A Convergence Theory for Deep Learning via Over-Parameterization

Zeyuan Allen-Zhu
zeyuan@csail.mit.edu
Microsoft Research AI

Yuanzhi Li
yuanzhil@stanford.edu
Stanford University

Zhao Song
zhaos@utexas.edu
UT-Austin

November 8, 2018*

Abstract

Deep neural networks (DNNs) have demonstrated dominating performance in many fields, e.g., computer vision, natural language progressing, and robotics. Since AlexNet, the neural networks used in practice are going wider and deeper. On the theoretical side, a long line of works have been focusing on why we can train neural networks when there is only one hidden layer. The theory of multi-layer neural networks remains somewhat unsettled.

We present a new theory to understand the convergence of training DNNs. We only make two assumptions: the inputs do not degenerate and the network is over-parameterized. The latter means the number of hidden neurons is sufficiently large: polynomial in n, the number of training samples and in L, the number of layers.

We show on the training dataset, starting from randomly initialized weights, simple algorithms such as stochastic gradient descent attain 100% accuracy in classification tasks, or minimize ℓ_2 regression loss in linear convergence rate, with a number of iterations that only scale polynomial in n and L. Our theory applies to the widely-used but non-smooth ReLU activation, and to any smooth and possibly non-convex loss functions. In terms of network architectures, our theory at least applies to fully-connected neural networks, convolutional neural networks (CNN), and residual neural networks (ResNet).

*This work was done when Yuanzhi Li and Zhao Song were 2018 summer interns at Microsoft Research Redmond. We would like to specially thank Greg Yang for very helpful conversations.

1 Introduction

Neural networks have demonstrated a great success in numerous machine-learning tasks [5, 23, 28, 31, 34, 44, 45]. One of the empirical findings is that neural networks, trained by first-order methods from random initialization, have a remarkable ability of fitting training data [55].

From a capacity perspective, the ability to fit training data may not surprising: modern neural networks are always heavily over-parameterized — they have (much) more parameters than the total number of training samples. Thus, in theory, there always exists parameter choices that achieve zero training error as long as the data does not degenerate.

Yet, from an optimization perspective, the fact that randomly initialized first-order methods can find such an optimal solution on the training data is *quite non-trivial*: neural networks used in practice are often equipped with the ReLU activation function, which makes the training objective not only non-convex, but even non-smooth. Even the general convergence for finding approximate first and second-order critical points of a non-convex, non-smooth function is not fully understood [11], and appears to be a challenging question on its own. This is in direct contrast to practice, in which ReLU networks trained by stochastic gradient descent (SGD) from random initialization *almost never* face the problem of non-smoothness or non-convexity, and can converge to even a *global minimal* over the training set quite easily.

Recently, there are quite a few papers trying to understand the success of neural networks from optimization perspective. Many of them focus on the case when the inputs are random Gaussian, and work only for two-layer neural networks [10, 17, 20, 33, 41, 48, 53, 57, 58].

In Li and Liang [32], it was shown that for a two-layer network with ReLU activation, SGD finds nearly-global optimal (say, 99% classification accuracy) solutions on the training data, as long as the network is *over-parameterized*, meaning that when the number of neurons is polynomially large comparing to the input size. Moreover, if the data is sufficiently structured (say, coming from mixtures of separable distributions), this perfect accuracy can be extended to *test data* as well. As a separate note, over-parameterization is suggested as the possible key to avoid bad local minima by Safran and Shamir [42] even for two-layer neural networks.

There are also results that go beyond two-layer neural networks but with limitations. Some consider deep *linear* neural networks without any activation functions [6, 8, 24, 29]. The result of Daniely [13] applies to multi-layer neural network with ReLU activation, but is about the *convex* training process only with respect to the last layer. Daniely worked in a parameter regime where the weight changes of all layers except the last one make negligible contribution to the final output (and they form the so-called conjugate kernel). The result of Soudry and Carmon [50] shows that under over-parameterization and under random input perturbation, there is bad local minima for multi-layer neural networks. Their work did not show any provable convergence rate.

In this paper, we study the following fundamental question

> *Can DNN be trained close to zero training error efficiently under mild assumptions?*
> *If so, can the convergence rate depend only polynomially in the number of layers?*

Motivation. In 2012 AlexNet [31] was born with 5 convolutional layers. Since then, the common trend in the deep learning community is to build network architectures that go deeper. In 2014, Simonyan and Zisserman [47] proposed a VGG network with 19 layers. Later, Szegedy et al. [52] proposed GoogleNet with 22 layers. In practice, we cannot make the network deeper by naively stacking layers together due to the so-called vanishing / exploding gradient issues. For this reason, in 2015, He et al. [28] proposed an ingenious deep network structure called Deep Residual Network (ResNet), with the capability of handling at least 152 layers. For more overview and variants of ResNet, we refer the readers to [19].

Compared to the practical neural networks that go much deeper, the existing theory has been mostly around two-layer (thus one-hidden-layer) networks even just for the training process alone. It is natural to ask if we can theoretically understand how the training process has worked for multi-layer neural networks.

Remark. In this paper, we do not cover the study of the generalization of neural networks. We refer interested readers to some practical evidence [51, 54] that deeper (and wider) neural networks generalize better.

1.1 Our Result

In this paper, we extend the over-parameterization theory to *multi-layer* neural networks. We show that over-parameterized neural networks can indeed be trained by regular first-order methods to *zero* training error, as as long as the dataset is non-degenerate. We say that the dataset is non-degenerate if the data points are distinct. This is a minimal requirement since a dataset $\{(x_1, y_1), (x_2, y_2)\}$ with the same input $x_1 = x_2$ and different labels $y_1 \neq y_2$ can not be trained to zero error.

For instance, consider the ℓ_2 regression task with an L-layer fully-connected feedforward neural network, each of m neurons equipped with ReLU activation. We show that, as long as $m \geq \mathsf{poly}(n, L, \delta^{-1})$ where n is the number of data points and δ is the minimum (relative) distance between two training data points, for every $\varepsilon > 0$, gradient descent (GD) and stochastic gradient descent (SGD) can find an ε-error solution with linear convergence rate, starting from random Gaussian initialized weights. Using the same network, if the task is multi-label classification, then GD and SGD finds an 100% accuracy classifier on the training set in $\mathsf{poly}(n, L, \delta^{-1})$ iterations. Our result also applies to other Lipschitz-smooth loss functions, and some other network architectures including convolutional neural networks (CNNs) and residual neural networks (ResNet).

Our result also gives the theoretical explanation about an important practical observation: when training an over-parameterized deep neural network with small learning rate, the model could rapidly memorize all the training examples and stop the learning process before a good generalization error is reached.

1.2 Other Related Works

Li and Liang [32] only proved their result for the cross-entropy loss, and the "training accuracy" part of this result was later extended to the ℓ_2 loss [18]. The result of [18] seems to have adopted a learning rate that is m times larger than [32], but that is only because they have re-scaled the network by a factor of $\sqrt{m}$.

Linear networks without activation functions are important subjects on its own. Besides the already cited references [6, 8, 24, 29], there are a number of works that study *linear dynamical systems*, which can be viewed as the linear version of recurrent neural networks or reinforcement learning. Recent works in this line of research include [1, 7, 15, 16, 25–27, 37, 40, 46].

There is sequence of work about one-hidden-layer (multiple neurons) CNN [10, 17, 22, 39, 57]. Whether the patches overlap or not plays a crucial role in analyzing algorithms for such CNN. One category of the results have required the patches to be disjoint [10, 17, 57]. The other category [22, 39] have figured out a weaker assumption or even removed that patch-disjoint assumption. On input data distribution, most relied on inputs being Gaussian [10, 17, 39, 57], and some assumed inputs to be symmetrically distributed with identity covariance and boundedness [22].

As for ResNet, Li and Yuan [33] proved that SGD learns one-hidden-layer residual neural networks under Gaussian input assumption. The techniques in [57, 58] can also be generalized to one-hidden-layer ResNet under the Gaussian input assumption; they can show that GD starting from good initialization point (via tensor initialization) learns ResNet. Hardt and Ma [24] deep *linear* residual networks have no spurious local optima.

If no assumption is allowed, neural networks have been shown hard in several different perspectives. Thirty years ago, Blum and Rivest [9] first proved that learning the neural network is NP-complete. Stronger hardness results have been proved over the last decade [12, 14, 21, 30, 35, 36, 49].

An Over-Parameterized RNN Theory. For experts in DNN theory, one may view this present paper as a deeply-simplified version of the recurrent neural network (RNN) paper [4] by the same set of authors. A recurrent neural network executed on input sequences with time horizon L is *very similar* to a feedforward neural network with L layers. The main difference between the two is that in feedforward neural networks, the weight matrices are different across layers, and thus independently randomly initialized; in contrast, in RNN, the same weight matrix is applied across the entire time horizon so we do not have fresh new randomness for proofs that involve in induction. This makes the over-parameterized convergence theory of DNN *much simpler* than that of RNN. We write this DNN result as a separate paper because: (1) we believe the convergence of DNN is important on its own, (2) the proof in this paper is much simpler (30 vs 80 pages) and could reach out to a wider audience, (3) the simplicity of this paper allows us to tighten many parameters in quite non-trivial ways, and (4) the simplicity of this paper allows us to also study convolutional networks, residual networks, as well as different loss functions (all of them were missing from [4]).

2 Preliminaries

We use $\|v\|$ to denote Euclidean norms of vectors v, and $\|\mathbf{M}\|_2, \|\mathbf{M}\|_F$ to denote spectral and Frobenius norms of matrices $\mathbf{M}$. For a tuple $\overrightarrow{\mathbf{W}} = (\mathbf{W}_1, \dots, \mathbf{W}_L)$ of matrices, we let $\|\overrightarrow{\mathbf{W}}\|_2 = \max_{\ell\in[L]} \|\mathbf{W}_\ell\|_2$ and $\|\overrightarrow{\mathbf{W}}\|_F = (\sum_{\ell=1}^L \|\mathbf{W}_\ell\|_F^2)^{1/2}$.

We use $\phi(x) = \max\{0, x\}$ to denote the ReLU function, and extend it to vectors $v \in \mathbb{R}^m$ by letting $\phi(v) = (\phi(v_1), \dots, \phi(v_m))$. We use $\mathbb{1}_{event}$ to denote the indicator function for *event*.

The training data consist of vector pairs $\{(x_i, y_i^*)\}_{i\in[n]}$, where each $x_i \in \mathbb{R}^{\mathfrak{d}}$ is the feature vector and y_i^* is the label of the i-th training sample. We assume for simplicity that data are normalized: $\|x_i\| = 1$. We make the following separable assumption on the training data (motivated by [32]):

Assumption 2.1. *For every pair $i, j \in [n]$, we have $\|x_i - x_j\| \geq \delta$.*

To present the simplest possible proof, the main body of this paper only focuses on depth-L feedforward fully-connected neural networks with an ℓ_2-regression task. Therefore, each $y_i^* \in \mathbb{R}^d$ is a target vector for the regression task. We explain how to extend it to more general settings in Section 3.3 and the Appendix. For notational simplicity, we assume all the hidden layers have the same number of neurons, and our results trivially generalize to each layer having different number of neurons. Specifically, we focus on the following network

$$
\begin{aligned}
g_{i,0} &= \mathbf{A}x_i & h_{i,0} &= \phi(\mathbf{A}x_i) & &\text{for } i \in [n] \\
g_{i,\ell} &= \mathbf{W}_\ell h_{i,\ell-1} & h_{i,\ell} &= \phi(\mathbf{W}_\ell h_{i,\ell-1}) & &\text{for } i \in [n] \text{ and } \ell \in [L] \\
y_i &= \mathbf{B}h_{i,L} & & & &\text{for } i \in [n]
\end{aligned}
$$

where $\mathbf{A} \in \mathbb{R}^{m\times\mathfrak{d}}$ is the weight matrix for the input layer, $\mathbf{W}_\ell \in \mathbb{R}^{m\times m}$ is the weight matrix for the ℓ-th hidden layer, and $\mathbf{B} \in \mathbb{R}^{d\times m}$ is the weight matrix for the output layer. For notational convenience in the proofs, we may also use $h_{i,-1}$ to denote x_i and $\mathbf{W}_0$ to denote $\mathbf{A}$.

Definition 2.2 (diagonal sign matrix). *For each $i \in [n]$ and $\ell \in \{0, 1, \ldots, L\}$, we denote by $\mathbf{D}_{i,\ell}$ the diagonal sign matrix where $(\mathbf{D}_{i,\ell})_{k,k} = \mathbb{1}_{(\mathbf{W}_\ell h_{i,\ell-1})_k \geq 0}$ for each $k \in [m]$.*

As a result, we have $h_{i,\ell} = \mathbf{D}_{i,\ell}\mathbf{W}_\ell h_{i,\ell-1} = \mathbf{D}_{i,\ell} g_{i,\ell}$ and $(\mathbf{D}_{i,\ell})_{k,k} = \mathbb{1}_{(g_{i,\ell})_k \geq 0}$.
We make the following standard choices of random initialization:

Definition 2.3. *We say that $\overrightarrow{\mathbf{W}} = (\mathbf{W}_1, \ldots, \mathbf{W}_L)$, $\mathbf{A}$ and $\mathbf{B}$ are at random initialization if*

- $[\mathbf{W}_\ell]_{i,j} \sim \mathcal{N}(0, \frac{2}{m})$ *for every $i, j \in [m]$ and $\ell \in [L]$;*
- $\mathbf{A}_{i,j} \sim \mathcal{N}(0, \frac{2}{m})$ *for every $(i,j) \in [m] \times [\mathfrak{d}]$; and*
- $\mathbf{B}_{i,j} \sim \mathcal{N}(0, \frac{1}{d})$ *for every $(i,j) \in [d] \times [m]$.*

Assumption 2.4. *Throughout this paper we assume $m \geq \Omega\big((\delta^{-1} n L \log m)^{30} \cdot d \cdot \log^2 \varepsilon^{-1}\big)$. To present the simplest proof, we did not try hard to improve such polynomial factors.*

2.1 Objective and Gradient

Our regression objective is

$$F(\overrightarrow{\mathbf{W}}) \stackrel{\text{def}}{=} \sum_{i=1}^n F_i(\overrightarrow{\mathbf{W}}) \quad \text{where} \quad F_i(\overrightarrow{\mathbf{W}}) \stackrel{\text{def}}{=} \frac{1}{2}\|\mathbf{B} h_{i,L} - y_i^*\|^2 \quad \text{for each } i \in [n]$$

We also denote by $\mathsf{loss}_i \stackrel{\text{def}}{=} \mathbf{B} h_{i,L} - y_i^*$ the *loss vector* for sample i. For simplicity, we only focus on training $\overrightarrow{\mathbf{W}}$ in this paper and thus leave $\mathbf{A}$ and $\mathbf{B}$ at random initialization. Our techniques can be extended to the case when $\mathbf{A}$, $\mathbf{B}$ and $\overrightarrow{\mathbf{W}}$ are jointly trained.

Definition 2.5. *For each $\ell \in \{1, 2, \cdots, L\}$, we define $\mathsf{Back}_{i,\ell} \stackrel{\text{def}}{=} \mathbf{B}\mathbf{D}_{i,L}\mathbf{W}_L \cdots \mathbf{D}_{i,\ell}\mathbf{W}_\ell \in \mathbb{R}^{d \times m}$ and for $\ell = L+1$, we define $\mathsf{Back}_{i,\ell} = \mathbf{B} \in \mathbb{R}^{d \times m}$.*

Using this notation, one can calculate the gradient of $F(\overrightarrow{\mathbf{W}})$ as follows.

Fact 2.6. *The gradient with respect to the k-th row of $\mathbf{W}_\ell \in \mathbb{R}^{m \times m}$ is*

$$\nabla_{[\mathbf{W}_\ell]_k} F(\overrightarrow{\mathbf{W}}) = \textstyle\sum_{i=1}^n (\mathsf{Back}_{i,\ell+1}^\top \mathsf{loss}_i)_k \cdot h_{i,\ell-1} \cdot \mathbb{1}_{\langle [\mathbf{W}_\ell]_k, h_{i,\ell-1}\rangle \geq 0}$$

The gradient with respect to $\mathbf{W}_\ell$ is

$$\nabla_{\mathbf{W}_\ell} F(\overrightarrow{\mathbf{W}}) = \textstyle\sum_{i=1}^n \mathbf{D}_{i,\ell}(\mathsf{Back}_{i,\ell+1}^\top \mathsf{loss}_i) h_{i,\ell-1}^\top$$

We denote by $\nabla F(\overrightarrow{\mathbf{W}}) = \big(\nabla_{\mathbf{W}_1} F(\overrightarrow{\mathbf{W}}), \ldots, \nabla_{\mathbf{W}_L} F(\overrightarrow{\mathbf{W}})\big)$.

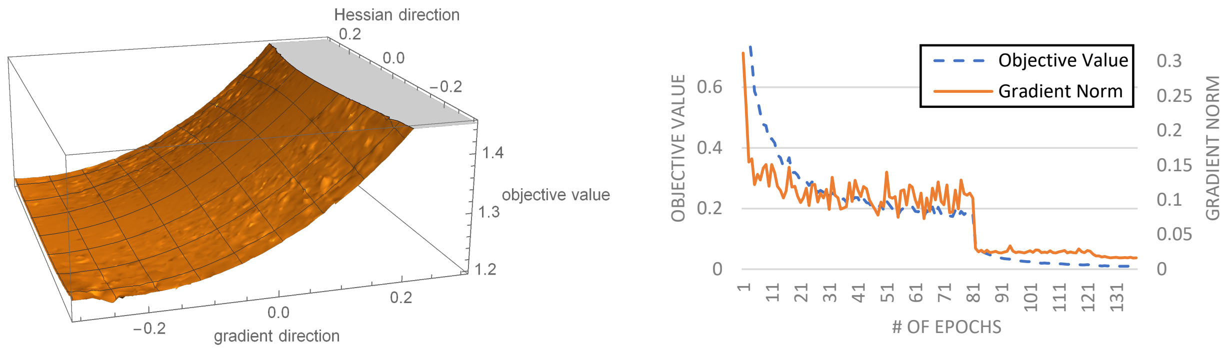
2.2 Probability

Fact 2.7. *Suppose $x \sim \mathcal{N}(0, \sigma^2)$ is a Gaussian random variable. For any $t \in (0, \sigma)$ we have*

$$\mathbf{Pr}[x \geq t] \in \left[\tfrac{1}{2}(1 - \tfrac{4}{5}\tfrac{t}{\sigma}), \tfrac{1}{2}(1 - \tfrac{2}{3}\tfrac{t}{\sigma})\right].$$

Similarly, if $x \sim \mathcal{N}(\mu, \sigma^2)$, for any $t \in (0, \sigma)$, we have

$$\mathbf{Pr}[|x| \geq t] \in \left[1 - \tfrac{4}{5}\tfrac{t}{\sigma}, 1 - \tfrac{2}{3}\tfrac{t}{\sigma}\right].$$



(a) (b)

Figure 1: (a) A typical landscape of the training objective $F(W)$ near a point $W = W_t$ that is on the SGD training trajectory. Here, the x and y axes represent the gradient direction $\nabla F(W_t)$ and the most negatively curved direction (found by Oja's method [2, 3]), and the z axis represents the objective value.
(b) A typical training curve for SGD, where the norm of (full) gradient decreases as objective value decreases. The gradient norm does tend to zero because we are using the cross-entropy loss for multi-label classification (see Section 3.3). The training accuracy already becomes 99.8%.
The used dataset is CIFAR10, and used the neural network is ResNet with 32 layers. Similar landscapes can also be spotted for AlexNet, VGG, DenseNet, etc.

3 Our Results and Techniques

To present our result in the simplest possible way, we choose to mainly focus on fully-connected L-layer neural networks with the ℓ_2 regression loss. We shall extend it to more general settings (such as convolutional and residual networks and other losses) in Section 3.3. Our main results can be stated as follows;

Theorem 1 (gradient descent). *Suppose $m \geq \widetilde{\Omega}\big((nL/\delta)^{30} \cdot d \cdot \log^2 \varepsilon^{-1}\big)$. Starting from random initialization, with probability at least $1-e^{-\Omega(\log^2 m)}$, gradient descent with learning rate $\eta = \Theta\big(\frac{d\delta}{n^4 L^2 m}\big)$ finds a point $F(\overrightarrow{\mathbf{W}}) \leq \varepsilon$ in $T = \Theta\big(\frac{n^6 L^2}{\delta^2} \log \frac{1}{\varepsilon}\big)$ iterations.*

This is known as the linear convergence rate because ε drops exponentially fast in T. We have not tried to improve the polynomial factors in m and T, and are aware of several ways to improve these factors (but at the expense of complicating the proof).

Theorem 2 (stochastic gradient descent). *Suppose $m \geq \widetilde{\Omega}\big(\frac{(nL/\delta)^{30} \cdot d \cdot \log^2 \varepsilon^{-1}}{b}\big)$ and $b \in [n]$. Starting from random initialization, with probability at least $1 - e^{-\Omega(\log^2 m)}$, stochastic gradient descent with learning rate $\eta = \Theta(\frac{b\delta d}{n^5 L^2 m \log^2 m})$ and mini-batch size b finds $F(\overrightarrow{\mathbf{W}}) \leq \varepsilon$ in $T = \Theta\big(\frac{n^7 L^2 \log^2 m}{b\delta^2} \log \frac{1}{\varepsilon}\big)$ iterations.*

This is a nearly-linear convergence rate because $T \propto \log \frac{1}{\varepsilon} \log^2 \log \frac{1}{\varepsilon}$. The reason for the additional $\log^2 \log \frac{1}{\varepsilon}$ factor is because we have a $1 - e^{-\Omega(\log^2 m)}$ high confidence bound.

Remark. For experts in optimization theory, one may immediately question the accuracy of Theorem 2, because SGD is known to converge at a slower rate $T \propto \frac{1}{\mathsf{poly}(\varepsilon)}$ even for convex functions. There is no contradiction here. Imaging a strongly convex function $f(x) = \sum_{i=1}^n f_i(x)$ that has a common minimizer $x^* \in \arg\min_x \{f_i(x)\}$ for every $i \in [n]$, then SGD is known to converge in a linear convergence rate.

3.1 Technical Theorems

The main difficulty of this paper is to prove the following two technical theorems. The first one is about the gradient bounds for points that are sufficiently close to the random initialization:

Theorem 3. *With probability at least* $1 - e^{-\Omega(m/\mathsf{poly}(n,L,\delta^{-1}))}$ *over the randomness of* $\overrightarrow{\mathbf{W}}^{(0)}, \mathbf{A}, \mathbf{B}$*, it satisfies for every* $\ell \in [L]$*, every* $i \in [n]$*, and every* $\overrightarrow{\mathbf{W}}$ *with* $\|\overrightarrow{\mathbf{W}} - \overrightarrow{\mathbf{W}}^{(0)}\|_2 \leq \frac{1}{\mathsf{poly}(n,L,\delta^{-1})}$*,*

$$\|\nabla F(\overrightarrow{\mathbf{W}})\|_F^2 \leq O\Big(F(\overrightarrow{\mathbf{W}}) \times \frac{Lnm}{d}\Big) \quad \text{and} \quad \|\nabla F(\overrightarrow{\mathbf{W}})\|_F^2 \geq \Omega\Big(F(\overrightarrow{\mathbf{W}}) \times \frac{\delta m}{dn^2}\Big) \ .$$

The second one is about a smoothness property that is different but analogous to the classical Lipschitz smoothness [38].

Theorem 4. *With probability at least* $1 - e^{-\Omega(m/\mathsf{poly}(L,\log m))}$ *over the randomness of* $\overrightarrow{\mathbf{W}}^{(0)}, \mathbf{A}, \mathbf{B}$*, we have for every* $\breve{\overrightarrow{\mathbf{W}}} \in (\mathbb{R}^{m\times m})^L$ *with* $\|\breve{\overrightarrow{\mathbf{W}}} - \overrightarrow{\mathbf{W}}^{(0)}\|_2 \leq \frac{1}{\mathsf{poly}(L,\log m)}$*, and for every* $\overrightarrow{\mathbf{W}}' \in (\mathbb{R}^{m\times m})^L$ *with* $\|\overrightarrow{\mathbf{W}}'\|_2 \leq \frac{1}{\mathsf{poly}(L,\log m)}$*,*

$$F(\breve{\overrightarrow{\mathbf{W}}} + \overrightarrow{\mathbf{W}}') \leq F(\breve{\overrightarrow{\mathbf{W}}}) + \langle \nabla F(\breve{\overrightarrow{\mathbf{W}}}), \overrightarrow{\mathbf{W}}' \rangle + \frac{\mathsf{poly}(L)\sqrt{nm\log m}}{\sqrt{d}} \cdot \|\overrightarrow{\mathbf{W}}'\|_2 \big(F(\breve{\overrightarrow{\mathbf{W}}})\big)^{1/2} + O\big(\frac{nL^2m}{d}\big)\|\overrightarrow{\mathbf{W}}'\|_2^2$$

Intuitively, the second property of Theorem 3 says that as long as the gradient is large, the objective value is also large. At the same time, Theorem 4 ensures that the objective is sufficiently smooth, and thus moving in the gradient direction can indeed decrease the objective. The derivation of our main Theorem 1 and 2 from technical Theorem 3 and 3 is quite straightforward, and can be found in Section 9 and 10.

Remark. In our proofs, we show that GD and SGD can converge fast enough and thus the weights stay close to random initialization, by a seemingly small spectral norm bound $\frac{1}{\mathsf{poly}(n,L,\delta^{-1})}$. In fact this bound is large enough to totally change the outputs and fit the training data, because weights are randomly initialized (per entry) at around $\frac{1}{\sqrt{m}}$ for m being large.

In practice, we acknowledge that one often goes beyond this theory-predicted spectral-norm boundary. However, quite interestingly, we still observe Theorem 3 and 4 happen in practice at least for vision tasks. In Figure 1(b), we show the typical landscape near a point $\overrightarrow{\mathbf{W}}$ on the SGD training trajectory. The gradient is sufficiently large and going in its direction can indeed decrease the objective; in contrast, though the objective is non-convex, the negative curvature of its "Hessian" is not significant comparing to gradient. From Figure 1(b) we also see that the objective function is sufficiently smooth (at least in the two interested dimensions that we plot).

3.2 Main Techniques

Our proof to the Theorem 3 and 4 mostly consist of the following steps.

Step 1: properties at random initialization. Let $\overrightarrow{\mathbf{W}} = \overrightarrow{\mathbf{W}}^{(0)}$ be at random initialization and $h_{i,\ell}$ and $\mathbf{D}_{i,\ell}$ be defined with respect to $\overrightarrow{\mathbf{W}}$. We first show that forward propagation neither explode or vanish. That is, $\|h_{i,\ell}\| \approx 1$ for all $i \in [n]$ and $\ell \in [L]$. This is basically because for a fixed y, we have $\|\mathbf{W}y\|^2$ is around 2, and if its signs are sufficiently random, then ReLU activation kills half of the norm, that is $\|\phi(\mathbf{W}y)\| \approx 1$. Then applying induction finishes the proof.

Analyzing forward propagation is not enough. We also need spectral norm bounds on the backward matrix $\|\mathbf{B}\mathbf{D}_{i,L}\mathbf{W}_L \cdots \mathbf{D}_{i,a}\mathbf{W}_a\|_2 \leq O(\sqrt{m/d})$, and on the intermediate matrix $\|\mathbf{D}_{i,a}\mathbf{W}_a \cdots \mathbf{D}_{i,b}\mathbf{W}_b\|_2 \leq O(\sqrt{L})$ for every $a, b \in [L]$. Note that if one naively bounds the spectral norm by induction, then

$\|\mathbf{D}_{i,a}\mathbf{W}_a\|_2 \approx 2$ and it will *exponentially blow up!* Our more careful analysis ensures that even when L layers are stacked together, there is no exponential blow up in L.

The final lemma in this step proves that, as long as $\|x_i - x_j\| \geq \delta$, then for each layer $\ell \in [L]$ it also satisfies $\|h_{i,\ell} - h_{j,\ell}\| \geq \Omega(\delta)$. This can be proved by a careful induction. Details are in Section 4.

Step 2: stability after adversarial perturbation. We show that for every $\overrightarrow{\mathbf{W}}$ that is "close" to initialization, meaning $\|\mathbf{W}_\ell - \mathbf{W}_\ell^{(0)}\|_2 \leq \omega$ for every ℓ and for some $\omega \leq \frac{1}{\mathsf{poly}(L)}$, then the number of sign changes $\|\mathbf{D}_{i,\ell} - \mathbf{D}_{i,\ell}^{(0)}\|_0$ is at most $O(m\omega^{2/3}L)$, and the perturbation amount $\|h_{i,\ell} - h_{i,\ell}^{(0)}\|$ is at most $O(\omega L^{5/2})$. We emphasize here that $\overrightarrow{\mathbf{W}}$ may depend on the randomness of $\mathbf{W}^{(0)}$ so one cannot use union bound. We call this "forward stability", and it is one of the most technical proof of this paper.

Another main result in this step is to show that the backward matrix $\mathbf{B}\mathbf{D}_{i,L}\mathbf{W}_L \cdots \mathbf{D}_{i,a}\mathbf{W}_a$ does not change by more than $O(\omega^{1/3}L^2\sqrt{m/d})$ in spectral norm. (Recall that in the Step 1 we shown that this matrix is of spectral norm $O(\sqrt{m/d})$; thus as long as $\omega^{1/3}L^2 \ll 1$, this change is somewhat negligible. Details are in Section 5.

Step 3: gradient bound. The hard part of Theorem 3 is to show gradient lower bound. For this purpose, recall from Fact 2.6 that each term in the gradient can be written as $\mathbf{D}_{i,\ell}(\mathsf{Back}_{i,\ell+1}^\top \mathsf{loss}_i)h_{i,\ell-1}^\top$ where the backward matrix is applied to a loss vector loss_i. To show that this is large, intuitively, one wishes to show $(\mathsf{Back}_{i,\ell+1}^\top \mathsf{loss}_i)$ and $h_{i,\ell-1}$ are both vectors with large Euclidean norm. However, the main difficulty is that in calculating gradient, different samples $i \in [n]$ may form different gradient matrices and, when summing together, they could in principle each other and possibly even form a zero matrix. To deal with this issue, we use $\|h_{i,\ell} - h_{j,\ell}\| \geq \Omega(\delta)$ from Step 1. In other words, even if the gradient matrix with respect to one sample is fixed, that with respect to other samples still have sufficient randomness so as the final gradient matrix will not be zero. This idea comes from the prior work [32] and helps us prove Theorem 3.[1] Details in Appendix 6 and 7.

Step 4: smoothness. In order to prove Theorem 4, one needs to argue, if we are currently at $\breve{\overrightarrow{\mathbf{W}}}$ and perturb it by $\overrightarrow{\mathbf{W}}'$, then how much does the objective change in second and higher order terms. This is different from our stability theory in Step 2, because Step 2 is regarding having a perturbation on $\overrightarrow{\mathbf{W}}^{(0)}$; in contrast, in Theorem 4 we need a (small) perturbation $\overrightarrow{\mathbf{W}}'$ on top of $\breve{\overrightarrow{\mathbf{W}}}$, which may already be a point perturbed from $\overrightarrow{\mathbf{W}}^{(0)}$. Nevertheless, we still manage to show that, if $\breve{h}_{i,\ell}$ is calculated on $\breve{\overrightarrow{\mathbf{W}}}$ and $h_{i,\ell}$ is calculated on $\breve{\overrightarrow{\mathbf{W}}} + \overrightarrow{\mathbf{W}}'$, then $\|h_{i,\ell} - \breve{h}_{i,\ell}\| \leq O(L^{1.5})\|\mathbf{W}'\|_2$. This is proportional to the small perturbation $\|\mathbf{W}'\|_2$ so, along with other properties to prove, ensures smoothness. This explains Theorem 4 and details are in Section 8.

3.3 Notable Extensions

Our Step 1 through Step 4 in Section 3.2 in fact give rise to a general plan for proving the training convergence of any neural network (at least with respect to the ReLU activation). Thus, it is expected that it can be generalized to many other settings. Not only we can have different number of neurons each layer, our theorems can be extended at least in the following three major directions.[2]

[1] This is the only technical idea that we borrowed from Li and Liang [32], which is the over-parameterization theory for 2-layer neural networks.

[2] In principle, each such proof may require a careful rewriting of the main body of this paper. We choose to sketch only the proof difference in order to keep this paper short. If there is sufficient interest from the readers, we can consider adding the full proofs in the future revision of this paper.

Different loss functions. There is absolutely no need to restrict our attention only to ℓ_2 regression loss. We prove in Appendix A that, for any Lipschitz-smooth loss function f):

- If f is cross-entropy for multi-label classification, then we achieve 100% training accuracy in at most $T = O(n^6 L^2/\delta^2)$ iterations.
- If f is gradient dominant (a.k.a. Polyak-Łojasiewicz) but possibly non-convex, we still have linear convergence.[3]
- If f is convex, then we have convergence rate $T \propto \frac{1}{\varepsilon}$.
- If f is non-convex, then we have convergence rate $T \propto \frac{1}{\varepsilon^2}$ for finding $\|\nabla f\| \le \varepsilon$.[4]

Convolutional neural networks (CNN). There are lots of different ways to design CNN and each of them may require somewhat different proofs. In Appendix B, we study the case when $\mathbf{A}, \mathbf{W}_1, \dots, \mathbf{W}_{L-1}$ are convolutional while $\mathbf{W}_L$ and $\mathbf{B}$ are fully connected. We assume for notational simplicity that each hidden layer has $\mathfrak{d}$ points each with m channels. (In vision tasks, a point is a pixel). In the most general setting, these values $\mathfrak{d}$ and m can vary across layers. Our Theorem 5 says that, as long as m is polynomially large, GD and SGD find an ε-error solution for ℓ_2 regression in $T = \frac{\mathsf{poly}(n,L,\mathfrak{d})}{\delta^2} \log \frac{1}{\varepsilon}$ iterations.

Residual neural networks (ResNet). There are lots of different ways to design ResNet and each of them may require somewhat different proofs. In symbols, between two layers, one may study $h_\ell = \phi(h_{\ell-1} + \mathbf{W} h_{\ell-1})$, $h_\ell = \phi(h_{\ell-1} + \mathbf{W}_2 \phi(\mathbf{W}_1 h_{\ell-1}))$, or even $h_\ell = \phi(h_{\ell-1} + \mathbf{W}_3 \phi(\mathbf{W}_2 \phi(\mathbf{W}_1 h_{\ell-1})))$. Since the main purpose here is to illustrate the generality of our techniques but not to attack each specific setting, in Appendix C, we choose to consider the simplest residual setting $h_\ell = \phi(h_{\ell-1} + \mathbf{W} h_{\ell-1})$ (that was also studied for instance by theoretical work [24]). With appropriately chosen random initialization, our Theorem C shows that one can also have linear convergence rate $T = O\big(\frac{n^6 L^2}{\delta^2} \log \frac{1}{\varepsilon}\big)$ in the over-parameterized setting.

4 Properties at Random Initialization

Throughout this section we assume $\overrightarrow{\mathbf{W}}, \mathbf{A}$ and $\mathbf{B}$ are randomly generated according to Def. 2.3. The diagonal sign matrices $\mathbf{D}_{i,\ell}$ are also determined according to this random initialization.

4.1 Forward Propagation

Lemma 4.1 (forward propagation). *If $\varepsilon \in (0,1]$, with probability at least $1 - nLe^{-\Omega(m\varepsilon^2/L)}$ over the randomness of $\mathbf{A} \in \mathbb{R}^{m\times\mathfrak{d}}$ and $\overrightarrow{\mathbf{W}} \in (\mathbb{R}^{m\times m})^L$, we have*

$$\forall i \in [n], \ell \in \{0,1,\dots,L\} \quad : \quad \|h_{i,\ell}\| \in [1-\varepsilon, 1+\varepsilon] \ .$$

Before proving Lemma 4.1 we note a simple mathematical fact:

Fact 4.2. *Let $h, q \in \mathbb{R}^p$ be fixed vectors, $\mathbf{W} \in \mathbb{R}^{m\times p}$ be random matrix with i.i.d. entries $\mathbf{W}_{i,j} \sim \mathcal{N}(0, \frac{2}{m})$, and vector $v \in \mathbb{R}^m$ defined as $v_i = \phi((\mathbf{W}h)_i) = \mathbb{1}_{(\mathbf{W}(h+q))_i \ge 0} (\mathbf{W}h)_i$. Then, the absolute values of coordinates $|v_i|$ follow from from i.i.d. folded Gaussian distributions $|\mathcal{N}(0, \frac{\|h\|^2}{m})|$. As a result, $\frac{m\|v\|^2}{\|h\|^2}$ is in distribution identical to chi-square distribution χ^2_m.*

[3] Note that the loss function when combined with the neural network together $f(\mathbf{B}h_{i,L})$ is *not* gradient dominant. Therefore, one cannot apply classical theory on gradient dominant functions to derive our same result.

[4] Again, this cannot be derived from classical theory of finding approximate saddle points for non-convex functions, because weights $\overrightarrow{\mathbf{W}}$ with small $\|\nabla f(\mathbf{B}h_{i,L})\|$ is a very different (usually much harder) task comparing to having small gradient with respect to $\overrightarrow{\mathbf{W}}$ for the entire composite function $f(\mathbf{B}h_{i,L})$.

Proof of Fact 4.2. We assume each vector $\mathbf{W}_i$ is generated by first generating a gaussian vector $g \sim \mathcal{N}(0, \frac{2\mathbf{I}}{m})$ and then setting $\mathbf{W}_i = \pm g$ where the sign is chosen with half-half probability. Now, $|\langle \mathbf{W}_i, h\rangle| = |\langle g, h\rangle|$ only depends on g, and is in distribution identical to $|\mathcal{N}(0, \frac{2\|h\|^2}{m})|$. Next, after the sign is determined, the indicator $\mathbb{1}_{\langle \mathbf{W}_i, h+q\rangle \geq 0}$ is 1 with half probability and 0 with another half. Therefore, $|v_i|$ is in distribution identical to $|\mathcal{N}(0, \frac{\|h\|^2}{m})|$. □

Proof of Lemma 4.1. We only prove Lemma 4.1 for a fixed $i \in [n]$ and $\ell \in \{0, 1, 2, \dots, L\}$ because we can apply union bound at the end. Below, we drop the subscript i for notational convenience, and write $h_{i,\ell}$ and x_i as h_ℓ and x respectively.

According to Fact 4.2, fixing any $h_{\ell-1}$ and letting $\mathbf{W}_\ell$ be the only source of randomness, we have $\Delta_\ell \stackrel{\text{def}}{=} \frac{\|h_\ell\|^2}{\|h_{\ell-1}\|^2}$ is distributed according to a χ^2_m random variable divided by m. Therefore, we have

$$\log \|h_{b-1}\|^2 = \log \|x\|^2 + \sum_{\ell=0}^{b-1} \log \Delta_\ell = \sum_{\ell=0}^{b-1} \log \Delta_\ell \ .$$

One can verify that $\mathbb{E}[\log \Delta_\ell] = \log \frac{2}{m} + \psi(\frac{m}{2})$ where $\psi(h) = \frac{\Gamma'(h)}{\Gamma(h)}$ is the digamma function. Using the bound $\log h - \frac{1}{h} \leq \psi(h) \leq \log h - \frac{1}{2h}$ of digamma function, we have

$$-\frac{2}{m} \leq \mathbb{E}[\log \Delta_\ell] \leq -\frac{1}{m}. \tag{4.1}$$

Let $X = \log \Delta_\ell$, the following three case analysis proves that X is a $\frac{m}{8}$-subgaussian random variable.

- Suppose $\lambda < 0$. Recall tail bound of chi-square distribution $\mathbf{Pr}[\Delta_\ell \leq h] \leq (he^{1-h})^{m/2}$ when $h \in [0, 1]$. Using this, we derive

$$\mathbf{Pr}[\log \Delta_\ell \leq \mathbb{E}[\log \Delta_\ell] - \lambda] \leq \mathbf{Pr}[\Delta_\ell \leq e^{-\lambda}] \leq (e^{1-\lambda-e^{-\lambda}})^{m/2} \leq e^{-\lambda^2 m/4}.$$

- Suppose $\lambda \in [0, \frac{4}{m}]$. Using $m \geq 4$, we have

$$\begin{aligned}\mathbf{Pr}[\log \Delta_\ell \geq \mathbb{E}[\log \Delta_\ell] + \lambda] &\leq \mathbf{Pr}[\log \Delta_\ell \geq -2/m] \leq \mathbf{Pr}[\Delta_\ell \geq e^{-2/m}] \leq \mathbf{Pr}[\Delta_\ell \geq 1 - 2/m] \\ &= \frac{\Gamma\left(\frac{m}{2}, \frac{m-2}{2}\right)}{\Gamma\left(\frac{m}{2}\right)} \leq 0.736 \leq e^{-1/m} \leq e^{-\lambda^2 m/16}.\end{aligned}$$

- Suppose $\lambda \geq \frac{4}{m}$. Recall tail bound of chi-square distribution $\mathbf{Pr}[\Delta_\ell \geq h] \leq (he^{1-h})^{m/2}$ when $h \geq 1$. Using this, we derive

$$\begin{aligned}\mathbf{Pr}[\log \Delta_\ell \geq \mathbb{E}[\log \Delta_\ell] + \lambda] \leq \mathbf{Pr}[\log \Delta_\ell \geq -\frac{2}{m} + \lambda] &\leq \mathbf{Pr}[\Delta_\ell \geq e^{\lambda - 2/m}] \\ &\leq (e^{1+\lambda-2/m-e^{\lambda-2/m}})^{m/2} \leq (e^{-(\lambda-2/m)^2/2})^{m/2} \\ &\leq (e^{-\lambda^2/8})^{m/2} = e^{-\lambda^2 m/16}.\end{aligned}$$

Using martingale on subgaussian variables (see for instance [43]), we have for $\varepsilon > 0$,

$$\mathbf{Pr}\left[\left|\sum_{\ell=0}^{b-1} \log \Delta_\ell - \mathbb{E}[\log \Delta_\ell]\right| > \varepsilon\right] \leq e^{-\Omega(\varepsilon^2 m/L)}.$$

Combining this with (4.1), we have $\|h_{b-1}\|^2 \in \left[1-\varepsilon, 1+\varepsilon\right]$ with probability at least $1 - e^{-\Omega(\varepsilon^2 m/L)}$. □

4.2 Intermediate Layers

Lemma 4.3 (intermediate layers)**.** *Suppose $m \geq \Omega(nL\log(nL))$. With probability at least $\geq 1 - e^{-\Omega(m/L)}$ over the randomness of $\overrightarrow{\mathbf{W}} \in (\mathbb{R}^{m\times m})^L$, for all $i \in [n], 1 \leq a \leq b \leq L$,*

(a) $\|\mathbf{W}_b \mathbf{D}_{i,b-1}\mathbf{W}_{b-1}\cdots\mathbf{D}_{i,a}\mathbf{W}_a\|_2 \leq O(\sqrt{L})$.

(b) $\|\mathbf{W}_b \mathbf{D}_{i,b-1}\mathbf{W}_{b-1}\cdots\mathbf{D}_{i,a}\mathbf{W}_a v\| \leq 2\|v\|$ *for all vectors v with* $\|v\|_0 \leq O\big(\frac{m}{L\log m}\big)$.

(c) $\|u^\top \mathbf{W}_b \mathbf{D}_{i,b-1}\mathbf{W}_{b-1}\cdots\mathbf{D}_{i,a}\mathbf{W}_a\| \leq O(1)\|u\|$ *for all vectors u with* $\|u\|_0 \leq O\big(\frac{m}{L\log m}\big)$.

For any integer s with $1 \leq s \leq O\big(\frac{m}{L\log m}\big)$, with probability at least $1 - e^{-\Omega(s\log m)}$ over the randomness of $\overrightarrow{\mathbf{W}} \in (\mathbb{R}^{m\times m})^L$:

(d) $|u^\top \mathbf{W}_b \mathbf{D}_{i,b-1}\mathbf{W}_{b-1}\cdots\mathbf{D}_{i,a}\mathbf{W}_a v| \leq \|u\|\|v\|\cdot O\big(\frac{\sqrt{s\log m}}{\sqrt{m}}\big)$ *for all vectors u, v with* $\|u\|_0, \|v\|_0 \leq s$.

Proof. Again we prove the lemma for fixed i, a and b because we can take a union bound at the end. We drop the subscript i for notational convenience.

(a) Let z_{a-1} be any fixed unit vector, and define $z_\ell = \mathbf{D}_\ell \mathbf{W}_\ell \cdots \mathbf{D}_a \mathbf{W}_a z_{a-1}$. According to Fact 4.2 again, fixing any $z_{\ell-1}$ and letting $\mathbf{W}_\ell$ be the only source of randomness, we have $\Delta_\ell \stackrel{\text{def}}{=} \frac{\|z_\ell\|^2}{\|z_{\ell-1}\|^2}$ is distributed according to a χ^2_m random variable divided by m. Therefore, we have

$$\log\|z_{b-1}\|^2 = \log\|z_{a-1}\|^2 + \sum_{\ell=a}^{b-1}\log\Delta_\ell = \sum_{\ell=a}^{b-1}\log\Delta_\ell \ .$$

Using exactly the same proof as Lemma 4.1, we have

$$\|z_{b-1}\|^2 = \|\mathbf{W}_b\mathbf{D}_{b-1}\mathbf{W}_{b-1}\cdots\mathbf{D}_a\mathbf{W}_a z_{a-1}\|^2 \in \big[1 - 1/3, 1 + 1/3\big]$$

with probability at least $1 - e^{-\Omega(m/L)}$. As a result, if we fix a subset $M \subseteq [m]$ of cardinality $|M| \leq O(m/L)$, taking ε-net, we know that with probability at least $e^{-\Omega(m/L)}$, it satisfies

$$\|\mathbf{W}_b\mathbf{D}_{b-1}\mathbf{W}_{b-1}\cdots\mathbf{D}_a\mathbf{W}_a u\| \leq 2\|u\| \tag{4.2}$$

for all vectors u whose coordinates are zeros outside M. Now, for an arbitrary unit vector $v \in \mathbb{R}^m$, we can decompose it as $v = u_1 + \cdots + u_N$ where $N = O(L)$, each u_j is non-zero only at $O(m/L)$ coordinates, and the vectors $u_1, \ldots, u_N$ are non-zeros on different coordinates. We can apply (4.2) for each each such u_j and triangle inequality. This gives

$$\|\mathbf{W}_b\mathbf{D}_{b-1}\mathbf{W}_{b-1}\cdots\mathbf{D}_a\mathbf{W}_a v\| \leq 2\sum_{j=1}^N\|u_j\| \leq 2\sqrt{N}\Big(\sum_{j=1}^N\|u_j\|^2\Big)^{1/2} \leq O(\sqrt{L})\cdot\|v\|.$$

(b) The proof of Lemma 4.3b is the same as Lemma 4.3a, except to take ε-net over all $O\big(\frac{m}{L\log m}\big)$-sparse vectors u and then applying union bound.

(c) Similar to the proof of Lemma 4.3a, for any fixed vector v, we have that with probability at least $1 - e^{-\Omega(m/L)}$ (over the randomness of $\mathbf{W}_{b-1}, \ldots, \mathbf{W}_1, \mathbf{A}$),

$$\|\mathbf{D}_{b-1}\mathbf{W}_{b-1}\cdots\mathbf{D}_a\mathbf{W}_a v\| \leq 2\|v\|.$$

Conditioning on this event happens, using the randomness of $\mathbf{W}_b$, we have for each fixed vector $u \in \mathbb{R}^m$, we have

$$\Pr_{\mathbf{W}_b}\left[\Big|u^\top\mathbf{W}_b\big(\mathbf{D}_{b-1}\mathbf{W}_{b-1}\cdots\mathbf{D}_a\mathbf{W}_a v\big)\Big| \geq \frac{4}{\sqrt{L}}\|u\|\|v\|\right] \leq e^{-\Omega(m/L)}.$$

Now consider the case that v is a sparse vector that is only non-zero over some fixed index set $M \subseteq [m]$ (with $|M| \le O(m/L)$), and that u is of sparsity $s = O\big(\frac{m}{L \log m}\big)$. Taking ε-net over all such possible vectors u and v, we have with probability at least $1 - e^{-\Omega(m/L)}$, for all vectors $u \in \mathbb{R}^m$ with $\|u\|_0 \le s$ and all vectors $v \in \mathbb{R}^m$ that have non-zeros only in M,

$$\Big| u^\top \mathbf{W}_b \big(\mathbf{D}_{b-1} \mathbf{W}_{b-1} \cdots \mathbf{D}_a \mathbf{W}_a v\big) \Big| \le \frac{8}{\sqrt{L}} \|u\| \|v\| \ . \tag{4.3}$$

Back to the case when v is an arbitrary vector, we can partition $[m]$ into N index sets $[m] = M_1 \cup M_2 \cup \cdots \cup M_N$ and write $v = v_1 + v_2 + \cdots + v_N$, where $N = O(L)$ and each v_j is non-zero only in M_j. By applying (4.3) for N times and using triangle inequality, we have

$$\begin{aligned} \Big| u^\top \mathbf{W}_b \big(\mathbf{D}_{b-1} \mathbf{W}_{b-1} \cdots \mathbf{D}_a \mathbf{W}_a v\big) \Big| &\le \sum_{j=1}^N \Big| u^\top \mathbf{W}_b \big(\mathbf{D}_{b-1} \mathbf{W}_{b-1} \cdots \mathbf{D}_a \mathbf{W}_a v_j\big) \Big| \\ &\le \frac{8}{\sqrt{L}} \|u\| \times \sum_{j=1}^N \|v_j\| \le O(1) \times \|u\| \|v\| \ . \end{aligned}$$

(d) We apply the same proof as Lemma 4.3c with minor changes to the parameters. We can show with probability at least $1 - e^{-\Omega(m/L)}$ (over the randomness of $\mathbf{W}_{b-1}, \dots, \mathbf{W}_1, \mathbf{A}$), for a fixed vector $v \in \mathbb{R}^m$:

$$\|\mathbf{D}_{b-1} \mathbf{W}_{b-1} \cdots \mathbf{D}_a \mathbf{W}_a v\| \le 2 \|v\| \ .$$

Further using the randomness of $\mathbf{W}_b$, we have that conditioning on the above event, fixing any $u \in \mathbb{R}^m$, with probability at least $1 - e^{-\Omega(s \log m)}$ over the randomness of $\mathbf{W}_b$:

$$\big| u \mathbf{W}_b \big(\mathbf{D}_{b-1} \mathbf{W}_{b-1} \cdots \mathbf{D}_a \mathbf{W}_a v\big) \big| \le \big(\frac{s \log m}{m}\big)^{1/2} \times O\big(\|v\| \|u\|\big) \ .$$

Finally, taking ε-net over all possible vectors u, v that are s sparse, we have the desired result.

□

4.3 Backward Propagation

Lemma 4.4 (backward propagation)**.** *Suppose $m \ge \Omega(nL \log(nL))$. If $s \ge \Omega\big(\frac{d}{\log m}\big)$ and $s \le O\big(\frac{m}{L \log m}\big)$, then with probability at least $1 - e^{-\Omega(s \log m)}$, for all $i \in [n]$, $a = 1, 2, \dots, L+1$,*

(a) $|v^\top \mathbf{B} \mathbf{D}_{i,L} \mathbf{W}_L \cdots \mathbf{D}_{i,a} \mathbf{W}_a u| \le O\big(\frac{\sqrt{s \log m}}{\sqrt{d}}\big) \|v\| \|u\|$ for all $v \in \mathbb{R}^d$ and all $u \in \mathbb{R}^m$ with $\|u\|_0 \le s$.

With probability at least $\ge 1 - e^{-\Omega(m/L)}$, for all $i \in [n], 1 \le a \le L$,

(b) $\|v^\top \mathbf{B} \mathbf{D}_{i,L} \mathbf{W}_L \cdots \mathbf{D}_{i,a} \mathbf{W}_a\| \le O(\sqrt{m/d}) \|v\|$ for all vectors $u \in \mathbb{R}^d$ if $d \le O\big(\frac{m}{L \log m}\big)$.

Proof. (a) The proof follows the same idea of Lemma 4.3 (but choosing $b = L$). Given any fixed vector u, we have with probability at least $1 - e^{-\Omega(m/L)}$ (over the randomness of $\mathbf{W}_L, \dots, \mathbf{W}_1, \mathbf{A}$),

$$\|\mathbf{D}_L \mathbf{W}_L \cdots \mathbf{D}_a \mathbf{W}_a u\| \le 2 \|u\| \ .$$

Conditioning on this event happens, using the randomness of $\mathbf{B}$ (recall each entry of $\mathbf{B}$ follows from $\mathcal{N}(0, \frac{1}{d})$), we have for each fixed vector $u \in \mathbb{R}^m$,

$$\Pr_{\mathbf{B}} \left[\Big| v^\top \mathbf{B} \big(\mathbf{D}_L \mathbf{W}_L \cdots \mathbf{D}_a \mathbf{W}_a u\big) \Big| \ge \frac{\sqrt{s \log m}}{\sqrt{d}} \cdot O(\|u\| \|v\|) \right] \le e^{-\Omega(s \log m)} \ .$$

Finally, one can take ε-net over all s-sparse vectors $u \in \mathbb{R}^m$ and all vectors $v \in \mathbb{R}^d$ and apply union bound.

(b) The proof is identical to Lemma 4.3c, except the fact that each entry of $\mathbf{B}$ follows from $\mathcal{N}(0, \frac{1}{d})$ instead of $\mathcal{N}(0, \frac{2}{m})$.

$\square$

4.4 δ-Separateness

Lemma 4.5 (δ-separateness)**.** *Let $m \geq \Omega\big(\frac{L\log(nL)}{\delta^6}\big)$. There exists some constant $C > 1$ so that, if $\delta \leq \frac{1}{CL}$, $\|x_1\| = \cdots = \|x_n\| = 1$ and $\|x_i - x_j\| \geq \delta$ for every pair $i, j \in [n]$, then with probability at least $1 - e^{-\Omega(\delta^6 m/L)}$, we have :*

$$\forall i \neq j \in [n], \quad \forall \ell \in \{0, 1, \dots, L\} \colon \|(\mathbf{I} - \frac{h_{i,\ell}h_{i,\ell}^\top}{\|h_{i,\ell}\|^2})h_{j,\ell}\| \geq \frac{\delta}{2}.$$

Proof of Lemma 4.5. We first apply Lemma 4.1 to show that $\|h_{i,\ell}\| \in [1 - \delta^3/10, 1 + \delta^3/10]$. Next we prove Lemma 4.5 by induction.

In the base case of $\ell = -1$, since $\|x_i - x_j\| \geq \delta$ by our assumption, we already have

$$\|(\mathbf{I} - \frac{h_{i,\ell}h_{i,\ell}^\top}{\|h_{i,\ell}\|^2})h_{j,\ell}\|^2 = \|(\mathbf{I} - \frac{x_i x_i^\top}{\|x_i\|^2})x_j\|^2 = \|x_j - x_i \cdot \langle x_i, x_j\rangle\|^2 = 1 - \big(\langle x_i, x_j\rangle\big)^2 \geq \frac{3}{4}\delta^2 \ .$$

Suppose $h_{i,\ell-1}$ and $h_{j,\ell-1}$ are fixed and satisfies $\|(\mathbf{I} - \frac{h_{i,\ell-1}h_{i,\ell-1}^\top}{\|h_{i,\ell-1}\|^2})h_{j,\ell-1}\|^2 \geq \delta_{\ell-1}^2$ for some $\delta_{\ell-1} \geq \delta/2$. We write $\mathbf{W}_\ell h_{i,\ell-1} = \vec{g}_1$ where $\vec{g}_1 \sim N(0, \frac{2\|h_{i,\ell-1}\|^2}{m}\mathbf{I})$.

Denoting by $\widehat{h} = h_{i,\ell-1}/\|h_{i,\ell-1}\|$, we can write $\mathbf{W}_\ell h_{j,\ell-1} = \mathbf{W}_\ell \widehat{h}\widehat{h}^\top h_{j,\ell-1} + \mathbf{W}_\ell\big(\mathbf{I} - \widehat{h}\widehat{h}^\top\big)h_{j,\ell-1}$ and the randomness of the two terms are independent. In particular, we can write

$$\mathbf{W}_\ell h_{j,\ell-1} = \frac{\langle h_{i,\ell-1}, h_{j,\ell-1}\rangle}{\|h_{i,\ell-1}\|^2} \cdot \vec{g}_1 + \|\big(\mathbf{I} - \widehat{h}\widehat{h}^\top\big)h_{j,\ell-1}\| \cdot \vec{g}_2 \tag{4.4}$$

where $\vec{g}_2 \sim \mathcal{N}(0, \frac{2}{m}\mathbf{I})$ is independent of g_1. Applying Claim 4.6 for each coordinate $k \in [m]$ (and re-scaling by $\frac{m}{\|h_{i,\ell-1}\|^2}$, we have

$$\mathbb{E}[(\phi(\mathbf{W}_\ell h_{i,\ell-1}) - \phi(\mathbf{W}_\ell h_{j,\ell-1}))_k^2] \geq \left(\frac{\delta_{\ell-1}}{\|h_{i,\ell-1}\|}\right)^2 \left(1 - \frac{\delta_{\ell-1}}{\|h_{i,\ell-1}\|}\right) \cdot \frac{\|h_{i,\ell-1}\|^2}{m} \geq \frac{\delta_{\ell-1}^2 \left(1 - O(\delta_{\ell-1})\right)}{m}$$

Applying Chernoff bound (on independent subgaussian random variables), we have with probability at least $1 - e^{-\Omega(\delta_{\ell-1}^4 m)}$,[5]

$$\|h_{i,\ell} - h_{j,\ell}\|^2 = \|\phi(\mathbf{W}_\ell h_{i,\ell-1}) - \phi(\mathbf{W}_\ell h_{j,\ell-1})\|^2 \geq \delta_{\ell-1}^2 \left(1 - O(\delta_{\ell-1})\right) \ .$$

Since $\|h_{i,\ell}\|$ and $\|h_{j,\ell}\|$ are close to 1, we have

$$\begin{aligned}\left\|\left(\mathbf{I} - \frac{h_{i,\ell}h_{i,\ell}^\top}{\|h_{i,\ell}\|^2}\right)h_{j,\ell}\right\|^2 &= \|h_{j,\ell}\|^2 - \frac{\langle h_{i,\ell}, h_{j,\ell}\rangle^2}{\|h_{i,\ell}\|^2} \\ &= \|h_{j,\ell}\|^2 + \frac{\|h_{i,\ell} - h_{j,\ell}\|^2 - \|h_{i,\ell}\|^2 - \|h_{j,\ell}\|^2}{2\|h_{i,\ell}\|^2} \geq \delta_{\ell-1}^2(1 - O(\delta_{\ell-1})) \ .\end{aligned}$$

$\square$

[5] More specifically, we can let $X_k = m\left(\phi(\mathbf{W}_\ell h_{i,\ell-1}) - \phi(\mathbf{W}_\ell h_{j,\ell-1})\right)_k^2$ which is $O(1)$-subgaussian and let $X = X_1 + \cdots + X_m$. We have $\mathbf{Pr}[X \geq \mathbb{E}[X](1 - \delta_{\ell-1})] \geq 1 - e^{-\Omega(\delta_{\ell-1}^2 \mathbb{E}[X])}$.

4.4.1 Auxiliary Claim

The following mathematical fact is needed in the proof of Lemma 4.5. Its proof is by carefully integrating the PDF of Gaussian distribution.

Claim 4.6. *Given $g_1, g_2 \sim \mathcal{N}(0,2)$, constant $\alpha \in \mathbb{R}$ and $\delta \in [0, \frac{1}{6}]$, we have*

$$\mathbb{E}_{g_1,g_2}\left[(\phi(g_1) - \phi(\alpha g_1 + \delta g_2))^2\right] \geq \delta^2(1-\delta) \enspace .$$

Proof of Claim 4.6. We first tackle two easy cases.

Suppose $a < \frac{3}{4}$. If so, then with probability at least 0.3 we have $g_1 > 1$. If this happens, then with probability at least $1/2$ we have $g_2 < 0$. If both happens, we have

$$\phi(g_1) - \phi(\alpha g_1 + \delta g_2) = g_1 - \phi(\alpha g_1 + \delta g_2) \geq g_1 - \alpha g_1 \geq \frac{1}{4} \enspace .$$

Therefore, we have if $a < \frac{3}{4}$ then the expectation is at least 0.03. For similar reason, if $a > \frac{5}{4}$ we also have the expectation is at least 0.03. In the remainder of the proof, we assume $\alpha \in \left[\frac{3}{4}, \frac{5}{4}\right]$.

If $g_1 \geq 0$, we have

$$\begin{aligned}
f(g_1) &\stackrel{\text{def}}{=} \mathbb{E}_{g_2}\left[(\phi(g_1) - \phi(\alpha g_1 + \delta g_2))^2 \,|\, g_1 \geq 0\right] \\
&= \int_0^\infty \frac{(x-g_1)^2 \exp\left(-\frac{(x-\alpha g_1)^2}{4\delta^2}\right)}{\sqrt{4\pi\delta^2}} \mathrm{d}x \\
&= \frac{(\alpha-2)\delta g_1 e^{-\frac{\alpha^2 g_1^2}{4\delta^2}}}{\sqrt{\pi}} + \frac{1}{2}\left((\alpha-1)^2 g_1^2 + 2\delta^2\right)\left(\operatorname{erf}\left(\frac{\alpha g_1}{2\delta}\right) + 1\right).
\end{aligned}$$

If $g_1 < 0$, we have

$$\begin{aligned}
f(g_1) &\stackrel{\text{def}}{=} \mathbb{E}_{g_2}\left[(\phi(g_1) - \phi(\alpha g_1 + \delta g_2))^2 \,|\, g_1 < 0\right] \\
&= \int_0^\infty \frac{x^2 \exp\left(-\frac{(x-\alpha g_1)^2}{4\delta^2}\right)}{\sqrt{4\pi\delta^2}} \mathrm{d}x \\
&= \frac{1}{2}\left(\alpha^2 g_1^2 + 2\delta^2\right)\left(\operatorname{erf}\left(\frac{\alpha g_1}{2\delta}\right) + 1\right) + \frac{\alpha\delta g_1 e^{-\frac{\alpha^2 g_1^2}{4\delta^2}}}{\sqrt{\pi}}.
\end{aligned}$$

Overall, we have

$$\begin{aligned}
&\mathbb{E}_{g_1,g_2}\left[(\phi(g_1) - \phi(\alpha g_1 + \delta g_2))^2\right] \\
&= \int_0^\infty \frac{f(g)\exp\left(-\frac{g^2}{4}\right)}{\sqrt{4\pi}} \mathrm{d}g + \int_{-\infty}^0 \frac{f(g)\exp\left(-\frac{g^2}{4}\right)}{\sqrt{4\pi}} \mathrm{d}g \\
&= \left(\frac{(\alpha-1)^2\alpha\delta}{\pi(\alpha^2+\delta^2)} + \frac{(\alpha-2)\delta^3}{\pi(\alpha^2+\delta^2)} + \frac{1}{2}\left((\alpha-1)^2+\delta^2\right) + \frac{1}{\pi}\left((\alpha-1)^2+\delta^2\right)\arctan\left(\frac{\alpha}{\delta}\right)\right) \\
&\quad + \frac{1}{2\pi}\left(\pi\left(\alpha^2+\delta^2\right) - 2\left(\alpha^2+\delta^2\right)\arctan\left(\frac{\alpha}{\delta}\right) - 2\alpha\delta\right) \\
&= \frac{\delta\left(-2\alpha^2+\alpha-2\delta^2\right)}{\pi\left(\alpha^2+\delta^2\right)} + \frac{(1-2\alpha)\arctan\left(\frac{\alpha}{\delta}\right)}{\pi} + (\alpha-1)\alpha + \delta^2 + \frac{1}{2} \\
&= \left(\alpha^2 - 2\alpha + 1\right) + \delta^2 + \frac{2}{\pi}\sum_{k=1}^{\infty}(-1)^k \frac{(\alpha+k)\delta^{2k+1}}{(2k+1)\alpha^{2k+1}}.
\end{aligned}$$

It is easy to see that, as long as $\delta \le \alpha$, we always have $\frac{(\alpha+k)\delta^{2k+1}}{(2k+1)\alpha^{2k+1}} \ge \frac{(\alpha+k+1)\delta^{2k+3}}{(2k+3)\alpha^{2k+3}}$. Therefore

$$\mathop{\mathbb{E}}_{g_1,g_2}\left[(\phi(g_1)-\phi(\alpha g_1+\delta g_2))^2\right] \ge \left(\alpha^2-2\alpha+1\right)+\delta^2-\frac{2}{\pi}\frac{(\alpha+1)\delta^3}{3\alpha^3} \ge \delta^2(1-\delta) \ . \qquad \square$$

5 Stability against Adversarial Weight Perturbations

Let $\mathbf{A}$, $\mathbf{B}$ and $\overrightarrow{\mathbf{W}}^{(0)} = (\mathbf{W}_1^{(0)},\dots,\mathbf{W}_L^{(0)})$ be matrices at random initialization (see Def. 2.3), and throughout this section, we consider (adversarially) perturbing $\overrightarrow{\mathbf{W}}$ by $\overrightarrow{\mathbf{W}}' = (\mathbf{W}'_1,\dots,\mathbf{W}'_L)$ satisfying $\|\overrightarrow{\mathbf{W}}'\|_2 \le \omega$ (meaning, $\|\mathbf{W}'_\ell\|_2 \le \omega$ for every $\ell \in [L]$). We stick to the following notations in this section

Definition 5.1.

$$\begin{aligned}
&g_{i,0}^{(0)} = \mathbf{A}x_i && g_{i,0} = \mathbf{A}x_i && \textit{for } i\in[n]\\
&h_{i,0}^{(0)} = \phi(\mathbf{A}x_i) && h_{i,0} = \phi(\mathbf{A}x_i) && \textit{for } i\in[n]\\
&g_{i,\ell}^{(0)} = \mathbf{W}_\ell^{(0)}h_{i,\ell-1} && g_{i,\ell} = (\mathbf{W}_\ell^{(0)}+\mathbf{W}'_\ell)h_{i,\ell-1} && \textit{for } i\in[n] \textit{ and } \ell\in[L]\\
&h_{i,\ell}^{(0)} = \phi(\mathbf{W}_\ell^{(0)}h_{i,\ell-1}) && h_{i,\ell} = \phi((\mathbf{W}_\ell^{(0)}+\mathbf{W}'_\ell)h_{i,\ell-1}) && \textit{for } i\in[n] \textit{ and } \ell\in[L]
\end{aligned}$$

Define diagonal matrices $\mathbf{D}_{i,\ell}^{(0)} \in \mathbb{R}^{m\times m}$ *and* $\mathbf{D}_{i,\ell} \in \mathbb{R}^{m\times m}$ *by letting* $(\mathbf{D}_{i,\ell}^{(0)})_{k,k} = \mathbb{1}_{(g_{i,\ell}^{(0)})_k \ge 0}$ *and* $(\mathbf{D}_{i,\ell})_{k,k} = \mathbb{1}_{(g_{i,\ell})_k\ge 0}, \forall k\in[m]$. *Accordingly, we let* $g'_{i,\ell} = g_{i,\ell} - g_{i,\ell}^{(0)}$, $h'_{i,\ell} = h_{i,\ell} - h_{i,\ell}^{(0)}$, *and diagonal matrix* $\mathbf{D}'_{i,\ell} = \mathbf{D}_{i,\ell} - \mathbf{D}_{i,\ell}^{(0)}$.

5.1 Forward Perturbation

Lemma 5.2 (forward perturbation)**.** *Suppose* $\omega \le \frac{1}{CL^{9/2}\log^3 m}$ *for some sufficiently large constant* $C>1$. *With probability at least* $1-e^{-\Omega(m\omega^{2/3}L)}$, *for every* $\overrightarrow{\mathbf{W}}'$ *satisfying* $\|\overrightarrow{\mathbf{W}}'\|_2 \le \omega$,

(a) $g'_{i,\ell}$ *can be written as* $g'_{i,\ell} = g'_{i,\ell,1} + g'_{i,\ell,2}$ *where* $\|g'_{i,\ell,1}\| \le O(\omega L^{3/2})$ *and* $\|g'_{i,\ell,2}\|_\infty \le O\left(\frac{\omega L^{5/2}\sqrt{\log m}}{\sqrt{m}}\right)$

(b) $\|\mathbf{D}'_{i,\ell}\|_0 \le O(m\omega^{2/3}L)$ *and* $\|\mathbf{D}'_{i,\ell}g_{i,\ell}\| \le O(\omega L^{3/2})$.

(c) $\|g'_{i,\ell}\|, \|h'_{i,\ell}\| \le O(\omega L^{5/2}\sqrt{\log m})$.

Proof of Lemma 5.2. In our proof below, we drop the subscript with respect to i for notational simplicity, and one can always take a union bound over all possible indices i at the end.

Using Lemma 4.1, we can first assume that $\|h_\ell^{(0)}\|, \|g_\ell^{(0)}\| \in [\frac{2}{3},\frac{4}{3}]$ for all ℓ. This happens with probability at least $1-e^{-\Omega(m/L)}$. We also assume $\left\|\prod_{b=\ell}^{a+1}\mathbf{W}_b^{(0)}\mathbf{D}_{b-1}^{(0)}\right\|_2 \le c_1\sqrt{L}$ where $c_1>0$ is the hidden constant in Lemma 4.3a.

We shall inductively prove Lemma 5.2. In the base case $\ell=0$, we have $g'_\ell = 0$ so all the statements holds. In the remainder of the proof, we assume that Lemma 5.2 holds for $\ell-1$ and we shall prove the three statements for layer ℓ. We first carefully rewrite:

$$
\begin{aligned}
g'_\ell &= (\mathbf{W}^{(0)}_\ell + \mathbf{W}'_\ell)(\mathbf{D}^{(0)}_{\ell-1} + \mathbf{D}'_{\ell-1})(g^{(0)}_{\ell-1} + g'_{\ell-1}) - \mathbf{W}^{(0)}_\ell \mathbf{D}^{(0)}_{\ell-1} g^{(0)}_{\ell-1} \\
&= \mathbf{W}'_\ell(\mathbf{D}^{(0)}_{\ell-1} + \mathbf{D}'_{\ell-1})(g^{(0)}_{\ell-1} + g'_{\ell-1}) + \mathbf{W}^{(0)}_\ell \mathbf{D}'_{\ell-1}(g^{(0)}_{\ell-1} + g'_{\ell-1}) + \mathbf{W}^{(0)}_\ell \mathbf{D}^{(0)}_{\ell-1} g'_{\ell-1} \\
&= \cdots \\
&= \sum_{a=1}^{\ell} \Big(\prod_{b=\ell}^{a+1} \mathbf{W}^{(0)}_b \mathbf{D}^{(0)}_{b-1} \Big) \Big(\underbrace{\mathbf{W}'_a(\mathbf{D}^{(0)}_{a-1} + \mathbf{D}'_{a-1})(g^{(0)}_{a-1} + g'_{a-1})}_{(\diamondsuit)} + \underbrace{\mathbf{W}^{(0)}_a \mathbf{D}'_{a-1}(g^{(0)}_{a-1} + g'_{a-1})}_{(\heartsuit)} \Big)
\end{aligned}
$$

For each term in $(\diamondsuit)$, we have

$$
\begin{aligned}
&\Big\| \Big(\prod_{b=\ell}^{a+1} \mathbf{W}^{(0)}_b \mathbf{D}^{(0)}_{b-1} \Big) \Big(\mathbf{W}'_a(\mathbf{D}^{(0)}_{a-1} + \mathbf{D}'_{a-1})(g^{(0)}_{a-1} + g'_{a-1}) \Big) \Big\| \\
&\leq \Big\| \prod_{b=\ell}^{a+1} \mathbf{W}^{(0)}_b \mathbf{D}^{(0)}_{b-1} \Big\|_2 \cdot \Big\| \mathbf{W}'_a \Big\|_2 \cdot \Big\| \mathbf{D}^{(0)}_{a-1} + \mathbf{D}'_{a-1} \Big\|_2 \cdot \Big\| g^{(0)}_{a-1} + g'_{a-1} \Big\| \\
&\overset{①}{\leq} c_1 \cdot \omega \cdot 1 \cdot \Big\| g^{(0)}_{a-1} + g'_{a-1} \Big\| \overset{②}{\leq} 2c_1 \sqrt{L}\omega + O\big(\omega^2 L^3 \sqrt{\log m}\big) \ .
\end{aligned}
$$

Above, inequality ① uses Lemma 4.3a and $\|\mathbf{D}^{(0)}_{a-1} + \mathbf{D}'_{a-1}\|_2 = \|\mathbf{D}_{a-1}\|_2 \leq 1$; and inequality ② has used $\|g^{(0)}_\ell\| \leq 2$ and our inductive assumption Lemma 5.2c. By triangle inequality, we have

$$
g'_\ell = \overrightarrow{\mathsf{err}}_1 + \sum_{a=1}^{\ell} \Big(\prod_{b=\ell}^{a+1} \mathbf{W}^{(0)}_b \mathbf{D}^{(0)}_{b-1} \Big) \Big(\underbrace{\mathbf{W}^{(0)}_a \mathbf{D}'_{a-1}(g^{(0)}_{a-1} + g'_{a-1})}_{(\heartsuit)} \Big)
$$

where $\|\overrightarrow{\mathsf{err}}_1\| \leq 2c_1 L^{1.5}\omega + O\big(\omega^2 L^4 \sqrt{\log m}\big)$. We next look at each term in $(\heartsuit)$. For each $a = 2, 3, \ldots, \ell$, we let

$$
x \overset{\text{def}}{=} \mathbf{D}'_{a-1}(g^{(0)}_{a-1} + g'_{a-1}) = \mathbf{D}'_{a-1}(\mathbf{W}^{(0)}_{a-1} h^{(0)}_{a-1} + g'_{a-1}) \ .
$$

If we re-scale x by $\frac{1}{\|h^{(0)}_{a-1}\|}$ (which is a constant in $[0.75, 1.5]$), we can apply Claim 5.3 (with parameter choices in Corollary 5.4) on x and this tells us, with probability at least $1 - e^{-\Omega(m\omega^{2/3}L)}$:

$$
\|x\|_0 \leq O(m\omega^{2/3}L) \quad \text{and} \quad \|x\| \leq O(\omega L^{3/2}). \tag{5.1}
$$

Next, each term in $(\heartsuit)$ contributes to g'_ℓ by

$$
y = \Big(\prod_{b=\ell}^{a+1} \mathbf{W}^{(0)}_b \mathbf{D}^{(0)}_{b-1} \Big) \mathbf{W}^{(0)}_a \Big(\mathbf{D}'_{a-1}(g^{(0)}_{a-1} + g'_{a-1}) \Big)
$$

using (5.1) and Claim 5.5 (with $s = O(m\omega^{2/3}L)$), we have with probability at least $1 - e^{-\Omega(s \log m)}$, one can write $y = y_1 + y_2$ for

$$
\|y_1\| \leq O\big(\omega L^{3/2} \cdot L^{1/2}\omega^{1/3} \log m\big) \quad \text{and} \quad \|y_2\|_\infty \leq O\Big(\omega L^{3/2} \cdot \frac{\sqrt{\log m}}{\sqrt{m}}\Big) \ .
$$

And therefore by triangle inequality we can write

$$
g'_\ell = \overrightarrow{\mathsf{err}}_1 + \overrightarrow{\mathsf{err}}_2 + \overrightarrow{\mathsf{err}}_3
$$

where $\|\overrightarrow{\mathsf{err}}_2\| \leq O\big(L \cdot \omega L^{3/2} \cdot L^{1/2}\omega^{1/3} \log m\big) = O\big(\omega^{4/3} L^3 \log m\big)$ and $\|\overrightarrow{\mathsf{err}}_3\|_\infty \leq O\Big(L \cdot \omega L^{3/2} \cdot \frac{\sqrt{\log m}}{\sqrt{m}}\Big)$.

Together with the upper bound on $\overrightarrow{\mathsf{err}}_1$, we have

$$\|\overrightarrow{\mathsf{err}}_1 + \overrightarrow{\mathsf{err}}_2\| \le 2c_1 L^{1.5}\omega + O\big(\omega^2 L^4 \sqrt{\log m} + \omega^{4/3} L^3 \log m\big) \enspace .$$

Therefore, when ω is sufficiently small, the above term is at most $4c_1 L^{1.5}\omega$. This finishes the proof of Lemma 5.2a for layer ℓ. Finally,

- Lemma 5.2b is due to (5.1),
- g'_ℓ part of Lemma 5.2c is a simple corollary of Lemma 5.2a, and
- h'_ℓ part of Lemma 5.2c is due to $h'_\ell = \mathbf{D}_\ell g'_\ell + \mathbf{D}'_\ell g_\ell$ together with the bound on $\|g'_\ell\|$ and the bound on $\mathbf{D}'_\ell g_\ell$ from Lemma 5.2b.

□

5.1.1 Auxiliary Claim

Claim 5.3. *Suppose $\mathbf{W}^{(0)} \in \mathbb{R}^{m\times m}$ is a random matrix with entries drawn i.i.d. from $\mathcal{N}\big(0, \frac{2}{m}\big)$, and suppose $\omega L^{3/2} \le O(1)$. With probability at least $1 - e^{-\Omega(m\omega^{2/3}L)}$, the following holds.*

For all unit vector $h^{(0)} \in \mathbb{R}^m$, and for all $g' \in \mathbb{R}^m$ that can be written as

$$g' = g'_1 + g'_2 \text{ where } \|g'_1\| \le O(1) \text{ and } \|g'_2\|_\infty \le \frac{1}{4\sqrt{m}} .$$

Let $\mathbf{D}' \in \mathbb{R}^{m\times m}$ be the diagonal matrix where $(\mathbf{D}')_{k,k} = \mathbb{1}_{(\mathbf{W}^{(0)}h^{(0)}+g')_k \ge 0} - \mathbb{1}_{(\mathbf{W}^{(0)}h^{(0)})_k \ge 0}, \forall k \in [m]$. Then, letting $x = \mathbf{D}'(\mathbf{W}^{(0)}h^{(0)} + g') \in \mathbb{R}^m$, we have

$$\|x\|_0 \le \|\mathbf{D}'\|_0 \le O(m\|g'_1\|^{2/3} + \|g'_2\|_\infty m^{3/2}) \quad \text{and} \quad \|x\| \le O(\|g'_1\| + \|g'_2\|_\infty^{3/2} m^{3/4}) \enspace .$$

Corollary 5.4. *In particular, if $\|g'_1\| \le O(\omega L^{3/2})$ and $\|g'_2\|_\infty \le O\big(\frac{\omega^{2/3}L}{m^{1/2}}\big)$, then*

$$\|x\|_0 \le O(m\omega^{2/3}L) \quad \text{and} \quad \|x\| \le O(\omega L^{3/2}) \enspace .$$

Proof of Claim 5.3. We first observe $g^{(0)} = \mathbf{W}^{(0)}h^{(0)}$ follows from $\mathcal{N}\big(0, \frac{2\mathbf{I}}{m}\big)$ regardless of the choice of $h^{(0)}$. Therefore, in the remainder of the proof, we just focus on the randomness of $g^{(0)}$.

We also observe that $(\mathbf{D}')_{j,j}$ is non-zero for some diagonal $j \in [m]$ only if

$$|(g'_1 + g'_2)_j| > |(g^{(0)})_j| \enspace . \tag{5.2}$$

Let $\xi \le \frac{1}{2\sqrt{m}}$ be a parameter to be chosen later. We shall make sure that $\|g'_2\|_\infty \le \xi/2$.

- We denote by $S_1 \subseteq [m]$ the index sets where j satisfies $|(g^{(0)})_j| \le \xi$. Since we know $(g^{(0)})_j \sim \mathcal{N}(0, 2/m)$, we have $\mathbf{Pr}[|(g^{(0)})_j| \le \xi] \le O\left(\xi\sqrt{m}\right)$ for each $j \in [m]$. Using Chernoff bound for all $j \in [m]$, we have with probability at least $1 - e^{-\Omega(m^{3/2}\xi)}$,

  $$|S_1| = \Big|\Big\{i \in [m] \colon |(g^{(0)})_j| \le \xi\Big\}\Big| \le O(\xi m^{3/2}) \enspace .$$

  Now, for each $j \in S_1$ such that $x_j \neq 0$, we must have $|x_j| = |(g^{(0)} + g'_1 + g'_2)_j| \le |(g'_1)_j| + 2\xi$ so we can calculate the ℓ_2 norm of x on S_1:

  $$\sum_{i\in S_1} x_j^2 \le O(\|g'_1\|^2 + \xi^2|S_1|) \le O(\|g'_1\|^2 + \xi^3 m^{3/2}) \enspace .$$

- We denote by $S_2 \subseteq [m] \setminus S_1$ the index set of all $j \in [m] \setminus S_1$ where $x_j \neq 0$. Using (5.2), we have for each $j \in S_2$:

  $$|(g'_1)_j| \ge |(g^{(0)})_j| - |(g'_2)_j| \ge \xi - \|g'_2\|_\infty \ge \xi/2 \enspace .$$

This means

$$|S_2| \le \frac{4\|g_1'\|^2}{\xi^2} \enspace .$$

Now, for each $j \in S_2$ where $x_j \ne 0$, we know that the signs of $(g^{(0)} + g_1' + g_2')_j$ and $(g^{(0)})_j$ are opposite. Therefore, we must have

$$|x_j| = |(g^{(0)} + g_1' + g_2')_j| \le |(g_1' + g_2')_j| \le |(g_1')_j| + \xi/2 \le 2|(g_1')_j|$$

and therefore

$$\sum_{j \in S_2} x_j^2 \le 4 \sum_{j \in S_2} (g_1')_j^2 \le 4\|g_1'\|^2 \enspace .$$

From above, we have $\|x\|_0 \le |S_1| + |S_2| \le O\big(\xi m^{3/2} + \frac{\|g_1'\|^2}{\xi^2}\big)$ and $\|x\|^2 \le O\big(\|g_1'\|^2 + \xi^3 m^{3/2}\big)$. Choosing $\xi = \max\{2\|g_2'\|_\infty, \Theta(\frac{\|g_1'\|^{2/3}}{m^{1/2}})\}$ for the former, and choosing $\xi = 2\|g_2'\|_\infty$ for the latter, we have the desired result. □

Claim 5.5. *For any $2 \le a \le b \le L$ and any positive integer $s \le O\big(\frac{m}{L \log m}\big)$, with probability at least $1 - e^{-\Omega(s \log m)}$, for all $x \in \mathbb{R}^m$ with $\|x\| \le 1$ and $\|x\|_0 \le s$, letting $y = \mathbf{W}_b^{(0)} \mathbf{D}_{b-1}^{(0)} \mathbf{W}_{b-1}^{(0)} \cdots \mathbf{D}_a^{(0)} \mathbf{W}_a^{(0)} x$, we can write $y = y_1 + y_2$ with*

$$\|y_1\| \le O\big(\sqrt{s/m} \log m\big) \quad \text{and} \quad \|y_2\|_\infty \le \frac{2\sqrt{\log m}}{\sqrt{m}} \enspace .$$

Proof of Claim 5.5. First of all, fix any x, we can let $u = \mathbf{D}_{b-1}^{(0)} \mathbf{W}_{b-1}^{(0)} \cdots \mathbf{D}_a^{(0)} \mathbf{W}_a^{(0)} x$ and the same proof of Lemma 4.3 implies that with probability at least $1 - e^{-\Omega(m/L)}$ we have $\|u\| \le O(\|x\|)$. We next condition on this event happens.

Let $\beta = \sqrt{\log m}/\sqrt{m}$. If u is fixed and using only the randomness of $\mathbf{W}_b$, we have $y_i \sim \mathcal{N}\big(0, \frac{2\|u\|^2}{m}\big)$ so for every $p \ge 1$, by Gaussian tail bound

$$\mathbf{Pr}[|y_i| \ge \beta p] \le e^{-\Omega(\beta^2 p^2 m / \|x\|^2)} \le e^{-\Omega(\beta^2 p^2 m)} \enspace .$$

As long as $\beta^2 p^2 m \ge \beta^2 m \ge \Omega(\log m)$, we know that if $|y_i| \ge \beta p$ occurs for q/p^2 indices i out of $[m]$, this cannot happen with probability more than

$$\binom{m}{q/p^2} \times \Big(e^{-\Omega(\beta^2 p^2 m)}\Big)^{q/p^2} \le e^{\frac{q}{p^2}\big(O(\log m) - \Omega(\beta^2 p^2 m)\big)} \le e^{-\Omega(\beta^2 q m)} \enspace .$$

In other words,

$$\mathbf{Pr}\left[|\{i \in [m] \colon |y_i| \ge \beta p\}| > q/p^2\right] \le e^{-\Omega(\beta^2 q m)} \enspace .$$

Finally, by applying union bound over $p = 1, 2, 4, 8, 16, \dots$ we have with probability $\ge 1 - e^{-\Omega(\beta^2 q m)} \cdot \log q$,

$$\sum_{i \colon |y_i| \ge \beta} y_i^2 \le \sum_{k=0}^{\lceil \log q \rceil} (2^{k+1}\beta)^2 \left|\left\{i \in [m] \colon |y_i| \ge 2^k \beta\right\}\right| \le \sum_{k=0}^{\lceil \log q \rceil} (2^{k+1}\beta)^2 \cdot \frac{q}{2^{2k}} \le O(q\beta^2 \log q) \tag{5.3}$$

In other words, vector y can be written as $y = y_1 + y_2$ where $\|y_2\|_\infty \le \beta$ and $\|y_1\|^2 \le O(q\beta^2 \log q)$.

Finally, we want to take ε-net over all s-sparse inputs x. This requires $\beta^2 q m \ge \Omega(s \log m)$, so we can choose $q = \Theta\big(\frac{s \log m}{m \beta^2}\big) = \Theta(s)$. □

5.2 Intermediate Layers

Lemma 5.6 (intermediate perturbation). *For any integer s with $1 \leq s \leq O\big(\frac{m}{L^3 \log m}\big)$, with probability at least $1 - e^{-\Omega(s \log m)}$ over the randomness of $\overrightarrow{\mathbf{W}}^{(0)}, \mathbf{A}$,*

- *for every $i \in [n], 1 \leq a \leq b \leq L$,*
- *for every diagonal matrices $\mathbf{D}''_{i,0}, \dots, \mathbf{D}''_{i,L} \in [-3,3]^{m \times m}$ with at most s non-zero entries.*
- *for every perturbation matrices $\mathbf{W}'_1, \dots, \mathbf{W}'_L \in \mathbb{R}^{m \times m}$ with $\|\overrightarrow{\mathbf{W}}'\|_2 \leq \omega \in [0,1]$.*

we have

(a) $\|\mathbf{W}_b^{(0)}(\mathbf{D}_{i,b-1}^{(0)} + \mathbf{D}''_{i,b-1}) \cdots (\mathbf{D}_{i,a}^{(0)} + \mathbf{D}''_{i,a})\mathbf{W}_a^{(0)}\|_2 \leq O(\sqrt{L})$.

(b) $\|(\mathbf{W}_b^{(0)} + \mathbf{W}'_b)(\mathbf{D}_{i,b-1}^{(0)} + \mathbf{D}''_{i,b-1}) \cdots (\mathbf{D}_{i,a}^{(0)} + \mathbf{D}''_{i,a})(\mathbf{W}_a^{(0)} + \mathbf{W}'_a)\|_2 \leq O(\sqrt{L})$ *if* $\omega \leq O(\frac{1}{L^{1.5}})$.

Proof. For notational simplicity we ignore subscripts in i in the proofs.

(a) Note that each $\mathbf{D}''_\ell$ can be written as $\mathbf{D}''_\ell = \mathbf{D}_\ell^{0/1}\mathbf{D}''_\ell\mathbf{D}_\ell^{0/1}$, where each $\mathbf{D}_\ell^{0/1}$ is a diagonal matrix satisfying

$$(\mathbf{D}_\ell^{0/1})_{k,k} = \begin{cases} 1, & (\mathbf{D}''_\ell)_{k,k} \neq 0; \\ 0, & (\mathbf{D}''_\ell)_{k,k} = 0. \end{cases} \quad \text{and} \quad \|\mathbf{D}_\ell^{0/1}\|_0 \leq s \ .$$

In order to bound the spectral norm of $\mathbf{W}_b^{(0)}(\mathbf{D}_{b-1}^{(0)} + \mathbf{D}''_{b-1})\mathbf{W}_{b-1}^{(0)} \cdots (\mathbf{D}_a^{(0)} + \mathbf{D}''_a)\mathbf{W}_a^{(0)}$, by triangle inequality, we can expend it into 2^{b-a} matrices and bound their spectral norms individually. Each such matrix can be written as (ignoring the subscripts)

$$(\mathbf{W}^{(0)}\mathbf{D}^{(0)} \cdots \mathbf{W}^{(0)}\mathbf{D}^{0/1})\mathbf{D}''(\mathbf{D}^{0/1}\mathbf{W}^{(0)}\mathbf{D}^{(0)} \cdots \mathbf{W}^{(0)}\mathbf{D}^{0/1})\mathbf{D}'' \cdots \mathbf{D}''(\mathbf{D}^{0/1}\mathbf{W}^{(0)}\mathbf{D}^{(0)} \cdots \mathbf{W}^{(0)}) \tag{5.4}$$

Therefore, it suffices for us to bound the spectral norm of the following four types of matrices:

- $\mathbf{W}^{(0)}\mathbf{D}^{(0)} \cdots \mathbf{W}^{(0)}\mathbf{D}^{0/1}$, such matrix has spectral norm at most 2 owing to Lemma 4.3b;
- $\mathbf{D}^{0/1}\mathbf{W}^{(0)}\mathbf{D}^{(0)} \cdots \mathbf{W}^{(0)}$, such matrix has spectral norm at most $O(1)$ owing to Lemma 4.3c;
- $\mathbf{D}^{0/1}\mathbf{W}^{(0)}\mathbf{D}^{(0)} \cdots \mathbf{W}^{(0)}\mathbf{D}^{0/1}$, such matrix has spectral norm at most $\frac{1}{100L^{1.5}}$ owing to Lemma 4.3d and our choice $s \leq O(\frac{m}{L^3 \log m})$;
- $\mathbf{D}''$, such matrix has spectral norm at most 3.

Together, we have

$$\begin{aligned} &\left\|\mathbf{W}_b^{(0)}(\mathbf{D}_{b-1}^{(0)} + \mathbf{D}''_{b-1})\mathbf{W}_{b-1}^{(0)} \cdots (\mathbf{D}_a^{(0)} + \mathbf{D}''_a)\mathbf{W}_a^{(0)}\right\| \\ &\leq O(\sqrt{L}) + \sum_{j=1}^{b-a} \binom{b-a}{j} \cdot O(1) \cdot \left(\frac{1}{100L^{1.5}}\right)^{j-1} \cdot 3^j \cdot O(1) \leq O(\sqrt{L}) \ . \end{aligned}$$

(b) In order to bound the spectral norm of $(\mathbf{W}_b^{(0)} + \mathbf{W}'_b)(\mathbf{D}_{b-1}^{(0)} + \mathbf{D}''_{b-1}) \cdots (\mathbf{D}_a^{(0)} + \mathbf{D}''_a)(\mathbf{W}_a^{(0)} + \mathbf{W}'_a)$, by triangle inequality, we can expend it into 2^{b-a+1} matrices in terms of $\mathbf{W}'$ and bound their spectral norms individually. Each such matrix can be written as (ignoring the subscripts, and denoting $\breve{\mathbf{D}} = \mathbf{D}^{(0)} + \mathbf{D}'$)

$$(\mathbf{W}^{(0)}\breve{\mathbf{D}} \cdots \mathbf{W}^{(0)}\breve{\mathbf{D}})\mathbf{W}'(\breve{\mathbf{D}}\mathbf{W}^{(0)} \cdots \mathbf{W}^{(0)}\breve{\mathbf{D}}) \cdots \mathbf{W}'(\breve{\mathbf{D}}\mathbf{W}^{(0)} \cdots \breve{\mathbf{D}}\mathbf{W}^{(0)})$$

Moreover, from Lemma 5.6a, we know the following three types of matrices

- $\mathbf{W}^{(0)}\breve{\mathbf{D}} \cdots \mathbf{W}^{(0)}\breve{\mathbf{D}}$,

- $\breve{\mathbf{D}}\mathbf{W}^{(0)}\cdots\mathbf{W}^{(0)}\breve{\mathbf{D}}$, and
- $\breve{\mathbf{D}}\mathbf{W}^{(0)}\cdots\breve{\mathbf{D}}\mathbf{W}^{(0)}$

all have spectral norm at most $O(\sqrt{L})$. Together, using $\|\mathbf{W}'_\ell\|_2 \le O(\frac{1}{L^{1.5}})$, we have

$$\begin{aligned}&\left\|(\mathbf{W}_b^{(0)}+\mathbf{W}'_b)(\mathbf{D}_{b-1}^{(0)}+\mathbf{D}''_{b-1})(\mathbf{W}_{b-1}^{(0)}+\mathbf{W}'_{b-1})\cdots(\mathbf{D}_a^{(0)}+\mathbf{D}''_a)(\mathbf{W}_a^{(0)}+\mathbf{W}'_a)\right\|\\ \le& \sum_{j=0}^{b-a+1}\binom{b-a+1}{j}\cdot\Big(O(\sqrt{L})\Big)^{j+1}\cdot\Big(O(\frac{1}{L^{1.5}})\Big)^j \le O(\sqrt{L})\ .\end{aligned}$$

□

5.3 Backward

Lemma 5.7 (backward perturbation)**.** *For any integer $s \in \big[\Omega(\frac{d}{\log m}), O(\frac{m}{L^3\log m})\big]$, for $d \le O(\frac{m}{L\log m})$, with probability at least $1-e^{-\Omega(s\log m)}$ over the randomness of $\overrightarrow{\mathbf{W}}^{(0)}, \mathbf{A}, \mathbf{B}$,*

- *for all $i\in[n]$, $a=1,2,\dots,L+1$,*
- *for every diagonal matrices $\mathbf{D}''_{i,0},\dots,\mathbf{D}''_{i,L}\in[-3,3]^{m\times m}$ with at most s non-zero entries,*
- *for every perturbation matrices $\mathbf{W}'_{i,1},\dots,\mathbf{W}'_{i,L}\in\mathbb{R}^{m\times m}$ with $\|\overrightarrow{\mathbf{W}}'\|_2\le\omega=O(\frac{1}{L^{1.5}})$,*

it satisfies $\|\mathbf{B}(\mathbf{D}_{i,L}^{(0)}+\mathbf{D}''_{i,L})(\mathbf{W}_L^{(0)}+\mathbf{W}'_L)\cdots(\mathbf{W}_{a+1}^{(0)}+\mathbf{W}'_{a+1})(\mathbf{D}_{i,a}^{(0)}+\mathbf{D}''_{i,a})-\mathbf{B}\mathbf{D}_{i,L}^{(0)}\mathbf{W}_L^{(0)}\cdots\mathbf{W}_{a+1}^{(0)}\mathbf{D}_{i,a}^{(0)}\|_2 \le O(\frac{\sqrt{L^3 s\log m+\omega^2L^3m}}{\sqrt{d}})$. Note that if $s=O(m\omega^{2/3}L)$, this upper bound becomes $O(\frac{\omega^{1/3}L^2\sqrt{m\log m}}{\sqrt{d}})$.

Proof. For notational simplicity we ignore subscripts in i in the proofs.

Ignoring the subscripts for cleanness, we have

$$\begin{aligned}&\big\|\mathbf{B}(\mathbf{D}_{i,L}^{(0)}+\mathbf{D}''_{i,L})(\mathbf{W}_L^{(0)}+\mathbf{W}'_L)\cdots(\mathbf{W}_{a+1}^{(0)}+\mathbf{W}'_{a+1})(\mathbf{D}_{i,a}^{(0)}+\mathbf{D}''_{i,a})-\mathbf{B}\mathbf{D}_{i,L}^{(0)}\mathbf{W}_L^{(0)}\cdots\mathbf{W}_{a+1}^{(0)}\mathbf{D}_{i,a}^{(0)}\big\|_2\\ \le& \sum_{\ell=a}^{L}\underbrace{\big\|\mathbf{B}\mathbf{D}_{i,L}^{(0)}\mathbf{W}_L^{(0)}\cdots\mathbf{W}_{\ell+1}^{(0)}\mathbf{D}_\ell^{0/1}\big\|_2}_{\text{Lemma 4.4a}}\|\mathbf{D}''_\ell\|_2\underbrace{\big\|\mathbf{D}_\ell^{0/1}(\mathbf{W}_\ell^{(0)}+\mathbf{W}'_\ell)\cdots(\mathbf{D}_{i,a}^{(0)}+\mathbf{D}''_{i,a})\big\|_2}_{\text{Lemma 5.6b}}\\ &+\sum_{\ell=a+1}^{L}\underbrace{\big\|\mathbf{B}\mathbf{D}_{i,L}^{(0)}\mathbf{W}_L^{(0)}\cdots\mathbf{W}_{\ell+1}^{(0)}\mathbf{D}_\ell^{(0)}\big\|_2}_{\text{Lemma 4.4b}}\|\mathbf{W}'_\ell\|_2\underbrace{\big\|(\mathbf{D}_{\ell-1}^{(0)}+\mathbf{D}''_{\ell-1})(\mathbf{W}_{\ell-1}^{(0)}+\mathbf{W}'_{\ell-1})\cdots(\mathbf{D}_{i,a}^{(0)}+\mathbf{D}''_{i,a})\big\|_2}_{\text{Lemma 5.6b}}\\ \le& L\cdot O\left(\frac{\sqrt{s\log m}}{\sqrt{d}}\right)\cdot O(\sqrt{L})+L\cdot O(\sqrt{m/d})\cdot\omega\cdot O(\sqrt{L})\end{aligned}$$

□

6 Gradient Bound at Random Initialization

Throughout this section we assume $\overrightarrow{\mathbf{W}}, \mathbf{A}$ and $\mathbf{B}$ are randomly generated according to Def. 2.3. The diagonal sign matrices $\mathbf{D}_{i,\ell}$ are also determined according to this random initialization.

Recall we have defined $\mathsf{Back}_{i,\ell}\stackrel{\text{def}}{=}\mathbf{B}\mathbf{D}_{i,L}\mathbf{W}_L\cdots\mathbf{D}_{i,\ell}\mathbf{W}_\ell\in\mathbb{R}^{d\times m}$. In this section, we introduce the following notion

Definition 6.1. *For any vector tuple $\vec{\mathsf{v}} = (\mathsf{v}_1, \dots, \mathsf{v}_n) \in (\mathbb{R}^d)^n$ (viewed as a fake loss vector), for each $\ell \in [L]$, we define*

$$\widehat{\nabla}^{\vec{\mathsf{v}}}_{[\mathbf{W}_\ell]_k} F(\overrightarrow{\mathbf{W}}) \stackrel{\text{def}}{=} \sum_{i=1}^n (\mathsf{Back}_{i,\ell+1}^\top \mathsf{v}_i)_k \cdot h_{i,\ell-1} \cdot \mathbb{1}_{\langle [\mathbf{W}_\ell]_k, h_{i,\ell-1}\rangle \geq 0}, \forall k \in [m]$$
$$\widehat{\nabla}^{\vec{\mathsf{v}}}_{\mathbf{W}_\ell} F(\overrightarrow{\mathbf{W}}) \stackrel{\text{def}}{=} \sum_{i=1}^n \widehat{\nabla}^{\vec{\mathsf{v}}}_{\mathbf{W}_\ell} F_i(\overrightarrow{\mathbf{W}}) \quad \textit{where} \quad \widehat{\nabla}^{\vec{\mathsf{v}}}_{\mathbf{W}_\ell} F_i(\overrightarrow{\mathbf{W}}) \stackrel{\text{def}}{=} \mathbf{D}_{i,\ell}(\mathsf{Back}_{i,\ell+1}^\top \mathsf{v}_i) h_{i,\ell-1}^\top$$

Remark 6.2. It is an easy exercise to check that, if letting $\vec{\mathsf{v}} = (\mathsf{v}_1, \dots, \mathsf{v}_n)$ where $\mathsf{v}_i = \mathbf{B} h_{i,L} - y_i^*$, then $\widehat{\nabla}^{\vec{\mathsf{v}}}_{[\mathbf{W}_\ell]_k} F(\overrightarrow{\mathbf{W}}) = \nabla_{[\mathbf{W}_\ell]_k} F(\overrightarrow{\mathbf{W}})$ and $\widehat{\nabla}^{\vec{\mathsf{v}}}_{\mathbf{W}_\ell} F_i(\overrightarrow{\mathbf{W}}) = \nabla_{\mathbf{W}_\ell} F_i(\overrightarrow{\mathbf{W}})$.

Our main lemma of this section is the following.

Lemma 6.3 (gradient bound at random initialization)**.** *Fix any $\vec{\mathsf{v}} \in (\mathbb{R}^d)^n$, with probability at least $1 - e^{-\Omega(\delta m/n)}$ over the randomness of $\mathbf{A}, \overrightarrow{\mathbf{W}}, \mathbf{B}$, it satisfies for every $\ell \in [L]$:*

$$\|\widehat{\nabla}^{\vec{\mathsf{v}}}_{\mathbf{W}_\ell} F_i(\overrightarrow{\mathbf{W}})\|_F^2 \leq O\Big(\frac{\|\mathsf{v}_i\|^2}{d} \times m\Big) \qquad \|\widehat{\nabla}^{\vec{\mathsf{v}}}_{\mathbf{W}_\ell} F(\overrightarrow{\mathbf{W}})\|_F^2 \leq O\Big(\frac{\|\vec{\mathsf{v}}\|^2}{d} \times mn\Big)$$
$$\|\widehat{\nabla}^{\vec{\mathsf{v}}}_{\mathbf{W}_L} F(\overrightarrow{\mathbf{W}})\|_F^2 \geq \Omega\Big(\frac{\max_{i\in[n]} \|\mathsf{v}_i\|^2}{dn/\delta} \times m\Big)$$

6.1 Proof of Lemma 6.3: Upper Bound

For each $i \in [n], \ell \in [L]$, we can calculate that

$$\begin{aligned}
\Big\|\widehat{\nabla}^{\vec{\mathsf{v}}}_{\mathbf{W}_\ell} F_i(\overrightarrow{\mathbf{W}})\Big\|_F &= \Big\|\mathbf{D}_{i,\ell}(\mathsf{Back}_{i,\ell+1}^\top \cdot \mathsf{v}_i) \cdot h_{i,\ell-1}^\top\Big\|_F \\
&= \Big\|\mathbf{D}_{i,\ell}(\mathsf{Back}_{i,\ell+1}^\top \cdot \mathsf{v}_i)\Big\|_2 \cdot \|h_{i,\ell-1}\|_2 \\
&\leq \|\mathsf{Back}_{i,\ell+1}\|_2 \cdot \|\mathsf{v}_i\|_2 \cdot \|h_{i,\ell-1}\|_2 \\
&\leq \|\mathbf{B}\mathbf{W}_L \mathbf{D}_{L-1} \cdots \mathbf{D}_{i,\ell+1} \mathbf{W}_{\ell+1}\|_2 \cdot \|\mathsf{v}_i\|_2 \cdot \|h_{i,\ell-1}\|_2 \\
&\stackrel{①}{\leq} O(\sqrt{m/d}) \cdot O(1) \cdot \|\mathsf{v}_i\|_2 \enspace .
\end{aligned}$$

where inequality ① uses Lemma 4.4b and Lemma 4.1 with high probability. Applying triangle inequality with respect to all $\ell \in [L]$, taking square on both sides, and summing up over all $i \in [n]$ finish the proof.

6.2 Proof of Lemma 6.3: Lower Bound

Let $i^* = \arg\max_{i\in[n]}\{\|\mathsf{v}_i\|\}$. Recall

$$\widehat{\nabla}^{\vec{\mathsf{v}}}_{[\mathbf{W}_L]_k} F(\overrightarrow{\mathbf{W}}) = \sum_{i=1}^n \langle \mathbf{B}_k, \mathsf{v}_i \rangle \cdot h_{i,L-1} \cdot \mathbb{1}_{(\mathbf{W}_L h_{i,L-1})_k \geq 0}$$

Let $\widehat{h} \stackrel{\text{def}}{=} \frac{h_{i^*,L-1}}{\|h_{i^*,L-1}\|}$. For analysis purpose, after $\widehat{h}$ is fixed (so after fixing the randomness of $\mathbf{A}, \mathbf{W}_1, \dots, \mathbf{W}_{L-1}$), we redefine $\mathbf{W}_L \widehat{h} = \sqrt{1-\theta^2}\widehat{g}_1 + \theta \widehat{g}_2$ where $\widehat{g}_1$ and $\widehat{g}_2$ are generated independently from $\mathcal{N}(0, \frac{2\mathbf{I}}{m})$. We can do so because the two sides are equal in distribution. In other words, we can set

$$\mathbf{W}_L' \stackrel{\text{def}}{=} \mathbf{W}_L\big(\mathbf{I} - \widehat{h}\widehat{h}^\top\big) - \sqrt{1-\theta^2}\widehat{g}_1 \widehat{h}^\top \text{ and } \mathbf{W}_L'' \stackrel{\text{def}}{=} \theta \widehat{g}_2 \widehat{h}^\top,$$

then we have $\mathbf{W}_L = \mathbf{W}_L' + \mathbf{W}_L''$. In particular, the randomness of $\mathbf{W}_L'$ and $\mathbf{W}_L''$ are *independent.*

In the remainder of the proof, let us choose $\theta \stackrel{\text{def}}{=} \frac{\delta}{5n} \leq \frac{1}{5}$.

We first make two technical claims, and the proof of the first one can be found in Section 6.2.1.

Claim 6.4. *We have* $\mathbf{Pr}_{\mathbf{W}'_L,\mathbf{W}_{L-1},\dots,\mathbf{W}_1,\mathbf{A}}\left[|N_2| \geq \frac{\delta}{40n}m\right] \geq 1 - e^{\Omega(\delta m/n)}$

$$N_2 \stackrel{\text{def}}{=} \left\{k \in [m] \colon \left(\left|(\mathbf{W}'_L h_{i^*,L-1})_k\right| \leq \frac{\delta}{10n\sqrt{m}}\right) \bigwedge \left(\forall i \in [n] \setminus \{i^*\}, \quad \left|\left(\mathbf{W}'_L h_{i,L-1}\right)_k\right| \geq \frac{\delta}{4n\sqrt{m}}\right)\right\}$$

Claim 6.5. *Given set* $N_2 \subset [m]$ *and* $\vec{\mathsf{v}}$*, we have*

$$\mathbf{Pr}_{\mathbf{B}_k}\left[\left|\left\{k \in N_2 \colon \left|\langle \mathbf{B}_k, \mathsf{v}_{i^*}\rangle\right| \geq \frac{\|\mathsf{v}_{i^*}\|}{\sqrt{d}}\right\}\right| \geq \frac{|N_2|}{2}\right] \geq 1 - e^{-\Omega(|N_2|)}$$

Proof of Claim 6.5. Observe that each $\langle \mathbf{B}_k, \mathsf{v}_{i^*}\rangle$ follows from $\mathcal{N}(0, \|\mathsf{v}_{i^*}\|^2/d)$, so with probability at least 0.68 it satisfies $|\langle \mathbf{B}_k, \mathsf{v}_{i^*}\rangle| \geq \frac{\|\mathsf{v}_{i^*}\|}{\sqrt{d}}$. Using Chernoff bound we have the desired claim. $\square$

Combining Claim 6.4 and Claim 6.5, we can obtain a set $N \subseteq [m]$ satisfying

$$N \stackrel{\text{def}}{=} \left\{k \in [m] \colon \left(\left|(\mathbf{W}'_L h_{i^*,L-1})_k\right| \leq \frac{\delta}{10n\sqrt{m}}\right) \bigwedge \left(\forall i \in [n] \setminus \{i^*\}, \quad \left|\left(\mathbf{W}'_L h_{i,L-1}\right)_k\right| \geq \frac{\delta}{4n\sqrt{m}}\right) \bigwedge \left|\langle \mathbf{B}_k, \mathsf{v}_{i^*}\rangle\right| \geq \frac{\|\mathsf{v}_{i^*}\|}{\sqrt{d}}\right\}$$

of cardinality $|N| \geq \frac{\delta}{100n}m$. Let us fix the randomness of $\mathbf{W}'_L$ so that N is fixed. Let k be any index in N. We can write

$$\widehat{\nabla}^{\vec{\mathsf{v}}}_{[\mathbf{W}_L]_k} F(\overrightarrow{\mathbf{W}}) = \sum_{i=1}^n \langle \mathbf{B}_k, \mathsf{v}_i\rangle \cdot h_{i,L-1} \cdot \mathbb{1}_{(\mathbf{W}'_L h_{i,L-1})_k + (\mathbf{W}''_L h_{i,L-1})_k \geq 0}.$$

The only remaining source of randomness comes from $\mathbf{W}''_L = \theta \widehat{g}_2 \widehat{h}^\top$.

Recalling that $\theta = \frac{1}{5n}$ and $\widehat{g}_2 \sim \mathcal{N}(0, \frac{2}{m}\mathbf{I})$, so since $\theta(\widehat{g}_2)_k \sim \mathcal{N}(0, \frac{2\theta^2}{m})$, using numerical values of Gaussian CDF, one can verify that

$$\mathbf{Pr}_{\widehat{g}_2}\left[|\theta(\widehat{g}_2)_k| \in \left(\frac{\delta}{9n\sqrt{m}}, \frac{\delta}{5n\sqrt{m}}\right)\right] \geq 0.2 \ .$$

Let us denote this event of $\widehat{g}_2$ as $\mathfrak{E}_k$. Conditioning on $\mathfrak{E}_k$ happens, recalling $\|h_{i,L-1}\| \in [0.9, 1.1]$ from Lemma 4.1,

- For every $i \in [n] \setminus \{i^*\}$, we have

  $$|(\mathbf{W}''_L h_{i,L-1})_k| = |(\theta \widehat{g}_2 \widehat{h}^\top h_{i,L-1})_k| \leq |(\theta \widehat{g}_2)_k| \cdot \|h_{i,L-1}\| < \frac{\delta}{5n\sqrt{m}} \cdot 1.1 < |(\mathbf{W}'_L h_{i,L-1})_k|$$

  and this means $\mathbb{1}_{(\mathbf{W}_L h_{i,L-1})_k \geq 0} = \mathbb{1}_{(\mathbf{W}'_L h_{i,L-1})_k \geq 0}$.
- For $i = i^*$, we have

  $$|(\mathbf{W}''_L h_{i^*,L-1})_k| = |(\theta \widehat{g}_2 \widehat{h}^\top h_{i^*,L-1})_k| = |(\theta \widehat{g}_2)_k| \cdot \|h_{i^*,L-1}\| > \frac{\delta}{9n\sqrt{m}} \cdot 0.9 > |(\mathbf{W}'_L h_{i^*,L-1})_k|$$

  and this means $\mathbb{1}_{(\mathbf{W}_L h_{i^*,L-1})_k \geq 0} \neq \mathbb{1}_{(\mathbf{W}'_L h_{i^*,L-1})_k \geq 0}$ with probability exactly $\frac{1}{2}$ — this is because, conditioning on event $\mathfrak{E}_k$, the sign of $(\theta \widehat{g}_2)_k$ is ± 1 each with half probability.

Recall that for every $k \in N$,

$$\widehat{\nabla}^{\vec{\mathsf{v}}}_{[\mathbf{W}_L]_k} F(\overrightarrow{\mathbf{W}}) = \underbrace{\langle \mathbf{B}_k, \mathsf{v}_{i^*}\rangle \cdot h_{i^*,L-1} \cdot \mathbb{1}_{(\mathbf{W}_L h_{i^*,L-1})_k \geq 0}}_{\spadesuit} + \sum_{i \in [n] \setminus \{i^*\}} \underbrace{\langle \mathbf{B}_k, \mathsf{v}_i\rangle \cdot h_{i,L-1} \cdot \mathbb{1}_{(\mathbf{W}_L h_{i,L-1})_k \geq 0}}_{\clubsuit}$$

Now, fix the randomness of $\mathbf{A}, \mathbf{B}, \mathbf{W}_1, \ldots, \mathbf{W}_{L-1}, \mathbf{W}'_L$ and let $\widehat{g}_2$ be the only randomness. Conditioning on $\mathfrak{E}_k$, we have that each term in ♣ is fixed (i.e., independent of $\widehat{g}_2$) because $\mathbb{1}_{(\mathbf{W}_L h_{i,L-1})_k \geq 0} = \mathbb{1}_{(\mathbf{W}'_L h_{i,L-1})_k \geq 0}$. In contrast, conditioning on $\mathfrak{E}_k$, the indicator $\mathbb{1}_{(\mathbf{W}_L h_{i^*,L-1})_k \geq 0}$ of the ♠ term may be 1 or 0 each with half probability. This means,

$$\mathop{\mathbf{Pr}}_{(\widehat{g}_2)_k}\left[\|\widehat{\nabla}^{\vec{\mathsf{v}}}_{[\mathbf{W}_L]_k} F(\overrightarrow{\mathbf{W}})\|^2 \geq |\langle \mathbf{B}_k, \mathsf{v}_{i^*}\rangle|^2 \cdot \|h_{i^*,L-1}\|^2 \,\Big|\, k \in N \wedge \mathfrak{E}_k\right] \geq \frac{1}{2} \ .$$

Taking into account the fact that $|\langle \mathbf{B}_k, \mathsf{v}_{i^*}\rangle| \geq \frac{\|\mathsf{v}_{i^*}\|}{\sqrt{d}}$ (by definition of N), the fact that $\|h_{i,L-1}\| \geq 0.9$, and the fact that $\mathbf{Pr}_{(\widehat{g}_2)_k}[\mathfrak{E}] \geq 0.2$, we have

$$\mathop{\mathbf{Pr}}_{(\widehat{g}_2)_k}\left[\|\widehat{\nabla}^{\vec{\mathsf{v}}}_{[\mathbf{W}_L]_k} F(\overrightarrow{\mathbf{W}})\|^2 \geq 0.8\frac{\|\mathsf{v}_{i^*}\|^2}{d} \,\Big|\, k \in N\right] \geq \frac{1}{10} \ .$$

Using the independence of $(\widehat{g}_2)_k$ with respect to different $k \in N$, we can apply Chernoff bound and derive:

$$\mathop{\mathbf{Pr}}_{\widehat{g}_2}\left[\sum_{k\in N} \|\widehat{\nabla}^{\vec{\mathsf{v}}}_{[\mathbf{W}_L]_k} F(\overrightarrow{\mathbf{W}})\|^2 \geq 0.8\frac{\|\mathsf{v}_{i^*}\|^2}{d} \cdot \frac{|N|}{15} \,\Big|\, N\right] \geq 1 - e^{-\Omega(|N|)} \ .$$

Finally, using and $|N| \geq \frac{\delta}{100n} m$, we have

$$\mathbf{Pr}\left[\|\widehat{\nabla}^{\vec{\mathsf{v}}}_{\mathbf{W}_L} F(\overrightarrow{\mathbf{W}})\|_F^2 \geq \frac{\|\mathsf{v}_{i^*}\|^2}{d}\frac{\delta}{2000n} m\right] \geq 1 - e^{-\Omega(\delta m/n)} \ .$$

We finish the upper bound proof of Lemma 6.3. ■

6.2.1 Proof of Claim 6.4

Claim 6.4. *We have* $\mathbf{Pr}_{\mathbf{W}'_L, \mathbf{W}_{L-1}, \ldots, \mathbf{W}_1, \mathbf{A}}\left[|N_2| \geq \frac{\delta}{40n} m\right] \geq 1 - e^{\Omega(\delta m/n)}$

$$N_2 \stackrel{\text{def}}{=} \left\{k \in [m] \colon \left(\left|(\mathbf{W}'_L h_{i^*,L-1})_k\right| \leq \frac{\delta}{10n\sqrt{m}}\right) \bigwedge \left(\forall i \in [n] \setminus \{i^*\}, \quad \left|\left(\mathbf{W}'_L h_{i,L-1}\right)_k\right| \geq \frac{\delta}{4n\sqrt{m}}\right)\right\}$$

Proof of Claim 6.4. Throughout the proof we assume $\mathbf{W}_{L-1}, \ldots, \mathbf{A}$ are good enough so that Lemma 4.1 holds (for $\varepsilon = 0.01$) and we fix their randomness. Define

$$N_1 \stackrel{\text{def}}{=} \left\{k \in [m] \colon \left|(\mathbf{W}'_L h_{i^*,L-1})_k\right| \leq \frac{\delta}{10n\sqrt{m}}\right\}$$

Since $\|h_{i^*,L-1}\|^2 \leq 1.1$ by Lemma 4.1, and since by definition of $\mathbf{W}'_L$ we have $(\mathbf{W}'_L h_{i^*,L-1})_k \sim \mathcal{N}(0, \frac{2(1-\theta^2)\|h_{i^*,L-1}\|^2}{m})$. By standard properties of Gaussian CDF (see Fact 2.7), we know $|(\mathbf{W}'_L h_{i^*,L-1})_k| \leq \frac{\delta}{10n\sqrt{m}}$ with probability at least $\frac{\delta}{25n}$ for each $k \in [m]$. By Chernoff bound,

$$\mathop{\mathbf{Pr}}_{\mathbf{W}'_L}\left[|N_1| \geq \frac{\delta}{30n} m\right] \geq 1 - e^{-\Omega(\delta m/n)}$$

Next, suppose we fix the randomness of $\mathbf{W}'_L \widehat{h}$. Define

$$N_2 \stackrel{\text{def}}{=} \left\{k \in N_1 \colon \forall i \in [n] \setminus \{i^*\}, \quad \left|\left(\mathbf{W}'_L h_{i,L-1}\right)_k\right| \geq \frac{\delta}{4n\sqrt{m}}\right\}$$

For each $k \in N_1$ and $i \in [n] \setminus \{i^*\}$, we can write

$$\mathbf{W}'_L h_{i,L-1} = \mathbf{W}'_L \widehat{h}(\widehat{h}^\top h_{i,L-1}) + \mathbf{W}'_L(\mathbf{I} - \widehat{h}\widehat{h}^\top) h_{i,L-1} \ .$$

Above, the first term on the right hand side is fixed (because we have fixed the randomness of

$\mathbf{W}'_L\widehat{h}$); however, $\mathbf{W}'_L(\mathbf{I}-\widehat{h}\widehat{h}^\top)h_{i,L-1}$ is still fresh new random Gaussian. In symbols,

$$\mathbf{W}'_L h_{i,L-1} \sim \mathcal{N}\left(\mathbf{W}'_L\widehat{h}\widehat{h}^\top h_{i,L-1}, \frac{2\|(\mathbf{I}-\widehat{h}\widehat{h}^\top)h_{i,L-1}\|^2}{m}\mathbf{I}\right) .$$

According to Lemma 4.5, the variance here is at least $\frac{2}{m}\|(\mathbf{I}-\widehat{h}\widehat{h}^\top)h_{i,L-1}\|^2 \geq \frac{\delta^2}{2m}$. Using standard properties of Gaussian CDF (see Fact 2.7), we know $|(\mathbf{W}'_L h_{i,L-1})_k| \geq \frac{\delta}{4n\sqrt{m}}$ with probability at least $1-\frac{1}{8n}$ for each $k\in[m]$. By union bound, for this $k\in[m]$, with probability at least $\frac{7}{8}$ we know $|(\mathbf{W}'_L h_{i,L-1})_k| \geq \frac{\delta}{4n\sqrt{m}}$ for all $i\in[n]\setminus\{i^*\}$. By Chernoff bound (over all $k\in N_1$), we conclude that

$$\mathop{\mathbf{Pr}}_{\mathbf{W}'_L}\left[|N_2| \geq \frac{3}{4}|N_1| \,\Big|\, N_1\right] \geq 1-e^{-\Omega(|N_1|)} = 1-e^{-\Omega(\delta m/n)} .$$

Combining the two bounds we finish the proof. □

7 Gradient Bound at After Perturbation

In this section we prove our main theorem on the gradient upper and lower bounds.

Theorem 3 (gradient bound at after perturbation)**.** *Let* $\omega \stackrel{\text{def}}{=} O\big(\frac{\delta^{3/2}}{n^{9/2}L^6\log^3 m}\big)$*. With probability at least* $1-e^{-\Omega(m\omega^{2/3}L)}$ *over the randomness of* $\overrightarrow{\mathbf{W}}^{(0)},\mathbf{A},\mathbf{B}$*, it satisfies for every* $\ell\in[L]$*, every* $i\in[n]$*, and every* $\overrightarrow{\mathbf{W}}$ *with* $\|\overrightarrow{\mathbf{W}}-\overrightarrow{\mathbf{W}}^{(0)}\|_2\leq\omega$*,*

$$\|\nabla_{\mathbf{W}_\ell}F_i(\overrightarrow{\mathbf{W}})\|_F^2 \leq O\Big(\frac{F_i(\overrightarrow{\mathbf{W}})}{d}\times m\Big) \qquad \|\nabla_{\mathbf{W}_\ell}F(\overrightarrow{\mathbf{W}})\|_F^2 \leq O\Big(\frac{F(\overrightarrow{\mathbf{W}})}{d}\times mn\Big)$$
$$\|\nabla_{\mathbf{W}_L}F(\overrightarrow{\mathbf{W}})\|_F^2 \geq \Omega\Big(\frac{\max_{i\in[n]}F_i(\overrightarrow{\mathbf{W}})}{dn/\delta}\times m\Big) .$$

Remark 7.1. Our Theorem 3 only gives gradient lower bound on $\|\nabla_{\mathbf{W}_L}F(\overrightarrow{\mathbf{W}})\|_F$. In principle, one can derive similar lower bounds on $\|\nabla_{\mathbf{W}_\ell}F(\overrightarrow{\mathbf{W}})\|_F$ for all $\ell=1,2,\dots,L-1$. However, the proof will be significantly more involved. We choose not to derive those bounds at the expense of losing a polynomial factor in L in the final running time. For readers interested in the techniques for obtaining those bounds, we refer to them to the "randomness decomposition" part of our separate paper [4].

Proof of Theorem 3. Again we denote by $\mathbf{D}^{(0)}_{i,\ell}$ and $\mathbf{D}_{i,\ell}$ respectively the sign matrix at the initialization $\overrightarrow{\mathbf{W}}^{(0)}$ and at the current point $\overrightarrow{\mathbf{W}}$; and by $h^{(0)}_{i,\ell}$ and $h_{i,\ell}$ respectively the forward vector at $\overrightarrow{\mathbf{W}}^{(0)}$ and at $\overrightarrow{\mathbf{W}}$. Let us choose $s=O(m\omega^{2/3}L)$ which bounds the sparsity of $\|\mathbf{D}_{i,\ell}-\mathbf{D}^{(0)}_{i,\ell}\|_0$ by Lemma 5.2b. Recall

$$\begin{aligned}&\widehat{\nabla}^{\vec{\mathsf{v}}}_{\mathbf{W}_\ell}F(\overrightarrow{\mathbf{W}}^{(0)}) - \widehat{\nabla}^{\vec{\mathsf{v}}}_{\mathbf{W}_\ell}F(\overrightarrow{\mathbf{W}})\\ &= \sum_{i=1}^n \Big(\big(\mathsf{v}_i^\top\mathbf{B}\mathbf{D}^{(0)}_{i,L}\mathbf{W}^{(0)}_L\cdots\mathbf{W}^{(0)}_{\ell+1}\mathbf{D}^{(0)}_{i,\ell}\big)^\top (h^{(0)}_{i,\ell-1})^\top - \big(\mathsf{v}_i^\top\mathbf{B}\mathbf{D}_{i,L}\mathbf{W}_L\cdots\mathbf{W}_{\ell+1}\mathbf{D}_{i,\ell}\big)^\top (h_{i,\ell-1})^\top\Big)\end{aligned}$$

By Lemma 5.7, we know that

$$\|\mathsf{v}_i^\top\mathbf{B}\mathbf{D}^{(0)}_{i,L}\mathbf{W}^{(0)}_L\cdots\mathbf{D}^{(0)}_{i,a}\mathbf{W}^{(0)}_a\mathbf{D}^{(0)}_{i,a-1} - \mathsf{v}_i^\top\mathbf{B}\mathbf{D}_{i,L}\mathbf{W}_L\cdots\mathbf{D}_{i,a}\mathbf{W}_a\mathbf{D}_{i,a-1}\| \leq O(\omega^{1/3}L^2\sqrt{m\log m}/\sqrt{d})\cdot\|\mathsf{v}_i\|$$

By Lemma 4.4b we know

$$\|\mathsf{v}_i^\top\mathbf{B}\mathbf{D}^{(0)}_{i,L}\mathbf{W}^{(0)}_L\cdots\mathbf{D}^{(0)}_{i,a}\mathbf{W}^{(0)}_a\mathbf{D}^{(0)}_{i,a-1}\| \leq O(\sqrt{m/d})\cdot\|\mathsf{v}_i\|$$

By Lemma 4.1 and Lemma 5.2c, we have

$$\|h_{i,\ell-1}\| \le 1.1 \quad \text{and} \quad \|h_{i,\ell-1} - h^{(0)}_{i,\ell-1}\| \le O(\omega L^{5/2}\sqrt{\log m})$$

Together, they imply

$$\begin{aligned}\left\|\widehat{\nabla}^{\vec{\mathsf{v}}}_{\mathbf{W}_\ell} F(\overrightarrow{\mathbf{W}}^{(0)}) - \widehat{\nabla}^{\vec{\mathsf{v}}}_{\mathbf{W}_\ell} F(\overrightarrow{\mathbf{W}})\right\|_F^2 &\le n\|\vec{\mathsf{v}}\|^2 \cdot O\left(\omega^{1/3}L^2\sqrt{m\log m}/\sqrt{d} + \sqrt{m/d}\times \omega L^{5/2}\sqrt{\log m}\right)^2 \\ &\le n\|\vec{\mathsf{v}}\|^2 \cdot O\left(\frac{m\log m}{d}\cdot \omega^{2/3}L^4\right)\end{aligned}$$

With our parameter assumption on ω, this together with Lemma 6.3 implies the same upper and lower bounds at point $\overrightarrow{\mathbf{W}} = \overrightarrow{\mathbf{W}}^{(0)} + \overrightarrow{\mathbf{W}}'$:

$$\|\widehat{\nabla}^{\vec{\mathsf{v}}}_{\mathbf{W}_\ell} F_i(\overrightarrow{\mathbf{W}}^{(0)} + \overrightarrow{\mathbf{W}}')\|_F^2 \le O\Big(\frac{\|\mathsf{v}_i\|^2}{d}\times m\Big) \quad \|\widehat{\nabla}^{\vec{\mathsf{v}}}_{\mathbf{W}_\ell} F(\overrightarrow{\mathbf{W}}^{(0)} + \overrightarrow{\mathbf{W}}')\|_F^2 \le O\Big(\frac{\|\vec{\mathsf{v}}\|^2}{d}\times mn\Big)$$

$$\|\widehat{\nabla}^{\vec{\mathsf{v}}}_{\mathbf{W}_L} F(\overrightarrow{\mathbf{W}}^{(0)} + \overrightarrow{\mathbf{W}}')\|_F^2 \ge \Omega\Big(\frac{\max_{i\in[n]}\|\mathsf{v}_i\|^2}{dn/\delta}\times m\Big) \ .$$

Finally, taking ε-net over all possible vectors $\vec{\mathsf{v}} = (\mathsf{v}_1,\dots,\mathsf{v}_n) \in (\mathbb{R}^d)^n$, we know that the above bounds hold not only for fixed $\vec{\mathsf{v}}$ but for all $\vec{\mathsf{v}}$. In particular, we can now plug in the choice of $\mathsf{v}_i = \mathsf{loss}_i = \mathbf{B}h_{i,L} - y_i^*$ and it implies our desired bounds on the true gradients. □

8 Objective Smoothness

The purpose of this section is to prove

Theorem 4 (objective smoothness, restated)**.** *Let* $\omega \in \left[\Omega(\frac{d^{3/2}}{m^{3/2}L^{3/2}\log^{3/2} m}), O(\frac{1}{L^{4.5}\log^{3/2} m})\right]$ *and* $\overrightarrow{\mathbf{W}}^{(0)}, \mathbf{A}, \mathbf{B}$ *be at random initialization. With probability at least* $1-e^{-\Omega(m\omega^{2/3}L)}$ *over the randomness of* $\overrightarrow{\mathbf{W}}^{(0)}, \mathbf{A}, \mathbf{B}$, *we have for every* $\breve{\overrightarrow{\mathbf{W}}} \in (\mathbb{R}^{m\times m})^L$ *with* $\|\breve{\overrightarrow{\mathbf{W}}} - \overrightarrow{\mathbf{W}}^{(0)}\|_2 \le \omega$, *and for every* $\overrightarrow{\mathbf{W}}' \in (\mathbb{R}^{m\times m})^L$ *with* $\|\overrightarrow{\mathbf{W}}'\|_2 \le \omega$, *we have*

$$F(\breve{\overrightarrow{\mathbf{W}}} + \overrightarrow{\mathbf{W}}') \le F(\breve{\overrightarrow{\mathbf{W}}}) + \langle \nabla F(\breve{\overrightarrow{\mathbf{W}}}), \overrightarrow{\mathbf{W}}'\rangle + \sqrt{nF(\breve{\overrightarrow{\mathbf{W}}})}\cdot\frac{\omega^{1/3}L^2\sqrt{m\log m}}{\sqrt{d}}\cdot O(\|\overrightarrow{\mathbf{W}}'\|_2) + O\big(\frac{nL^2m}{d}\big)\|\overrightarrow{\mathbf{W}}'\|_2^2$$

We introduce the following notations before we go to proofs.

Definition 8.1. *For* $i \in [n]$ *and* $\ell \in [L]$*:*

$$\begin{aligned} g^{(0)}_{i,0} &= \mathbf{A}x_i & \breve{g}_{i,0} &= \mathbf{A}x_i & g_{i,0} &= \mathbf{A}x_i \\ h^{(0)}_{i,0} &= \phi(\mathbf{A}x_i) & \breve{h}_{i,0} &= \phi(\mathbf{A}x_i) & h_{i,0} &= \phi(\mathbf{A}x_i) \\ g^{(0)}_{i,\ell} &= \mathbf{W}^{(0)}_\ell h^{(0)}_{i,\ell-1} & \breve{g}_{i,\ell} &= \breve{\mathbf{W}}_\ell \breve{h}_{i,\ell-1} & g_{i,\ell} &= (\breve{\mathbf{W}}_\ell + \mathbf{W}'_\ell)h_{i,\ell-1} \\ h^{(0)}_{i,\ell} &= \phi(\mathbf{W}^{(0)}_\ell h^{(0)}_{i,\ell-1}) & \breve{h}_{i,\ell} &= \phi(\breve{\mathbf{W}}_\ell \breve{h}_{i,\ell-1}) & h_{i,\ell} &= \phi((\breve{\mathbf{W}}_\ell + \mathbf{W}'_\ell)h_{i,\ell-1}) \end{aligned}$$

$$\breve{\mathsf{loss}}_i = B\breve{h}_{i,L} - y_i^*$$

Define diagonal matrices $\mathbf{D}^{(0)}_{i,\ell} \in \mathbb{R}^{m\times m}$ *and* $\breve{\mathbf{D}}_{i,\ell} \in \mathbb{R}^{m\times m}$ *respectively by letting*

$$(\mathbf{D}^{(0)}_{i,\ell})_{k,k} = \mathbb{1}_{(g^{(0)}_{i,\ell})_k \ge 0} \text{ and } (\breve{\mathbf{D}}_{i,\ell})_{k,k} = \mathbb{1}_{(\breve{g}_{i,\ell})_k\ge 0}, \forall k\in[m].$$

The following claim gives rise to a new recursive formula to calculate $h_{i,\ell} - \breve{h}_{i,\ell}$.

Claim 8.2. *There exist diagonal matrices $\mathbf{D}''_{i,\ell} \in \mathbb{R}^{m\times m}$ with entries in $[-1,1]$ such that,*

$$\forall i \in [n], \forall \ell \in [L]: \quad h_{i,\ell} - \breve{h}_{i,\ell} = \sum_{a=1}^{\ell} (\breve{\mathbf{D}}_{i,\ell} + \mathbf{D}''_{i,\ell}) \breve{\mathbf{W}}_\ell \cdots \breve{\mathbf{W}}_{a+1} (\breve{\mathbf{D}}_{i,a} + \mathbf{D}''_{i,a}) \mathbf{W}'_a h_{i,a-1} \tag{8.1}$$

Furthermore, we have $\|h_{i,\ell} - \breve{h}_{i,\ell}\| \le O(L^{1.5})\|\mathbf{W}'\|_2$, $\|\mathbf{B}h_{i,\ell} - \mathbf{B}\breve{h}_{i,\ell}\| \le O(L\sqrt{m/d})\|\mathbf{W}'\|_2$ and $\|\mathbf{D}''_{i,\ell}\|_0 \le O(m\omega^{2/3}L)$.

Proof of Theorem 4. First of all, since

$$\frac{1}{2}\|\mathbf{B}h_{i,L} - y_i^*\|^2 = \frac{1}{2}\|\breve{\mathsf{loss}}_i + \mathbf{B}(h_{i,L} - \breve{h}_{i,L})\|^2 = \frac{1}{2}\|\breve{\mathsf{loss}}_i\|^2 + \breve{\mathsf{loss}}_i^\top \mathbf{B}(h_{i,L} - \breve{h}_{i,L}) + \frac{1}{2}\|\mathbf{B}(h_{i,L} - \breve{h}_{i,L})\|^2 \tag{8.2}$$

we can write

$$\begin{aligned}
& F(\overrightarrow{\breve{\mathbf{W}}} + \overrightarrow{\mathbf{W}}') - F(\overrightarrow{\mathbf{W}}) - \langle \nabla F(\overrightarrow{\mathbf{W}}), \overrightarrow{\mathbf{W}}' \rangle \\
\overset{①}{=} & -\langle \nabla F(\overrightarrow{\breve{\mathbf{W}}}), \overrightarrow{\mathbf{W}}' \rangle + \frac{1}{2}\sum_{i=1}^n \|\mathbf{B}h_{i,L} - y_{i,L}^*\|^2 - \|\mathbf{B}\breve{h}_{i,L} - y_{i,L}^*\|^2 \\
\overset{②}{=} & -\langle \nabla F(\overrightarrow{\breve{\mathbf{W}}}), \overrightarrow{\mathbf{W}}' \rangle + \sum_{i=1}^n \breve{\mathsf{loss}}_i^\top \mathbf{B}(h_{i,L} - \breve{h}_{i,L}) + \frac{1}{2}\|\mathbf{B}(h_{i,L} - \breve{h}_{i,L})\|^2 \\
\overset{③}{=} & \sum_{i=1}^n \breve{\mathsf{loss}}_i^\top \mathbf{B} \left((h_{i,L} - \breve{h}_{i,L}) - \sum_{\ell=1}^L \breve{\mathbf{D}}_{i,L}\breve{\mathbf{W}}_L \cdots \breve{\mathbf{W}}_{\ell+1}\breve{\mathbf{D}}_{i,\ell}\mathbf{W}'_\ell \breve{h}_{i,\ell-1} \right) + \frac{1}{2}\|\mathbf{B}(h_{i,L} - \breve{h}_{i,L})\|^2 \\
\overset{④}{=} & \sum_{i=1}^n \breve{\mathsf{loss}}_i^\top \mathbf{B} \left(\sum_{\ell=1}^L (\breve{\mathbf{D}}_{i,L} + \mathbf{D}''_{i,L})\breve{\mathbf{W}}_L \cdots \breve{\mathbf{W}}_{\ell+1}(\breve{\mathbf{D}}_{i,\ell} + \mathbf{D}''_{i,\ell})\mathbf{W}'_\ell h_{i,\ell-1} - \breve{\mathbf{D}}_{i,L}\breve{\mathbf{W}}_L \cdots \breve{\mathbf{W}}_{\ell+1}\breve{\mathbf{D}}_{i,\ell}\mathbf{W}'_\ell \breve{h}_{i,\ell-1} \right) \\
& \quad + \frac{1}{2}\sum_{i=1}^n \|\mathbf{B}(h_{i,L} - \breve{h}_{i,L})\|^2
\end{aligned} \tag{8.3}$$

Above, ① is by the definition of $F(\cdot)$; ② is by (8.2); ③ is by the definition of $\nabla F(\cdot)$ (see Fact 2.6 for an explicit form of the gradient).

We next bound the RHS of (8.3). We first note that by Lemma 5.2b, we have $\|\breve{\mathbf{D}}_{i,\ell} + \mathbf{D}''_{i,\ell} - \mathbf{D}^{(0)}_{i,\ell}\|_0 \le s$ and $\|\breve{\mathbf{D}}_{i,\ell} - \mathbf{D}^{(0)}_{i,\ell}\|_0 \le s$ for $s = O(m\omega^{2/3}L)$.

We ignore subscripts in i for notational convenience. We first use Claim 8.2 to get

$$\|\mathbf{B}(h_L - \breve{h}_L)\| \le O(L\sqrt{m/d}) \cdot \|\overrightarrow{\mathbf{W}}'\|_2 \ . \tag{8.4}$$

Next we calculate that

$$\begin{aligned}
& \left| \breve{\mathsf{loss}}_i^\top \mathbf{B}(\breve{\mathbf{D}}_L + \mathbf{D}''_L)\breve{\mathbf{W}}_L \cdots (\breve{\mathbf{D}}_\ell + \mathbf{D}''_\ell)\mathbf{W}'_\ell h_{\ell-1} - \breve{\mathsf{loss}}_i^\top \mathbf{B}\breve{\mathbf{D}}_L\breve{\mathbf{W}}_L \cdots \breve{\mathbf{D}}_\ell \mathbf{W}'_\ell h_{\ell-1} \right| \\
\le & \|\breve{\mathsf{loss}}_i\| \cdot \underbrace{\left\| \mathbf{B}(\breve{\mathbf{D}}_L + \mathbf{D}''_L)\breve{\mathbf{W}}_L \cdots \breve{\mathbf{W}}_{\ell-1}(\breve{\mathbf{D}}_\ell + \mathbf{D}''_\ell) - \mathbf{B}\breve{\mathbf{D}}_L\breve{\mathbf{W}}_L \cdots \breve{\mathbf{W}}_{\ell-1}\breve{\mathbf{D}}_\ell \right\|_2}_{\text{Lemma 5.7 with } s = O(m\omega^{2/3}L)} \cdot \|\mathbf{W}'_\ell h_{\ell-1}\| \\
\le & \|\breve{\mathsf{loss}}_i\| \cdot O\left(\frac{\sqrt{L^3\omega^{2/3}Lm\log m}}{\sqrt{d}} \right) \cdot O(\|\mathbf{W}'_\ell\|_2) \ .
\end{aligned} \tag{8.5}$$

Finally, we also have

$$\left|\breve{\mathsf{loss}}_i^\top \mathbf{B}\breve{\mathbf{D}}_L\breve{\mathbf{W}}_L\cdots\breve{\mathbf{D}}_\ell\mathbf{W}'_\ell(h_{\ell-1}-\breve{h}_{\ell-1})\right|$$
$$\overset{①}{\leq} \|\breve{\mathsf{loss}}_i\|_2\cdot O\left(\sqrt{m/d}+\frac{\omega^{1/3}L^2\sqrt{m\log m}}{\sqrt{d}}\right)\cdot\|\mathbf{W}'_\ell\|_2\cdot\|h_\ell-\breve{h}_\ell\|_2$$
$$\overset{②}{\leq} O(L^{0.5}\sqrt{m/d})\cdot\|\breve{\mathsf{loss}}_i\|_2\cdot L^{1.5}\|\mathbf{W}'_\ell\|_2^2 \tag{8.6}$$

where ① uses Lemma 4.4b (and Lemma 5.7 for bounding the perturbation) and ② uses Claim 8.2 to bound $\|h_\ell-\breve{h}_\ell\|_2$ and our choice of ω.

Putting (8.4), (8.5) and (8.6) back to (8.3), and using triangle inequality, we have the desired result. □

8.1 Proof of Claim 8.2

We first present a simple proposition about the ReLU function.

Proposition 8.3. *Given vectors $a,b\in\mathbb{R}^m$ and $\mathbf{D}\in\mathbb{R}^{m\times m}$ the diagonal matrix where $\mathbf{D}_{k,k}=\mathbb{1}_{a_k\geq 0}$. Then, then there exists a diagonal matrix $\mathbf{D}''\in\mathbb{R}^{m\times m}$ with*

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

Proof. We verify coordinate by coordinate for each $k\in[m]$.

- If $a_k\geq 0$ and $b_k\geq 0$, then $(\phi(a)-\phi(b))_k=a_k-b_k=\big(\mathbf{D}(a-b)\big)_k$.
- If $a_k<0$ and $b_k<0$, then $(\phi(a)-\phi(b))_k=0-0=\big(\mathbf{D}(a-b)\big)_k$.
- If $a_k\geq 0$ and $b_k<0$, then $(\phi(a)-\phi(b))_k=a_k=(a_k-b_k)+\frac{b_k}{a_k-b_k}(a_k-b_k)=\big(\mathbf{D}(a-b)+\mathbf{D}''(a-b)\big)_k$, if we define $(\mathbf{D}'')_{k,k}=\frac{b_k}{a_k-b_k}\in[-1,0]$.
- If $a_k<0$ and $b_k\geq 0$, then $(\phi(a)-\phi(b))_k=-b_k=0\cdot(a_k-b_k)-\frac{b_k}{b_k-a_k}(a_k-b_k)=\big(\mathbf{D}(a-b)+\mathbf{D}''(a-b)\big)_k$, if we define $(\mathbf{D}'')_{k,k}=\frac{b_k}{b_k-a_k}\in[0,1]$. □

Proof of Claim 8.2. We ignore the subscript in i for cleanness, and calculate that

$$\begin{aligned}
h_\ell-\breve{h}_\ell &\overset{①}{=} \phi((\breve{\mathbf{W}}_\ell+\mathbf{W}'_\ell)h_{\ell-1})-\phi(\breve{\mathbf{W}}_\ell\breve{h}_{\ell-1})\\
&\overset{②}{=} (\breve{\mathbf{D}}_\ell+\mathbf{D}''_\ell)\Big((\breve{\mathbf{W}}_\ell+\mathbf{W}'_\ell)h_{\ell-1}-\breve{\mathbf{W}}_\ell\breve{h}_{\ell-1}\Big)\\
&= (\breve{\mathbf{D}}_\ell+\mathbf{D}''_\ell)\breve{\mathbf{W}}_\ell(h_{\ell-1}-\breve{h}_{\ell-1})+(\breve{\mathbf{D}}_\ell+\mathbf{D}''_\ell)\mathbf{W}'_\ell h_{\ell-1}\\
&\overset{③}{=} \sum_{a=1}^{\ell}(\breve{\mathbf{D}}_\ell+\mathbf{D}''_\ell)\breve{\mathbf{W}}_\ell\cdots\breve{\mathbf{W}}_{a+1}(\breve{\mathbf{D}}_a+\mathbf{D}''_a)\mathbf{W}'_a h_{a-1}
\end{aligned}$$

Above, ① is by the recursive definition of h_ℓ and $\breve{h}_\ell$; ② is by Proposition 8.3 and $\mathbf{D}''_\ell$ is defined according to Proposition 8.3; and inequality ③ is by recursively computing $h_{\ell-1}-\breve{h}_{\ell-1}$. As for the remaining properties:

- We have $\|\mathbf{D}''_\ell\|_0\leq O(m\omega^{2/3}L)$.

This is because, $(\mathbf{D}''_\ell)_{k,k}$ is non-zero only at the coordinates $k \in [m]$ where the signs of $\breve{g}_\ell$ and g_ℓ are opposite (by Proposition 8.3). Such a coordinate k must satisfy either $(\mathbf{D}^{(0)}_\ell)_{k,k} \neq (\breve{\mathbf{D}}_\ell)_{k,k}$ or $(\mathbf{D}^{(0)}_\ell)_{k,k} \neq (\mathbf{D}_\ell)_{k,k}$, and therefore by Lemma 5.2b there are at most $O(m\omega^{2/3}L)$ such coordinates k.

- We have $\|h_\ell - \breve{h}_\ell\| \leq O(L^{1.5})\|\overrightarrow{\mathbf{W}}'\|_2$.

  This is because we have $\big\|(\breve{\mathbf{D}}_\ell + \mathbf{D}''_\ell)\breve{\mathbf{W}}_\ell \cdots \breve{\mathbf{W}}_{a+1}(\breve{\mathbf{D}}_a + \mathbf{D}''_a)\big\|_2 \leq O(\sqrt{L})$ from Lemma 5.6b, we have $\|h_{a-1}\| \leq O(1)$ (by $\|h^{(0)}_{a-1}\| \leq O(1)$ from Lemma 4.1 and $\|h^{(0)}_{a-1} - h_{a-1}\| \leq o(1)$ from Lemma 5.2c); and and $\|\mathbf{W}'_a h_{a-1}\| \leq \|\mathbf{W}'_a\|_2 \|h_{a-1}\| \leq O(\|\overrightarrow{\mathbf{W}}'\|_2)$.

- We have $\|\mathbf{B}h_\ell - \mathbf{B}\breve{h}_\ell\| \leq O(L\sqrt{m/d})\|\overrightarrow{\mathbf{W}}'\|_2$.

  This is because we have $\big\|\mathbf{B}(\breve{\mathbf{D}}_\ell + \mathbf{D}''_\ell)\breve{\mathbf{W}}_\ell \cdots \breve{\mathbf{W}}_{a+1}(\breve{\mathbf{D}}_a + \mathbf{D}''_a)\big\|_2 \leq O(\sqrt{m/d})$ from Lemma 4.4b (along with perturbation bound Lemma 5.7), we have $\|h_{a-1}\| \leq O(1)$ (by $\|h^{(0)}_{a-1}\| \leq O(1)$ from Lemma 4.1 and $\|h^{(0)}_{a-1} - h_{a-1}\| \leq o(1)$ from Lemma 5.2c); and and $\|\mathbf{W}'_a h_{a-1}\| \leq \|\mathbf{W}'_a\|_2 \|h_{a-1}\| \leq O(\|\overrightarrow{\mathbf{W}}'\|_2)$.

□

9 Convergence Rate of GD

Theorem 1 (gradient descent, restated)**.** *For any $\varepsilon \in (0,1]$, $\delta \in \big(0, O(\frac{1}{L})\big]$. Let $m \geq \widetilde{\Omega}\big((nL/\delta)^{30} \cdot d \cdot \log^2 \varepsilon^{-1}\big)$, $\eta \stackrel{\text{def}}{=} \Theta\big(\frac{d\delta}{n^4 L^2 m}\big)$, and $\overrightarrow{\mathbf{W}}^{(0)}, \mathbf{A}, \mathbf{B}$ are at random initialization. Then, with probability at least $1 - e^{-\Omega(\log^2 m)}$, suppose we start at $\overrightarrow{\mathbf{W}}^{(0)}$ and for each $t = 0, 1, \dots, T-1$,*

$$\overrightarrow{\mathbf{W}}^{(t+1)} = \overrightarrow{\mathbf{W}}^{(t)} - \eta \nabla F(\overrightarrow{\mathbf{W}}^{(t)}) \enspace .$$

Then, it satisfies

$$F(\overrightarrow{\mathbf{W}}^{(T)}) \leq \varepsilon \quad \text{for} \quad T = \Theta\left(\frac{n^6 L^2}{\delta^2} \log \frac{1}{\varepsilon}\right) \enspace .$$

In other words, the training loss drops to ε in a linear convergence speed.

Proof of Theorem 1. Using Lemma 4.1 we have $\|h_{i,L}\|_2 \leq 1.1$ and then using the randomness of $\mathbf{B}$, it is easy to show that $\|\mathbf{B}h^{(0)}_{i,L} - y^*_i\|^2 \leq O(\log^2 m)$ with at least $1 - e^{-\Omega(\log^2 m)}$ (where $h^{(0)}_{i,L}$ is defined with respect to the random initialization $\overrightarrow{\mathbf{W}}^{(0)}$), and therefore

$$F(\overrightarrow{\mathbf{W}}^{(0)}) \leq O(n \log^2 m) \enspace .$$

Let us assume for every $t = 0, 1, \dots, T-1$, the following holds

$$\|\overrightarrow{\mathbf{W}}^{(t)} - \overrightarrow{\mathbf{W}}^{(0)}\|_F \leq \omega \stackrel{\text{def}}{=} O\left(\frac{n^3\sqrt{d}}{\delta\sqrt{m}} \log \frac{n \log m}{\varepsilon}\right) \enspace . \tag{9.1}$$

We shall prove the convergence of GD assuming (9.1) holds, so that previous statements such as Theorem 4 and Theorem 3 can be applied. At the end of the proof, we shall verify that (9.1) is satisfied.

Letting $\nabla_t = \nabla F(\overrightarrow{\mathbf{W}}^{(t)})$, we calculate that

$$
\begin{aligned}
& F(\overrightarrow{\mathbf{W}}^{(t+1)}) \\
& \overset{①}{\leq} F(\overrightarrow{\mathbf{W}}^{(t)}) - \eta\|\nabla F(\overrightarrow{\mathbf{W}}^{(t)})\|_F^2 + \eta\sqrt{nF(\overrightarrow{\mathbf{W}}^{(t)})} \cdot O\left(\frac{\omega^{1/3}L^2\sqrt{m\log m}}{\sqrt{d}}\right) \cdot \|\nabla_t\|_2 + O\big(\eta^2 \frac{nL^2m}{d}\big)\|\nabla_t\|_2^2 \\
& \overset{②}{\leq} F(\overrightarrow{\mathbf{W}}^{(t)}) - \eta\|\nabla F(\overrightarrow{\mathbf{W}}^{(t)})\|_F^2 + O\left(\frac{\eta n L^2 m\omega^{1/3}\sqrt{\log m}}{d} + \frac{\eta^2 n^2 L^2 m^2}{d^2}\right) \cdot F(\overrightarrow{\mathbf{W}}^{(t)}) \\
& \overset{③}{\leq} \left(1 - \Omega\Big(\frac{\eta\delta m}{dn^2}\Big)\right) F(\overrightarrow{\mathbf{W}}^{(t)}) \ . \qquad (9.2)
\end{aligned}
$$

Above, ① uses Theorem 4; ② uses Theorem 3 (which gives $\|\nabla_t\|_2^2 \leq \max_{\ell\in[L]} \|\nabla_{\mathbf{W}_\ell} F(\overrightarrow{\mathbf{W}}^{(t)})\|_F^2 \leq O\big(\frac{F(\overrightarrow{\mathbf{W}}^{(t)})}{d} \times mn\big)$); ③ use gradient lower bound from Theorem 3 and our choice of η. In other words, after $T = \Omega(\frac{dn^2}{\eta\delta m}) \log\frac{n\log m}{\varepsilon}$ iterations we have $F(\overrightarrow{\mathbf{W}}^{(T)}) \leq \varepsilon$.

We need to verify for each t, $\|\overrightarrow{\mathbf{W}}^{(t)} - \overrightarrow{\mathbf{W}}^{(0)}\|_F$ is small so that (9.1) holds. By Theorem 3,

$$
\begin{aligned}
\|\mathbf{W}_\ell^{(t)} - \mathbf{W}_\ell^{(0)}\|_F &\leq \sum_{i=0}^{t-1} \|\eta\nabla_{\mathbf{W}_\ell} F(\overrightarrow{\mathbf{W}}^{(i)})\|_F \leq O(\eta\sqrt{nm/d}) \cdot \sum_{i=0}^{t-1} \sqrt{F(\overrightarrow{\mathbf{W}}^{(i)})} \\
&\leq O(\eta\sqrt{nm/d}) \cdot O(T\cdot\sqrt{n}) \leq \eta T \cdot O(\sqrt{n^2m/d}) \leq O\left(\frac{n^3\sqrt{d}}{\delta\sqrt{m}} \log\frac{n\log m}{\varepsilon}\right) \ .
\end{aligned}
$$

where the last step follows by our choice of T. □

10 Convergence Rate of SGD

Theorem 2 (stochastic gradient descent, stated)**.** *For any $\varepsilon \in (0,1]$, $\delta \in \big(0, O(\frac{1}{L})\big]$, $b \in [n]$. Let $m \geq \widetilde{\Omega}\big(\frac{(nL/\delta)^{30}\cdot d\cdot\log^2\varepsilon^{-1}}{b}\big)$, $\eta \stackrel{\text{def}}{=} \Theta(\frac{b\delta d}{n^5L^2m\log^2 m})$, and $\overrightarrow{\mathbf{W}}^{(0)}, \mathbf{A}, \mathbf{B}$ are at random initialization. Suppose we start at $W^{(0)}$ and for each $t = 0, 1, \dots, T-1$,*

$$W^{(t+1)} = W^{(t)} - \eta\cdot\frac{n}{|S_t|}\sum_{i\in S_t}\nabla F(W^{(t)}) \qquad \text{(for a random subset } S_t \subseteq [n] \text{ of fixed cardinality } b.)$$

Then, it satisfies with probability at least $1 - e^{-\Omega(\log^2 m)}$ over the randomness of $S_1, \dots, S_T$:

$$F(W^{(T)}) \leq \varepsilon \quad \textit{for all} \quad T \stackrel{\text{def}}{=} \Theta\Big(\frac{dn^2}{\eta\delta m}\log\frac{n\log m}{\varepsilon}\Big) = \Theta\Big(\frac{n^7L^2\log^2 m}{b\delta^2}\log\frac{n\log m}{\varepsilon}\Big) \ .$$

Proof of Theorem 2. Using similar argument as the proof of Theorem 1, we have with at least $1 - e^{-\Omega(\log^2 m)}$ probability

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

Let us assume for every $t = 0, 1, \dots, T-1$, the following holds

$$\|\overrightarrow{\mathbf{W}}^{(t)} - \overrightarrow{\mathbf{W}}^{(0)}\|_F \leq \omega \stackrel{\text{def}}{=} O\left(\frac{n^{3.5}\sqrt{d}}{\delta\sqrt{bm}}\log\frac{n\log m}{\varepsilon}\right) \ . \qquad (10.1)$$

We shall prove the convergence of SGD assuming (10.1) holds, so that previous statements such as Theorem 4 and Theorem 3 can be applied. At the end of the proof, we shall verify that (10.1) is satisfied throughout the SGD with high probability.

For each $t = 0, 1, \ldots, T-1$, using the same notation as Theorem 1, except that we choose $\nabla_t = \frac{n}{|S_t|} \sum_{i \in S_t} \nabla F_i(\overrightarrow{\mathbf{W}}^{(t)})$. We have $\mathbb{E}_{S_t}[\nabla_t] = \nabla F(\overrightarrow{\mathbf{W}}^{(t)})$ and therefore

$$\begin{aligned}
&\mathop{\mathbb{E}}_{S_t}[F(\overrightarrow{\mathbf{W}}^{(t+1)})] \\
&\overset{①}{\leq} F(\overrightarrow{\mathbf{W}}^{(t)}) - \eta \|\nabla F(\overrightarrow{\mathbf{W}}^{(t)})\|_F^2 + \eta \sqrt{nF(\overrightarrow{\mathbf{W}}^{(t)})} \cdot O\left(\frac{\omega^{1/3} L^2 \sqrt{m \log m}}{\sqrt{d}}\right) \cdot \mathop{\mathbb{E}}_{S_t}[\|\nabla_t\|_2] \\
&\quad + O\big(\eta^2 \frac{nL^2 m}{d}\big) \mathop{\mathbb{E}}_{S_t}[\|\nabla_t\|_2^2] \\
&\overset{②}{\leq} F(\overrightarrow{\mathbf{W}}^{(t)}) - \eta \|\nabla_t\|_F^2 + O\left(\frac{\eta n L^2 m \omega^{1/3} \sqrt{\log m}}{d} + \frac{\eta^2 n^2 L^2 m^2}{d^2}\right) \cdot F(\overrightarrow{\mathbf{W}}^{(t)}) \\
&\overset{③}{\leq} \left(1 - \Omega\Big(\frac{\eta \delta m}{dn^2}\Big)\right) F(\overrightarrow{\mathbf{W}}^{(t)}) \ . \qquad (10.2)
\end{aligned}$$

Above, ① uses Theorem 4 and $\mathbb{E}_{S_t}[\nabla_t] = \nabla F(\overrightarrow{\mathbf{W}}^{(t)})$; ② uses Theorem 3 which give

$$\mathop{\mathbb{E}}_{S_t}\left[\|\nabla_t\|_2^2\right] \leq \frac{n^2}{b} \mathop{\mathbb{E}}_{S_t}\left[\sum_{i \in S_t} \max_{\ell \in [L]} \left\|\nabla_{\mathbf{W}_\ell} F_i(\overrightarrow{\mathbf{W}}^{(t)})\right\|_F^2\right] \leq O\Big(\frac{nmF(\overrightarrow{\mathbf{W}}^{(t)})}{d}\Big)$$

$$\mathop{\mathbb{E}}_{S_t}\left[\|\nabla_t\|_2\right] \leq \left(\mathop{\mathbb{E}}_{S_t}\left[\|\nabla_t\|_2^2\right]\right)^{1/2} \leq O\Big(\big(\frac{nmF(\overrightarrow{\mathbf{W}}^{(t)})}{d}\big)^{1/2}\Big) \ ;$$

③ use gradient lower bound from Theorem 3 and our choice of η.

At the same time, we also have the following absolute value bound:

$$\begin{aligned}
F(\overrightarrow{\mathbf{W}}^{(t+1)}) &\overset{①}{\leq} F(\overrightarrow{\mathbf{W}}^{(t)}) + \eta \|\nabla F(\overrightarrow{\mathbf{W}}^{(t)})\|_F \cdot \|\nabla_t\|_F \\
&\quad + \eta \sqrt{nF(\overrightarrow{\mathbf{W}}^{(t)})} \cdot O\left(\frac{\omega^{1/3} L^2 \sqrt{m \log m}}{\sqrt{d}}\right) \cdot \|\nabla_t\|_2 + O\Big(\eta^2 \frac{nL^2 m}{d}\Big) \cdot \|\nabla_t\|_2^2 \\
&\overset{②}{\leq} F(\overrightarrow{\mathbf{W}}^{(t)}) + \eta \cdot O\left(\sqrt{\frac{LF(\overrightarrow{\mathbf{W}}^{(t)})mn}{d}}\right) \cdot O\left(\sqrt{\frac{n^2 m L F(\overrightarrow{\mathbf{W}}^{(t)})}{bd}}\right) \\
&\quad + \eta \sqrt{nF(\overrightarrow{\mathbf{W}}^{(t)})} \cdot O\left(\frac{\omega^{1/3} L^2 \sqrt{m \log m}}{\sqrt{d}}\right) \cdot \frac{\sqrt{n^2 m F(\overrightarrow{\mathbf{W}}^{(t)})}}{\sqrt{bd}} \\
&\quad + O\Big(\eta^2 \frac{nL^2 m}{d}\Big) \cdot \frac{n^2}{b} O\Big(\frac{mF(\overrightarrow{\mathbf{W}}^{(t)})}{d}\Big) \\
&\overset{③}{\leq} \Big(1 + O\Big(\frac{\eta L m n^{1.5}}{\sqrt{b}d} + \frac{\eta n^{1.5} \omega^{1/3} L^2 m \sqrt{\log m}}{\sqrt{b}d} + \frac{\eta^2 n^3 L^2 m^2}{d^2 b}\Big)\Big) F(\overrightarrow{\mathbf{W}}^{(t)}) \ . \qquad (10.3)
\end{aligned}$$

Above, ① uses Theorem 4 and Cauchy-Schwarz $\langle A, B\rangle \leq \|A\|_F \|B\|_F$, and ② uses Theorem 3 which give

$$\|\nabla_t\|_2^2 \leq \frac{n^2}{b}\left[\sum_{i \in S_t} \max_{\ell \in [L]} \left\|\nabla_{\mathbf{W}_\ell} F_i(\overrightarrow{\mathbf{W}}^{(t)})\right\|_F^2\right] \leq \frac{n^2}{b} O\Big(\frac{mF(\overrightarrow{\mathbf{W}}^{(t)})}{d}\Big)$$

$$\|\nabla_t\|_F^2 \leq \frac{n^2}{b}\left[\sum_{i \in S_t} \sum_{\ell=1}^{L} \left\|\nabla_{\mathbf{W}_\ell} F_i(\overrightarrow{\mathbf{W}}^{(t)})\right\|_F^2\right] \leq \frac{Ln^2}{b} O\Big(\frac{mF(\overrightarrow{\mathbf{W}}^{(t)})}{d}\Big)$$

and the derivation from (10.2).

Next, taking logarithm on both sides of (10.2) and (10.3), and using Jensen's inequality $\mathbb{E}[\log X] \leq \log \mathbb{E}[X]$, we have

$$\mathbb{E}[\log F(\overrightarrow{\mathbf{W}}^{(t+1)})] \leq \log F(\overrightarrow{\mathbf{W}}^{(t)}) - \Omega\Big(\frac{\eta\delta m}{dn^2}\Big) \quad \text{and} \quad \log F(\overrightarrow{\mathbf{W}}^{(t+1)}) \leq \log F(\overrightarrow{\mathbf{W}}^{(t)}) + O\Big(\frac{\eta L m n^{1.5}}{\sqrt{b}d}\Big)$$

By (one-sided) martingale concentration, we have with probability at least $1 - e^{-\Omega(\log^2 m)}$, for every $t = 1, 2, \dots, T$:

$$\log F(\overrightarrow{\mathbf{W}}^{(t)}) - \mathbb{E}[\log F(\overrightarrow{\mathbf{W}}^{(t)})] \leq \sqrt{t} \cdot O\Big(\frac{\eta L m n^{1.5}}{\sqrt{b}d}\Big) \cdot \log m \ .$$

This implies two things.

- On one hand, if $T \geq \Omega\Big(\frac{L^2 n^7}{b\delta^2} \log^2 m\Big)$ and $T = \Omega(\frac{dn^2}{\eta\delta m} \log \frac{n \log m}{\varepsilon})$ iterations we have

$$\begin{aligned}\log F(\overrightarrow{\mathbf{W}}^{(T)}) &\leq \sqrt{T} \cdot O\Big(\frac{\eta L m n^{1.5}}{\sqrt{b}d}\Big) \cdot \log m + \log F(\overrightarrow{\mathbf{W}}^{(0)}) - \Omega\Big(\frac{\eta\delta m}{dn^2}\Big)T \\ &\leq \log F(\overrightarrow{\mathbf{W}}^{(0)}) - \Omega\Big(\frac{\eta\delta m}{dn^2}\Big)T \leq \log O(n\log^2 m) - \Omega\Big(\log\frac{n\log^2 m}{\varepsilon}\Big) \leq \log\varepsilon \ .\end{aligned}$$

  Therefore, we have $F(\overrightarrow{\mathbf{W}}^{(T)}) \leq \varepsilon$.

- On the other hand, for every $t = 1, 2, \dots, T$, we have

$$\begin{aligned}\log F(\overrightarrow{\mathbf{W}}^{(t)}) &\leq \sqrt{t} \cdot O\Big(\frac{\eta L m n^{1.5}}{\sqrt{b}d}\Big) \cdot \log m + \log F(\overrightarrow{\mathbf{W}}^{(0)}) - \Omega\Big(\frac{\eta\delta m}{dn^2}\Big)t \\ &\overset{①}{=} \log F(\overrightarrow{\mathbf{W}}^{(0)}) - \left(\sqrt{\frac{\eta\delta m}{dn^2}} \cdot \Omega(\sqrt{t}) - \sqrt{\frac{dn^2}{\eta\delta m}} \cdot O\Big(\frac{\eta L m n^{1.5}}{\sqrt{b}d} \log m\Big)\right)^2 \\ &\quad + O\Big(\frac{\eta L^2 m n^5}{b\delta d} \log^2 m\Big) \\ &\overset{②}{\leq} \log F(\overrightarrow{\mathbf{W}}^{(0)}) + 1\end{aligned}$$

  where in ① we have used $2a\sqrt{t} - b^2 t = -(b\sqrt{t} - a/b)^2 + a^2/b^2$, and in ② we have used our choice of η. This implies $F(\overrightarrow{\mathbf{W}}^{(t)}) \leq O(n)$. We can now verify for each t, $\|\overrightarrow{\mathbf{W}}^{(t)} - \overrightarrow{\mathbf{W}}^{(0)}\|_F$ is small so that (10.1) holds. By Theorem 3,

$$\begin{aligned}\|\mathbf{W}_\ell^{(t)} - \mathbf{W}_\ell^{(0)}\|_F &\leq \sum_{i=0}^{t-1} \Big\|\eta \frac{n}{|S_t|} \sum_{i\in S_t} \nabla_{\mathbf{W}_\ell} F_i(\overrightarrow{\mathbf{W}}^{(t)})\Big\|_F \leq O\left(\eta\sqrt{\frac{n^2 m}{bd}}\right) \cdot \sum_{i=0}^{t-1} \sqrt{F(\overrightarrow{\mathbf{W}}^{(i)})} \\ &\leq O\left(\eta\sqrt{\frac{n^2 m}{bd}}\right) \cdot O(T\sqrt{n}) \leq \eta T \cdot O(\sqrt{mn^3/bd}) = O\left(\frac{n^{3.5}\sqrt{d}}{\delta\sqrt{bm}} \log\frac{n\log m}{\varepsilon}\right) \ .\end{aligned}$$

  where the last step follows by our choice of T. □

Appendix

A Extension to Other Loss Functions

For simplicity, in them main body of this paper we have used the ℓ_2 regression loss. Our results generalize easily to other Lipschitz smooth (but possibly nonconvex) loss functions.

Suppose we are given loss function $f(z; y)$ that takes as input a neural-network output $z \in \mathbb{R}^d$ and a label y. Then, our training objective for the i-th training sample becomes $F_i(\mathbf{W}) = f(\mathbf{B}h_{i,L}; y_i^*)$. We redefine the loss vector $\mathsf{loss}_i \stackrel{\text{def}}{=} \nabla f(\mathbf{B}h_{i,L}; y_i^*) \in \mathbb{R}^d$ (where the gradient is with respect to z). Note that if $f(z; y) = \frac{1}{2}\|z - y\|^2$ is the ℓ_2 loss, then this notion coincides with Section 2. We assume that $f(z; y)$ is 1-Lipscthiz (upper) smooth with respect to z.[6]

All the results in Section 4, 5 and 6 remain unchanged. Section 7 also remains unchanged, except we need to restate Theorem 3 with respect to this new notation:

$$\|\nabla_{\mathbf{W}_\ell} F_i(\overrightarrow{\mathbf{W}})\|_F^2 \leq O\Big(\frac{\|\mathsf{loss}_i\|^2}{d} \times m\Big) \qquad \|\nabla_{\mathbf{W}_\ell} F(\overrightarrow{\mathbf{W}})\|_F^2 \leq O\Big(\frac{\|\mathsf{loss}\|^2}{d} \times mn\Big)$$

$$\|\nabla_{\mathbf{W}_L} F(\overrightarrow{\mathbf{W}})\|_F^2 \geq \Omega\Big(\frac{\max_{i\in[n]} \|\mathsf{loss}_i\|^2}{dn/\delta} \times m\Big) \ .$$

Section 8 also remains unchanged, except that we need to replace the precise definition of ℓ_2 loss in (8.2) with the smoothness condition:

$$\begin{aligned} F_i(\overrightarrow{\mathbf{W}}) = f(\mathbf{B}h_{i,L}; y_i^*) &\leq f(\mathbf{B}\breve{h}_{i,L}; y_i^*) + \langle \nabla f(\mathbf{B}\breve{h}_{i,L}, y_i^*), \mathbf{B}(h_{i,L} - \breve{h}_{i,L})\rangle + \frac{1}{2}\|\mathbf{B}(h_{i,L} - \breve{h}_{i,L})\|^2 \\ &= F_i(\breve{\overrightarrow{\mathbf{W}}}) + \langle \breve{\mathsf{loss}}_i, \mathbf{B}(h_{i,L} - \breve{h}_{i,L})\rangle + \frac{1}{2}\|\mathbf{B}(h_{i,L} - \breve{h}_{i,L})\|^2 \end{aligned} \tag{A.1}$$

and the rest of the proof remains unchanged.

As for the final convergence theorem of gradient descent, we can replace (9.2) with

$$F(\overrightarrow{\mathbf{W}}^{(t+1)}) \leq F(\overrightarrow{\mathbf{W}}^{(t)}) - \Omega\Big(\frac{\eta\delta m}{dn^2}\Big) \cdot \|\mathsf{loss}^{(t)}\|^2 \ . \tag{A.2}$$

This means many things:

- If the loss is bounded (say, $|f(z; y)| \leq O(1)$), then in $T = O\Big(\frac{dn^2}{\eta\delta m\varepsilon^2}\Big) = O\big(\frac{n^6L^2}{\delta^2\varepsilon^2}\big)$ iterations, we can find a point $\overrightarrow{\mathbf{W}}^{(T)}$ with $\|\mathsf{loss}^{(T)}\| \leq \varepsilon$. (We choose $m \geq \widetilde{\Omega}\big((nL/\delta)^{30} d\varepsilon^{-4}\big)$.)
- If the loss is cross entropy $f(z; y) = \frac{e^{z_y}}{\sum_{i=1}^d e^{z_i}}$ for classification, then $\|\nabla f(z; y)\| < 1/4$ implies perfect classification.[7] Thus, we have 100% training accuracy in $T = O\big(\frac{n^6L^2}{\delta^2}\big)$ iterations.
- If the loss satisfies the Polyak-Łojasiewicz condition $\|\nabla f(z; y)\|^2 \geq \sigma(f(z; y) - f(z^*; y))$, then in $T = \Omega\Big(\frac{dn^2}{\eta\delta m\sigma}\Big) = O\big(\frac{n^6L^2}{\delta^2\sigma}\log\frac{1}{\varepsilon}\big)$ iterations, we can find a point $\overrightarrow{\mathbf{W}}^{(T)}$ with $\|\mathsf{loss}^{(T)}\| \leq \varepsilon$. (We choose $m \geq \widetilde{\Omega}\big((nL/\delta)^{30} d\sigma^{-2}\log^2\varepsilon^{-1}\big)$.)
- If the loss is convex and its minimizer has bounded norm, meaning there exists z^* so that $f(z^*; y) = \min_z f(z; y)$ and $\|z - z^*\| \leq D$. Then, by convexity

$$f(z; y) - f(z^*; y) \leq \langle \nabla f(z; y), z - z^*\rangle \leq D\|\nabla f(z; y)\|$$

[6] That is, $f(z + z'; y) \leq f(z) + \langle \nabla f(z; y), z'\rangle + \frac{1}{2}\|z'\|^2$.

[7] Recall $\frac{\partial f(z;y)}{z_y} = p_y(1 - p_y)$ where $p_j = \frac{e^{z_j}}{\sum_{i=1}^d e^{z_i}}$. If $p_y > 1/2$, then z correctly predicts the target label y because $p_y > p_j$ for $j \neq z$.

Putting this into (A.2), we have (here $\overrightarrow{\mathbf{W}}^* = \arg\min_{\overrightarrow{\mathbf{W}}} F_i(\overrightarrow{\mathbf{W}})$ for all $i \in [n]$)

$$F(\overrightarrow{\mathbf{W}}^{(t+1)}) - F(\overrightarrow{\mathbf{W}}^*) \leq F(\overrightarrow{\mathbf{W}}^{(t)}) - F(\overrightarrow{\mathbf{W}}^*) - \Omega\Big(\frac{\eta\delta m}{dn^2 D^2}\Big) \cdot \sum_{i\in[n]} \big(F_i(\overrightarrow{\mathbf{W}}^{(t)}) - F_i(\overrightarrow{\mathbf{W}}^*)\big)^2$$
$$\leq F(\overrightarrow{\mathbf{W}}^{(t)}) - F(\overrightarrow{\mathbf{W}}^*) - \Omega\Big(\frac{\eta\delta m}{dn^3 D^2}\Big) \cdot \big(F(\overrightarrow{\mathbf{W}}^{(t)}) - F(\overrightarrow{\mathbf{W}}^*)\big)^2 \ .$$

This implies (see for instance the classical calculation steps in [38]) that after $T = O\Big(\frac{dn^3 D^2}{\eta\delta m\varepsilon}\Big) = O\big(\frac{n^7 L^2 D^2}{\delta^2 \varepsilon}\big)$ iterations, we can have $F(\overrightarrow{\mathbf{W}}^{(T)}) - F(\overrightarrow{\mathbf{W}}^*) \leq \varepsilon$.

B Extension to Convolutional Neural Networks

There are numerous versions of convolutional neural networks (CNNs) that are used in practice. To demonstrate the capability of applying our techniques to such convolutional settings, in this section, we study a simple enough CNN for the ℓ_2 regression task.

A Simple CNN Model. We assume that for the input layer (corresponding to $\mathbf{A}$) and for each hidden layer $\ell = 1, 2, \dots, L-1$ (corresponding to $\mathbf{W}_1, \dots, \mathbf{W}_{L-1}$), there are $\mathfrak{d}$ positions each consisting of m channels. (Each position can be thought as a pixel of an image in computer vision tasks.) We assume the last hidden layer $\ell = L$ (corresponding to $\mathbf{W}_L$) and the output layer (corresponding to $\mathbf{B}$) are fully connected. We assume for each $j \in [\mathfrak{d}]$, there exists a set $\mathcal{Q}_j \subseteq [\mathfrak{d}]$ of fixed cardinality $q \in [\mathfrak{d}]$ so that the value at position j in any convolutional layer is completely determined by positions $k \in \mathcal{Q}_j$ of the previous layer.

Assumption B.1. *We assume that $(\mathcal{Q}_1, \dots, \mathcal{Q}_{\mathfrak{d}})$ give rise to a q-regular bipartite graph: each $\mathcal{Q}_j$ has exactly q entries and each $k \in [\mathfrak{d}]$ appears in exactly q different sets $\mathcal{Q}_j$.*
(In vision tasks, if 3×3 kernels are used then $|\mathcal{Q}_j| = 9$. We ignore the padding issue for simplicity.)

The output of each convolutional layer $\ell = 0, 1, 2, \dots, L-1$ is represented by a $\mathfrak{d}m$-dimensional vector $h_\ell = (h_{\ell,1}, \dots, h_{\ell,\mathfrak{d}})$ where each $h_{\ell,j} \in \mathbb{R}^m, \forall j \in [\mathfrak{d}]$. In the input layer and each $j \in [\mathfrak{d}]$, we assume

$$h_{0,j} = \phi\big(\mathbf{A}_j x_{\mathcal{Q}_j}\big) \in \mathbb{R}^m$$

where $x_{\mathcal{Q}_j} \in \mathbb{R}^q$ denotes the concatenation of x_k for all $k \in \mathcal{Q}_j$ given input $x \in \mathbb{R}^{\mathfrak{d}}$, and $\mathbf{A}_j \in \mathbb{R}^{m\times q}$ is randomly initialized at $\mathcal{N}(0, \frac{2}{\sqrt{qm}})$ per entry. For notational simplicity, we define matrix $\mathbf{A} \in \mathbb{R}^{\mathfrak{d}m\times\mathfrak{d}}$ so that it satisfies $h_1 = \phi\big(\mathbf{A}x\big)$. Each row of $\mathbf{A}$ has q non-zero entries.

For each layer $\ell = 1, \dots, L-1$ and each $j \in [\mathfrak{d}]$, we assume

$$h_{\ell,j} = \phi\big(\mathbf{W}_{\ell,j} h_{\ell-1,\mathcal{Q}_j} + \tau \cdot \mathbf{b}_{\ell,j}\big) \in \mathbb{R}^m$$

where $h_{\ell-1,\mathcal{Q}_j} \in \mathbb{R}^{qm}$ denotes the concatenation of $h_{\ell-1,k}$ for all $k \in \mathcal{Q}_j$, the weights $\mathbf{W}_{\ell,j} \in \mathbb{R}^{m\times(qm)}$ and the bias the $\mathbf{b}_{\ell,j} \in \mathbb{R}^m$ are randomly initialized at $\mathcal{N}(0, \frac{2}{qm})$ per entry, and τ is a small parameter (say, $\tau = \frac{\delta^2}{10\mathfrak{d}L}$) for bias. For notational simplicity, we define matrix $\mathbf{W}_\ell \in \mathbb{R}^{\mathfrak{d}m\times\mathfrak{d}m}$ and vector $\mathbf{b}_\ell \in \mathbb{R}^{\mathfrak{d}m}$ so that it satisfies $h_\ell = \phi\big(\mathbf{W}_\ell h_{\ell-1} + \tau\mathbf{b}_\ell\big)$, and define vector $g_\ell \stackrel{\text{def}}{=} \mathbf{W}_\ell h_{\ell-1} + \tau\mathbf{b}_\ell \in \mathbb{R}^{\mathfrak{d}m}$. Note that each row of $\mathbf{W}_\ell$ has qm non-zero entries.

We assume the last layer $\mathbf{W}_L$ and the output layer $\mathbf{B}$ are simply fully connected (say without bias). That is, each entry of $\mathbf{W}_L \in \mathbb{R}^{\mathfrak{d}m\times\mathfrak{d}m}$ is from $\mathcal{N}(0, \frac{2}{qm})$, and of $\mathbf{B} \in \mathbb{R}^{d\times\mathfrak{d}m}$ is from $\mathcal{N}(0, \frac{1}{d})$.

We denote by $h_{i,\ell}$ the value of h_ℓ when the input vector is x_i, and define $g_{i,\ell}$, $\mathbf{D}_{i,\ell}$ in the same way as before.

B.1 Changes in the Proofs

If one is willing to loose polynomial factors in L and $\mathfrak{d}$ in the final complexity, then changes to each of the lemmas of this paper is very little.[8]

Changes to Section 4. The first main result is Lemma 4.1: $\|h_{i,\ell}\|$ is in $[1-\varepsilon, 1+\varepsilon]$ with high probability. In the CNN case, for every $j \in [\mathfrak{d}]$, recalling that $h_{i,\ell,j} = \phi(\mathbf{W}_{\ell,j} h_{i,\ell-1,\mathcal{Q}_j} + \tau \mathbf{b}_\ell)$. Applying Fact 4.2, we have that $\frac{qm\|h_{i,\ell,j}\|^2}{\|h_{i,\ell-1,\mathcal{Q}_j}\|^2+\tau^2}$ is distributed as χ-square distribution of order m. Due to concentration of χ-square distribution, $\|h_{i,\ell,j}\|^2$ is extremely close to its expectation $\frac{\|h_{i,\ell-1,\mathcal{Q}_j}\|^2+\tau^2}{q}$. Summing this up over all $j \in [\mathfrak{d}]$, and using Assumption B.1, we have $\|h_{i,\ell,j}\|^2$ is concentrated at $\|h_{i,\ell-1}\|^2 + \frac{\tau^2 \mathfrak{d}}{q}$. Applying induction, we have $\|h_{i,\ell}\|$ is in $[1-\varepsilon, 1+\varepsilon]$ with probability at least $1 - e^{-\Omega(m\varepsilon^2/L^2)}$, as long as $\tau^2 \leq \frac{\varepsilon q}{10 \mathfrak{d} L}$.[9]

The changes to Lemma 4.3 and Lemma 4.4 are the same as above, but we loose some polynomial factors in L (because we are not careful in the argument above). For instance, the intermediate bound in Lemma 4.3a becomes $\|\mathbf{W}_b \mathbf{D}_{i,b-1} \mathbf{W}_{b-1} \cdots \mathbf{D}_{i,a} \mathbf{W}_a\|_2 \leq O(L)$.

As for the δ-separateness Lemma 4.5, we need to redefine the notion of δ-separateness between $h_{i,\ell}$ and $h_{j,\ell}$:

$$\sum_{k \in [\mathfrak{d}]} \left\| (\mathbf{I} - \tfrac{h_{i,\ell,k} h_{i,\ell,k}^\top}{\|h_{i,\ell,k}\|^2}) h_{j,\ell,k} \right\|^2 \geq \Omega(\delta^2) \tag{B.1}$$

Then, denoting by $\widehat{h}_k = h_{i,\ell-1,k} / \|h_{i,\ell-1,k}\|$, we have

$$h_{j,\ell,k} = \phi(\mathbf{W}_{\ell,k} h_{j,\ell-1,\mathcal{Q}_j} + \tau \mathbf{b}_{\ell,k}) = \phi\Big(\vec{g}_1 + \Big(\sum_{z \in \mathcal{Q}_k} \|(\mathbf{I} - \widehat{h}_z \widehat{h}_z^\top) h_{j,\ell-1,z}\|^2\Big)^{1/2} \vec{g}_2\Big)$$

where $\vec{g}_2 \sim \mathcal{N}(0, \frac{2}{qm}\mathbf{I})$ is independent of the randomness of $h_{i,\ell,k}$ once $\mathbf{A}, \mathbf{W}_1, \ldots, \mathbf{W}_{\ell-1}$ are fixed. One can use this to replace (4.4) and the rest of the proof follows.

Changes to Section 5. The first main result is Lemma 5.2, and we discuss necessary changes here to make it work for CNN. The first change in the proof is to replace $2c_1 L^{1.5}$ with $2c_1 L^2$ due to the above additional factor from Lemma 4.3a. Next, call that the proof of Lemma 5.2 relied on Claim 5.3 and Claim 5.5:

- For Claim 5.3, we can replace the definition of x with $x = \mathbf{D}'(\mathbf{W}^{(0)} h^{(0)} + \tau \mathbf{b} + g')$ for $\mathbf{b} \in \mathcal{N}(0, \frac{2}{qm}\mathbf{I})$. This time, instead of using the randomness of $\mathbf{W}^{(0)}$ like in the old proof (because $\mathbf{W}^{(0)}$ is no longer a full matrix), we use the randomness of $\tau \mathbf{b}$. The new statement becomes

  $$\|x\|_0 \leq O\Big(\frac{\mathfrak{d} m}{\tau^{2/3}} \|g_1'\|^{2/3} + \frac{1}{\tau} \|g_2'\|_\infty (\mathfrak{d} m)^{3/2}\Big) \quad \text{and} \quad \|x\| \leq O\Big(\|g_1'\| + \frac{1}{\sqrt{\tau}} \|g_2'\|_\infty^{3/2} (\mathfrak{d} m)^{3/4}\Big) \ .$$

  and its proof is by re-scaling x by $\frac{1}{\tau}$ and then applying the old proof (with dimension m replaced with $\mathfrak{d} m$).
- For Claim 5.5, it becomes $\|y_1\| \leq O\big(\sqrt{qs/m} \log m\big)$ and $\|y_2\|_\infty \leq \frac{2\sqrt{\log m}}{\sqrt{qm}}$.

After making all of these changes, we loose at most some polynomial factors in L and $\mathfrak{d}$ for the new statement of Lemma 5.2:

[8] We acknowledge the existence of more careful modifications to avoid loosing too many such factors, but do not present such result for the simplicity of this paper.

[9] We note that in all of our applications of Lemma 4.1, the minimal choice of ε is around δ^3 from the proof of δ-separateness. Therefore, choosing $\tau = \frac{\delta^2}{10 \mathfrak{d} L}$ is safe. We are aware of slightly more involved proofs that are capable of handling much larger values of τ.

(a) $\|\mathbf{D}'_{i,\ell}\|_0 \le m\omega^{2/3}\mathsf{poly}(L,\mathfrak{d})$ and $\|\mathbf{D}'_{i,\ell} g_{i,\ell}\| \le \omega\mathsf{poly}(L,\mathfrak{d})$.

(b) $\|g'_{i,\ell}\|, \|h'_{i,\ell}\| \le \omega\mathsf{poly}(L,\mathfrak{d})\sqrt{\log m}$.

Finally, the statements of Lemma 5.6 and Lemma 5.7 only loose polynomial factors in L and $\mathfrak{d}$.

Changes to Section 6. The norm upper bound part is trivial to modify so we only focus on the gradient norm lower bound. Since we have assumed $\mathbf{W}_L$ to be fully connected, the gradient on $\mathbf{W}_L$ is the same as before:

$$\widehat{\nabla}^{\vec{\mathsf{v}}}_{[\mathbf{W}_L]_k} F(\overrightarrow{\mathbf{W}}) = \sum_{i=1}^n \langle \mathbf{B}_k, \mathsf{v}_i \rangle \cdot h_{i,L-1} \cdot \mathbb{1}_{(\mathbf{W}_L h_{i,L-1})_k \ge 0}$$

Since we still have δ-separateness (B.1), one can verify for $\ell = L-1$,

$$\|h_{i,\ell} - h_{j,\ell}\|^2 = \sum_{k\in[\mathfrak{d}]} \|h_{i,\ell,k} - h_{j,\ell,k}\|^2 \ge \sum_{k\in[\mathfrak{d}]} \left\|(\mathbf{I} - \frac{h_{i,\ell,k} h_{i,\ell,k}^\top}{\|h_{i,\ell,k}\|^2}) h_{j,\ell,k}\right\|^2 \ge \Omega(\delta^2) \ .$$

Since $\|h_{i,\ell}\| \approx 1$ and $\|h_{j,\ell}\| \approx 1$, this gives back the old definition of δ-separateness: $(\mathbf{I} - h_{i,\ell} h_{i,\ell}^\top / \|h_{i,\ell}\|^2) h_{j,\ell}$ has norm at least $\Omega(\delta)$. Therefore, the entire rest of Section 6 follows as before.

Final Theorem. Since Section 7 and 8 rely on previous sections, they do not need to be changed (besides some polynomial factor blowup in L and $\mathfrak{d}$). Our final theorem becomes

Theorem 5 (CNN)**.** *Let $m \ge \widetilde{\Omega}\big(\mathsf{poly}(n, L, \mathfrak{d}, \delta^{-1}) \cdot d \cdot \log^2 \varepsilon^{-1}\big)$. For the convolutional neural network defined in this section, with probability at least $1 - e^{-\Omega(\log^2 m)}$ over the random initialization, GD and SGD respectively need at most $T = \frac{\mathsf{poly}(n,L,\mathfrak{d})}{\delta^2} \log \frac{1}{\varepsilon}$ and $T = \frac{\mathsf{poly}(n,L,\mathfrak{d}) \cdot \log^2 m}{\delta^2} \log \frac{1}{\varepsilon}$ iterations to find a point $F(\overrightarrow{\mathbf{W}}) \le \varepsilon$.*

C Extension to Residual Neural Networks

Again as we have discussed in Section C, there are numerous versions of residual neural networks that are used in practice. To demonstrate the capability of applying our techniques to residual settings, in this section, we study a simple enough residual network for the ℓ_2 regression task (without convolutional layers).

A Simple Residual Model. We consider an input layer $h_0 = \phi(\mathbf{A}x)$, $L-1$ residual layers $h_\ell = \phi(h_{\ell-1} + \tau \mathbf{W}_\ell h_{\ell-1})$ for $\ell = 1, 2, \dots, L-1$, a fully-connected layer $h_L = \phi(\mathbf{W}_L h_{L-1})$ and an output layer $y = \mathbf{B} h_L$. We assume that $h_0, \dots, h_L \in \mathbb{R}^m$ and the entries of $\mathbf{W}_\ell \in \mathbb{R}^{m\times m}$ are from $\mathcal{N}(0, \frac{2}{m})$ as before. We choose $\tau = \frac{1}{\Omega(L \log m)}$ which is similar as previous work [56].

We denote by $g_0 = \mathbf{A}x$, $g_\ell = h_{\ell-1} + \tau \mathbf{W}_\ell h_{\ell-1}$ for $\ell = 1, 2, \dots, L-1$ and $g_L = \mathbf{W}_L h_{L-1}$. For analysis, we use $h_{i,\ell}$ and $g_{i,\ell}$ to denote the value of h_ℓ when the input vector is x_i, and $\mathbf{D}_{i,\ell}$ the diagonal sign matrix so that $[\mathbf{D}_{i,\ell}]_{k,k} = \mathbb{1}_{(g_{i,\ell})_k \ge 0}$.

C.1 Changes in the Proofs

Conceptually, we need to replace all the occurrences of $\mathbf{W}_\ell$ with $(\mathbf{I} + \mathbf{W}_\ell)$ for $\ell = 1, 2, \dots, L-1$. Many of the proofs in the residual setting becomes much simpler when residual links are present. The main property we shall use is that the spectral norm

$$\|(\mathbf{I} + \mathbf{W}_a)\mathbf{D}_{i,a+1} \cdots \mathbf{D}_{i,b}(\mathbf{I} + \mathbf{W}_b)\|_2 \le 1.01 \tag{C.1}$$

for any $L-1 \geq a \geq b \geq 1$ with our choice of τ.

Changes to Section 4. For Lemma 4.1, ignoring subscripts in i for simplicity, we can combine the old proof with (C.1) to derive that $\|h_\ell\| \leq 1.02$ for every i and ℓ. We also have $\|h_\ell\| \geq \frac{1}{2}$ by the following argument.

- Fact 4.2 says each coordinate of h_0 follows from an independent folded Gaussian distribution $|\mathcal{N}(0, \frac{1}{m})|$ and therefore, with high probability, at least $m/2$ of the coordinates $k \in [m]$ will satisfy $|(h_0)_k| \geq \frac{0.6}{\sqrt{m}}$. Denote this set as $M_0 \subseteq [m]$.
- In the following layer $\ell = 1$, $(h_\ell)_k \geq (h_{\ell-1})_k - \tau|(\mathbf{W}_\ell h_{\ell-1})_k|$. Since $\mathbf{W}_\ell h_{\ell-1} \sim \mathcal{N}(0, \frac{2\|h_{\ell-1}\|^2}{m}\mathbf{I})$ and $\|h_{\ell-1}\| \leq 1.02$, we know with high probability, at least $1-\frac{1}{10L}$ fraction of the coordinates in M_0 will satisfy $|(\mathbf{W}_\ell h_{\ell-1})_k| \leq O(\frac{\log L}{\sqrt{m}})$. Therefore, for each of these $(1-\frac{1}{10L})|M_0|$ coordinates, we have $(h_\ell)_k \geq (h_{\ell-1})_k - \frac{1}{10L}$ by our choice of τ. Denote this set as $M_1 \subseteq M_0$, then we have $(h_\ell)_k \geq \frac{0.6}{\sqrt{m}} - \frac{1}{10L\sqrt{m}}$ for each $k \in M_1$.
- Continuing this argument for $\ell = 2, 3, \ldots, L-1$, we know that every time we move from $M_{\ell-1}$ to M_ℓ, its size shrinks by a factor $1 - \frac{1}{10L}$, and the magnitude of $(h_\ell)_k$ for $k \in M_\ell$ decreases by $\frac{1}{10L\sqrt{m}}$. Putting this together, we know $\|h_\ell\|^2 \geq (\frac{0.6}{\sqrt{m}} - \frac{1}{10\sqrt{m}})^2 \cdot (1-\frac{1}{10L})^L \cdot \frac{m}{2} \geq \frac{1}{10}$ for all $\ell = 1, 2, \ldots, L-1$. The proof of the last layer h_L is the same as the old proof.

Lemma 4.3 is not needed anymore because of (C.1). Lemma 4.4 becomes trivial to prove using (C.1): for instance for Lemma 4.4a, we have $\|\mathbf{D}_{i,L}\mathbf{W}_L\mathbf{D}_{i,L-1}(\mathbf{I}+\mathbf{W}_{L-1})\cdots\mathbf{D}_{i,a}\mathbf{W}_a u\| \leq O(\|u\|)$ and thus $\|\mathbf{B}\mathbf{D}_{i,L}\mathbf{W}_L\mathbf{D}_{i,L-1}(\mathbf{I}+\mathbf{W}_{L-1})\cdots\mathbf{D}_{i,a}\mathbf{W}_a u\| \leq O\big(\frac{\sqrt{s\log m}}{\sqrt{d}}\big)\|u\|$ for all s-sparse vectors u.

Lemma 4.5 needs the following changes in the same spirit as our changes to Lemma 4.1. With probability at least $1 - e^{-\Omega(\log^2 m)}$ it satisfies $\|\mathbf{W}_\ell h_{i,\ell}\|_\infty \leq O(\frac{\log m}{\sqrt{m}})$ for all $i \in [n]$ and $\ell \in L$. In the following proof we condition on this event happens.[10] Consider $i, j \in [n]$ with $i \neq j$.

- In the input layer, since $\|x_i - x_j\| \geq \delta$, the same Claim 4.6 shows that, with high probability, there are at least $\frac{3}{4}m$ coordinates $k \in [m]$ with $|(h_{i,0} - h_{j,0})_k| \geq \frac{\delta}{10\sqrt{m}}$. At the same time, at least $\frac{3}{4}m$ coordinates $k \in [m]$ will satisfy $(h_{i,0})_k \geq \frac{1}{10\sqrt{m}}$ and $(h_{j,0})_k \geq \frac{1}{10\sqrt{m}}$. Denote $M_0 \subseteq [m]$ as the set of coordinates k satisfying both properties. We have $|M_0| \geq \frac{m}{2}$ and $\sum_{k\in M_0} |(h_{i,0} - h_{j,0})_k| \geq \frac{\delta}{20}\sqrt{m}$.
- In the following layer $\ell = 1$, we have

  $$(h_{i,\ell} - h_{j,\ell})_k = \phi((h_{i,\ell-1})_k + \tau(\mathbf{W}_\ell h_{i,\ell-1})_k) - \phi((h_{j,\ell-1})_k + \tau(\mathbf{W}_\ell h_{j,\ell-1})_k)$$

  Using $\|\mathbf{W}_\ell h_{i,\ell}\|_\infty \leq O(\frac{\log m}{\sqrt{m}})$ and our choice of τ, we know for every $k \in M_0$, it satisfies $(h_{i,\ell})_k \geq \frac{1}{10\sqrt{m}} - \frac{1}{100L\sqrt{m}}$ and $(h_{j,\ell})_k \geq \frac{1}{10\sqrt{m}} - \frac{1}{100L\sqrt{m}}$. Therefore, the ReLU activation becomes identity for such coordinates $k \in M_0$ and

  $$\Delta_k \stackrel{\text{def}}{=} (h_{i,\ell} - h_{j,\ell})_k = (h_{i,\ell-1} - h_{j,\ell-1})_k + \tau(\mathbf{W}_\ell(h_{i,\ell-1} - h_{j,\ell-1}))_k \ .$$

  Let $s_k = 1$ if $(h_{i,\ell-1} - h_{j,\ell-1})_k \geq 0$ and $s_k = -1$ otherwise. Then,

  $$\sum_{k\in M_0} |\Delta_k| \geq \sum_{k\in M_0} s_k \cdot \Delta_k = \sum_{k\in M_0} |(h_{i,\ell-1} - h_{j,\ell-1})_k| + \tau \cdot s_k(\mathbf{W}_\ell(h_{i,\ell-1} - h_{j,\ell-1}))_k$$

  Note that when $h_{i,\ell-1}$ and $h_{j,\ell-1}$ are fixed, the values $s_k(\mathbf{W}_\ell(h_{i,\ell-1} - h_{j,\ell-1}))_k$ are independent

[10] For simplicity, we only show how to modify Lemma 4.5 with success probability $1 - e^{-\Omega(\log^2 m)}$ because that is all we need to the downstream application of Lemma 4.5. If one is willing to be more careful, the success probability can be much higher.

Gaussian with mean zero. This means, with probability at least $1-e^{-\Omega(\log^2 m)}$, the summation $\sum_{k\in M_0} s_k(\mathbf{W}_\ell(h_{i,\ell-1}-h_{j,\ell-1}))_k$ is at most $O(\log m)$ in absolute value. Putting this into the above equation, we have

$$\sum_{k\in M_0}|\Delta_k| \geq \sum_{k\in M_0}|(h_{i,\ell-1}-h_{j,\ell-1})_k| - O(\tau\log m) \geq \frac{\delta}{20}\sqrt{m} - O(\tau\log m) \ .$$

- Continuing this process for $\ell = 2,3,\dots,L-1$, we can conclude that $\sum_{k\in M_0}|(h_{i,L-1}-h_{j,L-1})_k| \geq \frac{\delta}{30}\sqrt{m}$ and therefore $\|h_{i,L-1}-h_{j,L-1}\| \geq \Omega(\delta^2)$. This is the same statement as before that we shall need for the downstream application of Lemma 4.5.

Changes to Section 5. Lemma 5.2 becomes easy to prove with all the L factors disappear for the following reason. Fixing i and ignoring the subscript in i, we have for $\ell = 1,2,\dots,L-1$:

$$\begin{aligned} h'_\ell &= \mathbf{D}''_\ell\big((\mathbf{I}+\tau\mathbf{W}_\ell+\tau\mathbf{W}'_\ell)h_{\ell-1} - (\mathbf{I}+\tau\mathbf{W}_\ell)h^{(0)}_{\ell-1}\big) \\ &= \mathbf{D}''_\ell\big((\mathbf{I}+\tau\mathbf{W}_\ell)h'_{\ell-1} + \tau\mathbf{W}'_\ell h_{\ell-1}\big) \end{aligned}$$

For some diagonal matrix $\mathbf{D}''_\ell \in \mathbb{R}^{m\times m}$ with diagonal entries in $[-1,1]$ (see Proposition 8.3). By simple spectral norm of matrices bound we have

$$\|h'_\ell\| \leq (1+\tau\|\mathbf{W}_\ell\|_2+\tau\|\mathbf{W}'_\ell\|_2)\|h'_{\ell-1}\|+\tau\|\mathbf{W}'_\ell\|_2\|h^{(0)}_{\ell-1}\| \leq (1+\frac{1}{10L})\|h'_{\ell-1}\|+O(\tau\omega) \leq \cdots \leq O(\tau\omega)$$

This implies $\|h'_\ell\|, \|g'_\ell\| \leq O(\tau\omega)$ for all $\ell\in[L-1]$, and combining with the old proof we have $\|h'_L\|, \|g'_L\| \leq O(\omega)$.

As for the sparsity $\|\mathbf{D}'_\ell\|_0$, because $g^{(0)}_\ell = h^{(0)}_{\ell-1} + \tau\mathbf{W}^{(0)}_\ell h^{(0)}_{\ell-1} \sim \mathcal{N}\big(h^{(0)}_{\ell-1}, \frac{2\tau^2\|h^{(0)}_{\ell-1}\|^2}{m}\big)$ and $\|g'_\ell\| \leq O(\tau\omega)$, applying essentially the same Claim 5.3, we have $\|\mathbf{D}'_\ell\|_0 \leq O(m\omega^{2/3})$ for every $\ell = 1,2,\dots,L-1$. One can similarly argue that $\|\mathbf{D}'_L\|_0 \leq O(m\omega^{2/3})$.

Next, Lemma 5.6 and Lemma 5.7 become trivial to prove (recall we have to change $\mathbf{W}^{(0)}_\ell$ with $\mathbf{I}+\tau\mathbf{W}^{(0)}_\ell$ for $\ell<L$) and the L factor also gets improved.

Changes to Section 6. The proofs of this section require only notational changes.

Final Theorem. Since Section 7 and 8 rely on previous sections, they do not need to be changed (besides improving polynomial factors in L). Our final theorem becomes

Theorem 6 (ResNet)**.** *Let $m \geq \widetilde{\Omega}\big(\mathsf{poly}(n,L,\mathfrak{d},\delta^{-1})\cdot d\cdot\log^2\varepsilon^{-1}\big)$. For the residual neural network defined in this section, with probability at least $1-e^{-\Omega(\log^2 m)}$ over the random initialization, GD needs at most $T = O\big(\frac{n^6L^2}{\delta^2}\log\frac{1}{\varepsilon}\big)$ iterations and SGD needs at most $T = O\big(\frac{n^7L^2\log^2 m}{b\delta^2}\log\frac{1}{\varepsilon}\big)$ iterations to find a point $F(\overrightarrow{\mathbf{W}}) \leq \varepsilon$.*

References

[1] Atiye Alaeddini, Siavash Alemzadeh, Afshin Mesbahi, and Mehran Mesbahi. Linear model regression on time-series data: Non-asymptotic error bounds and applications. *arXiv preprint arXiv:1807.06611*, 2018.

[2] Zeyuan Allen-Zhu and Yuanzhi Li. Follow the Compressed Leader: Faster Online Learning of Eigenvectors and Faster MMWU. In *ICML*, 2017. Full version available at `http://arxiv.org/abs/1701.01722`.

[3] Zeyuan Allen-Zhu and Yuanzhi Li. Neon2: Finding Local Minima via First-Order Oracles. In *NIPS*, 2018. Full version available at `http://arxiv.org/abs/1711.06673`.

[4] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. On the convergence rate of training recurrent neural networks. *arXiv preprint arXiv:1810.12065*, 2018.

[5] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in English and Mandarin. In *International Conference on Machine Learning (ICML)*, pages 173–182, 2016.

[6] Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. *arXiv preprint arXiv:1810.02281*, 2018.

[7] Sanjeev Arora, Elad Hazan, Holden Lee, Karan Singh, Cyril Zhang, and Yi Zhang. Towards provable control for unknown linear dynamical systems. 2018.

[8] Peter Bartlett, Dave Helmbold, and Phil Long. Gradient descent with identity initialization efficiently learns positive definite linear transformations. In *International Conference on Machine Learning (ICML)*, pages 520–529, 2018.

[9] Avrim L Blum and Ronald L Rivest. Training a 3-node neural network is np-complete. In *Machine learning: From theory to applications (A preliminary version of this paper was appeared in NIPS 1989)*, pages 9–28. Springer, 1993.

[10] Alon Brutzkus and Amir Globerson. Globally optimal gradient descent for a convnet with gaussian inputs. In *International Conference on Machine Learning (ICML)*. http://arxiv.org/abs/1702.07966, 2017.

[11] James V Burke, Adrian S Lewis, and Michael L Overton. A robust gradient sampling algorithm for nonsmooth, nonconvex optimization. *SIAM Journal on Optimization*, 15(3):751–779, 2005.

[12] Amit Daniely. Complexity theoretic limitations on learning halfspaces. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing (STOC)*, pages 105–117. ACM, 2016.

[13] Amit Daniely. SGD learns the conjugate kernel class of the network. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2422–2430, 2017.

[14] Amit Daniely and Shai Shalev-Shwartz. Complexity theoretic limitations on learning dnfs. In *Conference on Learning Theory (COLT)*, pages 815–830, 2016.

[15] Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *arXiv preprint arXiv:1710.01688*, 2017.

[16] Sarah Dean, Stephen Tu, Nikolai Matni, and Benjamin Recht. Safely learning to control the constrained linear quadratic regulator. *arXiv preprint arXiv:1809.10121*, 2018.

[17] Simon S. Du, Jason D. Lee, Yuandong Tian, Barnabás Póczos, and Aarti Singh. Gradient descent learns one-hidden-layer CNN: don't be afraid of spurious local minima. In *International Conference on Machine Learning (ICML)*. http://arxiv.org/abs/1712.00779, 2018.

[18] Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient Descent Provably Optimizes Over-parameterized Neural Networks. *ArXiv e-prints*, 2018.

[19] Vincent Fung. An overview of resnet and its variants. *https://towardsdatascience.com/an-overview-of-resnet-and-its-variants-5281e2f56035*, 2017.

[20] Rong Ge, Jason D. Lee, and Tengyu Ma. Learning one-hidden-layer neural networks with landscape design. In *ICLR*, 2017. URL `http://arxiv.org/abs/1711.00501`.

[21] Surbhi Goel, Varun Kanade, Adam Klivans, and Justin Thaler. Reliably learning the ReLU in polynomial time. In *Conference on Learning Theory (COLT)*, 2017.

[22] Surbhi Goel, Adam Klivans, and Raghu Meka. Learning one convolutional layer with overlapping patches. In *International Conference on Machine Learning (ICML)*. arXiv preprint arXiv:1802.02547, 2018.

[23] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6645–6649. IEEE, 2013.

[24] Moritz Hardt and Tengyu Ma. Identity matters in deep learning. In *ICLR*, 2017. URL `http://arxiv.org/abs/1611.04231`.

[25] Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *Journal of Machine Learning Research (JMLR)*, 19(29):1–44, 2018.

[26] Elad Hazan, Karan Singh, and Cyril Zhang. Learning linear dynamical systems via spectral filtering. In *Advances in Neural Information Processing Systems (NIPS)*, pages 6702–6712, 2017.

[27] Elad Hazan, Holden Lee, Karan Singh, Cyril Zhang, and Yi Zhang. Spectral filtering for general linear dynamical systems. In *Advances in Neural Information Processing Systems (NINPS)*, 2018.

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

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

[30] Adam R Klivans and Alexander A Sherstov. Cryptographic hardness for learning intersections of halfspaces. *Journal of Computer and System Sciences*, 75(1):2–12, 2009.

[31] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

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

[33] Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with ReLU activation. In *Advances in Neural Information Processing Systems (NIPS)*. http://arxiv.org/abs/1705.09886, 2017.

[34] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[35] Roi Livni, Shai Shalev-Shwartz, and Ohad Shamir. On the computational efficiency of training neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 855–863, 2014.

[36] Pasin Manurangsi and Daniel Reichman. The computational complexity of training ReLU(s). *arXiv preprint arXiv:1810.04207*, 2018.

[37] Jakub Marecek and Tigran Tchrakian. Robust spectral filtering and anomaly detection. *arXiv preprint arXiv:1808.01181*, 2018.

[38] Yurii Nesterov. *Introductory Lectures on Convex Programming Volume: A Basic course*, volume I. Kluwer Academic Publishers, 2004. ISBN 1402075537.

[39] Samet Oymak. Learning compact neural networks with regularization. *arXiv preprint arXiv:1802.01223*, 2018.

[40] Samet Oymak and Necmiye Ozay. Non-asymptotic identification of LTI systems from a single trajectory. *arXiv preprint arXiv:1806.05722*, 2018.

[41] Rina Panigrahy, Ali Rahimi, Sushant Sachdeva, and Qiuyi Zhang. Convergence results for neural networks via electrodynamics. In *ITCS*, 2018.

[42] Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer ReLU neural networks. In *International Conference on Machine Learning (ICML)*. http://arxiv.org/abs/1712.08968, 2018.

[43] Ohad Shamir. A variant of azuma's inequality for martingales with subgaussian tails. *ArXiv e-prints*, abs/1110.2392, 10 2011.

[44] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529 (7587):484–489, 2016.

[45] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354, 2017.

[46] Max Simchowitz, Horia Mania, Stephen Tu, Michael I Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference on Learning Theory (COLT)*. arXiv preprint arXiv:1802.08334, 2018.

[47] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[48] Mahdi Soltanolkotabi. Learning ReLUs via gradient descent. *CoRR*, abs/1705.04591, 2017. URL `http://arxiv.org/abs/1705.04591`.

[49] Le Song, Santosh Vempala, John Wilmes, and Bo Xie. On the complexity of learning neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 5514–5522, 2017.

[50] Daniel Soudry and Yair Carmon. No bad local minima: Data independent training error guarantees for multilayer neural networks. *arXiv preprint arXiv:1605.08361*, 2016.

[51] Rupesh K Srivastava, Klaus Greff, and Jürgen Schmidhuber. Training very deep networks. In *Advances in neural information processing systems (NIPS)*, pages 2377–2385, 2015.

[52] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[53] Yuandong Tian. An analytical formula of population gradient for two-layered ReLU network and its applications in convergence and critical point analysis. In *International Conference on Machine Learning (ICML)*. http://arxiv.org/abs/1703.00560, 2017.

[54] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[55] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR)*, 2017.

[56] Jiong Zhang, Yibo Lin, Zhao Song, and Inderjit S Dhillon. Learning long term dependencies via Fourier recurrent units. In *International Conference on Machine Learning (ICML)*. arXiv preprint arXiv:1803.06585, 2018.

[57] Kai Zhong, Zhao Song, and Inderjit S Dhillon. Learning non-overlapping convolutional neural networks with multiple kernels. *arXiv preprint arXiv:1711.03440*, 2017.

[58] Kai Zhong, Zhao Song, Prateek Jain, Peter L Bartlett, and Inderjit S Dhillon. Recovery guarantees for one-hidden-layer neural networks. In *International Conference on Machine Learning (ICML)*. arXiv preprint arXiv:1706.03175, 2017.