbf{x}} \in \mathcal{N}$ and $l \in [L]$. Therefore we have

$$\langle \widetilde{\boldsymbol{\alpha}}_l, \widehat{\mathbf{x}}_l \rangle \geqslant \langle \widetilde{\boldsymbol{\alpha}}_{l-1}, \widehat{\mathbf{x}}_{l-1} \rangle - \gamma/(8L) \geqslant \cdots \geqslant \gamma/2 - \gamma/8 = (3/8)\gamma.$$

for all $l = 2, \ldots, L$. For any $\mathbf{x} \in S^{d-1}$, there exists $\widehat{\mathbf{x}} \in \mathcal{N}$ such that $\|\mathbf{x} - \widehat{\mathbf{x}}\|_2 \leqslant R = \gamma/(4 \cdot 2^L C_4 L)$. Hence we have

$$|\langle \widetilde{\boldsymbol{\alpha}}_l, \mathbf{x}_l \rangle - \langle \widetilde{\boldsymbol{\alpha}}_l, \widehat{\mathbf{x}}_l \rangle| \leqslant 2^{l-1} \|\mathbf{x}_l - \widehat{\mathbf{x}}_l\|_2 \leqslant 2^{l-1} \cdot C_4 L \cdot R \leqslant \gamma/8,$$

and therefore

$$\langle \widetilde{\boldsymbol{\alpha}}_l, \mathbf{x}_l \rangle \geqslant \gamma/4.$$

Setting $\boldsymbol{\alpha}_l = \widetilde{\boldsymbol{\alpha}}_l/\|\widetilde{\boldsymbol{\alpha}}_l\|_2$, we obtain

$$\langle \boldsymbol{\alpha}_l, \mathbf{x}_l \rangle \geqslant 2^{-(l-1)} \cdot \gamma/4 \geqslant 2^{-(l+1)}\gamma.$$

This completes the proof. □

Lemma C.3. For any $\delta > 0$, under the same assumptions as Lemma C.2, with probability at least $1-\delta$, the inequality

$$\sum_{j=1}^{m_l} \left\| \frac{1}{n}\sum_{i=1}^{n}[a(\boldsymbol{x}_i, y_i) \cdot y_i \cdot \sigma'(\mathbf{w}_{L,j}^\top \boldsymbol{x}_{L-1,i}) \cdot \boldsymbol{x}_{L-1,i}] \right\|_2^2 \geqslant 4^{-L}/8 \cdot m_l \cdot \gamma^2 \cdot \left[\frac{1}{n}\sum_{i=1}^{n} a(\boldsymbol{x}_i, y_i) \right]^2$$

holds for any function $a(\mathbf{x}, y) : S^{d-1} \times \{\pm 1\} \to \mathbb{R}^+$.

Proof of Lemma C.3. By Lemma C.2, with probability at least $1-\delta/2$, there exists $\alpha_{L-1} \in S^{m_{L-1}-1}$ such that $y \cdot \langle \alpha_{L-1}, \mathbf{x} \rangle \geqslant 2^{-L}\gamma$ for all $(\mathbf{x}, y) \in \operatorname{supp}(\mathcal{D})$. Moreover, since

$$\mathbb{E}[\sigma'(\mathbf{w}_{L,j}^\top \boldsymbol{x}_{L-1,i}) | \boldsymbol{x}_{L-1,i}] = 1/2,$$

by Hoeffding inequality, with probability at least $1-\delta/2$ we have

$$\frac{1}{m_L}\sum_{j=1}^{m_L} \sigma'(\mathbf{w}_{L,j}^\top \boldsymbol{x}_{L-1,i}) \geqslant \frac{1}{2} - C_1\sqrt{\frac{\log(n/\delta)}{m_L}} \geqslant \frac{1}{2\sqrt{2}} > 0$$

for all $i \in [n]$, where C_1 is an absolute constant. By Jensen's inequality, we have

$$\begin{aligned}
&\sum_{j=1}^{m_L} \left\| \frac{1}{n}\sum_{i=1}^{n}[a(\boldsymbol{x}_i, y_i) \cdot y_i \cdot \sigma'(\mathbf{w}_{L,j}^\top \boldsymbol{x}_{L-1,i}) \cdot \boldsymbol{x}_{L-1,i}] \right\|_2^2 \\
&\quad \geqslant m_L \left\| \frac{1}{n}\sum_{i=1}^{n} \left[a(\boldsymbol{x}_i, y_i) \cdot y_i \cdot \boldsymbol{x}_{L-1,i} \cdot \frac{1}{m_L}\sum_{j=1}^{m_L} \sigma'(\mathbf{w}_{L,j}^\top \boldsymbol{x}_{L-1,i}) \right] \right\|_2^2 \\
&\quad \geqslant m_L \left[\frac{1}{n}\sum_{i=1}^{n} \left\langle a(\boldsymbol{x}_i, y_i) \cdot y_i \cdot \boldsymbol{x}_{L-1,i} \cdot \frac{1}{m_L}\sum_{j=1}^{m_L} \sigma'(\mathbf{w}_{L,j}^\top \boldsymbol{x}_{L-1,i}), \boldsymbol{\alpha}_{L-1} \right\rangle \right]^2 \\
&\quad \geqslant 4^{-L}\gamma^2 \cdot m_L \left[\frac{1}{n}\sum_{i=1}^{n} a(\boldsymbol{x}_i, y_i) \cdot \frac{1}{m_L}\sum_{j=1}^{m_L} \sigma'(\mathbf{w}_{L,j}^\top \boldsymbol{x}_{L-1,i}) \right]^2 \\
&\quad \geqslant 4^{-L}/8 \cdot \gamma^2 \cdot m_L \left[\frac{1}{n}\sum_{i=1}^{n} a(\boldsymbol{x}_i, y_i) \right]^2.
\end{aligned}$$

This completes the proof. □

Lemma C.4. For any $\delta > 0$, under the same assumptions as Lemma C.3, if $\tau \leqslant \nu 8^{-(L+2)} L^{-2} \gamma^3$ for some small enough absolute constant ν, then with probability at least $1-\delta$, there exists an absolute constant C such that

$$\|\nabla_{\mathbf{W}_L} L_S(\widetilde{\mathbf{W}})\|_F^2 \geqslant C16^{-L} \cdot \gamma^4 m_L \cdot \bigg\{\frac{1}{n}\sum_{i=1}^n \ell'\big[y_i \cdot f_{\widetilde{\mathbf{W}}}(\boldsymbol{x}_i)\big]\bigg\}^2.$$

for all $\widetilde{\mathbf{W}} \in \mathcal{W}(\mathbf{W}, \tau)$.

Proof of Lemma C.4. For $i = 1, \ldots, n$, denote by $\widetilde{y}_i = f_{\widetilde{\mathbf{W}}}(\boldsymbol{x}_i)$ the output of the neural network with parameter matrices $\widetilde{\mathbf{W}}_1, \ldots, \widetilde{\mathbf{W}}_L$ and input $\boldsymbol{x}_i$, and define

$$\mathbf{g}_j = \frac{1}{n}\sum_{i=1}^n [\ell'(y_i\widetilde{y}_i) \cdot v_j \cdot y_i \cdot \sigma'(\mathbf{w}_{L,j}^\top \boldsymbol{x}_{L-1,i}) \cdot \boldsymbol{x}_{L-1,i}],$$

Since $0 \leqslant |v_j \ell'(y_i\widetilde{y}_i)| \leqslant 1$, by Lemma C.3, with probability at least $1 - \delta/2$,

$$\sum_{j=1}^{m_L} \|\mathbf{g}_j\|_2^2 \geqslant 4^{-L}/8 \cdot m_L \cdot \gamma^2 \cdot \Bigg[\frac{1}{n}\sum_{i=1}^n \ell'(y_i\widetilde{y}_i)\Bigg]^2.$$

Define

$$A = \Bigg\{j \in [m_L] : \|\mathbf{g}_j\|_2^2 \geqslant 4^{-(L+2)} \cdot \gamma^2 \cdot \Bigg[\frac{1}{n}\sum_{i=1}^n \ell'(y_i\widetilde{y}_i)\Bigg]^2\Bigg\}.$$

Since for any $j \in [m_{L-1}]$, by Jensen's inequality we have

$$\begin{aligned}
\|\mathbf{g}_j\|_2 &= \Bigg\|\frac{1}{n}\sum_{i=1}^n [\ell'(y_i\widetilde{y}_i) \cdot v_j \cdot y_i \cdot \sigma'(\mathbf{w}_{L,j}^\top \boldsymbol{x}_{L-1,i}) \cdot \boldsymbol{x}_{L-1,i}]\Bigg\|_2 \\
&\leqslant \frac{1}{n}\sum_{i=1}^n \big\|\ell'(y_i\widetilde{y}_i) \cdot v_j \cdot y_i \cdot \sigma'(\mathbf{w}_{L,j}^\top \boldsymbol{x}_{L-1,i}) \cdot \boldsymbol{x}_{L-1,i}\big\|_2 \\
&\leqslant -\frac{1}{n}\sum_{i=1}^n \ell'(y_i\widetilde{y}_i).
\end{aligned}$$

Therefore

$$\begin{aligned}
|A| \cdot \Bigg[-\frac{1}{n}\sum_{i=1}^n \ell'(y_i\widetilde{y}_i)\Bigg]^2 &\geqslant \sum_{j \in A} \|\mathbf{g}_j\|_2^2 + \sum_{j \in A^c} \|\mathbf{g}_j\|_2^2 - m_L \cdot 4^{-(L+2)} \cdot \gamma^2 \cdot \Bigg[-\frac{1}{n}\sum_{i=1}^n \ell'(y_i\widetilde{y}_i)\Bigg]^2 \\
&\geqslant 4^{-(L+2)} \cdot m_L \gamma^2 \cdot \Bigg[-\frac{1}{n}\sum_{i=1}^n \ell'(y_i\widetilde{y}_i)\Bigg]^2,
\end{aligned}$$

and

$$|A| \geqslant 4^{-(L+2)} \cdot m_L \gamma^2.$$

Define

$$A' = \big\{j \in [m_L] : \sigma'(\langle \widetilde{\mathbf{w}}_{L,j}, \widetilde{\mathbf{x}}_{L-1}\rangle) \neq \sigma'(\langle \mathbf{w}_{L,j}, \mathbf{x}_{L-1}\rangle)\big\}.$$

Then by Theorem 5.4 we have

$$|A'| \leqslant C_1 L^{4/3}\tau^{2/3} m_L,$$

where C_1 is an absolute constant. Therefore as long as $\tau \leqslant \nu 8^{-(L+2)} L^{-2}\gamma^3$ for some small enough absolute constant ν, we have

$$|A \backslash A'| \geqslant 4^{-(L+2)}/2 \cdot m_L \gamma^2.$$

Now by definition we have

$$\nabla_{\mathbf{W}_{L,j}} L_S(\widetilde{\mathbf{W}}) = \frac{1}{n}\sum_{i=1}^{n}[\ell'(y_i\widetilde{y}_i)\cdot v_j \cdot y_i \cdot \sigma'(\widetilde{\mathbf{w}}_{L,j}^\top \widetilde{\boldsymbol{x}}_{L-1,i})\cdot \widetilde{\boldsymbol{x}}_{L-1,i}].$$

For any $j \in A\backslash A'$, by Theorem 5.4, triangle inequality and Jensen's inequality we have

$$\begin{aligned}
\|\mathbf{g}_j\|_2 - \big\|\nabla_{\mathbf{W}_{L,j}} L_S(\widetilde{\mathbf{W}})\big\|_2 &\leqslant \Bigg\|\frac{1}{n}\sum_{i=1}^{n}[\ell'(y_i\widetilde{y}_i)\cdot v_j \cdot y_i \cdot \sigma'(\mathbf{w}_{L,j}^\top \boldsymbol{x}_{L-1,i})\cdot (\boldsymbol{x}_{L-1,i} - \widetilde{\boldsymbol{x}}_{L-1,i})]\Bigg\|_2\\
&\leqslant \frac{1}{n}\sum_{i=1}^{n}\big\|\ell'(y_i\widetilde{y}_i)\cdot v_j \cdot y_i \cdot \sigma'(\mathbf{w}_{L,j}^\top \boldsymbol{x}_{L-1,i})\cdot (\boldsymbol{x}_{L-1,i} - \widetilde{\boldsymbol{x}}_{L-1,i})\big\|_2\\
&\leqslant C_2 L\tau \cdot \Bigg[-\frac{1}{n}\sum_{i=1}^{n}\ell'(y_i\widetilde{y}_i)\Bigg],
\end{aligned}$$

where C_2 is an absolute constant. Therefore as long as $\tau \leqslant \nu 2^{-L}/(8L)\cdot\gamma$ for some small enough absolute constant ν,

$$\big\|\nabla_{\mathbf{W}_{L,j}} L_S(\widetilde{\mathbf{W}})\big\|_2 \geqslant \|\mathbf{g}_j\|_2 - C_2 L\tau \cdot \Bigg[-\frac{1}{n}\sum_{i=1}^{n}\ell'(y_i\widetilde{y}_i)\Bigg] \geqslant 2^{-L}/8\cdot\gamma\cdot\Bigg[-\frac{1}{n}\sum_{i=1}^{n}\ell'(y_i\widetilde{y}_i)\Bigg]$$

for all $j \in A\backslash A'$. Therefore

$$\begin{aligned}
\big\|\nabla_{\mathbf{W}_L} L_S(\widetilde{\mathbf{W}})\big\|_F^2 &\geqslant |A\backslash A'|\cdot 4^{-L}/64\cdot\gamma^2\cdot\Bigg[-\frac{1}{n}\sum_{i=1}^{n}\ell'(y_i\widetilde{y}_i)\Bigg]^2\\
&\geqslant 16^{-L}/2048\cdot\gamma^4 m_L\cdot\Bigg[-\frac{1}{n}\sum_{i=1}^{n}\ell'(y_i\widetilde{y}_i)\Bigg]^2.
\end{aligned}$$

This completes the proof. □

Lemma C.5. Suppose that $\mathbf{W}_1,\ldots,\mathbf{W}_L$ are generated via Gaussian initialization, and the results of Theorem 5.3 and Theorem 5.4 all hold. If $\tau \leqslant \nu L^{-5}[\log(m)]^{-3/2}$ for some small enough absolute constant $\nu > 0$, then

$$\big\|\nabla_{\mathbf{W}_l} L_S(\widetilde{\mathbf{W}})\big\|_F \leqslant CL\cdot\sqrt{m}\cdot\Bigg[-\frac{1}{n}\sum_{i=1}^{n}\ell'(y_i\widetilde{y}_i)\Bigg]$$

for all $\widetilde{\mathbf{W}} \in \mathcal{W}(\mathbf{W}, \tau)$ and $l \in [L]$, where $\widetilde{y}_i = f_{\widetilde{\mathbf{W}}}(\boldsymbol{x}_i)$, $i = 1, \ldots, n$, and C is an absolute constant.

Proof of Lemma C.5. By definition, we have

$$\nabla_{\mathbf{W}_l} L_S(\widetilde{\mathbf{W}}) = \frac{1}{n}\sum_{i=1}^n \ell'(y_i\widetilde{y}_i) \cdot y_i \cdot \widetilde{\boldsymbol{x}}_{l-1,i}\mathbf{v}^\top \Bigg(\prod_{r=l+1}^L \widetilde{\boldsymbol{\Sigma}}_r(\boldsymbol{x}_i)\widetilde{\mathbf{W}}_l^\top\Bigg)\widetilde{\boldsymbol{\Sigma}}_l(\boldsymbol{x}_i).$$

By Theorem 5.3 and Theorem 5.4, we have

$$\|\widetilde{\boldsymbol{x}}_{l-1,i}\|_2 \leqslant C_1, \quad \Bigg\|\prod_{r=l+1}^L \widetilde{\boldsymbol{\Sigma}}_r(\boldsymbol{x}_i)\widetilde{\mathbf{W}}_l^\top\Bigg\|_2 \leqslant C_2 L,$$

for all $i \in [n]$ and $l \in [L]$, where C_1, C_2 are absolute constants. By triangle inequality, we have

$$\begin{aligned}
\|\nabla_{\mathbf{W}_l} L_S(\widetilde{\mathbf{W}})\|_F &\leqslant \frac{1}{n}\sum_{i=1}^n \Bigg\|\ell'(y_i\widetilde{y}_i) \cdot y_i \cdot \widetilde{\boldsymbol{x}}_{l-1,i}\mathbf{v}^\top \Bigg(\prod_{r=l+1}^L \widetilde{\boldsymbol{\Sigma}}_r(\boldsymbol{x}_i)\widetilde{\mathbf{W}}_l^\top\Bigg)\widetilde{\boldsymbol{\Sigma}}_l(\boldsymbol{x}_i)\Bigg\|_F \\
&= \frac{1}{n}\sum_{i=1}^n \big\|\ell'(y_i\widetilde{y}_i) \cdot y_i \cdot \widetilde{\boldsymbol{x}}_{l-1,i}\big\|_2 \cdot \Bigg\|\mathbf{v}^\top\Bigg(\prod_{r=l+1}^L \widetilde{\boldsymbol{\Sigma}}_r(\boldsymbol{x}_i)\widetilde{\mathbf{W}}_l^\top\Bigg)\widetilde{\boldsymbol{\Sigma}}_l(\boldsymbol{x}_i)\Bigg\|_2 \\
&\leqslant C_3 L \cdot \sqrt{m} \cdot \Bigg[-\frac{1}{n}\sum_{i=1}^n \ell'(y_i\widetilde{y}_i)\Bigg]
\end{aligned}$$

for all $l \in [L]$, where C_3 is an absolute constant. This completes the proof. □

Lemma C.6. Suppose that $\mathbf{W}_1, \ldots, \mathbf{W}_L$ are generated via Gaussian initialization, and the results of Theorem 5.3 and Theorem 5.4 all hold. If $\tau \leqslant \nu L^{-5}[\log(m)]^{-3/2}$ for some small enough absolute constant $\nu > 0$, then

$$\begin{aligned}
L_S(\widehat{\mathbf{W}}) - L_S(\widetilde{\mathbf{W}}) &\leqslant C\sum_{l=1}^L L^{8/3}\tau^{1/3}\sqrt{m\log(m)} \cdot \big\|\widehat{\mathbf{W}}_l - \widetilde{\mathbf{W}}_l\big\|_2 \cdot \Bigg[-\frac{1}{n}\sum_{i=1}^n \ell'(y_i\widetilde{y}_i)\Bigg] \\
&\quad + C\sum_{l=1}^L mL^3 \cdot \|\widehat{\mathbf{W}}_l - \widetilde{\mathbf{W}}_l\|_2^2 + \sum_{l=1}^L \mathrm{Tr}[(\widehat{\mathbf{W}}_l - \widetilde{\mathbf{W}}_l)^\top \nabla_{W_l} L_S(\widetilde{\mathbf{W}})]
\end{aligned}$$

for all $\widetilde{\mathbf{W}}, \widehat{\mathbf{W}} \in \mathcal{W}(\mathbf{W}, \tau)$, where C is an absolute constant.

Proof of Lemma C.6. Denote by $\widehat{y}_i$, $\widetilde{y}_i$ the outputs of the network with input $\boldsymbol{x}_i$ and parameter matrices $\widehat{\mathbf{W}}$, $\widetilde{\mathbf{W}}$ respectively. Since $\ell''(z) \leqslant 1/2$ for all $z \in \mathbb{R}$, we have

$$L_S(\widehat{\mathbf{W}}) - L_S(\widetilde{\mathbf{W}}) = \frac{1}{n}\sum_{i=1}^n[\ell(y_i\widehat{y}_i) - \ell(y_i\widetilde{y}_i)] \leqslant \frac{1}{n}\sum_{i=1}^n[\ell'(y_i\widehat{y}_i) \cdot y_i \cdot (\widehat{y}_i - \widetilde{y}_i) + (\widehat{y}_i - \widetilde{y}_i)^2/4].$$

Denote $\Delta_i = \widehat{y}_i - \widetilde{y}_i$. Then

$$L_S(\widehat{\mathbf{W}}) - L_S(\widetilde{\mathbf{W}}) \leqslant \frac{1}{n}\sum_{i=1}^n \big[\ell'(y_i\widetilde{y}_i) \cdot y_i \cdot \Delta_i + (\Delta_i)^2/4\big]. \tag{C.1}$$

By Theorem 5.4, we have

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}\ell'(y_i\widetilde{y}_i)\cdot y_i\cdot\Delta_i \leqslant\ & C_1\sum_{l=1}^{L}L^{8/3}\tau^{1/3}\sqrt{m\log(m)}\cdot\big\|\widehat{\mathbf{W}}_l-\widetilde{\mathbf{W}}_l\big\|_2\cdot\Bigg[-\frac{1}{n}\sum_{i=1}^{n}\ell'(y_i\widetilde{y}_i)\Bigg] \\ & + C_1\sum_{l=1}^{L}L^3\cdot\|\widehat{\mathbf{W}}_l-\widetilde{\mathbf{W}}_l\|_2^2\cdot\Bigg[-\frac{1}{n}\sum_{i=1}^{n}\ell'(y_i\widetilde{y}_i)\Bigg] \\ & + \sum_{l=1}^{L}\mathrm{Tr}[(\widehat{\mathbf{W}}_l-\widetilde{\mathbf{W}}_l)^\top\nabla_{W_l}L_S(\widetilde{\mathbf{W}})], \end{aligned} \tag{C.2}$$

where C_1 is an absolute constant. Moreover, by Theorem 5.4, clearly we have

$$\Delta_i^2 = [\mathbf{v}^\top(\widehat{\boldsymbol{x}}_{L,i}-\widetilde{\boldsymbol{x}}_{L,i})]^2 \leqslant C_2\cdot m\cdot L^2\Bigg(\sum_{l=1}^{L}\|\widehat{\mathbf{W}}_l-\widetilde{\mathbf{W}}_l\|_2\Bigg)^2 \leqslant C_2\cdot m\cdot L^3\cdot\sum_{l=1}^{L}\|\widehat{\mathbf{W}}_l-\widetilde{\mathbf{W}}_l\|_2^2 \tag{C.3}$$

for some absolute constant C_2. Plugging (C.2) and (C.3) into (C.1), and using the fact $-\ell'(z)\leqslant 1$, $z\in\mathbb{R}$ completes the proof.

□

D Proof of Theorem 5.6

Lemma D.1. Let $\mathbf{W}_1^{(0)},\ldots,\mathbf{W}_L^{(0)}$ be generated via Gaussian initialization, and $\mathbf{W}^{(k)}=\{\mathbf{W}_l^{(k)}\}_{l=1}^L$ be the k-th iterate of gradient descent with step size η. There exists absolute constants $C, C', C'', C''', C'''' > 0$ such that, If $\mathbf{W}^{(k)}\in\mathcal{W}(\mathbf{W}^{(0)},\tau)$ for all $k=1,\ldots,K$, where $\tau\leqslant C'16^{-3L}L^{-14}\gamma^{12}[\log(m)]^{-3/2}$, and $\eta\leqslant C''\cdot 16^{-L}L^{-6}\gamma m^{-1}$, then

$$L_S(\mathbf{W}^{(k+1)})-L_S(\mathbf{W}^{(k)})\leqslant -C'''\eta\cdot 16^{-L}\cdot\gamma^4\cdot m\cdot\Bigg[-\frac{1}{n}\sum_{i=1}^{n}\ell'\big(y_iy_i^{(k)}\big)\Bigg]^2$$

for all $k=0,\ldots,K-1$. Moreover, within K iterations, gradient descent finds a point $\mathbf{W}^{(k)}$ with

$$-\frac{1}{n}\sum_{i=1}^{n}\ell'\big[y_i\cdot f_{\mathbf{W}^{(k)}}(\boldsymbol{x}_i)\big]\leqslant C''''(K\eta\cdot m)^{-1/2}\cdot[\log(n/\delta)]^{1/4}\cdot 4^L\gamma^{-2}.$$

Proof of Lemma D.1. Denote $y_i^{(k)} = f_{\mathbf{W}^{(k)}}(\boldsymbol{x}_i)$. Then by Theorem 5.5 and $\mathbf{W}_l^{(k+1)}-\mathbf{W}_l^{(k)} =$

$-\eta\nabla_{\mathbf{W}_l} L_S(\mathbf{W}^{(k)})$, we have

$$\begin{aligned} L_S(\mathbf{W}^{(k+1)}) - L_S(\mathbf{W}^{(k)}) &\leqslant C_1 \sum_{l=1}^{L} L^{8/3}\tau^{1/3}\sqrt{m\log(m)} \cdot \|\mathbf{W}_l^{(k+1)} - \mathbf{W}_l^{(k)}\|_2 \bigg[-\frac{1}{n}\sum_{i=1}^{n} \ell'\big(y_i y_i^{(k)}\big)\bigg] \\ &\quad + C_1 \sum_{l=1}^{L} mL^3 \cdot \|\mathbf{W}_l^{(k+1)} - \mathbf{W}_l^{(k)}\|_2^2 \\ &\quad + \sum_{l=1}^{L} \mathrm{Tr}[(\mathbf{W}_l^{(k+1)} - \mathbf{W}_l^{(k)})^\top \nabla_{W_l} L_S(\mathbf{W}^{(k)})] \\ &\leqslant C_1 \sum_{l=1}^{L} L^{8/3}\tau^{1/3}\sqrt{m\log(m)} \cdot \eta \|\nabla_{\mathbf{W}_l} L_S(\mathbf{W}^{(k)})\|_2 \bigg[-\frac{1}{n}\sum_{i=1}^{n} \ell'\big(y_i y_i^{(k)}\big)\bigg] \\ &\quad + C_1 \sum_{l=1}^{L} mL^3 \cdot \eta^2 \|\nabla_{\mathbf{W}_l} L_S(\mathbf{W}^{(k)})\|_2^2] - \eta \cdot \sum_{l=1}^{L} \|\nabla_{W_l} L_S(\mathbf{W}^{(k)})\|_F^2, \end{aligned}$$

where C_1 is an absolute constant. Applying the upper bounds of $\|\nabla_{\mathbf{W}_l} L_S(\mathbf{W}^{(k)})\|_F$, $l = 1, \ldots, L$ and lower bound of $\|\nabla_{\mathbf{W}_L} L_S(\mathbf{W}^{(k)})\|_F$ given in Theorem 5.5, we obtain

$$\begin{aligned} L_S(\mathbf{W}^{(k+1)}) - L_S(\mathbf{W}^{(k)}) &\leqslant C_2 \sum_{l=1}^{L} L^{11/3}\tau^{1/3} m\sqrt{\log(m)} \cdot \eta \cdot \bigg[-\frac{1}{n}\sum_{i=1}^{n} \ell'\big(y_i y_i^{(k)}\big)\bigg]^2 \\ &\quad + C_2 \sum_{l=1}^{L} L^5 \cdot m^2 \cdot \eta^2 \cdot \bigg[-\frac{1}{n}\sum_{i=1}^{n} \ell'\big(y_i \cdot y_i^{(k)}\big)\bigg]^2 \\ &\quad - C_3\eta \cdot 16^{-L} \cdot \gamma^4 \cdot m \cdot \bigg[-\frac{1}{n}\sum_{i=1}^{n} \ell'\big(y_i \cdot y_i^{(k)}\big)\bigg]^2, \end{aligned}$$

where C_2, C_3 are absolute constants. By assumption, we have

$$L^{14/3}\tau^{1/3}\sqrt{\log(m)}, L^6\eta \leqslant C \cdot 16^{-L}\gamma^4 m$$

for some small enough absolute constant C. Therefore by the fact that $\ell'(z) \in (-1, 0)$ for all $z \in \mathbb{R}$, we have

$$L_S(\mathbf{W}^{(k+1)}) - L_S(\mathbf{W}^{(k)}) \leqslant -C_4\eta \cdot 16^{-L} \cdot \gamma^4 \cdot m \cdot \bigg[-\frac{1}{n}\sum_{i=1}^{n} \ell'\big(y_i y_i^{(k)}\big)\bigg]^2,$$

where C_4 is an absolute constant. Let

$$k^* = \underset{k\in\{0,\ldots,K-1\}}{\mathrm{argmin}} \bigg[-\frac{1}{n}\sum_{i=1}^{n} \ell'\big(y_i y_i^{(k)}\big)\bigg]^2.$$

Then telescoping over k gives

$$\begin{aligned} L_S(\mathbf{W}^{(K)}) - L_S(\mathbf{W}^{(0)}) &\leqslant -C_4\eta \cdot 16^{-L} \cdot \gamma^4 \cdot m \cdot \sum_{k=1}^{K} \bigg[-\frac{1}{n}\sum_{i=1}^{n} \ell'\big(y_i y_i^{(k)}\big)\bigg]^2 \\ &\leqslant -C_4 K\eta \cdot 16^{-L} \cdot \gamma^4 \cdot m \cdot \bigg[-\frac{1}{n}\sum_{i=1}^{n} \ell'\big(y_i y_i^{(k^*)}\big)\bigg]^2. \end{aligned}$$

Therefore by (iv) in Theorem 5.3 we have

$$\begin{aligned} -\frac{1}{n}\sum_{i=1}^{n} \ell'\big(y_i y_i^{(k^*)}\big) &\leqslant C_5(K\eta \cdot m)^{-1/2} \cdot [L_S(\mathbf{W}^{(0)})]^{1/2} \cdot 4^L \gamma^{-2} \\ &\leqslant C_5(K\eta \cdot m)^{-1/2} \cdot [\log(n/\delta)]^{1/4} \cdot 4^L \gamma^{-2}, \end{aligned}$$

where C_5 is an absolute constant. This completes the proof. □

Lemma D.2. Suppose that $\mathbf{W}_1^{(0)}, \ldots, \mathbf{W}_L^{(0)}$ are generated via Gaussian initialization, and $\mathbf{W}_1^{(k)}, \ldots, \mathbf{W}_L^{(k)}$, $k = 1, \ldots, K$ are the iterates obtained via gradient descent starting at $\mathbf{W}^{(0)}$ with step size η. Let $\tau \leqslant C_0 16^{-3L} L^{-14} \gamma^{12} [\log(m)]^{-3/2}$, where C_0 is the same absolute constant introduced in Lemma D.1 in the rate of τ. If $\eta \leqslant C \min\{16^{-L} L^{-6} \gamma m^{-1}, L^{-1} m^{-1/2} \tau\}$ and $K \cdot \eta \leqslant C' 16^{-L} L^{-2} \gamma^4 \tau^2 [\log(n/\delta)]^{-1/2}$ for some small enough absolute constants C, C', then $\mathbf{W}^{(k)} \in \mathcal{W}(\mathbf{W}^{(0)}, \tau/2)$ for all $k = 1, \ldots, K$.

Proof of Lemma D.2. We prove the result by induction. Assume that $\mathbf{W}^{(k')} \in \mathcal{W}(\mathbf{W}^{(0)}, \tau/2)$ for all $k' = 1, \ldots, k-1$. Then by Theorem 5.5, we have

$$\big\|\nabla_{\mathbf{W}_l} L_S(\mathbf{W}^{(k')})\big\|_F \leqslant C_1 L \cdot \sqrt{m} \cdot \bigg\{ -\frac{1}{n}\sum_{i=1}^{n} \ell'[y_i \cdot f_{\mathbf{W}^{(k')}}(\boldsymbol{x}_i)]\bigg\}$$

for all $k' = 1, \ldots, k-1$ and $l = 1, \ldots, L$, where C_1 is an absolute constant. By the assumption that $\eta \leqslant C \min\{16^{-L} L^{-6} \gamma m^{-1}, L^{-1} m^{-1/2} \tau\}$ for some small enough absolute constant C and the fact that $\ell'(z) \in (-1, 0)$ for all $z \in \mathbb{R}$, we have

$$\begin{aligned} \|\mathbf{W}_l^{(k)} - \mathbf{W}_l^{(0)}\|_F &\leqslant \|\mathbf{W}_l^{(k)} - \mathbf{W}_l^{(k-1)}\|_F + \|\mathbf{W}_l^{(k-1)} - \mathbf{W}_l^{(0)}\|_F \\ &\leqslant \eta \cdot \big\|\nabla_{\mathbf{W}_l} L_S(\mathbf{W}^{(k')})\big\|_F + \tau/2 \\ &\leqslant C_1 \eta \cdot L \cdot \sqrt{m} + \tau/2 \\ &\leqslant \tau. \end{aligned}$$

Therefore $\mathbf{W}_l^{(k)} \in \mathcal{W}(\mathbf{W}_l^{(0)}, \tau)$, and all results given by Lemma D.1 holds for $\mathbf{W}_l^{(k)}$. Therefore we have

$$
\begin{aligned}
\|\mathbf{W}_l^{(k)} - \mathbf{W}_l^{(0)}\|_F &\leqslant \eta \cdot \sum_{k'=0}^{k-1} \|\nabla_{\mathbf{W}_l} L_S(\mathbf{W}^{(k')})\|_F \\
&\leqslant \eta \cdot L \cdot \sqrt{m} \cdot \sum_{k'=0}^{k-1} \bigg\{ -\frac{1}{n} \sum_{i=1}^{n} \ell'\big[y_i \cdot f_{\mathbf{W}^{(k')}}(\boldsymbol{x}_i)\big] \bigg\} \\
&\leqslant \eta \cdot L \cdot \sqrt{m} \cdot \sqrt{k} \cdot \sqrt{\sum_{k'=0}^{k-1} \bigg\{ -\frac{1}{n} \sum_{i=1}^{n} \ell'\big[y_i \cdot f_{\mathbf{W}^{(k')}}(\boldsymbol{x}_i)\big] \bigg\}^2} \\
&\leqslant \eta \cdot L \cdot \sqrt{k} \cdot \sqrt{(\eta \cdot 16^{-L} \cdot \gamma^4)^{-1} \cdot \sum_{k'=0}^{k-1} \big[L_S(\mathbf{W}^{(k)}) - L_S(\mathbf{W}^{(k+1)})\big]} \\
&\leqslant \sqrt{k\eta} \cdot 4^L \cdot L \cdot \gamma^{-2} \cdot \sqrt{L_S(\mathbf{W}^{(0)})} \\
&\leqslant \sqrt{k\eta} \cdot 4^L \cdot L \cdot \gamma^{-2} \cdot [\log(n/\delta)]^{1/4} \\
&\leqslant \tau/2.
\end{aligned}
$$

This completes the proof. □

Proof of Theorem 5.6. By Theorem 5.5, Lemma E.2 and Lemma E.1, if

$$
\begin{aligned}
&\tau \leqslant C_1 16^{-3L} L^{-14} \gamma^{12} [\log(m)]^{-3/2}, && \text{(D.1)} \\
&C_2 (K\eta \cdot m)^{-1/2} \cdot [\log(n/\delta)]^{1/4} \cdot 4^L \gamma^{-2} \leqslant \epsilon/4, && \text{(D.2)} \\
&m \geqslant C_3 \max\{dL^2 \log(m/\delta), L^{-8/3} \tau^{-4/3} \log[m/(\tau\delta)]\}, && \text{(D.3)} \\
&K\eta \leqslant C_4 16^{-L} L^{-2} \gamma^4 \tau^2 [\log(n/\delta)]^{-1/2}, && \text{(D.4)} \\
&\eta \leqslant C_5 \min\{16^{-L} L^{-6} \gamma m^{-1}, L^{-1} m^{-1/2} \tau\}, && \text{(D.5)}
\end{aligned}
$$

where $C_1, \ldots, C_5$ are absolute constants, then gradient descent can find $\mathbf{W}^{(k)}$ within K iterations that gives a population error less than ϵ. Set

$$\tau = C_6 \cdot 16^L \cdot \gamma^{-4} \cdot \sqrt{\log(n/\delta)} \cdot \epsilon^{-1} m^{-1/2}$$

to be the solution of equations

$$
\begin{aligned}
&C_2 (K\eta \cdot m)^{-1/2} \cdot [\log(n/\delta)]^{1/4} \cdot 4^L \gamma^{-2} = \epsilon/4, \\
&K\eta = C_4 16^{-L} L^{-2} \gamma^4 \tau^2 [\log(n/\delta)]^{-1/2},
\end{aligned}
$$

where we cancel $K\eta$ to get dependency of τ on m. Moreover, let $\widehat{m} = \widetilde{\Omega}(\mathrm{poly}(2^L, \gamma^{-1})) \cdot \epsilon^{-2}$ be the solution of the equation

$$C_6 \cdot 16^L \cdot \gamma^{-4} \cdot \sqrt{\log(n/\delta)} \cdot \epsilon^{-1} m^{-1/2} = C_1 16^{-3L} L^{-14} \gamma^{12} [\log(m)]^{-3/2},$$

and set

$$
\begin{aligned}
m^* &= \max\{\widehat{m}, dL^2\log(m/\delta), L^{-8/3}\tau^{-4/3}\log[m/(\tau\delta)], \tau^{-2}\} \\
&= \widetilde{\Omega}(\mathrm{poly}(2^L, \gamma^{-1})) \cdot \max\{d, \epsilon^{-2}\}\log(1/\delta).
\end{aligned}
$$

Then it is clear that as long as $m \geqslant m^*$, there exist

$$
K = \widetilde{O}(\mathrm{poly}(2^L, \gamma^{-1})) \cdot \epsilon^{-2},\ \ \eta = O(16^{-L}L^{-6}\gamma m^{-1})
$$

such that inequalities (D.1), (D.2), (D.3), (D.4), (D.5) all hold. This completes the proof. □

E Proof of Theorem 5.7

Lemma E.1. Suppose that $\mathbf{W}^{(0)} = (\mathbf{W}_1^{(0)}, \ldots, \mathbf{W}_L^{(0)})$ are generated via Gaussian initialization and the results of Theorem 5.4 and Theorem 5.5 hold with $\mathbf{W} = \mathbf{W}^{(0)}$. Then there exists $m^* = \widetilde{O}(\mathrm{poly}(2^L, \gamma^{-1})) \cdot \epsilon^{-2}$ such that, for any $\delta > 0$, if $m \geqslant m^*$, then with probability at least $1-\delta$,

$$
\sup_{\mathbf{W}\in\mathcal{W}(\mathbf{W}^{(0)},\tau)} |\mathcal{E}_S(\mathbf{W}) - \mathcal{E}_{\mathcal{D}}(\mathbf{W})| \leqslant C\frac{16^L \cdot L^2 \cdot \gamma^{-4} \cdot \sqrt{\log(n/\delta)} \cdot \epsilon^{-1}}{\sqrt{n}} + \frac{\epsilon}{2},
$$

where C is an absolute constant, and $\tau = O(16^L \cdot \gamma^{-4} \cdot \sqrt{\log(n/\delta)} \cdot \epsilon^{-1} m^{-1/2})$ is the neighborhood radius given in Theorem 5.6.

Proof of Lemma E.1. Let $\mathcal{F}(\mathbf{W}^{(0)}, \tau) = \{f_{\mathbf{W}}(\mathbf{x}) : \mathbf{W} \in \mathcal{W}(\mathbf{W}^{(0)}, \tau)\}$. We consider the empirical Rademacher complexity (Bartlett and Mendelson, 2002; Mohri et al., 2018; Shalev-Shwartz and Ben-David, 2014) of $\mathcal{F}(\mathbf{W}^{(0)}, \tau)$ defined as follows

$$
\widehat{\mathfrak{R}}_n[\mathcal{F}(\mathbf{W}^{(0)}, \tau)] = \mathbb{E}_{\boldsymbol{\xi}}\Bigg[\sup_{f\in\mathcal{F}(\mathbf{W}^{(0)},\tau)} \frac{1}{n}\sum_{i=1}^n \xi_i f(\boldsymbol{x}_i)\Bigg] = \mathbb{E}_{\boldsymbol{\xi}}\Bigg[\sup_{\mathbf{W}\in\mathcal{W}(\mathbf{W}^{(0)},\tau)} \frac{1}{n}\sum_{i=1}^n \xi_i f_{\mathbf{W}}(\boldsymbol{x}_i)\Bigg],
$$

where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)^\top$ is an n-dimensional vector consisting of independent Rademacher random variables $\xi_1, \ldots, \xi_n$. Since $y \in \{+1, 1\}$, $|\ell'(z)| \leqslant 1$ and $\ell'(z)$ is 1-Lipschitz continuous, by symmetrization and the standard uniform convergence results in terms of empirical Rademacher complexity (Mohri et al., 2018; Shalev-Shwartz and Ben-David, 2014), with probability at least $1-\delta$ we have

$$
\begin{aligned}
\sup_{\mathbf{W}\in\mathcal{W}(\mathbf{W}^{(0)},\tau)} |\mathcal{E}_S(\mathbf{W}) - \mathcal{E}_{\mathcal{D}}(\mathbf{W})| &= \sup_{\mathbf{W}\in\mathcal{W}(\mathbf{W}^{(0)},\tau)} \Bigg|\frac{1}{n}\sum_{i=1}^n \ell'[y_i \cdot f_{\mathbf{W}}(\boldsymbol{x}_i)] - \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\ell'[y \cdot f_{\mathbf{W}}(\mathbf{x})]\Bigg| \\
&\leqslant 2\widehat{\mathfrak{R}}_n[\mathcal{F}(\mathbf{W}^{(0)}, \tau)] + C_1\sqrt{\frac{\log(1/\delta)}{n}},
\end{aligned}
$$

where C_1 is an absolute constant. We now bound the empirical Rademacher complexity term $\widehat{\mathfrak{R}}_n[\mathcal{F}(\mathbf{W}^{(0)}, \tau)]$. By definition, we have

$$
\widehat{\mathfrak{R}}_n[\mathcal{F}(\mathbf{W}^{(0)}, \tau)] = \mathbb{E}_{\boldsymbol{\xi}}\Bigg[\sup_{\mathbf{W}\in\mathcal{W}(\mathbf{W}^{(0)},\tau)} \frac{1}{n}\sum_{i=1}^n \xi_i f_{\mathbf{W}}(\boldsymbol{x}_i)\Bigg] \leqslant I_1 + I_2, \tag{E.1}
$$

where

$$I_1 = \mathbb{E}_{\boldsymbol{\xi}}\bigg\{ \sup_{\mathbf{W}\in\mathcal{W}(\mathbf{W}^{(0)},\tau)} \frac{1}{n}\sum_{i=1}^n \xi_i\big[f_{\mathbf{W}}(\boldsymbol{x}_i) - F_{\mathbf{W}^{(0)},\mathbf{W}}(\boldsymbol{x}_i)\big]\bigg\},$$
$$I_2 = \mathbb{E}_{\boldsymbol{\xi}}\bigg\{ \sup_{\mathbf{W}\in\mathcal{W}(\mathbf{W}^{(0)},\tau)} \frac{1}{n}\sum_{i=1}^n \xi_i \sum_{l=1}^L \mathrm{Tr}\big[(\mathbf{W}_l - \mathbf{W}_l^{(0)})^\top \nabla_{\mathbf{W}_l} f_{\mathbf{W}^{(0)}}(\boldsymbol{x}_i)\big]\bigg\},$$

and

$$F_{\mathbf{W}^{(0)},\mathbf{W}}(\mathbf{x}) = f_{\mathbf{W}^{(0)}}(\mathbf{x}) + \sum_{l=1}^L \mathrm{Tr}\big[(\mathbf{W}_l - \mathbf{W}_l^{(0)})^\top \nabla_{\mathbf{W}_l} f_{\mathbf{W}^{(0)}}(\mathbf{x})\big].$$

For I_1, by Theorem 5.4, we have

$$\big|f_{\mathbf{W}}(\mathbf{x}) - F_{\mathbf{W}^{(0)},\mathbf{W}}(\mathbf{x})\big| \leqslant C_2\big[L^{11/3}\tau^{4/3}\cdot\sqrt{m\log(m)} + L^4\tau^2\sqrt{m}\big] \leqslant R\cdot\epsilon^{-4/3}m^{-1/6},$$

where $R = \widetilde{O}(\mathrm{poly}(2^L,\gamma^{-1}))$. Therefore, there exists $m^* = \widetilde{O}(\mathrm{poly}(2^L,\gamma^{-1}))\cdot\epsilon^{-14}$ such that

$$\big|f_{\mathbf{W}}(\mathbf{x}) - F_{\mathbf{W}^{(0)},\mathbf{W}}(\mathbf{x})\big| \leqslant \epsilon/2$$

for $m \geqslant m^*$. Then by triangle inequality we have $I_1 \leqslant \epsilon/2$. For I_2, first, by Theorem 5.3 we have

$$\|\nabla_{\mathbf{W}_l} f_{\mathbf{W}^{(0)}}(\mathbf{x})\|_F = \bigg\|\mathbf{x}_{l-1}\mathbf{v}^\top\bigg[\prod_{r=l+1}^L \boldsymbol{\Sigma}_r(\mathbf{x})\mathbf{W}_r^\top\bigg]\boldsymbol{\Sigma}_l(\mathbf{x})\bigg\|_F \leqslant C_2 L\sqrt{m}$$

for all $l = 1,\ldots,L$, where C_2 is an absolute constant. Therefore

$$I_2 = \frac{1}{n}\sum_{l=1}^L \mathbb{E}_{\boldsymbol{\xi}}\bigg\{\sup_{\|\widetilde{\mathbf{W}}_l\|_F\leqslant\tau} \mathrm{Tr}\bigg[\widetilde{\mathbf{W}}_l^\top \sum_{i=1}^n \xi_i \nabla_{\mathbf{W}_l} f_{\mathbf{W}^{(0)}}(\boldsymbol{x}_i)\bigg]\bigg\} \leqslant \frac{\tau}{n}\sum_{l=1}^L \mathbb{E}_{\boldsymbol{\xi}}\bigg[\bigg\|\sum_{i=1}^n \xi_i\nabla_{\mathbf{W}_l} f_{\mathbf{W}^{(0)}}(\boldsymbol{x}_i)\bigg\|_F\bigg].$$

By Jensen's inequality, we have

$$I_2 \leqslant \frac{\tau}{n}\sum_{l=1}^L \sqrt{\mathbb{E}_{\boldsymbol{\xi}}\bigg[\bigg\|\sum_{i=1}^n \xi_i\nabla_{\mathbf{W}_l} f_{\mathbf{W}^{(0)}}(\boldsymbol{x}_i)\bigg\|_F^2\bigg]} = \frac{\tau}{n}\sum_{l=1}^L \sqrt{\sum_{i=1}^n \|\nabla_{\mathbf{W}_l} f_{\mathbf{W}^{(0)}}(\boldsymbol{x}_i)\|_F^2} \leqslant \frac{C_2 L^2\tau\cdot\sqrt{m}}{\sqrt{n}}.$$

Plugging in the value of τ gives

$$I_2 \leqslant \frac{C_3 16^L\cdot L^2\cdot\gamma^{-4}\cdot\sqrt{\log(n/\delta)}\cdot\epsilon^{-1}}{\sqrt{n}}.$$

Finally, plugging in the bounds of I_1 and I_2 into (E.1) completes the proof. □

Lemma E.2. For any parameter matrices $\mathbf{W}_1,\ldots,\mathbf{W}_l$, the following inequality holds:

$$\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big[y\cdot f_{\mathbf{W}}(\mathbf{x}) < 0\big] \leqslant 2\cdot\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\ell'\big[y\cdot f_{\mathbf{W}}(\mathbf{x})\big].$$

Proof of Lemma E.2. The result directly follows by Markov's inequality:

$$\begin{aligned}\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\big\{-\ell'\big[y\cdot f_{\mathbf{W}}(\mathbf{x})\big]\big\} &\geqslant \frac{1}{2}\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big\{-\ell'\big[y\cdot f_{\mathbf{W}}(\mathbf{x})\big]\geqslant 1/2\big\}\\ &= \frac{1}{2}\mathbb{P}_{(\mathbf{x},y)\sim\mathcal{D}}\big[y\cdot f_{\mathbf{W}}(\mathbf{x})<0\big].\end{aligned}$$

This completes the proof. □

Proof of Theorem 5.7. The result of Theorem 5.7 directly follows by Lemma E.1 and Lemma E.2. □

References

Allen-Zhu, Z., Li, Y. and Liang, Y. (2018a). Learning and generalization in overparameterized neural networks, going beyond two layers. *arXiv preprint arXiv:1811.04918* .

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

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

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

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

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* .

Arora, S., Ge, R., Neyshabur, B. and Zhang, Y. (2018c). Stronger generalization bounds for deep nets via a compression approach. *arXiv preprint arXiv:1802.05296* .

Bartlett, P. L., Foster, D. J. and Telgarsky, M. J. (2017). Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*.

Bartlett, P. L. and Mendelson, S. (2002). Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research* **3** 463–482.

Belkin, M., Ma, S. and Mandal, S. (2018). To understand deep learning we need to understand kernel learning. *arXiv preprint arXiv:1802.01396* .

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

Brutzkus, A., Globerson, A., Malach, E. and Shalev-Shwartz, S. (2017). Sgd learns over-parameterized networks that provably generalize on linearly separable data. *arXiv preprint arXiv:1710.10174* .

Chizat, L. and Bach, F. (2018). On the global convergence of gradient descent for over-parameterized models using optimal transport. *arXiv preprint arXiv:1805.09545* .

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

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* .

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

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

Du, S. S., Lee, J. D., Tian, Y., Poczos, B. and Singh, A. (2017b). Gradient descent learns one-hidden-layer cnn: Don't be afraid of spurious local minima. *arXiv preprint arXiv:1712.00779* .

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

Dziugaite, G. K. and Roy, D. M. (2017). Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008* .

Freeman, C. D. and Bruna, J. (2016). Topology and geometry of half-rectified network optimization. *arXiv preprint arXiv:1611.01540* .

Golowich, N., Rakhlin, A. and Shamir, O. (2017). Size-independent sample complexity of neural networks. *arXiv preprint arXiv:1712.06541* .

Gunasekar, S., Lee, J., Soudry, D. and Srebro, N. (2018a). Characterizing implicit bias in terms of optimization geometry. *arXiv preprint arXiv:1802.08246* .

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

Gunasekar, S., Woodworth, B. E., Bhojanapalli, S., Neyshabur, B. and Srebro, N. (2017). Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*.

Haeffele, B. D. and Vidal, R. (2015). Global optimality in tensor factorization, deep learning, and beyond. *arXiv preprint arXiv:1506.07540* .

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

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

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

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

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N. et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine* **29** 82–97.

Jacot, A., Gabriel, F. and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572* .

Ji, Z. and Telgarsky, M. (2018). Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300* .

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

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

Li, X., Lu, J., Wang, Z., Haupt, J. and Zhao, T. (2018a). On tighter generalization bound for deep neural networks: Cnns, resnets, and beyond. *arXiv preprint arXiv:1806.05159* .

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

Li, Y., Ma, T. and Zhang, H. (2018b). Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory.*

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

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

Liang, T. and Rakhlin, A. (2018). Just interpolate: Kernel" ridgeless" regression can generalize. *arXiv preprint arXiv:1808.00387* .

Lin, H. and Jegelka, S. (2018). Resnet with one-neuron hidden layers is a universal approximator. In *Advances in Neural Information Processing Systems.*

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

Mei, S., Montanari, A. and Nguyen, P.-M. (2018). A mean field view of the landscape of two-layers neural networks. *arXiv preprint arXiv:1804.06561* .

Mendelson, S. (2014). Learning without concentration. In *Conference on Learning Theory.*

Mohri, M., Rostamizadeh, A. and Talwalkar, A. (2018). *Foundations of machine learning.* MIT press.

Nacson, M. S., Srebro, N. and Soudry, D. (2018). Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. *arXiv preprint arXiv:1806.01796* .

Neyshabur, B., Bhojanapalli, S., McAllester, D. and Srebro, N. (2017). A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1707.09564* .

Neyshabur, B., Li, Z., Bhojanapalli, S., LeCun, Y. and Srebro, N. (2018). Towards understanding the role of over-parametrization in generalization of neural networks. *arXiv preprint arXiv:1805.12076* .

Neyshabur, B., Tomioka, R. and Srebro, N. (2015). Norm-based capacity control in neural networks. In *Conference on Learning Theory.*

Nguyen, Q. and Hein, M. (2017). The loss surface of deep and wide neural networks. *arXiv preprint arXiv:1704.08045* .

Rahimi, A. and Recht, B. (2009). Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Advances in neural information processing systems.*

Rotskoff, G. M. and Vanden-Eijnden, E. (2018). Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. *arXiv preprint arXiv:1805.00915* .

Safran, I. and Shamir, O. (2016). On the quality of the initial basin in overspecified neural networks. In *International Conference on Machine Learning.*

Safran, I. and Shamir, O. (2017). Spurious local minima are common in two-layer relu neural networks. *arXiv preprint arXiv:1712.08968* .

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms.* Cambridge university press.

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

Sirignano, J. and Spiliopoulos, K. (2018). Mean field analysis of neural networks: A central limit theorem. *arXiv preprint arXiv:1808.09372* .

Soltanolkotabi, M. (2017). Learning relus via gradient descent. *arXiv preprint arXiv:1705.04591* .

Soltanolkotabi, M., Javanmard, A. and Lee, J. D. (2017). Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *arXiv preprint arXiv:1707.04926* .

Soudry, D. and Carmon, Y. (2016). No bad local minima: Data independent training error guarantees for multilayer neural networks. *arXiv preprint arXiv:1605.08361* .

Soudry, D. and Hoffer, E. (2017). Exponentially vanishing sub-optimal local minima in multilayer neural networks. *arXiv preprint arXiv:1702.05777* .

Soudry, D., Hoffer, E. and Srebro, N. (2017). The implicit bias of gradient descent on separable data. *arXiv preprint arXiv:1710.10345* .

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

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

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

Venturi, L., Bandeira, A. and Bruna, J. (2018). Neural networks with finite intrinsic dimension have no spurious valleys. *arXiv preprint arXiv:1802.06384* .

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

Wei, C., Lee, J. D., Liu, Q. and Ma, T. (2018). On the margin theory of feedforward neural networks. *arXiv preprint arXiv:1810.05369* .

Xie, B., Liang, Y. and Song, L. (2017). Diverse neural network learns true target functions. In *Artificial Intelligence and Statistics*.

Yang, G. (2019). Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation. *arXiv preprint arXiv:1902.04760* .

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

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

Yun, C., Sra, S. and Jadbabaie, A. (2017). Global optimality conditions for deep neural networks. *arXiv preprint arXiv:1707.02444* .

Yun, C., Sra, S. and Jadbabaie, A. (2018a). A critical view of global optimality in deep learning. *arXiv preprint arXiv:1802.03487* .

Yun, C., Sra, S. and Jadbabaie, A. (2018b). Small nonlinearities in activation functions create bad local minima in neural networks .

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

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

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* .

Zhou, Y. and Liang, Y. (2017). Critical points of neural networks: Analytical forms and landscape properties. *arXiv preprint arXiv:1710.11205* .

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