On the Convergence Rate of Training Recurrent Neural Networks

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

Yuanzhi Li
yuanzhil@princeton.edu
Princeton University

Zhao Song
zhaos@utexas.edu
UT-Austin

October 28, 2018

(version 2)*

Abstract

Despite the huge success of deep learning, our understanding to how the non-convex neural networks are trained remains rather limited. Most of existing theoretical works only tackle neural networks with one hidden layer, and little is known for multi-layer neural networks.

Recurrent neural networks (RNNs) are special multi-layer networks extensively used in natural language processing applications. They are particularly hard to analyze, comparing to feedforward networks, because the weight parameters are reused across the entire time horizon.

We provide arguably the first theoretical understanding to the convergence speed of training RNNs (when activation functions are present). Specifically, when the weights are randomly initialized, and when the number of neurons is sufficiently large —meaning polynomial in the training data size and the time horizon— we show that gradient descent and stochastic gradient descent both minimize the regression loss in a linear convergence rate, that is, $\varepsilon \propto e^{-\Omega(T)}$.

*Version 1 appears on this date. Version 2 adds some missing citations of prior work. 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, Sébastien Bubeck and Ofer Dekel for useful conversations.

1 Introduction

Neural networks have been one of the most powerful tools in machine learning over the past a few decades [2, 34, 42, 50, 55, 83, 84]. The multi-layer structure of neural network gives it supreme power in expressibility and learning performance. However, it also raises complexity concerns: the training objective involving multi-layer (or even just two-layer) neural networks, equipped with non-linear activation functions, is generally neither convex nor concave. However, it is still observed that in practice, local-search algorithms such as stochastic gradient descent (SGD) is capable of finding globally optimal solutions, at least on the training data set [102].

In recent years, there have been a number of theoretical results aiming at a better understanding of this phenomenon. Many of them focus on two-layer (thus one-hidden-layer) neural networks and assume that the inputs are random Gaussian vectors [11, 25, 30, 54, 68, 86, 94, 105, 106]. Some study deep neural networks but assuming the activation function is linear [6, 9, 38]. On the technique side, some of these results try to understand the gradient dynamics [11, 12, 20, 25, 54, 68, 86, 94, 96, 105, 106], while others focus on the geometry properties of the training objective [28, 30, 38, 75, 107].

More recently, Safran and Shamir [75] provided evidence that, even when inputs are standard Gaussians, two-layer neural networks can indeed have spurious local minima, and suggested that *over-parameterization* (i.e., increasing the number of neurons) may be the key in avoiding spurious local minima. Later, Li and Liang [53] showed that, for two-layer networks with the cross-entropy loss, in the over-parametrization regime, gradient descent is capable of finding nearly-global optimal solutions on the training data. This result was later extended to the ℓ_2 loss by Du et al. [26].

Recurrent Neural Networks. Among different variations of neural networks, one of the *least* theoretically-understood structure is the recurrent one [27]. A recurrent neural network (RNN) recurrently applies the same network unit to a sequence of input points, such as a sequence of words in a language sentence. RNN is particularly useful when there are long-term, non-linear interactions between input points in the same sequence. These networks are widely used in practice for natural language processing, language generation, machine translation, speech recognition, video and music processing, and many other tasks [14, 16, 46, 61, 62, 76, 90, 92]. On the theory side, while there are some attempts to show that an RNN is more expressive than a feedforward neural network [48], when and how an RNN can be efficiently learned has nearly-zero theoretical explanation.

In practice, RNN is usually trained by simple local-search algorithms such as stochastic gradient descent, via back-propagation through time (BPTT) [98]. However, unlike shallow networks, the training process of RNN often runs into the trouble of vanishing or exploding gradient [91]. That is, the value of the gradient becomes exponentially small or large in the time horizon, even when the training objective is still constant.[1] In practice, one of the most popular way to resolve this issue is by the long short term memory (LSTM) structure [43]. However, one can also use rectified linear units (ReLUs) as activation functions to avoid vanishing or exploding gradient [77]. In fact, one of the earliest adoptions of ReLUs was on applications of RNNs for this purpose twenty years ago [37, 78].

Since the RNN structure was proposed, a large number of variations have been designed over

[1]Intuitively, an RNN recurrently applies the same network unit for L times if the input sequence is of length L. When this unit has "operator norm" larger than one or smaller than one, the final output can possibly exponentially explode or vanish in L. More importantly, when BPTT back propagates through time —which intuitively corresponds to applying the reverse unit multiple times— the gradient can also vanish or explode. Controlling the operator norm of a non-linear operator can be quite challenging.

past decades, including LSTM [43], Bidirectional RNN [80], Bidirectional LSTM [33], gate recurrent unit [15], statistical recurrent unit [66], Fourier recurrent unit [103]. For a more detailed survey, we refer the readers to Salehinejad et al. [77].

1.1 Our Question

In this paper, we study the following general question

- *Can ReLU provably stabilize the training process and avoid vanishing/exploding gradient?*
- *Can RNN be trained close to zero training error* efficiently *under mild assumptions?*

Remark 1.1. When there is no activation function, RNN is known as *linear dynamical system.* Hardt, Ma and Recht [39] first proved the convergence of finding global minima for such linear dynamical systems. Followups in this line of research include [40, 41].

Motivation. Indeed, the ultimate question in this line would be whether RNN can be trained close to zero generalization error. However, unlike feedforward neural networks, for RNNs, the training error, or the ability to memorize examples may actually be desirable. After all, many tasks involving RNN are related to memories, and certain RNN units are also referred to memory cells. Since RNN applies the same network unit to all the input points in a sequence, the following question can possibly of its own interest:

- *How does RNN learn a new mapping (say from one input point to its output in 3 time steps), without destroying other mappings?*

Another motivation is the following. An RNN can be viewed as a space constraint, differentiable Turing machine, except that the input is only allowed to be read in a fixed order. It was shown in Siegelmann and Sontag [82] that all Turing machines can be simulated by fully-connected recurrent networks built of neurons with non-linear activation functions. In practice, RNN is also used as a tool to build the neural Turing machine [35], equipped with a grand goal of automatically learning an algorithm based on the observation of the inputs and outputs. To this extent, we believe the task of understanding the *trainability* as a first step towards understanding RNN can be meaningful on its own.

Our Result. We answer this question positively in this paper. To present the simplest result, we focus on the classical Elman network with ReLU activation:

$$h_\ell = \phi(W \cdot h_{\ell-1} + Ax_\ell) \in \mathbb{R}^m \qquad \text{where } W \in \mathbb{R}^{m\times m},\ A \in \mathbb{R}^{m\times d_x}$$
$$y_\ell = B \cdot h_\ell \in \mathbb{R}^m \qquad \text{where } B \in \mathbb{R}^{d\times m}$$

We let $h_0 = 0$ and denote by ϕ the ReLU activation function: $\phi(x) = \max(x, 0)$.

We consider a regression task where each sequence of inputs consists of vectors $x_1, \dots, x_L \in \mathbb{R}^{d_x}$ and we perform regression with respect to $y^*_1, \dots, y^*_L \in \mathbb{R}^d$. The least-square regression loss with respect to this input is $\frac{1}{2}\sum_{\ell=1}^{L} \|Bh_\ell - y^*_\ell\|^2$. We assume there are n training sequences, each of length L. We assume the training sequences are δ-separable (say for instance vectors x_1 are different by relative distance $\delta > 0$ for every pairs of training sequences). Our main theorem can be stated as follows

Theorem. *If the number of neurons $m \geq \mathsf{poly}(n, d, L, \delta^{-1}, \log \varepsilon^{-1})$ is polynomially large, we can find weight matrices W, A, B where the RNN gives ε training error*

- *if gradient descent (GD) is applied for $T = \Omega\big(\frac{\mathsf{poly}(n,d,L)}{\delta^2}\log\frac{1}{\varepsilon}\big)$ iterations, starting from random Gaussian initializations; or*

- *if (mini-batch or regular) stochastic gradient descent (SGD) is applied for* $T = \Omega\big(\frac{\mathsf{poly}(n,d,L)}{\delta^2}\log\frac{1}{\varepsilon}\big)$ *iterations, starting from random Gaussian initializations.*[2]

(To present the simplest possible result, we have not tried to tighten the polynomial dependency with respect to n, d and L. We only tightened the dependency with respect to δ and ε.)

We believe this is the first proof of convergence of GD or SGD for recurrent neural networks with activation functions, and possibly the first proof of finding approximate *global optima* on the RNN training objective when activation functions are present.

Remark 1.2. Our theorem does not exclude the possible existence of (spurious) local minima.

Extension: Deep RNN. One may also wish to study the convergence behavior of RNNs with multiple layers of hidden neurons. This is sometimes referred to as deep RNN or RNN with multi-layer perceptron [77]. Our theorem also generalizes to deep RNNs, but we do not include it in this version of the paper, because the convergence of Elman RNN is already quite involved to prove.

1.2 Other Related Works

Another relevant work is Brutzkus et al. [12] where the authors studied over-paramterization in the case of two-layer neural network under a linear-separable assumption.

Some other works learn neural networks using more delicate algorithms —such as tensor decomposition [5, 17, 36, 63, 95] or tailored algorithm for the problem structure [4, 32, 104]— or design more special activation functions [3, 64, 68, 87]. In this paper, instead, we focus on why basic algorithms such as GD or SGD can already almost surely find global optima of the objective.

Instead of using randomly initialized weights like this paper, there is a line of work proposing algorithms using weights generated from some "tensor initialization" process [4, 45, 81, 94, 106].

There is huge literature on using the mean-field theory to study neural networks [13, 22, 52, 60, 69–71, 74, 79, 99–101]. At a high level, they study the network dynamics at random initialization when the number of hidden neurons grow to infinity, and use such initialization theory to predict performance after training. However, they do not provide theoretical convergence rate for the training process (at least when the number of neurons is finite).

Some other works focus on the expressibility of neural networks, that is, to show that there exist certain weights so that the network calculates interesting functions. These results do not usually cover the theoretical issue of how to find such weight parameters [8, 18, 22, 29, 44, 65, 72, 74, 93].

There is a long line of research focusing on the hardness result for neural network [10, 19, 21, 31, 47, 49, 56, 58, 88, 97].

Linear dynamical systems is an important topic on its own, and are also related to linear variants of reinforcement learning. Some recent works on this topic include [1, 7, 23, 24, 59, 67, 85].

---

[2] At a first glance, one may question how it is possible for SGD to enjoy a logarithmic time dependency in ε^{-1}; after all, even when minimizing strongly-convex and Lipschitz-smooth functions, the typical convergence rate of SGD is $T \propto 1/\varepsilon$ as opposed to $T \propto \log(1/\varepsilon)$.

We quickly point out there is no contradiction here if the stochastic pieces of the objective enjoy a *common* global minimizer. In math terms, suppose we want to minimize some function $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$, and suppose x^* is the global minimizer of convex functions $f_1(x), \dots, f_n(x)$. Then, if $f(x)$ is σ-strongly convex, and each each $f_i(x)$ is L-Lipschitz smooth, then SGD —moving in negative direction of $\nabla f_i(x)$ for a random $i \in [n]$ per step— can find ε-minimizer of this function in $O\big(\frac{L^2}{\sigma^2}\log\frac{1}{\varepsilon}\big)$ iterations.

2 Notations and Preliminaries

Notations. We denote by $\|\cdot\|_2$ (or sometimes $\|\cdot\|$) the Euclidean norm of vectors or spectral norm of matrices. We denote by $\|\cdot\|_\infty$ the infinite norm of vectors, $\|\cdot\|_0$ the sparsity of vectors or diagonal matrices, and $\|\cdot\|_F$ the Frobenius norm of matrices. We use $\mathcal{N}(\mu,\sigma)$ to denote Gaussian distribution with mean μ and variance σ; or $\mathcal{N}(\mu,\Sigma)$ to denote Gaussian vector with mean μ and covariance Σ.

We use $\mathbf{1}_{event}$ to denote the indicator function of whether *event* is true. We denote by $\mathbb{1}_k$ the k-th standard basis vector. We use $\phi(\cdot)$ to denote the ReLU function, namely $\phi(x)=\max\{x,0\}=\mathbf{1}_{x\geq 0}\cdot x$. Given univariate function $f\colon \mathbb{R}\to\mathbb{R}$, slightly abusing notation, we also use f to denote the same function over vectors: $f(x)=(f(x_1),\dots,f(x_m))$ if $x\in\mathbb{R}^m$.

Given matrix W, we denote by W_k the k-th column vector of W. Therefore, $(Wx)_k=\langle W_k,x\rangle$.

Given vectors $v_1,\dots,v_n\in\mathbb{R}^m$, we define $U=\mathsf{GS}(v_1,\dots,v_n)$ as their Gram-Schmidt orthonormalization. Namely, $U=[\widehat{v}_1,\dots,\widehat{v}_n]\in\mathbb{R}^{m\times n}$ where

$$\widehat{v}_1=\frac{v_1}{\|v_1\|}\quad\text{and}\quad\text{for } i\geq 2\text{:}\quad \widehat{v}_i=\frac{\prod_{j=1}^{i-1}(I-\widehat{v}_j\widehat{v}_j^\top)v_i}{\|\prod_{j=1}^{i-1}(I-\widehat{v}_j\widehat{v}_j^\top)v_i\|}.$$

Note that in the occasion that $\prod_{j=1}^{i-1}(I-\widehat{v}_j\widehat{v}_j^\top)v_i$ is the zero vector, we let $\widehat{v}_i$ be an arbitrary unit vector that is orthogonal to $\widehat{v}_1,\dots,\widehat{v}_{i-1}$.

2.1 Elman Recurrent Neural Network

We assume that n training inputs are given, in the form of $(x_{i,1},x_{i,2},\dots,x_{i,L})\in\left(\mathbb{R}^{d_x}\right)^L$ for each input $i\in[n]$. We assume n training labels are given, in the form of $(y^*_{i,1},y^*_{i,2},\dots,y^*_{i,L})\in\left(\mathbb{R}^{d}\right)^L$ for each input $i\in[n]$. Without loss of generality, we assume $\|x_{i,1}\|=1$ and $\|x_{i,\ell}\|\leq 1$ for every $i\in[n]$ and $\ell\in\{2,3,\dots,L\}$. We make the following assumption on the input data (see Footnote 6 for how to relax it):

Assumption 2.1. *$\|x_{i,1}-x_{j,1}\|\geq\delta$ for some parameter $\delta\in(0,1]$ and every pair of $i\neq j\in[n]$.*

Given weight matrices $W\in\mathbb{R}^{m\times m}$, $A\in\mathbb{R}^{m\times d_x}$, $B\in\mathbb{R}^{d\times m}$, we introduce the following notations to describe the evaluation of RNN on the input sequences. For each $i\in[n]$ and $j\in[L]$:

$$h_{i,0}=0\in\mathbb{R}^m \qquad g_{i,\ell}=W\cdot h_{i,\ell-1}+Ax_{i,\ell}\in\mathbb{R}^m$$
$$y_{i,\ell}=B\cdot h_{i,\ell}\in\mathbb{R}^d \qquad h_{i,\ell}=\phi(W\cdot h_{i,\ell-1}+Ax_{i,\ell})\in\mathbb{R}^m$$

A very important notion that this entire paper relies on is the following:

Definition 2.2. *For each $i\in[n]$ and $\ell\in[L]$, let $D_{i,\ell}\in\mathbb{R}^{m\times m}$ be the diagonal matrix where*

$$(D_{i,\ell})_{k,k}=\mathbf{1}_{(W\cdot h_{i,\ell-1}+Ax_{i,\ell})_k\geq 0}=\mathbf{1}_{(g_{i,\ell})_k\geq 0}\ .$$

As a result, we can write $h_{i,\ell}=D_{i,\ell}Wh_{i,\ell-1}$.

We consider the following random initialization distributions for W, A and B.

Definition 2.3. *We say that W,A,B are at random initialization, if the entries of W and A are i.i.d. generated from $\mathcal{N}(0,\frac{2}{m})$, and the entries of $B_{i,j}$ are i.i.d. generated from $\mathcal{N}(0,\frac{1}{d})$.*

Throughout this paper, for notational simplicity, we refer to index ℓ as the ℓ-th *layer* of RNN, and $h_{i,\ell}$, $x_{i,\ell}$, $y_{i,\ell}$ respectively as the hidden neurons, input, output on the ℓ-th layer. We acknowledge that in certain literatures, one may regard Elman network as a three-layer neural network.

Assumption 2.4. *We assume* $m \geq \mathsf{poly}(n, d, L, \frac{1}{\delta}, \log \frac{1}{\varepsilon})$ *for some sufficiently large polynomial.*

Without loss of generality, we assume $\delta \leq \frac{1}{CL^2 \log^3 m}$ for some sufficiently large constant C (if this is not satisfied one can decrease δ). Throughout the paper except the detailed proofs in the appendix, we use $\widetilde{O}$, $\widetilde{\Omega}$ and $\widetilde{\Theta}$ notions to hide polylogarithmic dependency in m. To simplify notations, we denote by

$$\rho \stackrel{\text{def}}{=} nLd \log m \quad \text{and} \quad \varrho \stackrel{\text{def}}{=} nLd\delta^{-1} \log(m/\varepsilon) \ .$$

2.2 Objective and Gradient

For simplicity, we only optimize over the weight matrix $W \in \mathbb{R}^{m \times m}$ and let A and B be at random initialization. As a result, our ℓ_2-regression objective is a function over W:[3]

$$f(W) \stackrel{\text{def}}{=} \sum_{i=1}^{n} f_i(W) \quad \text{and} \quad f_i(W) \stackrel{\text{def}}{=} \frac{1}{2} \sum_{\ell=2}^{L} \| \mathsf{loss}_{i,\ell} \|_2^2 \quad \text{where} \quad \mathsf{loss}_{i,\ell} \stackrel{\text{def}}{=} Bh_{i,\ell} - y^*_{i,\ell} \ . \tag{2.1}$$

Using chain rule, one can write down a closed form of the (sub-)gradient:

Fact 2.5. *For $k \in [m]$, the gradient with respect to W_k (denoted by ∇_k) and the full gradient are*

$$\nabla_k f(W) = \sum_{i=1}^{n} \sum_{a=2}^{L} \sum_{\ell=1}^{a-1} (\mathsf{Back}_{i,\ell+1 \to a}^\top \cdot \mathsf{loss}_{i,a})_k \cdot h_{i,\ell} \cdot \mathbf{1}_{\langle W_k, h_{i,\ell} \rangle + \langle A_k, x_{i,\ell+1} \rangle \geq 0}$$

$$\nabla f(W) = \sum_{i=1}^{n} \sum_{a=2}^{L} \sum_{\ell=1}^{a-1} D_{i,\ell+1} \big(\mathsf{Back}_{i,\ell+1 \to a}^\top \cdot \mathsf{loss}_{i,a} \big) \cdot h_{i,\ell}^\top$$

where for every $i \in [n]$, $\ell \in [L]$, and $a = \ell+1, \ell+2, \dots, L$:

$$\mathsf{Back}_{i,\ell \to \ell} \stackrel{\text{def}}{=} B \in \mathbb{R}^{d \times m} \quad \text{and} \quad \mathsf{Back}_{i,\ell \to a} \stackrel{\text{def}}{=} BD_{i,a}W \cdots D_{i,\ell+1}W \in \mathbb{R}^{d \times m} \ .$$

3 Our Results

Our main results can be formally stated as follows.

Theorem 1 (GD)**.** *Suppose* $\eta = \widetilde{\Theta}\big(\frac{\delta}{m} \mathsf{poly}(n, d, L)\big)$ *and* $m \geq \mathsf{poly}(n, d, L, \delta^{-1}, \log \varepsilon^{-1})$*. Let* $W^{(0)}, A, B$ *be at random initialization. With high probability over the randomness of* $W^{(0)}, A, B$*, if we apply gradient descent for* T *steps* $W^{(t+1)} = W^{(t)} - \eta \nabla f(W^{(t)})$*, then it satisfies*

$$f(W^{(T)}) \leq \varepsilon \text{ for } T = \widetilde{\Omega}\left(\frac{\mathsf{poly}(n, d, L)}{\delta^2} \log \frac{1}{\varepsilon} \right).$$

Theorem 2 (SGD)**.** *Suppose* $\eta = \widetilde{\Theta}\big(\frac{\delta}{m} \mathsf{poly}(n, d, L)\big)$ *and* $m \geq \mathsf{poly}(n, d, L, \delta^{-1}, \log \varepsilon^{-1})$*. Let* $W^{(0)}, A, B$ *be at random initialization. If we apply stochastic gradient descent for* T *steps* $W^{(t+1)} = W^{(t)} - \eta \nabla f_i(W^{(t)})$ *for a random index* $i \in [n]$ *per step, then with high probability (over* $W^{(0)}, A, B$ *and the randomness of SGD), it satisfies*

$$f(W^{(T)}) \leq \varepsilon \text{ for } T = \widetilde{\Omega}\left(\frac{\mathsf{poly}(n, d, L)}{\delta^2} \log \frac{1}{\varepsilon} \right).$$

[3]The index ℓ starts from 2, because $Bh_{i,1} = B\phi(Ax_{i,1})$ remains constant if we are not optimizing over A and B.

In both cases, we essentially have *linear convergence rates.*[4] Notably, our results show that the dependency of the number of layers L, is *polynomial.* Thus, even when RNN is applied to sequences of long input data, it does not suffer from exponential gradient explosion or vanishing (e.g., $2^{\Omega(L)}$ or $2^{-\Omega(L)}$) through the entire training process.

4 Main Technical Theorems

Our main Theorem 1 and Theorem 2 are in fact natural consequences of the following two technical theorems. They both talk about the first-order behavior of RNNs when the weight matrix W is sufficiently close to some random initialization.

The first theorem is similar to the classical Polyak-Łojasiewicz condition [57, 73], and says that $\|\nabla f(W)\|_F^2$ is at least as large as the objective value.

Theorem 3. *With high probability over random initialization $\widetilde{W}, A, B$, it satisfies*

$$\forall W \in \mathbb{R}^{m\times m} \text{ with } \|W-\widetilde{W}\|_2 \le \frac{\mathsf{poly}(\varrho)}{\sqrt{m}}: \qquad \|\nabla f(W)\|_F^2 \ge \frac{\delta}{\mathsf{poly}(\rho)} \times m \times f(W) \ ,$$
$$\|\nabla f(W)\|_F^2, \|\nabla f_i(W)\|_F^2 \le \mathsf{poly}(\rho) \times m \times f(W) \ .$$

(Only the first statement is the Polyak-Łojasiewicz condition; the second is a simple-to-proof gradient upper bound.) The second theorem shows a special smoothness property of the objective.

Theorem 4. *With high probability over random initialization $\widetilde{W}, A, B$, it satisfies for every $\breve{W} \in \mathbb{R}^{m\times m}$ with $\|\breve{W}-\widetilde{W}\| \le \frac{\mathsf{poly}(\varrho)}{\sqrt{m}}$, and for every $W' \in \mathbb{R}^{m\times m}$ with $\|W'\| \le \frac{\tau_0}{\sqrt{m}}$,*

$$f(\breve{W}+W') \le f(\breve{W}) + \langle \nabla f(\breve{W}), W'\rangle + \mathsf{poly}(\varrho) m^{1/3} \cdot \sqrt{f(\breve{W})} \cdot \|W'\|_2 + \mathsf{poly}(\rho) m \|W'\|_2^2 \ .$$

At a high level, the convergence of GD and SGD are careful applications of Theorem 3 and 4 stated above (see Appendix H and Appendix I respectively).

The main difficulty of this paper is to prove Theorem 3 and 4, and we shall sketch the proof ideas in Section 5 through 8. In such high-level discussions, we shall put our emphasize on

- how to avoid exponential blow up in L, and
- how to deal with the issue of randomness dependence across layers.

5 Basic Properties at Random Initialization

In this section we derive basic properties of the RNN when the weight matrices W, A, B are all *at random initialization.* The corresponding precise statements and proofs are in Appendix B.

The first one says that the forward propagation neither explodes or vanishes, that is,

$$\frac{1}{2} \le \|h_{i,\ell}\|_2, \|g_{i,\ell}\|_2 \le O(L) \ . \tag{5.1}$$

Intuitively, (5.1) very reasonable. Since the weight matrix W is randomly initialized with entries i.i.d. from $\mathcal{N}\left(0, \frac{2}{m}\right)$, the norm $\|Wz\|_2$ is around $\sqrt{2}$ for any fixed vector z. Equipped with ReLU activation, it "shuts down" roughly half of the coordinates of Wz and reduces the norm $\|\phi(Wz)\|$ to one. Since in each layer ℓ, there is an additional unit-norm signal $x_{i,\ell}$ coming in, we should expect the final norm of hidden neurons to be at most $O(L)$.

[4] We remark here that the $\widetilde{O}$ notation may hide additional polynomial dependency in $\log\log\varepsilon^{-1}$. This may not be necessary, and we have not tried to tighten such log-log dependency in ε^{-1}.

Unfortunately, the above argument cannot be directly applied since the weight matrix W is reused for L times so there is no fresh new randomness across layers. Let us explain how we deal with this issue carefully, because it is at the heart of *all* of our proofs in this paper. Recall, each time W is applied to some vector $h_{i,\ell}$, it only uses "one column of randomness" of W. Mathematically, letting $U_\ell \in \mathbb{R}^{m\times n\ell}$ denote the column orthonormal matrix using Gram-Schmidt

$$U_\ell \stackrel{\text{def}}{=} \mathsf{GS}\left(h_{1,1},\dots,h_{n,1},h_{1,2},\dots,h_{n,2},\dots,h_{1,\ell},\dots,h_{n,\ell}\right) \enspace ,$$

we have $Wh_{i,\ell} = WU_{\ell-1}U_{\ell-1}^\top h_{i,\ell} + W(I-U_{\ell-1}U_{\ell-1}^\top)h_{i,\ell}$.

- The second term $W(I-U_{\ell-1}U_{\ell-1}^\top)h_{i,\ell}$ has new randomness independent of the previous layers.[5]
- The first term $WU_{\ell-1}U_{\ell-1}^\top h_{i,\ell}$ relies on the randomness of W in the directions of $h_{i,a}$ for $a<\ell$ of the previous layers. We cannot rely on the randomness of this term, because when applying inductive argument till layer ℓ, the randomness of $WU_{\ell-1}$ is already used.

  Fortunately, $WU_{\ell-1} \in \mathbb{R}^{m\times n(\ell-1)}$ is a rectangular matrix with $m \gg n(\ell-1)$ (thanks to overparameterization!) so one can bound its spectral norm by roughly $\sqrt{2}$. This ensures that no matter how $h_{i,\ell}$ behaves (even arbitrarily correlated with $WU_{\ell-1}$), the norm of the first term cannot be too large. It is crucial here that $WU_{\ell-1}$ is a *rectangular* matrix, because for a square random matrix such as W, its spectral norm is 2 and using that, the forward propagation bound will exponentially blow up.

This summarizes the main idea for proving $\|h_{i,\ell}\| \leq O(L)$ in (5.1); the lower bound $\frac{1}{2}$ can be similarly argued.

Our next property says that in each layer, the amount of "fresh new randomness" is non-negligible, that is,

$$\|(I-U_{\ell-1}U_{\ell-1}^\top)h_{i,\ell}\|_2 \geq \widetilde{\Omega}(\frac{1}{L^2}) \enspace . \tag{5.2}$$

This relies on a more involved inductive argument than (5.1). At high level, one needs to show that in each layer, the amount of "fresh new randomness" reduces only by a factor at most $1-\frac{1}{L}$.

Using (5.1) and (5.2), we obtain the following interesting property about the data separability:

$$(I-U_\ell U_\ell^\top)h_{i,\ell+1} \text{ and } (I-U_\ell U_\ell^\top)h_{j,\ell+1} \text{ are } (\delta/2)\text{-separable}, \forall i,j\in[n] \text{ with } i\neq j \tag{5.3}$$

Here, we say two vectors x and y are δ-separable if $\left\|(I-yy^\top/\|y\|_2^2)x\right\| \geq \delta$ and vice versa.

We prove (5.3) by induction. In the first layer $\ell=1$ we have $h_{i,1}$ and $h_{j,1}$ are δ-separable which is a consequence of Assumption 2.1. If having fresh new randomness, given two δ separable vectors x,y, one can show that $\phi(Wx)$ and $\phi(Wy)$ are also $\delta(1-o(\frac{1}{L}))$-separable. Again, in RNN, we do not have fresh new randomness, so we rely on (5.2) to give us reasonably large fresh new randomness. Applying a careful induction helps us to derive that (5.3) holds for all layers.[6]

Intermediate Layers and Backward Propagation. Training RNN (or any neural network) is not only about forward propagation. We also have to bound its behavior in intermediate layers and in backward propagation.

[5] More precisely, letting $v=(I-U_{\ell-1}U_{\ell-1}^\top)h_{i,\ell}$, we have $W(I-U_{\ell-1}U_{\ell-1}^\top)h_{i,\ell} = \left(W\frac{v}{\|v\|}\right)\|v\|$. Here, $W\frac{v}{\|v\|}$ is a random Gaussian vector in $\mathcal{N}(0,\frac{2}{m}I)$ and is *independent* of all $\{h_{i,a} \mid i\in[n], a<\ell\}$.

[6] This is the only place that we rely on Assumption 2.1. This assumption is somewhat necessary in the following sense. If $x_{i,\ell}=x_{j,\ell}$ for some pair $i\neq j$ for all the first ten layers $\ell=1,2,\dots,10$, and if $y^*_{i,\ell}\neq y^*_{i,\ell}$ for even just one of these layers, then there is no hope in having the training objective decrease to zero. Of course, one can make more relaxed assumption on the input data, involving both $x_{i,\ell}$ and $y^*_{i,\ell}$. While this is possible, it complicates the statements so we do not present such results in this paper.

The first two results we derive are the following. For every $\ell_1 \geq \ell_2$ and diagonal matrices D' of sparsity $s \in [\rho^2, m^{0.49}]$:

$$\|WD_{i,\ell_1} \cdots D_{i,\ell_2}W\|_2 \leq O(L^3) \tag{5.4}$$

$$\left\|D'WD_{i,\ell_1} \cdots D_{i,\ell_2}WD'\right\|_2 \leq \widetilde{O}(\sqrt{s}/\sqrt{m}) \tag{5.5}$$

Intuitively, one cannot use spectral bound argument to derive (5.4) or (5.5): the spectral norm of W is 2, and even if ReLU activations (i.e., the $D_{i,\ell}$ matrices) cancel half of its ℓ_2^2 mass, the spectral norm remains to be $\sqrt{2}$. When multiplied together, this grows exponential in L.

Instead, we use an analogous argument to (5.1) to show that, for each fixed vector z, the norm of $\|WD_{i,\ell_1} \cdots D_{i,\ell_2}Wz\|_2$ is at most $O(1)$ with extremely high probability $1 - e^{-\Omega(m/L^2)}$. By standard ε-net argument, $\|WD_{i,\ell_1} \cdots D_{i,\ell_2}Wz\|_2$ is at most $O(1)$ for all $\frac{m}{L^3}$-sparse vectors z. Finally, for a possible dense vector z, we can divide it into L^3 chunks each of sparsity $\frac{m}{L^3}$. Finally, we apply the upper bound for L^3 times. This proves Lemma 5.4. One can use similar argument to prove (5.5).

Remark 5.1. We did not try to tighten the polynomial factor here in L. We conjecture that proving an $O(1)$ bound may be possible, but that question itself may be a sufficiently interesting random matrix theory problem on its own.

The next result is for back propagation. For every $\ell_1 \geq \ell_2$ and diagonal matrices D' of sparsity $s \in [\rho^2, m^{0.49}]$:

$$\left\|BD_{i,\ell_1} \cdots D_{i,\ell_2}WD'\right\|_2 \leq \widetilde{O}(\sqrt{s}) \tag{5.6}$$

Its proof is in the same spirit as (5.5), with the only difference being the spectral norm of B is around $\sqrt{m/d}$ as opposed to $O(1)$.

6 Stability After Adversarial Perturbation

In this section we study the behavior of RNN after adversarial perturbation. The corresponding precise statements and proofs are in Appendix C.

Letting $\widetilde{W}, A, B$ be at random initialization, we consider some matrix $W = \widetilde{W} + W'$ for $\|W'\|_2 \leq \frac{\mathsf{poly}(\varrho)}{\sqrt{m}}$. Here, W' may depend on the randomness of $\widetilde{W}, A$ and B, so we say it can be *adversarially* chosen. The results of this section will later be applied essentially twice:

- Once for those updates generated by GD or SGD, where W' is how much the algorithm has moved away from the random initialization.
- The other time (see Section 7.3) for a technique that we call "randomness decomposition" where we decompose the true random initialization W into $W = \widetilde{W} + W'$, where $\widetilde{W}$ is a "fake" random initialization but identically distributed as W. Such technique at least traces back to smooth analysis [89].

To illustrate our high-level idea, from this section on (so in Section 6, 7 and 8)

$$\boxed{\text{we ignore the polynomial dependency in } \varrho \text{ and } \textit{hide it in the big-O notion.}}$$

We denote by $\widetilde{D}_{i,\ell}$, $\widetilde{g}_{i,\ell}$, $\widetilde{h}_{i,\ell}$ respectively the values of $D_{i,\ell}$, $g_{i,\ell}$ and $h_{i,\ell}$ determined by $\widetilde{W}$ and A at random initialization; and by $D_{i,\ell} = \widetilde{D}_{i,\ell} + D'_{i,\ell}$, $g_{i,\ell} = \widetilde{g}_{i,\ell} + g'_{i,\ell}$ and $h_{i,\ell} = \widetilde{h}_{i,\ell} + h'_{i,\ell}$ respectively those determined by $W = \widetilde{W} + W'$ after the adversarial perturbation.

Forward Stability. Our first, and most technical result is the following:

$$\|g'_{i,\ell}\|_2, \|h'_{i,\ell}\|_2 \leq O(m^{-1/2}) \ , \quad \|D'_{i,\ell}\|_0 \leq O(m^{2/3}) \quad \text{and} \quad \|D'_{i,\ell}g_{i,\ell}\|_2 \leq O(m^{-1/2}) \ . \tag{6.1}$$

Intuitively, one may hope for proving (6.1) by simple induction, because we have (ignoring subscripts in i)

$$g'_{\ell'} = \underbrace{W' D_{\ell'-1} g_{\ell'-1}}_{①} + \underbrace{\widetilde{W} D'_{\ell'-1} g_{\ell'-1}}_{②} + \underbrace{\widetilde{W} \widetilde{D}_{\ell'-1} g'_{\ell'-1}}_{③} \enspace.$$

The main issue here is that, the spectral norm of $\widetilde{W}\widetilde{D}_{\ell'-1}$ in ③ is greater than 1, so we cannot apply naive induction due to exponential blow up in L. Neither can we apply techniques from Section 5, because the changes such as $g_{\ell'-1}$ can be *adversarial*.

In our actual proof of (6.1), instead of applying induction on ③, we recursively expand ③ by the above formula. This results in a total of L terms of ① type and L terms of ② type. The main difficulty is to bound a term of ② type, that is:

$$\big\|\widetilde{W}\widetilde{D}_{\ell_1} \cdots \widetilde{D}_{\ell_2+1}\widetilde{W} D'_{\ell_2} g_{\ell_2}\big\|_2$$

Our argument consists of two conceptual steps.

(1) Suppose $g_{\ell_2} = \widetilde{g}_{\ell_2} + g'_{\ell_2} = \widetilde{g}_{\ell_2} + g'_{\ell_2,1} + g'_{\ell_2,2}$ where $\|g'_{\ell_2,1}\|_2 \leq m^{-1/2}$ and $\|g'_{\ell_2,2}\|_\infty \leq m^{-1}$, then we argue that $\|D'_{\ell_2} g_{\ell_2}\|_2 \leq O(m^{-1/2})$ and $\|D'_{\ell_2} g_{\ell_2}\|_0 \leq O(m^{2/3})$.

(2) Suppose we have $x \in \mathbb{R}^m$ with $\|x\|_2 \leq m^{-1/2}$ and $\|x\|_0 \leq m^{2/3}$, then we show that $y = \widetilde{W}\widetilde{D}_{\ell_1} \cdots \widetilde{D}_{\ell_2+1}\widetilde{W} x$ can be written as $y = y_1 + y_2$ with $\|y_1\|_2 \leq O(m^{-2/3})$ and $\|y_2\|_\infty \leq O(m^{-1})$.

The two steps above enable us to perform induction without exponential blow up. Indeed, they together enable us to go through the following logic chain:

$$\left.\begin{array}{lc} \|\cdot\|_2 \leq m^{-1/2} \text{ and } \|\cdot\|_\infty \leq m^{-1} & \stackrel{(1)}{\Longrightarrow} \\ \|\cdot\|_2 \leq m^{-2/3} \text{ and } \|\cdot\|_\infty \leq m^{-1} & \underset{(2)}{\Longleftarrow} \end{array}\right\} \|\cdot\|_2 \leq m^{-1/2} \text{ and } \|\cdot\|_0 \leq m^{2/3}$$

Since there is a gap between $m^{-1/2}$ and $m^{-2/3}$, we can make sure that all blow-up factors are absorbed into this gap, using the property that m is polynomially large. This enables us to perform induction to prove (6.1) without exponential blow-up.

Intermediate Layers and Backward Stability. Using (6.1), and especially using the sparsity $\|D'\|_0 \leq m^{2/3}$ from (6.1), one can apply the results in Section 5 to derive the following stability bounds for intermediate layers and backward propagation:

$$\big\|D_{i,\ell_1} W \cdots D_{i,\ell_2} W - \widetilde{D}_{i,\ell_1}\widetilde{W} \cdots \widetilde{D}_{i,\ell_2}\widetilde{W}\big\|_2 \leq O(L^7) \tag{6.2}$$

$$\big\|B D_{i,\ell_1} W \cdots D_{i,\ell_2} W - B\widetilde{D}_{i,\ell_1}\widetilde{W} \cdots \widetilde{D}_{i,\ell_2}\widetilde{W}\big\|_2 \leq O\big(m^{1/3}\big) \enspace. \tag{6.3}$$

Special Rank-1 Perturbation. For technical reasons, we also need two bounds in the special case of $W' = yz^\top$ for some unit vector z and sparse y with $\|y\|_0 \leq \mathsf{poly}(\varrho)$. We prove that, for this type of rank-one adversarial perturbation, it satisfies for every $k \in [m]$:

$$|((\widetilde{W} + W') h'_{i,\ell})_k| \leq O\big(m^{-2/3}\big) \tag{6.4}$$

$$\big\|B D_{i,\ell_1} W \cdots D_{i,\ell_2} W \mathbb{1}_k - B\widetilde{D}_{i,\ell_1}\widetilde{W} \cdots \widetilde{D}_{i,\ell_2}\widetilde{W}\mathbb{1}_k\big\|_2 \leq O\big(m^{-1/6}\big) \tag{6.5}$$

7 Proof Sketch of Theorem 3: Polyak-Łojasiewicz Condition

The upper bound in Theorem 3 is easy to prove (based on Section 5 and 6), but the lower bound (a.k.a. the Polyak-Łojasiewicz condition) is the most technically involved result to prove in this paper. We introduce the notion of "fake gradient". Given *fixed* vectors $\{\mathsf{loss}_{i,a}\}_{i\in[n],a\in\{2,\dots,L\}}$, we define

$$\widehat{\nabla}_k f(W) \stackrel{\text{def}}{=} \sum_{i=1}^{n}\sum_{\ell=1}^{L-1} (u_{i,\ell})_k \cdot h_{i,\ell} \cdot \mathbf{1}_{(g_{i,\ell+1})_k \geq 0} \tag{7.1}$$

where $u_{i,\ell} \stackrel{\text{def}}{=} \sum_{a=\ell+1}^{L} \mathsf{Back}_{i,\ell+1\to a}^{\top} \cdot \mathsf{loss}_{i,a}$. Note that if $\mathsf{loss}_{i,a} = Bh_{i,a} - y^*_{i,a}$ is the *true* loss vector, then $\widehat{\nabla}_k f(W)$ will be identical to $\nabla_k f(W)$ by Fact 2.5. Our main technical theorem is the following:

Theorem 5. *For every fixed vectors $\{\mathsf{loss}_{i,a}\}_{i\in[n],a\in\{2,\dots,L\}}$, if W, A, B are at* random initialization*, then with high probability*

$$\|\widehat{\nabla} f(W)\|_F^2 \geq \widetilde{\Omega}\left(\frac{\delta m}{\mathsf{poly}(\rho)}\right) \times \max_{i,\ell}\{\|\,\mathsf{loss}_{i,\ell}\,\|^2\} \enspace .$$

There are only two conceptually simple steps from Theorem 5 to Theorem 3 (see Appendix F).

- First, one can use the stability lemmas in Section 6 to show that, the fake gradient $\|\widehat{\nabla} f(W + W')\|_F$ after adversarial perturbation W' (with $\|W'\|_2 \leq \frac{1}{\sqrt{m}}$) is also large.
- Second, one can apply ε-net and union bound to turn "fixed loss" into "for all loss". This allows us to turn the lower bound on the fake gradient into a lower bound on the true gradient $\|\nabla_k f(W + W')\|_F$.

Therefore, in the rest of this section, we only sketch the ideas behind proving Theorem 5.

Let $(i^*, \ell^*) = \arg\max_{i,\ell}\{\|\,\mathsf{loss}_{i,\ell}\,\|_2\}$ be the sample and layer corresponding to the largest loss. Recall W, A, B are at random initialization.

7.1 Indicator and Backward Coordinate Bounds

There are three factors in the notion of fake gradient (7.1): $(u_{i,\ell})_k \in \mathbb{R}$ the backward coordinate, $h_{i,\ell} \in \mathbb{R}^m$ the forward vector, and $\mathbf{1}_{(g_{i,\ell+1})_k \geq 0} \in \{0, 1\}$ the indicator coordinate. We already know very well how the forward vector $h_{i,\ell}$ behaves from the previous sections. Let us provide bounds on the other two factors at the random initialization. (Details in Appendix D.)

Our "backward coordinate bound" controls the value of $(u_{i,\ell})_k$: at random initialization,

$$\left|(u_{i^*,\ell^*})_k\right| \geq \frac{\|\,\mathsf{loss}_{i^*,\ell^*}\,\|_2}{\rho} \quad \text{for at least } 1 - o(1) \text{ fraction of } k \in [m] \tag{7.2}$$

The main idea behind proving (7.2) is to use the randomness of B. For a fixed $k \in [m]$, it is in fact not hard to show that $\left|(u_{i^*,\ell^*})_k\right|$ is large with high probability. Unfortunately, the randomness of B are shared for different coordinates k. We need to also bound the correlation between pairs of coordinates $k_1, k_2 \in [m]$, and resort to MiDiarmid inequality to provide a high concentration bound with respect to all the coordinates.

Our indicator coordinate bound controls the value $(g_{i,\ell+1})_k$ inside the indicator functions $\mathbf{1}_{(g_{i,\ell+1})_k \geq 0}$. It says, letting $\beta_+ \stackrel{\text{def}}{=} \frac{\delta}{\rho^2}$ and $\beta_- \stackrel{\text{def}}{=} \frac{\delta}{\rho^{10}}$, then at random initialization, for at least $\frac{\delta}{\mathsf{poly}(\rho)}$ *fraction* of the coordinates $k \in [m]$,

$$|(g_{i^*,\ell^*+1})_k| \leq \frac{\beta_-}{\sqrt{m}} \quad \text{and} \quad |(g_{i,\ell+1})_k| \geq \frac{\beta_+}{\sqrt{m}} \quad \begin{array}{l}\text{for } i \neq i^* \text{ and } \ell = \ell^*, \text{ or}\\ \text{for } i \in [n] \text{ and } \ell > \ell^*.\end{array} \tag{7.3}$$

This should be quite intuitive to prove, in the following two steps.

- First, there are $\frac{\delta}{\mathsf{poly}(\rho)} m$ coordinates $k \in [m]$ with $|(g_{i^*,\ell^*+1})_k| \leq \frac{\beta_-}{\sqrt{m}}$.

  To show this, we write $(g_{i^*,\ell^*+1})_k = (WU_{\ell^*}U_{\ell^*}h_{i^*,\ell^*} + Ax_{i^*,\ell^*+1})_k$, and prove that for every $z \in \mathbb{R}^{n\ell^*}$ with bounded norms (by ε-net), there are at least $\frac{\delta}{\mathsf{poly}(\rho)} m$ coordinates $k \in [m]$ with

$|(WU_{\ell^*} z + Ax_{i^*,\ell^*+1})_k| \leq \frac{\beta_-}{\sqrt{m}}$. This is possible using the independence between WU_{ℓ^*} and A.

- Then, *conditioning* on the first event happens, we look at $|(g_{i,\ell+1})_k|$ for (1) each $i \neq i^*$ and $\ell = \ell^*$, or (2) each $i \in [n]$ and $\ell > \ell^*$. In both cases, even though the value of g_{i^*,ℓ^*+1} is fixed, we still have sufficient fresh new randomness (by invoking (5.3) for case (1) and (5.2) for case (2)). Such additional randomness can make sure that, with high probability (over the fresh new randomness), the value of $(g_{i,\ell+1})_k$ is larger than $\frac{\beta_+}{\sqrt{m}}$.

7.2 Thought Experiment: Adding Small Rank-One Perturbation

We now focus on a fixed coordinate $k \in [m]$ satisfying (7.3) and (7.2) in the Section 7.1, and denote by $v_{i^*,\ell^*} \stackrel{\text{def}}{=} \frac{(I - U_{\ell^*-1} U_{\ell^*-1}^\top) h_{i^*,\ell^*}}{\|(I - U_{\ell^*-1} U_{\ell^*-1}^\top) h_{i^*,\ell^*}\|}$.

For analysis purpose, imagine that we apply a small *random* perturbation W'_k to the already-randomly initialized matrix W, in the rank-one direction $\mathbb{1}_k v_{i^*,\ell^*}^\top$. Namely, we set $W'_k = g \cdot \mathbb{1}_k v_{i^*,\ell^*}^\top$ where $g \sim \mathcal{N}(0, \theta^2)$ and θ is a parameter satisfying $\frac{\beta_-}{\sqrt{m}} \ll \theta \ll \frac{\beta_+}{\sqrt{m}}$. Using the fact that k satisfies (7.3), one can show that[7]

(a) $\mathbf{1}_{(g_{i,\ell+1})_k \geq 0}$ stays the same with respect to perturbation W'_k, except for $i = i^*$ and $\ell = \ell^*$; and

(b) $\mathbf{1}_{(g_{i^*,\ell^*+1})_k \geq 0}$ can be 0 or 1 each with at least constant probability over W'_k.

At the same time, using the fact that k satisfies (7.2), one can show that

(c) $h_{i,\ell}$ and $(u_{i,\ell})_k$ does not change by much (owing to Section 6); and

(d) $|(u_{i^*,\ell^*+1})_k| \geq \frac{\|\mathsf{loss}_{i^*,\ell^*}\|_2}{\rho}$ is large (owing to (7.2)).

Putting (a), (b), (c), and (d) together, we know for such specially chosen k, at least with constant probability over the random perturbation of W'_k,

$$\|\widehat{\nabla}_k f(W + W'_k)\|_F \geq \Omega\Big(\frac{\|\mathsf{loss}_{i^*,\ell^*}\|_2}{\rho}\Big) \ . \tag{7.4}$$

7.3 Real Proof: Randomness Decomposition and McDiarmid's Inequality

There are only two main differences between (7.4) and our desired Theorem 5. First, (7.4) gives a gradient lower bound at $W + W'_k$, while in Theorem 5 we need a gradient lower bound at random initialization W. Second, (7.4) gives a lower bound on $\widehat{\nabla}_k f(\cdot)$ with *constant probability* for a small fraction of good coordinates k, but in Theorem 5 we need a lower bound for the entire $\widehat{\nabla} f(\cdot)$.

Randomness Decomposition. To fix the first issue, we resort to a randomness decomposition technique at least tracing back to the smooth analysis of Spielman and Teng [89]:

Proposition 7.1. *Given small constant $\theta \in (0,1)$ and m-dimensional random $g \sim \mathcal{N}(0, \frac{1}{m} I)$, we can rewrite $g = g_1 + g_3$ where g_1 follows from $\mathcal{N}(0, \frac{1}{m} I)$ and g_3 is very close to $\mathcal{N}(0, \frac{\theta^2}{m} I)$.*

[7]Specifically,

- For every $\ell < \ell^*$, we have $(g_{i,\ell+1})_k$ is unchanged with respect to W'_k. This is because $g_{i,\ell+1}$ only depends on the randomness of WU_ℓ, but v_{i^*,ℓ^*} is orthogonal to the columns of U_ℓ.
- For $i = i^*$ and $i = \ell^*$, we have $(g_{i,\ell+1})_k$ will change sign with constant probability with respect to W'_k. This is because $\|W'_k\|$ is above $\frac{\beta_-}{\sqrt{m}}$ — the original $(g_{i,\ell+1})_k$ before perturbation.
- For any other i and ℓ, we have $(g_{i,\ell+1})_k$ will not change sign with high probability with respect to W'_k. This is because $\|W'_k\|$ is below $\frac{\beta_+}{\sqrt{m}}$ — the original $(g_{i,\ell+1})_k$ before perturbation.

(Note that there is no contradiction here because g_1 and g_3 shall be correlated.)

Proof. Let $g_1, g_2 \in \mathbb{R}^m$ be two independent random vectors sampled from $\mathcal{N}(0, \frac{1}{m}I)$. We can couple them and make sure $g = \sqrt{1-\theta^2}g_1 + \theta g_2$. We now choose $g_3 = \theta g_2 - (1-\sqrt{1-\theta^2})g_1$. Since $(1-\sqrt{1-\theta^2}) \le O(\theta^2)$, when θ is sufficiently small, we know that g_3 is close to being generated from distribution $\mathcal{N}(0, \frac{\theta^2}{m}I)$. $\square$

Using Proposition 7.1, for each good coordinate k, instead of "adding" perturbation W'_k to W, we can instead decompose W into $W = W_0 + W'_k$, where W_0 is distributed in the same way as W. In other words, W_0 is also at random initialization. If this idea is carefully implemented, one can immediately turn (7.4) into

$$\|\widehat{\nabla}_k f(W)\|_F = \|\widehat{\nabla}_k f(W_0 + W'_k)\|_F \ge \Omega\Big(\frac{\|\mathsf{loss}_{i^*,\ell^*}\|_2}{\rho}\Big) \ . \tag{7.5}$$

Extended McDiarmid's Inequality. To fix the second issue, one may wish to consider all the indices $k \in [m]$ satisfying (7.3) and (7.2). Since there are at least $\frac{\delta}{\mathsf{poly}(\rho)}$ fraction of such coordinates, if all of them satisfied (7.5), then we would have already proved Theorem 5. Unfortunately, neither can we apply Chernoff bound (because the events with different $k \in [m]$ are correlated), nor can we apply union bound (because the event occurs only with constant probability).

Our technique is to resort to (an extended probabilistic variant of) McDiarmid's inequality (see Appendix A.6) in a very non-trivial way to boost the confidence.

Give any fixed subset N of cardinality $|N| = (\mathsf{poly}(\rho)/\delta)^2$, one can show that there again exists $\frac{\delta}{\mathsf{poly}(\rho)}$ fraction of coordinate $k \in N$ satisfying (7.3) and (7.2).[8] Now, instead of decomposing $W = W_0 + W'_k$, we decompose it as $W = W_1 + W'_N$ for $W'_N = u_N v_{i^*,\ell^*}^\top$ where u_N is only supported on coordinates $k \in N$. In other words, we *simultaneously* perturb in the directions of W'_k for all $k \in N$. Since this perturbation is still small enough —i.e., $\|W'_N\|_2 \le \frac{\mathsf{poly}(\varrho)}{\sqrt{m}}$— one can show that (7.5) remains true, that is, for a large fraction of $k \in N$, with at least constant probability over W'_N:

$$\|\widehat{\nabla}_k f(W_1 + W'_N)\|_F \ge \Omega\Big(\frac{\|\mathsf{loss}_{i^*,\ell^*}\|_2}{\rho}\Big) \ . \tag{7.6}$$

In order to apply McDiarmid's, we next need to bound on the difference between $\|\widehat{\nabla}_k f(W_1 + W'_N)\|_F$ and $\|\widehat{\nabla}_k f(W_1 + W'_N + W''_j)\|_F$ for an arbitrary (but small) perturbation W''_j (in the direction of $\mathbb{1}_j v_{i^*,\ell^*}^\top$). We show that,

$$\sum_{k \in [N]} \Big| \|\widehat{\nabla}_k (W_1 + W'_N)\|_2^2 - \|\widehat{\nabla}_k (W_1 + W'_N + W''_j)\|_2^2 \Big| \le O\left(\rho^8 + \frac{|N|}{m^{1/6}}\right) \ . \tag{7.7}$$

In other words, although there are $|N|$ difference terms, their total summation only grows in rate $\frac{|N|}{m^{1/6}}$ according to (7.7). After applying a variant of McDiarmid's inequality, we derive that with *high probability* over W'_N, it satisfies

$$\sum_{k \in N} \|\widehat{\nabla}_k f(W)\|_F^2 = \sum_{k \in N} \|\widehat{\nabla}_k f(W_1 + W'_N)\|_F^2 \ge \Omega\Big(\frac{\|\mathsf{loss}_{i^*,\ell^*}\|_2^2}{\mathsf{poly}(\rho)}|N|\Big) \ . \tag{7.8}$$

Finally, by sampling sufficiently many random sets N to cover the entire space $[m]$, we can show

$$\|\widehat{\nabla} f(W)\|_F^2 = \sum_{k \in [m]} \|\widehat{\nabla}_k f(W)\|_F^2 \ge \Omega\Big(\frac{\|\mathsf{loss}_{i^*,\ell^*}\|_2^2}{\mathsf{poly}(\rho)} m\Big) \ .$$

[8] We use $(\mathsf{poly}(\rho))^2$ to emphasize that the first polynomial needs to be bigger than the second $\mathsf{poly}(\rho)$.

and therefore Theorem 5 holds.

8 Proof Sketch of Theorem 4: Objective Smoothness

The objective smoothness Theorem 4 turns out to be much simpler to prove than Theorem 3. It only relies on Section 5 and 6, and does not need randomness decomposition or McDiarmid's inequality. (Details in Appendix G.)

Recall that in Theorem 4, $\widetilde{W}, A, B$ are at random initialization. $\breve{W}$ is an adversarially chosen matrix with $\|\breve{W} - \widetilde{W}\| \leq \frac{\mathsf{poly}(\varrho)}{\sqrt{m}}$, and W' is some other adversarial perturbation on top of $\breve{W}$, satisfying $\|W'\| \leq \frac{\tau_0}{\sqrt{m}}$. We denote by

- $\widetilde{D}_{i,\ell}$, $\widetilde{g}_{i,\ell}$, $\widetilde{h}_{i,\ell}$ respectively the values of $D_{i,\ell}$, $g_{i,\ell}$, $h_{i,\ell}$ determined by weight matrix $\widetilde{W}$;
- $\breve{D}_{i,\ell}$, $\breve{g}_{i,\ell}$, $\breve{h}_{i,\ell}$, $\breve{\mathsf{loss}}_{i,\ell}$ respectively those of $D_{i,\ell}$, $g_{i,\ell}$, $h_{i,\ell}$ and $\mathsf{loss}_{i,\ell}$ at weight matrix $\breve{W}$; and
- $D_{i,\ell}$, $g_{i,\ell}$, $h_{i,\ell}$ respectively the values of $D_{i,\ell}$, $g_{i,\ell}$, $h_{i,\ell}$ at weight matrix $W = \breve{W} + W'$.

Our main tool is to derive the following strong formula for $h_{i,\ell} - \breve{h}_{i,\ell}$: there exist diagonal matrices $D''_{i,\ell} \in \mathbb{R}^{m\times m}$ with entries in $[-1,1]$ and sparsity $\|D''_{i,\ell}\|_0 \leq O(m^{2/3})$ such that,

$$h_{i,\ell} - \breve{h}_{i,\ell} = \sum_{a=1}^{\ell-1} (\breve{D}_{i,\ell} + D''_{i,\ell})\breve{W} \cdots \breve{W}(\breve{D}_{i,a+1} + D''_{i,a+1})W' h_{i,a} \tag{8.1}$$

In particular, (8.1) implies $\|h_{i,\ell} - \breve{h}_{i,\ell}\| \leq O(L^9)\|W'\|_2$ after careful linear-algebraic manipulations (esp. using (5.5)). The main take-away message from (8.1) is that, this difference $h_{i,\ell} - \breve{h}_{i,\ell}$ is proportional to the norm of the perturbation, $\|W'\|_2$, no matter how small it is. In contrast, in (5.1), we only derived a weak upper bound of the form $\|h_{i,\ell} - \widetilde{h}_{i,\ell}\| \leq O(m^{-1/2})$. Nevertheless, the proof of (8.1) relies on (5.1), so we are not duplicating proofs.

Finally, we carefully derive that

$$\begin{aligned} & f(\breve{W} + W') - f(W) - \langle \nabla f(W), W' \rangle \\ = & \sum_{i=1}^{n}\sum_{\ell=2}^{L} \breve{\mathsf{loss}}_{i,\ell}^\top B \left((h_{i,\ell} - \breve{h}_{i,\ell}) - \sum_{a=1}^{\ell-1} \breve{D}_{i,\ell}\breve{W} \cdots \breve{W}\breve{D}_{i,a+1}W'\breve{h}_{i,a} \right) + \frac{1}{2}\|B(h_{i,\ell} - \breve{h}_{i,\ell})\|^2 \end{aligned} \tag{8.2}$$

and plug (8.1) into (8.2) to derive our final Theorem 4.

9 Appendix Roadmap

- Appendix A recalls some old lemmas and derives some new lemmas in probability theory.
- Appendix B serves for Section 5, the basic properties at random initialization.
- Appendix C serves for Section 6, the stability after adversarial perturbation.
- Appendix D, E and F together serve for Section 7 and prove Theorem 3, the Polyak-Łojasiewicz condition and gradient upper bound. In particular:
  - Appendix D serves for Section 7.1 (the indicator and backward coordinate bounds).
  - Appendix E serves for Section 7.3 (the randomness decomposition and McDiarmid's inequality) and proves Theorem 5.
  - Appendix F shows how to go from Theorem 5 to Theorem 3.

- Appendix G serves for Section 8, the proof of Theorem 4, the objective smoothness.
- Appendix H gives the final proof for Theorem 1, the GD convergence theorem.
- Appendix I gives the final proof for Theorem 2, the SGD convergence theorem.

Parameters. We also summarize a few parameters we shall use in the proofs.

- In Definition D.1, we shall introduce two parameters
$$\beta_+ \stackrel{\text{def}}{=} \frac{\delta}{\rho^2} \text{ and } \beta_- \stackrel{\text{def}}{=} \frac{\delta}{\rho^{10}}$$
to control the thresholds of indicator functions (recall Section 7.1).
- In Definition E.3, we shall introduce parameter
$$\theta \in \left[\rho^4 \cdot \beta_-, \rho^{-3} \cdot \beta_+\right]$$
to describe how much randomness we want to decompose out of W (recall Section 7.2).
- In (E.13) of Appendix E.4, we shall choose
$$N = \frac{\rho^{22}}{\beta_-^2}$$
which controls the size of the set N where we apply McDiarmid's inequality (recall Section 7.3).

Appendix

Table of Contents

A Preliminaries on Probability Theory

The goal of this section is to present a list of probability tools.

- In Section A.1, we recall how to swap randomness.
- In Section A.2, we recall concentration bounds for the chi-square distribution.
- In Section A.3, we proved a concentration bound of sum of squares of ReLU of Gaussians.
- In Section A.4 and Section A.5, we show some properties for random Gaussian vectors.
- In Section A.6, we recall the classical McDiarmid's inequality and then prove a general version of it.

A.1 Swapping Randomness

Fact A.1 (probability splitting)**.** *If $f(X,Y)$ holds with probability at least $1-\varepsilon$, then*

- *with probability at least $1-\sqrt{\varepsilon}$ (over randomness of X), the following event holds,*
  - *$f(X,Y)$ holds with probability at least $1-\sqrt{\varepsilon}$ (over randomness of Y).*

In other words,

$$\mathbf{Pr}_X\left[\mathbf{Pr}_Y[f(X,Y)] \geq 1-\sqrt{\varepsilon}\right] \geq 1-\sqrt{\varepsilon}.$$

Fact A.2 (swapping probability and expectation)**.** *If $\mathbf{Pr}_{X,Y}[f(X,Y)\geq a]\geq \varepsilon$, then*

$$\mathop{\mathbf{Pr}}_{X}\left[\mathop{\mathbb{E}}_{Y}[f(X,Y)] \geq a\varepsilon/2\right] \geq \varepsilon/2.$$

Proof. We prove it by making a contradiction,

Suppose

$$\mathop{\mathbf{Pr}}_{X}\left[\mathop{\mathbb{E}}_{Y}[f(X,Y)] \leq a\varepsilon/2\right] \geq 1-\varepsilon/2,$$

which implies that

$$\mathop{\mathbf{Pr}}_{X,Y}[f(X,Y)\geq a|X] \leq \varepsilon.$$

□

A.2 Concentration of Chi-Square Distribution

Lemma A.3 (Lemma 1 on page 1325 of Laurent and Massart [51])**.** *Let $X\sim \mathcal{X}_k^2$ be a chi-squared distributed random variable with k degrees of freedom. Each one has zero mean and σ^2 variance. Then*

$$\mathbf{Pr}[X-k\sigma^2 \geq (2\sqrt{kt}+2t)\sigma^2] \leq \exp(-t)$$
$$\mathbf{Pr}[k\sigma^2 - X \geq 2\sqrt{kt}\sigma^2] \leq \exp(-t)$$

One straightforward application is

Lemma A.4. *Let $x_1, x_2, \cdots, x_n$ denote i.i.d. samples from $\mathcal{N}(0,\sigma^2)$. For any $b\geq 1$, we have*

$$\mathbf{Pr}\left[|\|x\|_2^2 - n\sigma^2| \geq \frac{n}{b}\sigma^2\right] \leq 2\exp(-n/(8b^2)).$$

Proof. We choose $t=k/(8b^2)$ in Lemma A.3,

$$\mathbf{Pr}\left[\left|\|x\|_2^2 - n\sigma^2\right| \geq (\frac{2n}{\sqrt{8}b}+\frac{2n}{8b^2})\sigma^2\right] \leq 2\exp(-n/(8b^2)).$$

Since $\frac{2n}{\sqrt{8b}} + \frac{2n}{8b^2} \le \frac{2n}{\sqrt{8b}} + \frac{2n}{8b} \le n/b$. Thus,

$$\mathbf{Pr}\left[\left|\|x\|_2^2 - n\sigma^2\right| \ge \frac{n}{b}\sigma^2\right] \le 2\exp(-n/(8b^2)),$$

which completes the proof. $\square$

Lemma A.5. *Let $x_1, x_2, \cdots, x_m$ denote i.i.d. samples from $\mathcal{N}(0,1)$, and $y_i = \max\{x_i^2 - \log m, 0\}$. We have*

$$\mathbf{Pr}\left[\sum_{i=1}^{m} y_i \ge 2\sqrt{m}\right] \le e^{-\Omega(\sqrt{m})} \ .$$

Proof. First of all, letting $g = \sqrt{\log m}$,

$$\mathbb{E}[y_i] = \int_g^{\infty} \frac{\exp\left(-\frac{x^2}{2}\right)(x-g)^2}{\sqrt{2\pi}} dx = \frac{1}{2}\left(g^2+1\right)\operatorname{erfc}\left(\frac{g}{\sqrt{2}}\right) - \frac{e^{-\frac{g^2}{2}} g}{\sqrt{2\pi}} \le \frac{1}{e^{\frac{g^2}{2}}} = \frac{1}{\sqrt{m}} \ .$$

On the other hand, each random variable y_i is $O(1)$-subgaussian. By subgaussian concentration,

$$\mathbf{Pr}\left[\sum_{i=1}^{m} y_i \ge 2\sqrt{m}\right] \le e^{-\Omega(\sqrt{m})}$$

$\square$

A.3 Concentration of Sum of Squares of ReLU of Gaussians

Lemma A.6 (Upper bound)**.** *Given n i.i.d. Gaussian random variables $x_1, x_2, \cdots, x_n \sim \mathcal{N}(0, \sigma^2)$, we have*

$$\mathbf{Pr}\left[\left(\sum_{i=1}^{n} \max(x_i, 0)^2\right)^{1/2} < (1+\varepsilon)\sqrt{n/2}\sigma\right] \ge 1 - \exp(-\varepsilon^2 n/100).$$

Proof. Using Chernoff bound, we know that with probability $1 - \exp(-\varepsilon^2 n/6)$, $\sum_{i=1}^{n} \max(x_i, 0)^2$ is a at most degree-$(1+\varepsilon)\frac{n}{2}$ Chi-square random variable. Let us say this is the first event.

Using Lemma A.3, we have

$$\begin{aligned}
& \mathbf{Pr}[X \ge k\sigma^2 + (2\sqrt{kt} + 2t)\sigma^2] \le \exp(-t) \\
\implies & \mathbf{Pr}[X \ge k\sigma^2 + (2\varepsilon k + 2\varepsilon^2 k)\sigma^2] \le \exp(-\varepsilon^2 k) && \text{by choosing } t = \varepsilon^2 k \\
\implies & \mathbf{Pr}[X \ge k(1+4\varepsilon)\sigma^2] \le \exp(-\varepsilon^2 k) \\
\implies & \mathbf{Pr}\left[X \ge (1+\varepsilon)(1+4\varepsilon)\frac{n}{2}\sigma^2\right] \le \exp(-\varepsilon^2(1+\varepsilon)n/2) && \text{by } k = (1+\varepsilon)n/2
\end{aligned}$$

Thus, we have with probability at least $1 - \exp(-\varepsilon^2 n/2)$,

$$X \le (1+4\varepsilon)^2 \frac{n}{2}\sigma^2.$$

Let the above event denote the second event.

By taking the union bound of two events, we have with probability $1 - 2\exp(-\varepsilon^2 n/6)$

$$\left(\sum_{i=1}^{n} \max(x_i, 0)^2\right)^{1/2} \le X \le (1+4\varepsilon)\frac{n}{2}\sigma$$

Then rescaling the ε, we get the desired result. $\square$

Lemma A.7 (Lower bound)**.** *Given n i.i.d. Gaussian random variables $x_1, x_2, \cdots, x_n \sim \mathcal{N}(0, \sigma^2)$, we have*

$$\mathbf{Pr}\left[\left(\sum_{i=1}^n \max(x_i, 0)^2 \right)^{1/2} > (1-\varepsilon)\sqrt{n/2}\sigma \right] \geq 1 - \exp(\varepsilon^2 n/100)$$

Proof. Using Chernoff bound, we know that with probability $1 - \exp(-\varepsilon^2 n/6)$, $\sum_{i=1}^n \max(x_i, 0)^2$ is a at most degree-$(1-\varepsilon)\frac{n}{2}$ Chi-square random variable. Let us say this is the first event.

Using Lemma A.3, we have

$$\begin{aligned}
& \mathbf{Pr}[X \leq k\sigma^2 - 2\sqrt{kt}\sigma^2] \leq \exp(-t) & \\
\implies & \mathbf{Pr}[X \leq k\sigma^2 - 2\varepsilon k\sigma^2] \leq \exp(-\varepsilon^2 k) & \text{by choosing } t = \varepsilon^2 k \\
\implies & \mathbf{Pr}[X \leq k\sigma^2(1-2\varepsilon)] \leq \exp(-\varepsilon^2 k) & \\
\implies & \mathbf{Pr}\left[X \leq (1-\varepsilon)(1-2\varepsilon)\frac{n}{2}\sigma^2\right] \leq \exp(-\varepsilon^2(1-\varepsilon)n/2) & \text{by } k = (1-\varepsilon)n/2
\end{aligned}$$

Thus, we have with probability at least $1 - \exp(-\varepsilon^2 n/4)$,

$$X \geq (1-2\varepsilon)^2 \frac{n}{2}\sigma^2.$$

Let the above event denote the second event.

By taking the union bound of two events, we have with probability at least $1 - 2\exp(-\varepsilon^2 n/6)$,

$$\left(\sum_{i=1}^n \max(x_i, 0)^2 \right)^{1/2} \geq X \geq (1-2\varepsilon)\frac{n}{2}\sigma.$$

Then rescaling the ε, we get the desired result. $\square$

Combining Lemma A.7 and Lemma A.6, we have

Lemma A.8 (Two sides bound)**.** *Given n i.i.d. Gaussian random variables $x_1, x_2, \cdots, x_n \sim \mathcal{N}(0, \sigma^2)$, let $\phi(a) = \max(a, 0)^2$. We have*

$$\mathbf{Pr}_x\left[\|\phi(x)\|_2 \in ((1-\varepsilon)\sqrt{n/2}\sigma, (1+\varepsilon)\sqrt{n/2}\sigma) \right] \geq 1 - 2\exp(-\varepsilon^2 n/100)$$

Corollary A.9 (Two sides bound for single matrix)**.** *Let $x \in \mathbb{R}^n$ denote a fixed vector. Given a random Gaussian matrix $A \in \mathbb{R}^{m\times n}$ where each entry is i.i.d. sampled from $\mathcal{N}(0, 2\sigma^2/m)$.*

$$\mathbf{Pr}_A\left[\|\phi(Ax)\|_2 \in ((1-\varepsilon)\|x\|_2\sigma, (1+\varepsilon)\|x\|_2\sigma) \right] \geq 1 - 2\exp(-\varepsilon^2 m/100).$$

Proof. For each $i \in [m]$, let $y_i = (Ax)_i$. Then $y_i \sim \mathcal{N}(0, \widetilde{\sigma}^2)$, where $\widetilde{\sigma}^2 = 2\sigma^2/m \cdot \|x\|_2^2$. Using Corollary A.8, we have

$$\mathbf{Pr}_y\left[\|\phi(y)\|_2 \in ((1-\varepsilon)\sqrt{m/2}\widetilde{\sigma}, (1+\varepsilon)\sqrt{m/2}\widetilde{\sigma}) \right] \geq 1 - 2\exp(-\varepsilon m/100)$$

which implies

$$\mathbf{Pr}_y\left[\|\phi(y)\|_2 \in ((1-\varepsilon)\|x\|_2\sigma, (1+\varepsilon)\|x\|_2\sigma) \right] \geq 1 - 2\exp(-\varepsilon^2 m/100).$$

Since $y = Ax$, thus we complete the proof. $\square$

Corollary A.10 (Two sides bound for multiple matrices)**.** *Let $x_1, x_2, \cdots, x_k$ denote k fixed vectors where $x_i \in \mathbb{R}^{n_i}$. Let $x = [x_1^\top \ x_2^\top \ \cdots \ x_k^\top]^\top \in \mathbb{R}^n$ where $n = \sum_{i=1}^k n_i$. Let $A_1, A_2, \cdots A_k$ denote*

k independent random Gaussian matrices where each entry of $A_i \in \mathbb{R}^{m \times n_i}$ is i.i.d. sampled from $\mathcal{N}(0, 2\sigma_i^2/m)$ for each $i \in [k]$. Let $A = [A_1 \ A_2 \ \cdots \ A_k]$. We have

$$\Pr_A \left[\|\phi(Ax)\|_2 \in \left((1-\varepsilon) \left(\sum_{i=1}^k \|x_i\|_2^2 \sigma_i^2 \right)^{1/2}, (1+\varepsilon) \left(\sum_{i=1}^k \|x_i\|_2^2 \sigma_i^2 \right)^{1/2} \right) \right] \geq 1 - 2\exp(-\varepsilon^2 m/100).$$

Proof. It is similar as Corollary A.9. □

Lemma A.11. *Let $h, q \in \mathbb{R}^p$ be fixed vectors, $W \in \mathbb{R}^{m \times p}$ be random matrix with i.i.d. entries $W_{i,j} \sim \mathcal{N}(0, \frac{2}{m})$, and vector $v \in \mathbb{R}^m$ defined as $v_i = \mathbf{1}_{\langle W_i, h+q \rangle \geq 0} \langle W_i, h \rangle$. Then, the absolute values of coordinates $|v_i|$ follow from i.i.d. folded Gaussian distributions $|\mathcal{N}(0, \frac{\|h\|_2^2}{m})|$. In particular, $\frac{m\|v\|_2^2}{\|h\|_2^2}$ is in distribution identical to Chi-square distribution $\mathcal{X}_m^2$.*

A.4 Gaussian Vector Percentile: Center

Fact A.12. *Suppose $x \sim \mathcal{N}(0, \sigma^2)$ is a Gaussian random variable. For any $t \in (0, \sigma]$ we have*

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

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

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

Lemma A.13. *Let $x \sim \mathcal{N}(0, \sigma^2 I)$. For any $\alpha \in (0, 1/2)$, we have with probability at least $1 - \exp(-\alpha^2 m/100)$,*

- *there exists at least $\frac{1}{2}(1-\alpha)$ fraction of i such that $x_i \geq 5\alpha\sigma/16$, and*
- *there exists at least $\frac{1}{2}(1-\alpha)$ fraction of i such that $x_i \leq -5\alpha\sigma/16$.*

Proof. Let $c_1 = 4/5$. For each $i \in [m]$, we define random variable y_i as

$$y_i = \begin{cases} 1, & \text{if } x_i \geq \alpha\sigma/(4c_1); \\ 0, & \text{otherwise.} \end{cases}$$

Let $p = \Pr[y_i = 1]$. Using Fact A.12 with $t = \alpha\sigma/(4c_1)$, we know that $p \geq \frac{1}{2}(1 - \alpha/4)$. Letting $Y = \sum_{i=1}^m y_i$, we have $\mu = \mathbb{E}[Y] = mp \geq \frac{1}{2}(1-\alpha/4)m$. Using Chernoff bound, we have

$$\Pr[Y \leq (1-\delta)\mu] \leq \exp(-\delta^2\mu/2).$$

Choosing $\delta = \frac{1}{4}\alpha$, we have

$$\Pr[Y \leq (1-\alpha/4)\mu] \leq \exp(-\alpha^2\mu/32)$$

Further we have

$$\begin{aligned}
\mathbf{Pr}[Y \leq \frac{1}{2}(1-\alpha)m] \leq\ & \mathbf{Pr}[Y \leq (1-\alpha/4)\frac{1}{2}(1-\alpha/4)m] && \text{by } (1-\alpha) \leq (1-\alpha/4)^2 \\
\leq\ & \mathbf{Pr}[Y \leq (1-\alpha/4)\mu] && \text{by } \mu \geq \frac{1}{2}(1-\alpha/4)m \\
=\ & \mathbf{Pr}[Y \leq (1-\delta)\mu] && \text{by } \delta = \alpha/4 \\
\leq\ & \exp(-\alpha^2\mu/32) \\
\leq\ & \exp(-\alpha^2\frac{1}{2}(1-\alpha/4)m/32) && \text{by } \mu \geq \frac{1}{2}(1-\alpha/4)m \\
=\ & \exp(-\alpha^2(1-\alpha/4)m/64)\ .
\end{aligned}$$

□

We provide the definition of (α,σ)-good. Note that this definition will be used often in the later proof.

Definition A.14 ((α,σ)-good)**.** *Given $w \in \mathbb{R}^m$, we say w is (α,σ)-good if the following two conditions holding:*

- *there are at least $\frac{1}{2}(1-\alpha)$ fraction coordinates satisfy that $w_i \geq \alpha\sigma$; and*
- *there are at least $\frac{1}{2}(1-\alpha)$ fraction coordinates satisfy that $w_i \leq -\alpha\sigma$.*

Lemma A.13 gives the following immediate corollary:

Corollary A.15 (random Gaussian is $(\alpha,\sigma/4)$-good)**.** *Let $x \sim \mathcal{N}(0,\sigma^2 I)$. For any $\alpha \in (0,1/2)$, we have with probability at least $1-\exp(-\alpha^2 m/100)$ that x is $(\alpha,\sigma/4)$-good.*

Corollary A.16. *Let $x_1, x_2, \cdots, x_k$ be k fixed vectors where $x_i \in \mathbb{R}^{n_i}$, and $A_1, A_2, \cdots, A_k$ be k independent random Gaussian matrices where each entry of $A_i \in \mathbb{R}^{m\times n_i}$ is i.i.d. sampled from $\mathcal{N}(0,\sigma_i^2)$ for each $i \in [k]$. Denote by $x = (x_1,\ldots,x_k) \in \mathbb{R}^n$ for $n = \sum_{i=1}^k n_i$, $A = [A_1, A_2, \cdots, A_k] \in \mathbb{R}^{m\times n}$, and $\sigma = \left(\sum_{i=1}^k \sigma_i^2 \|x_i\|_2^2\right)^{1/2}$. For any fixed parameter $\alpha \in (0,1/2)$, we have*

$$Ax = A_1x_1 + \cdots A_kx_k \text{ is } (\alpha,\sigma/4)\text{-good with probability at least } 1-\exp(-\alpha^2 m/100).$$

Proof. It is clear that Ax follows from a Gaussian distribution $\mathcal{N}\big(0, \big(\sum_{i=1}^k \sigma_i^2\|x_i\|_2^2\big)I\big)$, so we can directly apply Corollary A.15. □

A.5 Gaussian Vector Percentile: Tail

Lemma A.17. *Suppose $W \in \mathbb{R}^{m\times n}$ is a random matrix with entries drawn i.i.d. from $\mathcal{N}\big(0,\frac{2}{m}\big)$. Given $d \leq s \leq m$, with probability at least $1-e^{-\Omega(s\log^2 m)}$, for all $x \in \mathbb{R}^n$, letting $y = Wx$, we can write $y = y_1 + y_2$ with*

$$\|y_1\| \leq \frac{\sqrt{s}\log^2 m}{\sqrt{m}} \cdot \|x\| \quad \text{and} \quad \|y_2\|_\infty \leq \frac{\log m}{\sqrt{m}} \cdot \|x\|\ .$$

Proof of Lemma A.17. Without loss of generality we only prove the result for $\|x\| = 1$.

Fixing any such x and letting $\beta = \frac{\log m}{2\sqrt{m}}$, we have $y_i \sim \mathcal{N}\big(0,\frac{2}{m}\big)$ so for every $p \geq 1$, by Gaussian tail bound

$$\mathbf{Pr}[|y_i| \geq \beta p] \leq e^{-\Omega(\beta^2 p^2 m)}\ .$$

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

$$\binom{m}{q/p^2} \times \left(e^{-\Omega(\beta^2 p^2 m)}\right)^{q/p^2} \leq e^{\frac{q}{p^2}\left(O(\log m) - \Omega(\beta^2 p^2 m)\right)} \leq e^{-\Omega(\beta^2 q m)} \ .$$

In other words,

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

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

$$\sum_{i \colon |y_i| \geq \beta} y_i^2 \leq \sum_{k=0}^{\lceil \log q \rceil} (2^{k+1}\beta)^2 \left|\left\{i \in [m] \colon |y_i| \geq 2^k \beta\right\}\right| \leq \sum_{k=0}^{\lceil \log q \rceil} (2^{k+1}\beta)^2 \cdot \frac{q}{2^{2k}} \leq 4q\beta^2 \log q \ . \quad \text{(A.1)}$$

In other words, vector y can be written as $y = y_1 + y_2$ where $\|y_2\|_\infty \leq \beta$ and $\|y_1\|^2 \leq 4q\beta^2 \log q$.

At this point, we can choose $q = \frac{s \log^2 m}{m\beta^2} = 4s$ so the above event happens with probability at least $1 - e^{-\Omega(\beta^2 q m)} \geq 1 - e^{-\Omega(s \log^2 m)}$. Finally, applying standard ε-net argument over all unit vectors $x \in \mathbb{R}^n$, we have for each such $x \in \mathbb{R}^n$, we can decompose $y = Wx$ into $y = y_1 + y_2$ where

$$\|y_2\|_\infty \leq 2\beta = \frac{\log m}{\sqrt{m}} \text{ and } \|y_1\|^2 \leq \frac{8s \log^3 m}{m} \ .$$

□

A.6 McDiarmid's Inequality and An Extension

We state the standard McDiarmid's inequality,

Lemma A.18 (McDiarmid's inequality). *Consider independent random variables $x_1, \cdots, x_n \in \mathcal{X}$ and a mapping $f : \mathcal{X}^n \to \mathbb{R}$. If, for all $i \in [n]$, and for all $y_1, \cdots, y_n, y_i' \in \mathcal{X}$, then function f satisfies*

$$|f(y_1, \cdots, y_{i-1}, y_i, y_{i+1}, \cdots, y_n) - f(y_1, \cdots, y_{i-1}, y_i', y_{i+1}, \cdots, y_n)| \leq c_i.$$

Then

$$\mathbf{Pr}[f(x_1, \cdots, x_n) - \mathbb{E} f \geq t] \geq \exp\left(\frac{-2t^2}{\sum_{i=1}^n c_i^2}\right),$$
$$\mathbf{Pr}[f(x_1, \cdots, x_n) - \mathbb{E} f \leq -t] \geq \exp\left(\frac{2t^2}{\sum_{i=1}^n c_i^2}\right).$$

We prove a more general version of McDiarmid's inequality,

Lemma A.19 (McDiarmid extension). *Let $w_1, \dots, w_N$ be independent random variables and $f \colon (w_1, \dots, w_N) \mapsto [0, 1]$. Suppose it satisfies:*

- $\mathbb{E}_{w_1, \dots, w_N}[f(w_1, \dots, w_N)] \geq \mu$, *and*
- *With probability at least $1 - p$ over $w_1, \dots, w_N$, it satisfies*

$$\forall k \in [N], \forall w_k'' \colon \left|f(w_{-k}, w_k) - f(w_{-k}, w_k'')\right| \leq c \quad \text{(A.2)}$$

Then, $\mathbf{Pr}[f(w_1, \dots, w_N) \geq \mu/2] \geq 1 - N^2\sqrt{p} - e^{\Omega(\frac{-\mu^2}{N(c^2+p)})}$.

Proof of Lemma A.19. For each $t \in [N]$, we have with probability at least $1 - \sqrt{p}$ over $w_1, \dots, w_t$, it satisfies

$$\mathop{\mathbf{Pr}}_{w_{t+1}, \dots, w_N}\left[\forall k \in [N], \forall w_k'' \colon \left|f(w_{-k}, w_k) - f(w_{-k}, w_k'')\right| \leq c\right] \geq 1 - \sqrt{p} \ .$$

Define those $(w_1, \ldots, w_t)$ satisfying the above event to be $K_{\le t}$.

Define random variable X_t (which depends only on $w_1, \ldots, w_t$) as

$$X_t := \mathop{\mathbb{E}}_{w_{>t}} [f(\vec{w}) \mid w_{\le t}] \, \mathbf{1}_{(w_{\le 1}, \ldots, w_{\le t}) \in K_{\le 1} \times \cdots \times K_{\le t}} + N(1 - \mathbf{1}_{(w_{\le 1}, \ldots, w_{\le t}) \in K_{\le 1} \times \cdots \times K_{\le t}})$$

For every t and fixed $w_1, \ldots, w_{t-1}$.

- If $(w_{\le 1}, \ldots, w_{<t}) \notin K_{\le 1} \times \cdots \times K_{<t}$, then $X_t = X_{t-1} = N$.
- If $(w_{\le 1}, \ldots, w_{<t}) \in K_{\le 1} \times \cdots \times K_{<t}$,
  - If $w_{\le t} \notin K_{\le t}$, then $X_t - X_{t-1} = N - \cdots \ge 0$.
  - If $w_{\le t} \in K_{\le t}$, then

$$X_t - X_{t-1} = \mathop{\mathbb{E}}_{w_{>t}} [f(w_{<t}, w_t, w_{>t}) \mid w_{\le t}] - \mathop{\mathbb{E}}_{w_{\ge t}} [f(w_{<t}, w_t, w_{>t}) \mid w_{<t}]$$

    Recall from our assumption that, with probability at least $1 - \sqrt{p}$ over w_t and $w_{>t}$, it satisfies

$$\forall w_t'' \colon \left| f(w_{<t}, w_t'', w_{>t}) - f(w_{<t}, w_t, w_{>t}) \right| \le c$$

    Taking expectation over w_t and $w_{>t}$, we have

$$\forall w_t'' \colon \mathop{\mathbb{E}}_{w_{>t}} \left[f(w_{<t}, w_t'', w_{>t}) \right] - \mathop{\mathbb{E}}_{w_{\ge t}} [f(w_{<t}, w_t, w_{>t})] \ge -(c + \sqrt{p})$$

    This precisely means $X_t - X_{t-1} \ge c + \sqrt{p}$.

In sum, we have just shown that $X_t - X_{t-1} \ge -(c + \sqrt{p})$ always holds. By applying martingale concentration (with one-sided bound),

$$\mathbf{Pr}[X_N - X_0 \le -t] \le \exp\left(\frac{-t^2}{N(c + \sqrt{p})^2} \right)$$

Notice that $X_0 = \mu$ so if we choose $t = \mu/2$, we have

$$\mathbf{Pr}[X_N \ge \mu/2] \ge 1 - \exp\left(\frac{-\mu^2}{N(c + \sqrt{p})^2} \right)$$

Recalling

$$X_N := f(\vec{w}) \mathbf{1}_{(w_{\le 1}, \ldots, w_{\le t}) \in K_{\le 1} \times \cdots \times K_{\le t}} + N(1 - \mathbf{1}_{(w_{\le 1}, \ldots, w_{\le t}) \in K_{\le 1} \times \cdots \times K_{\le t}})$$

and we have $X_N = f(w_1, \ldots, w_N)$ with probability at least $1 - N\sqrt{p}$ (and $X_N = N$ with the remaining probabilities). Together, we have the desired theorem.

$\square$

B Basic Properties at Random Initialization

Recall that the recursive update equation of RNN can be described as follows

$$\begin{aligned}
h_{i,0} &= 0 && \forall i \in [n] \\
h_{i,\ell} &= \phi(W \cdot h_{i,\ell-1} + Ax_{i,\ell}) && \forall i \in [n], \forall \ell \in [L] \\
y_{i,\ell} &= B \cdot h_{i,\ell} && \forall i \in [n], \forall \ell \in [L]
\end{aligned}$$

Throughout this section, we assume that matrices $W \in \mathbb{R}^{m\times m}$, $A \in \mathbb{R}^{m\times d_x}$, $B \in \mathbb{R}^{d\times m}$ are at their random initialization position: each entry of W and A is sampled i.i.d. from $\mathcal{N}(0, \frac{2}{m})$ and each entry of B is sampled i.i.d. from $\mathcal{N}(0, \frac{1}{d})$. We recall

Definition 2.2. *For each $i \in [n]$ and $\ell \in [L]$, let $D_{i,\ell} \in \mathbb{R}^{m\times m}$ be the diagonal matrix where*

$$(D_{i,\ell})_{k,k} = \mathbf{1}_{(W\cdot h_{i,\ell-1}+Ax_{i,\ell})_k\geq 0} = \mathbf{1}_{(g_{i,\ell})_k\geq 0} \enspace .$$

As a result, we can write $h_{i,\ell} = D_{i,\ell}Wh_{i,\ell-1}$.

We introduce two notations that shall repeatedly appear in our proofs.

Definition B.1 (U_ℓ)**.** *Let $U_\ell \in \mathbb{R}^{m\times n\ell}$ denote the column orthonormal matrix using Gram-Schmidt*

$$U_\ell \stackrel{\text{def}}{=} \mathsf{GS}\left(h_{1,1}, \dots, h_{n,1}, h_{1,2}, \dots, h_{n,2}, \dots, h_{1,\ell}, \dots, h_{n,\ell}\right)$$

Definition B.2 ($v_{i,\ell}$)**.** *For each $i \in [n]$, $\ell \in [L]$, we define vector $v_{i,\ell} \in \mathbb{R}^m$ as*

$$v_{i,\ell} = \frac{(I - U_{\ell-1}U_{\ell-1}^\top)h_{i,\ell}}{\|(I - U_{\ell-1}U_{\ell-1}^\top)h_{i,\ell}\|_2}$$

Roadmap.

- Section B.1 proves that the forward propagation $\|h_{i,\ell}\|_2$ neither vanishes nor explodes.
- Section B.2 gives a lower bound on the projected forward propagation $\|(I - U_\ell U_\ell^\top)h_{i,\ell+1}\|_2$.
- Section B.3 proves that for two data points $i, j \in [n]$, the projected forward propagation $(I - U_\ell U_\ell^\top)h_{i,\ell+1}$ and $(I - U_\ell U_\ell^\top)h_{j,\ell+1}$ are separable from either other.
- Section B.4 and Section B.5 prove that the consecutive intermediate layers, in terms of spectral norm, do not explode (for full and sparse vectors respectively).
- Section B.6 proves that the backward propagation does not explode.

B.1 Forward Propagation

Our first result of this section is on showing upper and lower bounds on the forward propagation.

Lemma B.3 (c.f. (5.1))**.** *With probability at least $1 - \exp(-\Omega(m/L^2))$ over the random initialization W, A (see Definition 2.3), it satisfies*

$$\forall i \in [n], \quad \forall \ell \in \{0, 1, \dots, L-1\} \quad : \quad (1 - 1/(4L))^\ell \leq \|h_{i,\ell+1}\|_2 \leq 2\ell + 4.$$
$$\|g_{i,\ell+1}\|_2 \leq 4\ell + 8.$$

We prove Lemma B.3 by induction on $\ell \in [L]$. We only prove the $h_{i,\ell+1}$ part and the $g_{i,\ell+1}$ part is completely analogous.

For the base case $\ell = 0$, we have $\|x_{i,1}\|_2 = 1$ and $h_{i,1} = \phi(Ax_{i,1})$. Using Corollary A.10 we have $1 - 1/4L \leq \|h_{i,1}\|_2 \leq 1 + 1/4L$ with probability at least $1 - \exp(-\Omega(-m/L^2))$. We proceed the proof for the case of $\ell \geq 1$, assuming that Lemma B.3 already holds for $0, 1, \ldots, \ell - 1$. We first show

Claim B.4. *With probability at least* $1 - \exp(-\Omega(m/L^2))$ *over* W *and* A*,*

$$\forall i \in [n]\colon \|h_{i,\ell+1}\|_2 \leq (1 + 1/L)(\|h_{i,\ell}\|_2 + 1)$$

Note if Claim B.4 holds for all $\ell = 0, 1, \ldots, \ell-1$, then we must have $\|h_{i,\ell}\|_2 \leq (1+1/L)^\ell \|h_{i,0}\|_2 + \sum_{i=1}^{\ell}(1 + 1/L)^i = (1 + 1/L)^\ell + \sum_{i=1}^{\ell}(1 + 1/L)^i \leq 2\ell + 4$.

Proof of Claim B.4. Recall the definition of $h_{i,\ell} \in \mathbb{R}^m$, we have

$$\|h_{i,\ell+1}\|_2 = \|\phi(W \cdot h_{i,\ell} + Ax_{i,\ell+1})\|_2$$

We can rewrite vector $Wh_{i,\ell} \in \mathbb{R}^m$ as follows

$$\begin{aligned} Wh_{i,\ell} &= WU_{\ell-1}U_{\ell-1}^\top h_{i,\ell} + W(I - U_{\ell-1}U_{\ell-1}^\top)h_{i,\ell} \\ &= WU_{\ell-1}U_{\ell-1}^\top h_{i,\ell} + W\frac{(I - U_{\ell-1}U_{\ell-1}^\top)h_{i,\ell}}{\|(I - U_{\ell-1}U_{\ell-1}^\top)h_{i,\ell}\|_2} \cdot \|(I - U_{\ell-1}U_{\ell-1}^\top)h_{i,\ell}\|_2 \\ &= WU_{\ell-1}U_{\ell-1}^\top h_{i,\ell} + Wv_{i,\ell} \cdot \|(I - U_{\ell-1}U_{\ell-1}^\top)h_{i,\ell}\|, \end{aligned}$$

where the last step follows by definition of $v_{i,\ell}$ (See Definition B.2). Define $z_1 \in \mathbb{R}^{n(\ell-1)}, z_2 \in \mathbb{R}, z_3 \in \mathbb{R}^d$ as follows

$$z_1 = U_{\ell-1}^\top h_{i,\ell}, \quad z_2 = \|(I - U_{\ell-1}U_{\ell-1}^\top)h_{i,\ell}\|_2, \quad z_3 = x_{i,\ell+1}. \tag{B.1}$$

Then

$$\begin{aligned} \|z_1\|_2^2 + z_2^2 + \|z_3\|_2^2 &= \|U_{\ell-1}^\top h_{i,\ell}\|_2^2 + \|(I - U_{\ell-1}U_{\ell-1}^\top)h_{i,\ell}\|_2^2 + \|x_{i,\ell+1}\|_2^2 \\ &= \|h_{i,\ell}\|_2^2 + \|x_{i,\ell+1}\|_2^2 \leq \|h_{i,\ell}\|_2^2 + 1 \ . \end{aligned} \tag{B.2}$$

We can thus rewrite

$$\begin{aligned} Wh_{i,\ell} + Ax_{i,\ell+1} &= WU_{\ell-1}z_1 + Wv_{i,\ell}z_2 + Az_3 \\ &= \begin{bmatrix} M_1 & M_2 & M_3 \end{bmatrix} \cdot \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} \\ &= M \cdot z, \end{aligned}$$

where the last step follows by defining $M \in \mathbb{R}^{m \times (n(\ell-1)+1+d)}$ as follows,

$$M = \begin{bmatrix} M_1 & M_2 & M_3 \end{bmatrix} = \begin{bmatrix} WU_{\ell-1} & Wv_{i,\ell} & A \end{bmatrix} \tag{B.3}$$

We stress here that the entries of M_1, M_2, M_3 are i.i.d. from $\mathcal{N}(0, \frac{2}{m})$.[9] We have $\|h_{i,\ell+1}\| = \|\phi(Wh_{i,\ell} + Ax_{i,\ell+1})\| = \|\phi(M \cdot z)\|$.

Next, applying Corollary A.10, we know that if z_1, z_2, z_3 are fixed (instead of defined as in (B.1)), then, letting Choosing $\varepsilon = 1/2L$, we have

$$\Pr_M\left[\|h_{i,\ell+1}\| \leq (1 + \varepsilon) \cdot \sqrt{\|z_1\|_2^2 + z_2^2 + \|z_3\|_2^2}\right] \geq 1 - \exp(-\Omega(\varepsilon^2 m)) \ .$$

[9] Indeed, after Gram-Schmidt we can write $U_{\ell-1} = [\widehat{h}_1, \ldots, \widehat{h}_{n(\ell-1)}]$, where each $\widehat{h}_j$ only depends on the randomness of A and $W[\widehat{h}_1, \ldots, \widehat{h}_{j-1}]$. In other words, conditioning on any choice of A and $W[\widehat{h}_1, \ldots, \widehat{h}_{j-1}]$, we still have $W\widehat{h}_j$ is an independent Gaussian vector from $\mathcal{N}(0, \frac{2}{m}I)$. Similarly, $v_{i,\ell}$ may depend on the randomness of A and $WU_{\ell-1}$, but conditioning on any choice of A and $WU_{\ell-1}$, we still have $Wv_{i,\ell}$ follows from $\mathcal{N}(0, \frac{2}{m}I)$. This proves that the entries of M_1, M_2, M_3 are independent.

To move from fixed choices of z_1 and z_2 to *all* choices of z_1 and z_2, we perform a standard ε-net argument. Since the dimension of z_1 and z_2 are respectively $n(\ell-1)$ and 1, the size of ε-net for z_1 and z_2 is at most $e^{O(nL\log L)}$. Thus, with probability at least $1-e^{O(nL\log L+\log n)}\cdot\exp(-\Omega(\varepsilon^2 m))\geq 1-\exp(-\Omega(\varepsilon^2 m))$ we have: for all $z_1\in\mathbb{R}^{n(\ell-1)}$ and $z_2\in\mathbb{R}$ satisfying $\|z_1\|\leq 2L+4$ and $0\leq z_2\leq 2L+4$, and fixed $z_3=x_{i,\ell+1}$,

$$\|h_{i,\ell+1}\|\leq(1+2\varepsilon)\cdot\sqrt{\|z_1\|_2^2+z_2^2+\|z_3\|_2^2}.$$

In particular, since we have "for all" quantifies on z_1 and z_2 above, we can substitute the choice of z_1 and z_2 in (B.1) (which may depend on the randomness of W and A). This, together with (B.2), gives

$$\mathbf{Pr}\left[\|h_{\ell+1}\|_2\leq(1+\varepsilon)\cdot\sqrt{\|h_{i,\ell}\|_2^2+1}\right]\geq 1-\exp(-\Omega(\varepsilon^2 m))\ .$$ □

Similarly, we can prove a lower bound

Claim B.5. *With probability at least* $1-\exp(-\Omega(m/L^2))$ *over* W *and* A*,*

$$\|h_{i,\ell+1}\|_2\geq(1-\frac{1}{4L})\|h_{i,\ell}\|_2.$$

Proof. We can define z_1,z_2,z_3 in the same way as (B.1). This time, we show a lower bound

$$\|z_1\|_2^2+z_2^2+\|z_3\|_2^2\geq\|z_1\|_2^2+z_2^2=\|h_{i,\ell}\|_2^2. \tag{B.4}$$

Applying Corollary A.10, we know if z_1,z_2,z_3 are fixed (instead of defined as in (B.1)), then choosing $\varepsilon=1/8L$,

$$\mathbf{Pr}\left[\|h_{i,\ell+1}\|\geq(1-\varepsilon)\cdot\sqrt{\|z_1\|_2^2+z_2^2+\|z_3\|_2^2}\right]\geq 1-\exp(-\Omega(\varepsilon^2 m))\ ..$$

Again, after applying ε-net, we know with probability at least $1-e^{O(nL\log L+\log n)}\cdot\exp(-\Omega(\varepsilon^2 m))\geq 1-\exp(-\Omega(\varepsilon^2 m))$, for all $z_1\in\mathbb{R}^{n(\ell-1)}$ and $z_2\in\mathbb{R}$ satisfying $\|z_1\|\leq 2L+4$ and $0\leq z_2\leq 2L+4$, and fixed $z_3=x_{i,\ell+1}$,

$$\|h_{i,\ell+1}\|_2\geq(1-\varepsilon)\cdot\sqrt{\|z_1\|_2^2+z_2^2+\|z_3\|_2^2}.$$

Substituting the choice of z_1,z_2,z_3 in (B.1), and the lower bound (B.4), we have

$$\mathbf{Pr}\left[\|h_{i,\ell+1}\|\geq(1-\varepsilon)\cdot\|h_{i,\ell}\|_2\right]\geq 1-\exp(-\Omega(\varepsilon^2 m)).$$

Choosing $\varepsilon=1/(4L)$ gives the desired statement. □

Finally, recursively applying Claim B.4 and Claim B.5 for all $\ell=0,1,\ldots,L-1$, we have with probability at least $1-L^2\exp(-\Omega(m/L^2))\geq 1-\exp(-\Omega(m/L^2))$, we have

$$\|h_{i,\ell}\|_2\geq(1-\frac{1}{4L})^\ell,\qquad\|h_{i,\ell}\|_2\leq 2\ell+4\ .$$

This finishes the proof of Lemma B.3. ■

B.2 Forward Correlation

This subsection proves the following lemma which, as discussed in Section 5, bounds how much "fresh new randomness" is left after propagating to layer ℓ.

Lemma B.6 (c.f. (5.2))**.** *With probability at least* $1 - e^{-\Omega(\sqrt{m})}$ *over the random initialization* W, A *in Definition 2.3, letting* U_ℓ *be defined in Definition B.1, we have*

$$\forall i \in [n], \forall \ell \in \{0, 1, \cdots, L-1\} \quad : \quad \|(I - U_\ell U_\ell^\top) h_{i,\ell+1}\|_2 \geq \frac{1}{2 \cdot 10^6 L^2 \log^3 m} \enspace .$$

To prove Lemma B.6, we inductively (with the increasing order of ℓ) show for each $i \in [n]$, for each $\ell \in \{0, 1, \cdots, L-1\}$, we have

$$\|(I - U_\ell U_\ell^\top) h_{i,\ell+1}\|_2 \geq \xi_\ell \stackrel{\text{def}}{=} \frac{1}{10^6 L^2 \log^3 m} (1 - 2\alpha - \frac{1}{4L})^\ell \quad \text{where} \quad \alpha \stackrel{\text{def}}{=} \frac{1}{2 \cdot 10^4 L^2 \log^2 m} \enspace . \tag{B.5}$$

We first show (B.5) in the base case $\ell = 0$. Since U_0 is an empty matrix, we have $(I - U_0 U_0^\top) h_{i,1} = h_{i,1}$ so according to Lemma B.3 we have $\|h_{i,1}\|_2 \geq (1 - 1/4L)$.

The remainder of the proof assumes (B.5) already holds for $\ell - 1$. We can write

$$h_{i,\ell+1} = \phi(W h_{i,\ell} + A x_{i,\ell+1}) = \phi(M_1 z_1 + M_2 z_2 + M_3 z_3).$$

where $M_1 \in \mathbb{R}^{m \times n(\ell-1)}$, $M_2 \in \mathbb{R}^{m \times 1}$, $M_3 \in \mathbb{R}^{m \times d}$ and $z_1 \in \mathbb{R}^{n(\ell-1)}$, $z_2 \in \mathbb{R}$, $z_3 \in \mathbb{R}^d$ are defined as

$$\begin{aligned} M_1 &= W U_{\ell-1} & M_2 &= W v_{i,\ell} & M_3 &= A \\ z_1 &= U_{\ell-1}^\top h_{i,\ell} & z_2 &= \|(I - U_{\ell-1} U_{\ell-1}^\top) h_{i,\ell}\|_2 & z_3 &= x_{i,\ell+1} \enspace . \end{aligned} \tag{B.6}$$

in the same way as (B.1) and (B.3) as in the proof of Lemma B.3. We again have the entries of M_1, M_2, M_3 are i.i.d. from $\mathcal{N}(0, \frac{2}{m})$ (recall Footnote 9). For the quantity $M_2 z_2$, we can further decompose it as follows

$$M z_2 = \nu \cdot (z_2 - c_5 \alpha)_+ + \nu' z_2' \tag{B.7}$$

where $c_5 = \frac{1}{16 \log m}$ is some fixed parameter, ν and ν' denote two vectors that are independently generated from $\mathcal{N}(0, \frac{2I}{m})$, and

$$(z_2 - c_5\alpha)_+ = \begin{cases} 0, & \text{if } z_2 < c_5\alpha; \\ \sqrt{z_2^2 - c_5^2 \alpha^2}, & \text{if } z_2 \geq c_5 \alpha. \end{cases} \qquad z_2' = \begin{cases} z_2, & \text{if } z_2 < c_5\alpha; \\ c_5\alpha, & \text{if } z_2 \geq c_5\alpha. \end{cases}$$

It is clear that the two sides of (B.7) are identical in distribution (because $M_2 \sim \mathcal{N}(0, \frac{2I}{m})$ and $(z_2 - c_5\alpha)_+^2 + (z_2')^2 = z_2^2$).

Next, suppose z_1 and z_2 are fixed (instead of depending on the randomness of W and A) and satisfies[10]

$$\frac{1}{\sqrt{2}} \leq \|z_1\| \leq 2L + 6 \text{ and } |z_2| \leq 2L + 6. \tag{B.8}$$

We can apply Corollary A.16 to obtain the following statement: with probability at least $1 - \exp(-\Omega(\alpha^2 m))$,

$$w \stackrel{\text{def}}{=} M_1 z_1 + \nu (z_2 - c_5\alpha)_+ + M_3 z_3$$

is $(\alpha, \sigma/4)$-good where $\sigma = \left(\frac{2}{m}\|z_1\|_2^2 + \frac{2}{m}((z_2 - c_5\alpha)_+)^2 + \frac{2}{m}\|z_3\|_2^2\right)^{1/2}$. Using $\|z_1\|_2^2 \geq 1/2$, we can lower bound σ^2 as

$$\sigma^2 \geq \frac{2}{m}(\|z_1\|_2^2 + z_2^2 - c_5^2\alpha^2) \geq \frac{2}{m}(\frac{1}{2} - c_5^2 \alpha^2) \geq \frac{1}{2m} \enspace .$$

In other words, w is $(\alpha, \frac{1}{4\sqrt{2}\sqrt{m}})$-good. Next, applying standard ε-net argument, we have with probability at least $1 - \exp(-\Omega(\alpha^2 m))$, for all vectors $z_1 \in \mathbb{R}^{n(\ell-1)}$ and $z_2 \in \mathbb{R}$ satisfying (B.8), it

[10]Note that if z_1 and z_2 are random, then they satisfy such constraints by Lemma B.3.

satisfies w is $(\alpha, \frac{1}{8\sqrt{m}})$-good. This allows us to plug in the random choice of z_1 and z_2 in (B.6).

We next apply Lemma B.7 with

$$w = M_1 z_1 + \nu(z_2 - c_5\alpha)_+ + M_3 z_3, \quad r = z_2', \quad v = \nu, \quad U = U_\ell.$$

(We can do so because the randomness of v is independent of the randomness of U_ℓ and w.) Lemma B.7 tells us that, with probability at least $1 - \exp(-\Omega(\sqrt{m}))$ over the randomness of v,

$$\|(I - U_\ell U_\ell^\top) \cdot \phi(w + rv)\| \geq r(1 - 2\alpha) - \frac{\alpha^{1.5}}{4} \ .$$

By induction hypothesis, we know

$$r = z_2' \geq \min\{z_2, c_5\alpha\} \geq \min\{\xi_{\ell-1}, c_5\alpha\} \geq \xi_{\ell-1} \ ,$$

where the last step follows by $\xi_{\ell-1} \leq c_5\alpha$. Thus, we have

$$\|(I - U_\ell U_\ell^\top) \cdot \phi(w + rv)\| \geq r(1 - 2\alpha) - \frac{\alpha^{1.5}}{4} \geq \xi_{\ell-1}(1 - 2\alpha) - \frac{\alpha^{1.5}}{4} \geq \xi_{\ell-1}(1 - 2\alpha - \frac{1}{4L}) = \xi_\ell \ ,$$

where the last inequality uses $\alpha^{1.5} \leq \frac{\xi_{\ell-1}}{L}$.

■

B.2.1 Tools

Lemma B.7. *Suppose $\alpha \in [\frac{1}{100L^4}, 1/2)$, $m \geq 4\widetilde{d}/\alpha$ and $r \in (0, \frac{\alpha}{16\log m}]$. Suppose w is a fixed vector that is $(\alpha, \frac{1}{8\sqrt{m}})$-good (see Definition A.14), and $U \in \mathbb{R}^{m\times\widetilde{d}}$ is a fixed column orthonormal matrix. Then, if $v \sim \mathcal{N}(0, \frac{2}{m}I)$, with probability at least $1 - \exp(-\Omega(\sqrt{m}))$, it satisfies*

$$\|(I - UU^\top) \cdot \phi(w + rv)\| \geq r(1 - 2\alpha) - \frac{\alpha^{1.5}}{4} \ .$$

Proof. We split $w \in \mathbb{R}^m$ into three pieces $w = (w_1, w_2, w_3) \in \mathbb{R}^m$ with disjoint support such that:

- w_1 corresponds to the coordinates that are $\geq \alpha/8\sqrt{m}$,
- w_2 is the remaining, which corresponds to the coordinates that are within $(-\alpha/8\sqrt{m}, \alpha/8\sqrt{m})$.
- w_3 corresponds to the coordinates that are $\leq -\alpha/8\sqrt{m}$, and

We write $v = (v_1, v_2, v_3)$ according to the same partition. We consider the following three cases.

- For each index k in the first block, we have $\big(\phi(w_1 + rv_1)\big)_k \neq \big(w_1 + rv_1\big)_k$ only if $(rv_1)_k \leq -\alpha/8\sqrt{m}$. However, if this happens, we have

  $$0 \leq \big(\phi(w_1 + rv_1)\big)_k - \big(w_1 + rv_1\big)_k \leq \max\big\{\big(rv_1\big)_k - \alpha/8\sqrt{m}, 0\big\} \ .$$

  Applying Lemma A.5, we know with probability at least $1 - e^{-\Omega(\sqrt{m})}$,

  $$\delta_1 \stackrel{\text{def}}{=} \phi(w_1 + rv_1) - \big(w_1 + rv_1\big) \quad \text{satisfies} \quad \|\delta_1\|^2 \leq 2\sqrt{m}\frac{2r^2}{m} \leq 4\alpha^2/\sqrt{m} \ .$$

- Similarly, for each index k in the third block, we have $\big(\phi(w_1 + rv_1)\big)_k \neq 0$ only if $(rv_1)_k \geq \alpha/8\sqrt{m}$. Therefore, we can similarly derive that

  $$\delta_3 \stackrel{\text{def}}{=} \phi(w_3 + rv_3) \quad \text{satisfies} \quad \|\delta_3\|^2 \leq 4\alpha^2/\sqrt{m} \ .$$

- For the second block, we claim that

  $$\delta_2 \stackrel{\text{def}}{=} \phi(w_2 + rv_2) \quad \text{satisfies} \quad \|\delta_2\|_2 \leq \alpha^{3/2}/4 \ . \tag{B.9}$$

To prove (B.9), we use triangle inequality,

$$\|\phi(w_2 + rv_2)\|_2 \leq \|\phi(w_2)\|_2 + \|\phi(rv_2)\|_2$$

Since $\|\phi(w_2)\|_\infty \leq \alpha/8\sqrt{m}$ and the size of support of $\phi(w_2)$ is at most αm, we have

$$\|\phi(w_2)\|_2 \leq ((\alpha/8\sqrt{m})^2\alpha m)^{1/2} \leq \frac{\alpha^{3/2}}{8}.$$

Since $v \sim \mathcal{N}(0, \frac{2}{m}I)$, and since the size of support of v_2 is at most αm, we have $\|v_2\| \leq 2\sqrt{\alpha}$ with probability at least $1 - e^{-\Omega(\alpha m)}$ (due to chi-square distribution concentration). Thus

$$\|\phi(rv_2)\|_2 = r \cdot \|\phi(v_2)\|_2 \leq \frac{\alpha}{16} \cdot 2\sqrt{\alpha} = \frac{\alpha^{3/2}}{8} \ .$$

Together, by triangle inequality we have $\|\phi(w_2 + rv_2)\| \leq \frac{\alpha^{3/2}}{4}$. This finishes the proof of (B.9).

Denoting by $\delta = (\delta_1, \delta_2, \delta_3)$, we have

$$\phi(w + rv) = w_1 + rv_1 + \delta.$$

Taking the norm on both sides,

$$\begin{aligned}
\|(I - UU^\top)\phi(w + rv)\| &= \|(I - UU^\top)(w_1 + rv_1 + \delta)\| \\
&\geq \|(I - UU^\top)(w_1 + rv_1)\| - \|(I - UU^\top)\delta\| && \text{(by triangle inequality)} \\
&\geq \|(I - UU^\top)(w_1 + rv_1)\| - \|\delta\| && \text{(by } \|(I - UU^\top)\delta\| \leq \|\delta\|\text{)} \\
&\geq \|(I - UU^\top)(w_1 + rv_1)\| - \frac{\alpha^{3/2}}{4} - \frac{4\alpha}{m^{1/4}} \\
&\geq \|(I - UU^\top)(w_1 + rv_1)\| - \frac{\alpha^{3/2}}{2} \\
&\geq \|(I - UU^\top)w_1 + rv_1\| - r\|U^\top v_1\| - \frac{\alpha^{3/2}}{2} \ .
\end{aligned}$$

Now, since $U \in \mathbb{R}^{m \times \widetilde{d}}$, it is not hard to show that $\|U^\top v_1\|_2^2 \leq \frac{2\widetilde{d}}{m}$ with probability at least $1 - \exp(-\Omega(m))$. Using $\frac{2\widetilde{d}}{m} \leq \frac{\alpha}{2}$ (owing to our assumption $m \geq 4\widetilde{d}/\alpha$), we have $r\|U^\top v_1\| \leq \frac{r\alpha}{2}$.

On the other hand, the random vector $z = (I - UU^\top)w_1 + rv_1$ follows from distribution $\mathcal{N}(\mu, \frac{2r^2}{m})$ for some fixed vector $\mu = (I - UU^\top)w_1$ and has at least $\frac{m}{2}(1-\alpha)$ dimensions. By chi-square concentration, we have $\|z\| \geq r(1 - 3\alpha/2)$ with probability at least $1 - e^{-\Omega(\alpha^2 m)}$. Putting these together, we have

$$\|(I - UU^\top)\phi(w + rv)\| \geq r(1 - 3\alpha/2) - \frac{r\alpha}{2} - \frac{\alpha^{3/2}}{2} = r(1 - 2\alpha) - \frac{\alpha^{3/2}}{2} \ .$$

□

B.3 Forward δ-Separateness

We first give the definition of δ-Separable,

Definition B.8 (δ-separable vectors)**.** *For any two vectors x, y, we say x and y are δ-separable if*

$$\left\|(I - \frac{yy^\top}{\|y\|_2^2})x\right\| \geq \delta \quad \textit{and} \quad \left\|(I - \frac{xx^\top}{\|x\|_2^2})y\right\| \geq \delta \ ..$$

We say a finite set X is δ-separable if for any two vectors $x, y \in X$, x and y are δ-separable.

The goal of this subsection is to prove the δ-separateness over all layers ℓ.

Lemma B.9 (c.f. (5.3))**.** *Let $\{x_{i,1}\}_{i\in[n]}$ be δ-separable with $\delta \le \frac{1}{10^6 L^2 \log^3 m}$. With probability at least $1 - e^{-\Omega(\sqrt{m})}$, for all $\ell = 0, 1, \ldots, L-1$, for all $i, j \in [n]$ with $i \ne j$, we have*

$$(I - U_\ell U_\ell^\top) h_{i,\ell+1} \text{ and } (I - U_\ell U_\ell^\top) h_{j,\ell+1} \text{ are } \frac{\delta}{2}\text{-separable.}$$

To prove Lemma B.9, we inductively (with the increasing order of ℓ) show for each $i \in [n]$, for each $\ell \in \{0, 1, \cdots, L-1\}$, we have

$$(I - U_\ell U_\ell^\top) h_{i,\ell+1} \text{ and } (I - U_\ell U_\ell^\top) h_{j,\ell+1} \text{ are } \delta_\ell \text{ separable for } \delta_\ell \stackrel{\text{def}}{=} \delta(1 - 2\alpha - \frac{1}{4L})^\ell \tag{B.10}$$

where $\alpha \stackrel{\text{def}}{=} 16\delta \log m$.

We skip the base case and only prove (B.10) for $\ell \ge 1$ by assuming (B.10) already holds for $\ell - 1$.[11]

Claim B.10. *For $i \ne j$, if $(I - U_{\ell-1} U_{\ell-1}^\top) h_{i,\ell}$ and $(I - U_{\ell-1} U_{\ell-1}^\top) h_{j,\ell}$ are $\delta_{\ell-1}$-separable, then letting $U = [U_\ell, \widehat{h}]$ where $\widehat{h} = \frac{(I - U_\ell U_\ell^\top) h_{j,\ell+1}}{\|(I - U_\ell U_\ell^\top) h_{j,\ell+1}\|_2}$, we have*

$$\|(I - UU^\top) h_{i,\ell+1}\| \ge \delta_\ell = \delta_{\ell-1}(1 - 3\alpha) \ .$$

holds with probability at least $1 - \exp(-\Omega(\sqrt{m}))$.

Since it is easy to verify that

$$\left\| \left(I - \widehat{h}\widehat{h}^\top \right) (I - U_\ell U_\ell^\top) h_{i,\ell+1} \right\| = \|(I - UU^\top) h_{i,\ell+1}\|_2 \ ,$$

Claim B.10 immediately implies (B.10) for layer ℓ, and thus finishes the proof of Lemma B.9. Therefore, we only need to prove Claim B.10 below.

Proof of Claim B.10. Let $x = (I - U_{\ell-1} U_{\ell-1}^\top) h_{j,\ell}$ and $y = (I - U_{\ell-1} U_{\ell-1}^\top) h_{i,\ell}$. If x and y are $\delta_{\ell-1}$-separable, we split y into two parts where y_1 is parallel to x and y_2 is orthogonal to x

$$y = y_1 + y_2, \quad y_1 = \frac{\langle x, y\rangle x}{\|x\|_2^2}, \quad y_2 = (I - xx^\top / \|x\|_2^2) y$$

This also implies that the randomness in y_1 is independent of the randomness in y_2.

It is easy to see that $\|y_1\|_2 = \frac{\langle x, y\rangle}{\|x\|_2}$, so we can rewrite Wy as follows

$$Wy = Wy_1 + Wy_2 = W \frac{\langle x, y\rangle x}{\|x\|_2^2} + Wy_2 = (\|y_1\|_2 / \|x\|_2) Wx + Wy_2.$$

Therefore,

$$\begin{aligned}
Wh_{i,\ell} + Ax_{i,\ell+1} &= WU_{\ell-1} U_{\ell-1}^\top h_{i,\ell} + W(I - U_{\ell-1} U_{\ell-1}^\top) h_{i,\ell} + Ax_{i,\ell+1} \\
&= WU_{\ell-1} U_{\ell-1}^\top h_{i,\ell} + Wy + Ax_{i,\ell+1} && \text{(by definition of } y\text{)} \\
&= WU_{\ell-1} U_{\ell-1}^\top h_{i,\ell} + (\|y_1\|_2 / \|x\|_2) Wx + Ax_{i,\ell+1} + Wy_2 && \text{(by rewritting } Wy\text{)} \\
&= M_1 z_1 + M_2 z_2 + M_3 z_3 + M_4 z_4
\end{aligned}$$

[11] The proof of the base case is a replication of Claim B.10 and only simpler. Indeed, for the base case of $\ell = 0$, one can view $h_{i,\ell} = x_{i,1}$ and $h_{j,\ell} = x_{j,1}$, and view U_ℓ as an empty matrix. Then, the same analysis of Claim B.10 but with slightly different notations will apply.

where

$$
\begin{aligned}
&M_1 = WU_{\ell-1} && M_2 = W\frac{x}{\|x\|_2} && M_3 = A && M_4 = W\frac{y_2}{\|y_2\|_2} \\
&z_1 = U_{\ell-1}^\top h_{i,\ell} && z_2 = \|y_1\|_2 && z_3 = x_{i,\ell+1} && z_4 = \|y_2\|_2 \ ,
\end{aligned}
\tag{B.11}
$$

and we know the entries of M_1, M_2, M_3, M_4 are i.i.d. from $\mathcal{N}(0, \frac{2}{m})$, owing to a similar treatment as Footnote 9. For the vector $M_4 z_4$, we further rewrite it as

$$M_4 z_4 = \nu \cdot (z_4 - c_5\alpha)_+ + \nu' z_4'$$

where $c_5 = \frac{1}{16 \log m}$ is a fixed parameter, and ν and ν' denote two vectors that are independently generated from the same distribution $\mathcal{N}(0, \frac{2}{m}I)$ as vector $M_4 \in \mathbb{R}^m$, and

$$
(z_4 - c_5\alpha)_+ = \begin{cases} 0, & \text{if } z_4 < c_5\alpha; \\ \sqrt{z_4^2 - c_5^2\alpha^2}, & \text{if } z_4 \geq c_5\alpha. \end{cases} \qquad z_4' = \begin{cases} z_4, & \text{if } z_4 < c_5\alpha; \\ c_5\alpha, & \text{if } z_4 \geq c_5\alpha. \end{cases}
$$

Together, we can write

$$
\begin{aligned}
(I - UU^\top)h_{i,\ell+1} &= (I - UU^\top)\phi(Wh_{i,\ell} + Ax_{i,\ell+1}) \\
&= (I - UU^\top)\phi(M_1 z_1 + M_2 z_2 + M_3 z_3 + M_4 z_4) \\
&= (I - UU^\top)\phi(M_1 z_1 + M_2 z_2 + M_3 z_3 + \nu \cdot (z_4 - c_5\alpha)_+ + \nu' z_4') \\
&= (I - UU^\top)\phi(w + \nu' z_4')
\end{aligned}
$$

where the last step follows by defining

$$w = M_1 z_1 + M_2 z_2 + M_3 z_3 + \nu \cdot (z_4 - c_5\alpha)_+ \ .$$

Our plan is to first use the randomness in w to argue that w is $(\alpha, \gamma/4\sqrt{m})$-good. Then we conditioned w is good, and prove that the norm of $(I - UU^\top)\phi(w + \nu' z_4')$ is lower bounded (using (B.7)).

Now, suppose that $z_1 \in \mathbb{R}^{n(\ell-1)}$, $z_2 \in \mathbb{R}$, and $z_4 \in \mathbb{R}$ are fixed (instead of computed based on the randomness of W and A), and satisfies[12]

$$\frac{1}{\sqrt{2}} \leq \|z_1\| \leq 2L + 6 \text{ and } |z_2|, |z_4| \leq 2L + 6. \tag{B.12}$$

we can use Corollary A.16 to obtain the following statement: w is $(\alpha, \sigma/4)$-good with probability at least $1 - \exp(-\Omega(\alpha^2 m))$ where $\sigma = \left(\frac{2}{m}\|z_1\|_2^2 + \frac{2}{m}z_2^2 + \frac{2}{m}\|z_3\|_2^2 + \frac{2}{m}(z_4^2 - c_5^2\alpha^2)\right)^{1/2}$. Using $\|z_1\|_2^2 \geq 1/2$, we can lower bound σ^2 as

$$\sigma^2 \geq \frac{2}{m}(\|z_1\|_2^2 + z_2^2 + z_4^2 - c_5^2\alpha^2) \geq \frac{2}{m}(\frac{1}{2} - c_5^2\alpha^2)$$

In other words, w is $(\alpha, \frac{1}{4\sqrt{2}\sqrt{m}})$-good. Next, applying standard ε-net argument, we have with probability at least $1 - \exp(-\Omega(\alpha^2 m))$, for all z_1, z_2, z_4 satisfying (B.12), it satisfies w is $(\alpha, \frac{1}{8\sqrt{m}})$-good. This allows us to plug in the random choice of z_1, z_2, z_4 in (B.11).

We apply Lemma B.7 with the following setting

$$U = \left[U_\ell, \frac{(I - U_\ell U_\ell^\top)h_{j,\ell+1}}{\|(I - U_\ell U_\ell^\top)h_{j,\ell+1}\|_2}\right], \quad w = w, \quad v = \nu', \quad r = z_4' \in [0, c_5\alpha]$$

where w is $(\alpha, \gamma/8\sqrt{m})$-good. (We can do so because the randomness of v is independent of the randomness of U and w.) Lemma B.7 tells us that, with probability at least $1 - \exp(-\Omega(\sqrt{m}))$

[12]Note that if z_1, z_2 and z_4 are random, then they satisfy such constraints by Lemma B.3.

over the randomness of v,

$$\|(I - UU^\top) \cdot \phi(w + rv)\| \geq r(1-2\alpha) - \frac{\alpha^{1.5}}{4} \enspace .$$

By induction hypothesis, we know

$$r = z_4' \geq \min\{z_4, c_5\alpha\} \geq \min\{\delta_{\ell-1}, c_5\alpha\} \geq \delta_{\ell-1} \enspace ,$$

where the last step follows by $\delta_{\ell-1} \leq c_5\alpha$. Thus, we have

$$\|(I - UU^\top) \cdot \phi(w + rv)\| \geq r(1-2\alpha) - \frac{\alpha^{1.5}}{4} \geq \delta_{\ell-1}(1-2\alpha) - \frac{\alpha^{1.5}}{4} \geq \delta_{\ell-1}(1 - 2\alpha - \frac{1}{4L}) = \delta_\ell \enspace ,$$

where the last inequality uses $\alpha^{1.5} \leq \frac{\delta_{\ell-1}}{L}$. □

B.4 Intermediate Layers: Spectral Norm

The following lemma bounds the spectral norm of (consecutive) intermediate layers.

Lemma B.11 (c.f. (5.4))**.** *With probability at least $1 - \exp(-\Omega(m/L^2))$, we have for all $L \geq \ell_2 \geq \ell_1 \geq 0$ and $i \in [n]$*

$$\left\|\prod_{\ell=\ell_2}^{\ell_1} D_{i,\ell} W\right\|_2 \leq O(L^3),$$

We start with an important claim whose proof is almost identical to Lemma B.3.

Claim B.12. *Given $\ell > b \geq 1$, given $i \in [n]$, and given $z_{b-1} \in \mathbb{R}^m$ a fixed vector, letting*

$$z_\ell = D_{i,\ell} W D_{i,\ell-1} \cdots D_{i,b} W z_{b-1} \enspace ,$$

we have with probability at least $1 - \exp(-\Omega(m/L^2))$, it satisfies $\|z_\ell\|_2 \leq (1 + 1/L)^{\ell-b+1}\|z_{b-1}\|_2$.

Proof. Let U denote the following column orthonormal matrix using Gram-Schmidt (its first $n(\ell-1)$ columns coincide with $U_{\ell-1}$):

$$U \stackrel{\text{def}}{=} \mathsf{GS}\left(h_{1,1}, \dots, h_{n,1}, h_{1,2}, \dots, h_{n,2}, \dots, h_{1,\ell-1}, \dots, h_{n,\ell-1}, z_{b-1}, \cdots z_{\ell-1}\right) \enspace .$$

We can rewrite $\|z_\ell\|_2$ as follows:

$$\begin{aligned}
\|z_\ell\|_2 &= \|D_{i,\ell} W z_{\ell-1}\|_2 \\
&= \left\|\mathbf{1}_{W h_{i,\ell-1} + A x_{i,\ell} \geq 0} \cdot W z_{\ell-1}\right\|_2 \\
&= \left\|\mathbf{1}_{W U U^\top h_{i,\ell-1} + A x_{i,\ell} \geq 0} \cdot W U U^\top z_{\ell-1}\right\|_2 \\
&= \|\mathbf{1}_{My + Ax \geq 0} M z\|_2 \, .
\end{aligned}$$

where in the last step we have defined $M = WU$, $y = U^\top h_{i,\ell-1}$, $x = x_{i,\ell}$ and $z = U^\top z_{\ell-1}$. We stress here that the entries of M and A are i.i.d. from $\mathcal{N}(0, \frac{2}{m}I)$.[13]

Now, Claim B.13 tells us for fixed $x \in \mathbb{R}^d$ and fixed $y, z \in \mathbb{R}^{n\ell+\ell-b}$, we have with probability $1 - \exp(-\Omega(m/L^2))$ (over the randomness of M),

$$\|\mathbf{1}_{My + Ax \geq 0} M z\|_2 \leq \|z\|_2 (1 + 1/2L) \enspace .$$

After taking ε-net over all possible y, z, we have that for fixed $x \in \mathbb{R}^d$ but all $y, z \in \mathbb{R}^{n\ell+\ell-b}$:

$$\|\mathbf{1}_{My + Ax \geq 0} M z\|_2 \leq \|z\|_2 (1 + 1/L) \enspace .$$

[13] This follows from a similar argument as Footnote 9, taking into account the additional fact that each z_j may only depend on the randomness of A, $WU_{\ell-1}$, and $Wz_{b-1}, \dots Wz_{j-1}$.

We can thus plug in the choice $y = U^\top h_{i,\ell-1}$ and $z = U^\top z_{\ell-1}$ (both of which may depend on the randomness of W and A). Using $\|z\|_2 \le \|z_{\ell-1}\|_2$, we have

$$\|z_\ell\|_2 \le \|z_{\ell-1}\|_2(1+1/L) \ .$$

Finally, taking union bound over all possible ℓ and applying induction, we have

$$\|z_\ell\|_2 \le (1+1/L)^{\ell-b+1}\|z_{b-1}\|_2. \qquad \square$$

Now, to prove the spectral norm bound in Lemma B.11, we need to go from "for each z_{b-1} (see Claim B.12)" to "for all z_{b-1}". Since z_{b-1} has m dimensions, we cannot afford taking union bound over all possible z_{b-1} (or its ε-net).

To bypass this issue, we partition the coordinates of z_{b-1} into L^3 trunk (each of length m/L^3). We write $z_{b-1} = \sum_{j=1}^{L^3}(z_{b-1})_j$ where each $(z_{b-1})_j \in \mathbb{R}^m$ denotes a vector that only has non-zero entries on m/L^3 coordinates. For each $j \in [L^3]$, we have with probability $1-\exp(-m/L^2)$,

$$\|(z_\ell)_j\|_2 \le (1+1/L)^{\ell-b}\cdot\|(z_{b-1})_j\|_2 \le 2\|(z_{b-1})_j\|_2 \ .$$

By applying an ε-net argument over all such possible (but sparse) $(z_{b-1})_j$, we have with probability at least $1-2^{O(m/L^3)}\exp(-\Omega(m/L^2)) \ge 1-\exp(-\Omega(m/L^2))$, the above equation holds for all possible $(z_{b-1})_j$.

Next, taking a union bound over all $j \in [L^3]$, we have with probability at least $1-\exp(-\Omega(m/L^2))$:

$$\forall z_{b-1} \in \mathbb{R}^m \colon \|z_\ell\|_2 \le O(L^3)\|z_{b-1}\|_2 \ .$$

Taking a union bound over all ℓ, b, we complete the proof of Lemma B.11. ∎

B.4.1 Tools

Claim B.13. *For fixed $x \in \mathbb{R}^d, y, z \in \mathbb{R}^k$. Let $M \in \mathbb{R}^{m\times k}$ and $A \in \mathbb{R}^{m\times d}$ denote random Gaussian matrices where each entry is i.i.d. sampled from $\mathcal{N}(0, 2/m)$. We have with probability at least $1-\exp(-\Omega(m/L^2))$*

$$\|\mathbf{1}_{My+Ax\ge 0}\cdot Mz\|_2 \le \|z\|_2(1+1/2L) \ .$$

Proof. Without loss of generality we assume $\|z\|_2 = 1$. We can rewrite My as follows

$$\begin{aligned} My &= M(zz^\top)y + M(I-zz^\top)y \\ &= Mz\cdot z^\top y + \frac{M(I-zz^\top)y}{\|(I-zz^\top)y\|_2}\cdot\|(I-zz^\top)y\|_2 \\ &= M_1z_1 + M_2z_2, \end{aligned}$$

where $M_1, M_2 \in \mathbb{R}^m$ and $z_1, z_2 \in \mathbb{R}$ are defined as follows

$$\begin{aligned} M_1 = Mz, &\quad M_2 = \frac{M(I-zz^\top)y}{\|(I-zz^\top)y\|_2} \\ z_1 = z^\top y, &\quad z_2 = \|(I-zz^\top)y\|_2 \end{aligned}$$

It is easy to see that M_1 is independent of M_2. We can rewrite

$$\|\mathbf{1}_{My+Ax\ge 0}\cdot Mz\|_2 = \|\mathbf{1}_{M_1z_1+M_2z_2+Ax\ge 0}\cdot M_1\|_2$$

Using Lemma A.11, we have with probability at least $1-\exp(-\Omega(m/L^2))$,

$$\|\mathbf{1}_{M_1z_1+M_2z_2+Ax\ge 0}\cdot M_1\|_2 \le \|z\|_2(1+1/2L).$$

Thus, we complete the proof. $\square$

B.5 Intermediate Layers: Sparse Spectral Norm

This section proves two results corresponding to the spectral norm of intermediate layers with respect to *sparse* vectors. We first show Lemma B.14 and our Corollary B.15 and B.16 shall be direct applications of Lemma B.14.

Lemma B.14. *For every $k \in [m]$ and $t \geq 2$, with probability at least*

$$1 - nL^2 e^{O(k\log(m))}(e^{-\Omega(m/L^2)} + e^{-\Omega(nLt^2)})$$

it satisfies, for all $i \in [n]$, for all $L > \ell_2 \geq \ell_1 \geq 1$, for all k-sparse vectors $z, y \in \mathbb{R}^m$

$$\left| y^\top W \left(\prod_{\ell=\ell_2}^{\ell_1} D_{i,\ell} W \right) z \right| \leq \frac{5t\sqrt{nL}}{m^{1/2}} \cdot \|y\|_2 \cdot \|z\|_2 .$$

Proof. Without loss of generality we assume $\|y\| = \|z\| = 1$. Fixing $\ell_2 \geq \ell_1$, fixing i, and fixing $z_{\ell_1 - 1} = z$, we have according to Claim B.12, letting $z_{\ell_2} = D_{i,\ell_2} W D_{i,\ell_2-1} W \cdots D_{i,\ell_1} z_{\ell_1-1}$, then with probability at least $1 - e^{-\Omega(m/L^2)}$

$$\|z_{\ell_2}\|_2 \leq (1+1/L)^{\ell_2-\ell_1-1} \|z_{\ell_1-1}\|_2 \leq 2 \ .$$

Applying ε-net over all k-sparse vectors z, we have with probability at least $1 - e^{O(k\log m)} e^{-\Omega(m/L^2)}$, it satisfies $\|z_{\ell_2}\|_2 \leq 3$ for all k-sparse z:

Similar to the proof of Lemma B.11, we let U denote the following column orthonormal matrix using Gram-Schmidt:

$$U \stackrel{\text{def}}{=} \mathsf{GS}\left(h_{1,1}, \dots, h_{n,1}, h_{1,2}, \dots, h_{n,2}, \dots, h_{1,\ell_2-1}, \dots, h_{n,\ell_2-1}, z_{\ell_1-1}, \cdots z_{\ell_2}\right) \ ,$$

and we have

$$\left| y^\top W \left(\prod_{\ell=\ell_2}^{\ell_1} D_{i,\ell} W \right) z \right| = \left| y^\top W U U^\top z_{\ell_2} \right| \leq \left\| y^\top W U \right\| \cdot \|U^\top z_{\ell_2}\| \leq 3 \left\| y^\top W U \right\| \ .$$

Next, observe the entries of $WU \in \mathbb{R}^{m \times n(\ell_2-1)+(\ell_2-\ell_1+2)}$ are i.i.d. drawn from $\mathcal{N}(0, \frac{2}{m})$ (following a similar argument as Footnote 13). Therefore, if y is a fixed vector, then $y^\top WU$ is in distribution identical to a Gaussian vector $\mathcal{N}(0, \frac{2}{m} I)$ of $n(\ell_2-1)+(\ell_2-\ell_1+2) \leq 2nL$ dimensions. By chi-square distribution tail bound (see Lemma A.3), we have for $t \geq 2$:

$$\mathbf{Pr}[\|y^\top W U\|^2 \geq \frac{nLt^2}{m}] \leq e^{-\Omega(nLt^2)} \ .$$

Applying ε-net over all k-sparse vectors y, we have with probability at least $1 - e^{O(k\log m)} e^{-\Omega(nLt^2)}$, it satisfies $\|y^\top WU\|^2 \leq \frac{2nLt^2}{m}$ for all k-sparse vectors y.

Conditioning on both events happen, we have

$$\left| y^\top W \left(\prod_{\ell=\ell_2}^{\ell_1} D_{i,\ell} W \right) z \right| \leq 3 \frac{\sqrt{2nL} t}{\sqrt{m}} \ .$$

Taking union bound over all possible i, ℓ_1, ℓ_2 we finish the proof. □

Choosing $k = s^2 m^{2/3}$ and $t = \frac{s m^{1/3} \log m}{5\sqrt{nL}}$ in Lemma B.14, we have

Corollary B.15 (c.f. (5.5)). *Let $s \in [m^{-1/4}, m^{1/6}]$ be a fixed real. With probability at least $1 - e^{-\Omega(s^2 m^{2/3} \log^2 m)}$, we have for all $i \in [n]$, for all $L > \ell_2 \geq \ell_1 \geq 1$, for all $s^2 m^{2/3}$-sparse vectors $y, z \in \mathbb{R}^m$,*

$$\left| z^\top W \left(\prod_{\ell=\ell_2}^{\ell_1} D_{i,\ell} W \right) y \right| \leq \frac{s \log m}{m^{1/6}} \cdot \|z\|_2 \cdot \|y\|_2.$$

Choosing $k = 1$ and $t = \frac{\sqrt{nLd} \log m}{5}$ in Lemma B.14, we have

Corollary B.16 (c.f. (5.5)). *Let $\rho = nLd \log m$. With probability at least $1 - \exp(-\Omega(\rho^2))$, it satisfies for all $i \in [n]$, for all $L > \ell_2 \geq \ell_1 \geq 1$, and for all 1-sparse vectors $y, z \in \mathbb{R}^m$,*

$$\left| z^\top W \left(\prod_{\ell=\ell_2}^{\ell_1} D_{i,\ell} W \right) y \right| \leq \frac{\rho}{m^{1/2}} \cdot \|z\|_2 \cdot \|y\|_2 \ .$$

B.6 Backward Propagation

This section proves upper bound on the backward propagation against *sparse* vectors. We first show Lemma B.17 and our Corollary B.18 and B.19 shall be immediate corollaries.

Lemma B.17. *For any $k \in [m]$, $t \geq 2$ and any $a \in \mathbb{R}^d$, with probability at least*

$$1 - nL^2 e^{O(k \log m)} (e^{-\Omega(m/L^2)} + e^{-\Omega(t^2)}),$$

over the randomness of W, A, B, we have for all $i \in [n]$, for all $L \geq \ell_2 \geq \ell_1 \geq 1$, and for all k-sparse $y \in \mathbb{R}^m$,

$$\left| a^\top B \left(\prod_{\ell=\ell_2}^{\ell_1} D_{i,\ell} W \right) y \right| \leq \frac{t}{\sqrt{d}} \cdot \|a\|_2 \cdot \|y\|_2.$$

Proof of Lemma B.17. Using Claim B.12, we know that for fixed $z_{\ell_1 - 1} = y$, letting

$$z_{\ell_2} = D_{i,\ell_2} W D_{i,\ell_2 - 1} W \cdots D_{i,\ell_1} W z_{\ell_1 - 1} \ ,$$

with probability at least $1 - \exp(-\Omega(m/L^2))$ we have

$$\|z_{\ell_2}\|_2 \leq (1 + 1/L)^{\ell_2 - \ell_1 + 1} \|z_{\ell_1 - 1}\|_2 \leq 3 \|z_{\ell_1 - 1}\|_2.$$

Next, fixing vectors a and z_{ℓ_2} and letting B be the only source of randomness, we know $a^\top B z_{\ell_2}$ follows from $\mathcal{N}(0, \|a\|_2^2 \|z_{\ell_2}\|_2^2 / d)$. Then with probability at least $1 - \exp(-\Omega(t^2))$ over B,

$$|a^\top B z_{\ell_2}| \leq t \cdot \|a\|_2 \|z_{\ell_2}\|_2 / \sqrt{d}.$$

Taking a union of the above two events, we get with probability $1 - \exp(-\Omega(m/L^2)) - \exp(-\Omega(t^2))$,

$$|a^\top B z_{\ell_2}| \leq \frac{3t}{\sqrt{d}} \cdot \|a\|_2 \cdot \|y\|_2.$$

At this point, we apply ε-net argument for all k-sparse vectors $y \in \mathbb{R}^m$ (the size of which is at most $e^{O(k \log(m/k))}$). Taking a union bound over all such vectors in the ε-net, we have with probability at least

$$1 - \exp(O(d + k \log(m/k)))(\exp(-m/L^2) + \exp(-\Omega(t^2))) \ ,$$

for all k-sparse vector $y \in \mathbb{R}^m$,

$$|a^\top B z_{\ell_2}| \leq \frac{3t}{\sqrt{d}} \cdot \|a\|_2 \cdot \|y\|_2 \ .$$

Finally, we also take a union bound over all ℓ_2, ℓ_1 and i, and there are at most $O(nL^2)$ choices. $\square$

Using Lemma B.17 with $k = s^2 m^{2/3}$ and $t = sm^{1/3}\log m$, and taking union bound over all $a \in \mathbb{R}^d$, give

Corollary B.18 (c.f. (5.6))**.** *Let $s \in [m^{-1/4}, m^{1/6}]$ be a fixed real. With probability at least $1 - \exp(-\Omega(s^2 m^{2/3}\log^2 m))$, we have for all $i \in [n]$, for all $L \geq \ell_2 \geq \ell_1 \geq 1$, for all $(s^2 \cdot m^{2/3})$-sparse $y \in \mathbb{R}^m$, for all $a \in \mathbb{R}^d$,*

$$\left| a^\top B \left(\prod_{\ell=\ell_2}^{\ell_1} D_{i,\ell} W \right) y \right| \leq (sm^{1/3}\log m) \cdot \|a\|_2 \cdot \|y\|_2 \ .$$

Using Lemma B.17 with $k = 1$ and $t = nLd\log m$, and taking union bound over all $a \in \mathbb{R}^d$, give

Corollary B.19 (c.f. (5.6))**.** *Let $\rho = nLd\log m$. With probability at least $1 - \exp(-\Omega(\rho^2))$, we have for all $i \in [n]$, for all $L \geq \ell_2 \geq \ell_1 \geq 1$, for all 1-sparse vector $y \in \mathbb{R}^m$, and for all $a \in \mathbb{R}^d$,*

$$\left| a^\top B \left(\prod_{\ell=\ell_2}^{\ell_1} D_{i,\ell} W \right) y \right| \leq \rho \cdot \|a\|_2 \cdot \|y\|_2.$$

C Stability After Adversarial Perturbation

Throughout this section, we consider some random initialization $\widetilde{W}, A, B$, and some adversarially chosen perturbation $W' \in \mathbb{R}^{m\times m}$ which may depend on the randomness of $\widetilde{W}, A, B$. We introduce the following notations in this section

Definition C.1.

$$\begin{array}{lll} \widetilde{g}_{i,0} = \widetilde{h}_{i,0} = 0 & g_{i,0} = h_{i,0} = 0 & \textit{for } i \in [n] \\ \widetilde{g}_{i,\ell} = \widetilde{W}\widetilde{h}_{i,\ell-1} + Ax_{i,\ell} & g_{i,\ell} = (\widetilde{W} + W')h_{i,\ell-1} + Ax_{i,\ell} & \textit{for } i \in [n] \textit{ and } \ell \in [L] \\ \widetilde{h}_{i,\ell} = \phi(\widetilde{W}\widetilde{h}_{i,\ell-1} + Ax_{i,\ell}) & h_{i,\ell} = \phi((\widetilde{W} + W')h_{i,\ell-1} + Ax_{i,\ell}) & \textit{for } i \in [n] \textit{ and } \ell \in [L] \\ h'_{i,\ell} = h_{i,\ell} - \widetilde{h}_{i,\ell} & g'_{i,\ell} = g_{i,\ell} - \widetilde{g}_{i,\ell} & \textit{for } i \in [n] \textit{ and } \ell \in [L] \end{array}$$

Define diagonal matrices $\widetilde{D}_{i,\ell}$ and $D_{i,\ell}$ by letting

$$(\widetilde{D}_{i,\ell})_{k,k} = \mathbf{1}_{(\widetilde{g}_{i,\ell})_k \geq 0} \textit{ and } (D_{i,\ell})_{k,k} = \mathbf{1}_{(g_{i,\ell})_k \geq 0}.$$

Accordingly, we let diagonal matrix $D'_{i,\ell} = D_{i,\ell} - \widetilde{D}_{i,\ell}$.

Roadmap.

- Section C.1 proves the stability at forward when the perturbation matrix W' has small spectral norm. It gives bounds on $\|g'_{i,\ell}\|_2$, $\|h'_{i,\ell}\|_2$, $\|D'_{i,\ell}\|_0$, and $\|D'_{i,\ell}g_{i,\ell}\|_2$.
- Section C.2 proves the stability of intermediate layers. It gives bound on $\|\prod_{\ell=\ell_2}^{\ell_1} D_{i,\ell}\widetilde{W}\|_2$ and $\|\prod_{\ell=\ell_2}^{\ell_1} D_{i,\ell}W\|_2$.
- Section C.3 analyzes the stability for backward. It bounds the difference $\|a^\top B(\prod_{\ell=\ell_2}^{\ell_1} \widetilde{D}_{i,\ell}\widetilde{W} - \prod_{\ell=\ell_2}^{\ell_1} D_{i,\ell}\widetilde{W})\|_2$.
- Section C.4 considers a special type of rank-one perturbation matrix W', and provides stability bounds on the forward and backward propagation.

As discussed in Section 6, the results of Section C.1, C.2 and C.3 shall be used twice, once for the final training updates (see Appendix G), and once for the randomness decomposition (see Appendix E). In contrast, the results of Section C.4 shall only be used once in Appendix E.

C.1 Forward

The goal of this section is to prove Lemma C.2,

Lemma C.2 (forward stability, c.f. (6.1))**.** *Letting $\varrho = nLd\delta^{-1}\log(m/\varepsilon)$, for any $\tau_0 \in [\varrho^{-100}, \varrho^{100}]$, with probability at least $1 - e^{-\Omega(L^6\tau_0^{4/3}m^{1/3})}$ over the randomness of $\widetilde{W}, A, B$, for every $W' \in \mathbb{R}^{m\times m}$ with $\|W'\|_2 \leq \frac{\tau_0}{\sqrt{m}}$, for every $i \in [n]$ and $\ell \in [L]$, we have*

(a) $\|g'_{i,\ell}\|_2, \|h'_{i,\ell}\|_2 \leq O(L^6\tau_0)/m^{1/2}$,

(b) $\|D'_{i,\ell}\|_0 \leq O(L^{10/3}\tau_0^{2/3}) \cdot m^{2/3}$, *and*

(c) $\|D'_{i,\ell}g_{i,\ell}\|_2 \leq O(L^5\tau_0)/m^{1/2}$.

Proof of Lemma C.2. Suppose $C > 1$ is a large enough constant so that the hidden constant in Lemma B.11 can be C. We inductively prove that one can write $g'_{i,\ell} = g'_{i,\ell,1} + g'_{i,\ell,2}$ where

$$\begin{aligned} &\text{I:} \quad \|g'_{i,\ell,1}\|_2 \le \tau_1 \cdot \frac{1}{m^{1/2}} \quad &&\text{II:} \quad \|g'_{i,\ell,2}\|_\infty \le \tau_2 \cdot \frac{1}{m} \quad &&\text{III:} \quad \|g'_{i,\ell}\|_2 \le \tau_3 \cdot \frac{1}{m^{1/2}} \\ &\text{IV:} \quad \|D'_{i,\ell}\|_0 \le \tau_4 \cdot m^{2/3} \quad &&\text{V:} \quad \|D'_{i,\ell} g_{i,\ell}\|_2 \le \tau_5 \cdot \frac{1}{m^{1/2}} \quad &&\text{VI:} \quad \|h'_{i,\ell}\|_2 \le (\tau_3 + \tau_5) \frac{1}{m^{1/2}} \ . \end{aligned} \tag{C.1}$$

Above, we choose parameters

$$\tau_1 = 5CL^4(L+2)\tau_0 \qquad \tau_2 = 4L\tau_5 \log m \qquad \tau_3 = \tau_1 + \tau_2 \qquad \tau_4 = 10(\tau_1)^{2/3} \qquad \tau_5 = 3\tau_1 \ .$$

We emphasize that all these parameters are polynomial in ϱ so negligible when comparing to m.

Throughout the proof, we focus on some fixed $i \in [n]$ without loss of generality, and one can always take a union bound at the end. We drop the subscript i for notational simplicity. In order to prove (C.1), we first assume that it holds for all $1, 2, \dots, \ell - 1$. In particular, we assume for all $\ell' \le \ell - 1$,

$$\|g'_{\ell',1}\|_2 \le \tau_1 \cdot \frac{1}{m^{1/2}} \quad \text{and} \quad \|g'_{\ell',2}\|_\infty \le \tau_2 \cdot \frac{1}{m} \ .$$

A useful observation is $g'_{\ell'}$ can be split into three terms

$$g'_{\ell'} = \underbrace{W' D_{\ell'-1} g_{\ell'-1}}_{z_{\ell'-1,1}} + \underbrace{\widetilde{W} D'_{\ell'-1} g_{\ell'-1}}_{z_{\ell'-1,2}} + \underbrace{\widetilde{W} \widetilde{D}_{\ell'-1} g'_{\ell'-1}}_{z_{\ell'-1,3}} \ . \tag{C.2}$$

After recursively applying (C.2), we can write

$$g'_\ell = g'_{\ell,1} = \sum_{\ell_a=1}^{\ell-1} \left(\widetilde{W}\widetilde{D}_{\ell-1} \cdots \widetilde{W}\widetilde{D}_{\ell-\ell_a+1}\right) z_{\ell-\ell_a,1} + \left(\widetilde{W}\widetilde{D}_{\ell-1} \cdots \widetilde{W}\widetilde{D}_{\ell-\ell_a+1}\right) z_{\ell-\ell_a,2} \ . \tag{C.3}$$

Applying Claim C.3, with probability at least $1 - e^{-\Omega(\tau_1^{4/3} m^{1/3})}$, we have for all $\ell' \le \ell - 1$,

$$\|z_{\ell',1}\|_2 \le \frac{\tau_0}{\sqrt{m}} \left(4\ell' + 8 + \frac{\tau_1 + \tau_2}{\sqrt{m}}\right) \tag{C.4}$$

Applying Claim C.6, we have with probability at least $1 - e^{-\Omega(\tau_1^{4/3} m^{1/3})}$, one can write

$$\left(\widetilde{W}\widetilde{D}_{\ell-1} \cdots \widetilde{W}\widetilde{D}_{\ell-\ell_a+1}\right) z_{\ell-\ell_a,2} = z_{\ell-\ell_a,2^\sharp} + z_{\ell-\ell_a,2^\flat}$$

where

$$\|z_{\ell-\ell_a,2^\sharp}\|_2 \le \frac{3\tau_5\sqrt{\tau_4}\log^2 m}{m^{2/3}}, \quad \|z_{\ell-\ell_a,2^\flat}\|_\infty \le \frac{4\tau_5 \log m}{m} \ . \tag{C.5}$$

As a result, we can define $g'_\ell = g'_{\ell,1} + g'_{\ell,2}$ for

$$g'_{\ell,1} = \sum_{\ell_a=1}^{\ell-1} \left(\widetilde{W}\widetilde{D}_{\ell-1} \cdots \widetilde{W}\widetilde{D}_{\ell-\ell_a+1}\right) z_{\ell-\ell_a,1} + z_{\ell-\ell_a,2^\sharp} \quad \text{and} \quad g'_{\ell,2} = \sum_{\ell_a=1}^{\ell-1} z_{\ell-\ell_a,2^\flat} \ .$$

We first bound $\|g'_{\ell,1}\|_2$,

$$
\begin{aligned}
\|g'_{\ell,1}\|_2 \le& \sum_{\ell_a=1}^{\ell-1} \|\widetilde{W}\widetilde{D}_{i,\ell-1}\cdots\widetilde{W}\widetilde{D}_{i,\ell-\ell_a+1}\|_2 \cdot \|z_{\ell-\ell_a,1}\|_2 + \|z_{\ell-\ell_a,2^\sharp}\|_2 \\
&\overset{①}{\le} \sum_{\ell_a=1}^{\ell} (CL^3\|z_{\ell-\ell_a,1}\|_2 + \|z_{\ell-\ell_a,2^\sharp}\|_2) \\
&\overset{②}{\le} CL^4\frac{\tau_0}{\sqrt{m}}\left(4L+8+\frac{\tau_1+\tau_2}{\sqrt{m}}\right) + L\frac{3\tau_5\sqrt{\tau_4}\log^2 m}{m^{2/3}} \\
&\le 5CL^4\cdot\frac{\tau_0(L+2)}{\sqrt{m}} \overset{③}{\le} \tau_1\frac{1}{\sqrt{m}} \enspace .
\end{aligned}
$$

Above, inequality ① follows from Lemma B.11, inequality ② follows from (C.4) and (C.5), and ③ follows from our choice of $\tau_1 = 5CL^4(L+2)\tau_0$. We next bound $\|g'_{\ell,2}\|_\infty$,

$$
\|g'_{\ell,2}\|_\infty \le \sum_{\ell_a=1}^{\ell-1} \|z_{\ell-\ell_a,2^\flat}\|_\infty \overset{①}{\le} L\cdot\frac{4\tau_5\log m}{m} \overset{②}{\le} \tau_2\frac{1}{m}
$$

where inequality ① is due to (C.5), and ② follows from our choice of $\tau_2 = 4L\tau_5\log m$. Thus, we have showed I and II of (C.1):

$$
\|g'_{\ell,1}\|_2 \le \tau_1\cdot\frac{1}{m^{1/2}} \quad \text{and} \quad \|g'_{\ell,2}\|_\infty \le \tau_2\cdot\frac{1}{m} \enspace .
$$

They together further imply

$$
\|g'_\ell\|_2 \le (\tau_1+\tau_2)\frac{1}{m^{1/2}} = \tau_3\frac{1}{m^{1/2}}
$$

so III of (C.1) holds. IV and V of (C.1) are implied by Claim C.4, and VI is implied because

$$
\|h'_\ell\|_2 = \|\phi(g_\ell) - \phi(\widetilde{g}_\ell)\|_2 \le \|g_\ell - \widetilde{g}_\ell\|_2 \le \|g'_\ell\|_2 \enspace . \qquad \square
$$

C.1.1 Tools

Claim C.3. *Suppose* $g_{\ell-1} = \widetilde{g}_{\ell-1} + g'_{\ell-1} = \widetilde{g}_{\ell-1} + g'_{\ell-1,1} + g'_{\ell-1,2}$ *where*

$$
\|g'_{\ell-1,1}\|_2 \le \frac{\tau_1}{\sqrt{m}} \quad \textit{and} \quad \|g'_{\ell-1,2}\|_\infty \le \frac{\tau_2}{m} \enspace .
$$

Then, we have with probability at least $1 - e^{-\Omega(m/L^2)}$

$$
\|W'D_{\ell-1}g_{\ell-1}\|_2 \le \frac{\tau_0}{\sqrt{m}}\left(4\ell+8+\frac{\tau_1+\tau_2}{\sqrt{m}}\right).
$$

Proof of Claim C.3. Using triangle inequality, we can calculate

$$
\|W'D_{\ell-1}g_{\ell-1}\|_2 \le \|W'\|_2\cdot\|D_{\ell-1}\|_2\cdot(\|\widetilde{g}_{\ell-1}\|_2 + \|g'_{\ell-1,1}\|_2 + \|g'_{\ell-1,2}\|_2)
$$

Using Lemma B.3, we have $\|\widetilde{g}_{\ell-1}\| \le 4\ell+8$. Using elementary calculation, we have

$$
\|g'_{\ell-1,1}\|_2 + \|g'_{\ell-1,2}\|_2 \le \frac{\tau_1}{\sqrt{m}} + \sqrt{m}\frac{\tau_2}{m} \le \frac{\tau_1+\tau_2}{\sqrt{m}}
$$

Together we finish the proof. $\square$

Claim C.4. *With probability at least* $1 - e^{-\Omega(\tau_1^{4/3} m^{1/3})}$ *the following holds. Whenever*

$$\|g'_{\ell-1,1}\|_2 \le \tau_1 \cdot \frac{1}{m^{1/2}}, \quad \|g'_{\ell-1,2}\|_\infty \le \tau_2 \cdot \frac{1}{m} \enspace ,$$

then letting $\tau_4 = 10(\tau_1)^{2/3}$ *and* $\tau_5 = 3\tau_1$*, we have*

$$\|D'_{\ell-1} g_{\ell-1}\|_2 \le \tau_5 \cdot \frac{1}{m^{1/2}}, \quad \|D'_{\ell-1}\|_0 \le \tau_4 \cdot m^{2/3}$$

Proof of Claim C.4. We choose parameters $\xi = \frac{(\tau_1)^{2/3}}{10^{1/3} m^{5/6}}$ and $\alpha = 10\xi\sqrt{m}$ in the proof. We have $\|g'_{\ell-1,2}\|_\infty \le \xi$.

First of all, using similar (but simpler) proof as Lemma B.6, one can show with probability at least $1 - \exp(-\Omega(\alpha^2 m))$, the vector $\widetilde{g}_{\ell-1}$ is $(\alpha, \frac{1}{5\sqrt{m}})$-good (recall Definition A.14).[14] This implies that $\widetilde{g}_{\ell-1}$ has at most $\alpha m = 10\xi m^{3/2}$ coordinates j satisfying $|(\widetilde{g}_{\ell-1})_j| \le \frac{\alpha}{5\sqrt{m}} = 2\xi$.

For each $j \in [m]$, if it satisfies $(D'_{\ell-1})_{j,j} \ne 0$, then the sign of $\widetilde{g}_{\ell-1}$ and $\widetilde{g}_{\ell-1} + g'_{\ell-1,1} + g'_{\ell-1,2}$ must differ on coordinate j. As a result:

$$|(g'_{\ell-1,1} + g'_{\ell-1,2})_j| > |(\widetilde{g}_{\ell-1})_j| \enspace .$$

Define $y = D'_{\ell-1} g_{\ell-1}$. There are two possibilities for such j with $(D'_{\ell-1})_{j,j} \ne 0$.

- Case 1: $|(\widetilde{g}_{\ell-1})_j| \le 2\xi$. Let such coordinates be $S_1 \subset [m]$, and we have $|S_1| \le 10\xi m^{3/2}$ using the above argument.

  Next, for each such $j \in S_1$, we must have $|y_j| = |(\widetilde{g}_{\ell-1} + g'_{\ell-1,1} + g'_{\ell-1,2})_j| \le |(g'_{\ell-1,1} + g'_{\ell-1,2})_j| \le |(g'_{\ell-1,1})_j| + \xi$ so we can calculate the ℓ_2 norm of y on S_1:

  $$\sum_{i \in S_1} y_j^2 \le 2\|g'_{\ell-1,1}\|^2 + 2\xi^2 |S_1| \le \frac{2\tau_1^2}{m} + 20\xi^3 m^{3/2} = \frac{4\tau_1^2}{m} \enspace .$$

- Case 2: $|(\widetilde{g}_{\ell-1})_j| > 2\xi$. Let such coordinates be $S_2 \subset [m] \setminus S_1$. In this case we must have $|(g'_{\ell-1,1})_j| \ge |(\widetilde{g}_{\ell-1})_j| - |(g'_{\ell-1,2})_j| > 2\xi - \xi = \xi$. Therefore, $|S_2| \le \frac{\|g'_{\ell-1,1}\|_2^2}{\xi^2} \le \frac{\tau_1^2}{m\xi^2}$.

  Next, for each $j \in S_2$, we must have

  $$|y_j| = |(\widetilde{g}_{\ell-1} + g'_{\ell-1,1} + g'_{\ell-1,2})_j| \le |(g'_{\ell-1,1} + g'_{\ell-1,2})_j| \le |(g'_{\ell-1,1})_j| + \xi/2 \le \frac{5}{4} |(g'_{\ell-1,1})_j| \enspace .$$

  and therefore

  $$\sum_{j \in S_2} y_j^2 \le 2 \sum_{j \in S_2} (g'_{\ell-1,1})_j^2 \le \frac{2\tau_1^2}{m} \enspace .$$

[14] Indeed,

$$\widetilde{g}_{\ell-1} = \widetilde{W}\widetilde{h}_{\ell-2} + A x_{\ell-1} = M_1 z_1 + M_2 z_2 .$$

where $M_1 \in \mathbb{R}^{m \times n(\ell-2)}$, $M_3 \in \mathbb{R}^{m \times d}$ and $z_1 \in \mathbb{R}^{n(\ell-1)}$, $z_3 \in \mathbb{R}^d$ are defined as

$$\begin{aligned} M_1 &= \widetilde{W} U_{\ell-2} & \qquad M_2 &= A \\ z_1 &= U_{\ell-2}^\top \widetilde{h}_{\ell-2} & \qquad z_2 &= x_{\ell-1} \enspace . \end{aligned} \tag{C.6}$$

The entries of M_1 and M_2 are i.i.d. from $\mathcal{N}(0, \frac{2}{m})$ (recall Footnote 9). Suppose z_1 and z_2 are fixed (instead of depending on the randomness of $\widetilde{W}$ and A) and satisfies $\frac{1}{\sqrt{2}} \le \|z_1\| \le 2L + 6$. We can apply Corollary A.16 to obtain the following statement: with probability at least $1 - \exp(-\Omega(\alpha^2 m))$, $\widetilde{g}_{\ell-1}$ is $(\alpha, \sigma/4)$-good where $\sigma^2 = \frac{2}{m}\|z_1\|_2^2 + \frac{2}{m}\|z_3\|_2^2 \ge \frac{1}{m}$. In other words, w is $(\alpha, \frac{1}{4\sqrt{m}})$-good. Applying standard ε-net argument, we have with probability at least $1 - \exp(-\Omega(\alpha^2 m))$, for all vectors $z_1 \in \mathbb{R}^{n(\ell-2)}$ satisfying $\frac{1}{\sqrt{2}} \le \|z_1\| \le 2L + 6$, it satisfies $\widetilde{g}_{\ell-1}$ is $(\alpha, \frac{1}{5\sqrt{m}})$-good. This allows us to plug in the random choice of z_1 in (C.6), to conclude that $\widetilde{g}_{\ell-1}$ is $(\alpha, \frac{1}{5\sqrt{m}})$-good.

In sum, we conclude that

$$\|D'_{\ell-1}\|_0 \le |S_1| + |S_2| \le 10\xi m^{3/2} + \frac{\tau_1^2}{m\xi^2} < 10(\tau_1)^{2/3} m^{2/3} = \tau_4 \cdot m^{2/3}$$

and

$$\|y\|_2 \le \frac{\sqrt{4\tau_1^2 + 2\tau_1^2}}{\sqrt{m}} \le \frac{3\tau_1}{\sqrt{m}} = \tau_5 \cdot \frac{1}{m^{1/2}} \ . \qquad \square$$

Claim C.5. *With probability at least $1 - e^{-\Omega(\tau_4 m^{2/3} \log^2 m)}$, for all $L \ge \ell + 1 \ge b \ge 1$, for all*

$$x \in \mathbb{R}^m \quad \text{with} \quad \|x\|_2 \le \frac{\tau_5}{\sqrt{m}} \quad \text{and} \quad \|x\|_0 \le \tau_4 \cdot m^{2/3}$$

we have that the vector $y = \widetilde{W} \widetilde{D}_\ell \widetilde{W} \cdots \widetilde{D}_b \widetilde{W} x$ can be written as

$$y = y_1 + y_2 \quad \text{where} \quad \|y_1\|_2 \le \frac{3\tau_5 \sqrt{\tau_4} \log^2 m}{m^{2/3}} \quad \text{and} \quad \|y_2\|_\infty \le \frac{4\tau_5 \log m}{m}$$

Proof of Claim C.5. Let $s = \tau_4 \cdot m^{2/3}$ for notational simplicity. Let $z_{b-1} = x$ and assume for now that x is a fixed vector. Letting $z_\ell = \widetilde{D}_\ell \widetilde{W} \cdots \widetilde{D}_b z_{b-1}$, we have according to Claim B.12, with probability at least $1 - \exp(-\Omega(m/L^2))$,

$$\|z_\ell\|_2 \le (1 + 1/L)^{\ell-b-1} \|z_{b-1}\|_2 \le 3\|z_{b-1}\|_2.$$

Similar to the proof of Lemma B.11, we let U denote the following column orthonormal matrix using Gram-Schmidt:

$$U \stackrel{\text{def}}{=} \mathsf{GS}\left(h_{1,1}, \dots, h_{n,1}, h_{1,2}, \dots, h_{n,2}, \dots, h_{1,\ell-1}, \dots, h_{n,\ell-1}, z_{b-1}, \cdots z_\ell\right) \ ,$$

and we have

$$\widetilde{W} \widetilde{D}_\ell \widetilde{W} \cdots \widetilde{D}_b \widetilde{W} x = \widetilde{W} U U^\top z_{\ell_2} \ .$$

Observe the entries of $\widetilde{W} U \in \mathbb{R}^{m \times n(\ell-1)+(\ell-b+2)}$ are i.i.d. drawn from $\mathcal{N}(0, \frac{2}{m})$ (following a similar argument as Footnote 13). Therefore, according to Lemma A.17, we have can write $y = \widetilde{W} \widetilde{D}_\ell \widetilde{W} \cdots \widetilde{D}_b \widetilde{W} x$ as $y = y_1 + y_2$ with

$$\|y_1\| \le \frac{\sqrt{s} \log^2 m}{\sqrt{m}} \cdot \|U^\top z_{\ell_2}\| \quad \text{and} \quad \|y_2\|_\infty \le \frac{\log m}{\sqrt{m}} \cdot \|U^\top z_{\ell_2}\| \ .$$

Plugging in $\|U^\top z_\ell\|_2 \le \|z_\ell\|_2 \le 3\|z_{b-1}\|_2 = 3\|x\|_2 \le \frac{3\tau_5}{\sqrt{m}}$, we have

$$\|y_1\| \le \frac{3\tau_5 \sqrt{s} \log^2 m}{m} \quad \text{and} \quad \|y_2\|_\infty \le \frac{3\tau_5 \log m}{m} \ .$$

Finally, taking ε-net over all s-sparse vectors x, we have the desired result. $\square$

Combining Claim C.4 and Claim C.5, we have

Claim C.6. *With probability at least $1 - e^{-\Omega(\tau_1^{4/3} m^{1/3})}$, whenever*

$$\|g'_{\ell-1,1}\|_2 \le \frac{\tau_1}{\sqrt{m}}, \qquad \|g'_{\ell-1,2}\|_\infty \le \frac{\tau_2}{m} \ ,$$

we have the vector $y = \widetilde{W} \widetilde{D}_\ell \widetilde{W} \cdots \widetilde{D}_b \widetilde{W} D'_{\ell-1} g_{\ell-1}$ can be written as

$$y = y_1 + y_2 \quad \text{where} \quad \|y_1\|_2 \le \frac{3\tau_5 \sqrt{\tau_4} \log^2 m}{m^{2/3}} \quad \text{and} \quad \|y_2\|_\infty \le \frac{4\tau_5 \log m}{m}$$

C.2 Intermediate Layers

The goal of this subsection is to prove Lemma C.7.

Lemma C.7 (intermediate stability, c.f. (6.2))**.** *Letting $\varrho = nLd\log m$, for any $\tau_0 \in \left[\varrho^{-100}, \varrho^{100}\right]$, with probability at least $1-e^{-\Omega(L^6\tau_0^{4/3}m^{1/3})}$ over the randomness of $\widetilde{W}, A, B$, for every $W' \in \mathbb{R}^{m\times m}$ with $\|W'\|_2 \leq \frac{\tau_0}{\sqrt{m}}$, for every $i \in [n]$, for every $L \geq \ell_2 \geq \ell_1 \geq 1$, the following holds*

- $\left\|\prod_{\ell=\ell_2}^{\ell_1}(\widetilde{D}_{i,\ell} + D'_{i,\ell})\widetilde{W}\right\|_2 \leq O(L^7)$.
- $\left\|\prod_{\ell=\ell_2}^{\ell_1}(\widetilde{D}_{i,\ell} + D'_{i,\ell})(\widetilde{W} + W')\right\|_2 \leq O(L^7)$.

We show Claim C.8 and then use it to prove Corollary C.8.

Claim C.8. *Let $s \in [m^{-1/4}, m^{1/8}]$ be any fixed real. With probability at least $1 - e^{-\Omega(s^2m^{2/3}\log^2 m)}$, it satisfies that for every $i \in [n]$, for every $L \geq \ell_2 \geq \ell_1 \geq 1$, if $D'_{i,\ell} \in \mathbb{R}^{m\times m}$ is an arbitrary diagonal matrix with entries in $\{-1, 0, 1\}$ of sparsity $\|D'_{i,\ell}\|_0 \leq s^2m^{2/3}$ (for each $\ell = \ell_1, \ldots, \ell_2$), then*

$$\left\|\prod_{\ell=\ell_2}^{\ell_1}(\widetilde{D}_{i,\ell} + D'_{i,\ell})\widetilde{W}\right\|_2 \leq O(L^7) \ .$$

Proof of Claim C.8. We define set $\mathcal{C}$ as follows

$$\mathcal{C} = \left\{ \prod_{\ell=\ell_2}^{\ell_1}(D'_{i,\ell})^{c_\ell}(\widetilde{D}_{i,\ell})^{1-c_\ell}\widetilde{W} \;\middle|\; c_\ell \in \{0,1\}, \forall \ell \in [\ell_1, \ell_2] \right\}.$$

For ℓ, we define set $\mathcal{C}_\ell$ as follows

$$\mathcal{C}_\ell = \left\{ C \in \mathcal{C} \;\middle|\; D' \text{ appears } \ell \text{ times in } C \right\}$$

From Lemma B.11, we can bound the $\ell = 0$ term

$$\left\|\sum_{C\in\mathcal{C}_\ell} C\right\|_2 \leq O(L^3).$$

For each $C \in \mathcal{C}_\ell$ with $\ell > 0$, we have

$$\|C\|_2 \leq O(L^3)\cdot\left(\frac{s\log m}{m^{1/6}}\right)^{\ell-1}\cdot O(L^3)$$

where we apply Lemma B.11 twice(one on the left, one on the right) and apply Corollary B.15 $\ell - 1$ times. The probability of the above event is at least $1 - e^{\Omega(s^2\cdot m^{2/3}\log^2 m)}$. Therefore, for each $\ell \in [\ell_2 - \ell_1]$, we can bound

$$\left\|\sum_{C\in\mathcal{C}_\ell} C\right\|_2 \leq O(L^6)\binom{\ell_2-\ell_1}{\ell}\left(\frac{s\log m}{m^{1/6}}\right)^{\ell-1}$$

which is at most $O(L^7)$ when $\ell = 1$ and $o(L^7)$ for $\ell > 1$ by our parameter choices. Putting it altogether,

$$\text{LHS} = \left\|\sum_{\ell=0}^{\ell_2-\ell_1}\sum_{C\in\mathcal{C}_\ell} C\right\|_2 \leq \sum_{\ell=0}^{\ell_2-\ell_1}\left\|\sum_{C\in\mathcal{C}_\ell} C\right\|_2 \leq O(L^3+L^7) \qquad \square$$

Proof of Lemma C.7. Applying Lemma C.2c, we know with probability $\geq 1 - e^{-\Omega(L^6 \tau_0^{4/3} m^{1/3})}$ it satisfies $\|D'_{i,\ell}\|_0 \leq s^2 m^{2/3}$ for $s^2 = O(L^{10/3}\tau_0^{2/3})$. Therefore, applying Claim C.8 we have

$$\left\| \prod_{\ell=\ell_2}^{\ell_1} (\widetilde{D}_{i,\ell} + D'_{i,\ell})\widetilde{W} \right\|_2 \leq O(L^7) \ .$$

On the other hand, since $\|W'\| \leq \frac{\tau_0}{\sqrt{m}}$, we also have

$$\left\| \prod_{\ell=\ell_2}^{\ell_1} (\widetilde{D}_{i,\ell} + D'_{i,\ell})(\widetilde{W} + W') \right\|_2 \leq O(L^7) \ .$$

by expanding out the $2^{\ell_2 - \ell_1 + 1}$ terms with respect to W'. $\square$

C.3 Backward

We state the main result in this section.

Lemma C.9 (backward stability, c.f. (6.3))**.** *Letting $\varrho = nLd \log m$, for any $\tau_0 \in [\varrho^{-100}, \varrho^{100}]$, with probability at least $1 - e^{-\Omega(L^6 \tau_0^{4/3} m^{1/3})}$ over the randomness of $\widetilde{W}, A, B$, for every $W' \in \mathbb{R}^{m \times m}$ with $\|W'\|_2 \leq \frac{\tau_0}{\sqrt{m}}$, for every $1 \leq \ell_1 \leq \ell_2 \leq L$, for every $i \in [n]$, the followings hold*

(a) $\left\| a^\top B \prod_{\ell=\ell_2}^{\ell_1} \widetilde{D}_{i,\ell}\widetilde{W} - a^\top B \prod_{\ell=\ell_2}^{\ell_1} D_{i,\ell}\widetilde{W} \right\|_2 \leq O\big(\tau_0^{1/3} L^6 \log m \cdot m^{1/3}\big) \cdot \|a\|_2.$

(b) $\left\| a^\top B \prod_{\ell=\ell_2}^{\ell_1} D_{i,\ell}\widetilde{W} - a^\top B \prod_{\ell=\ell_2}^{\ell_1} D_{i,\ell} W \right\|_2 \leq O(\tau_0 L^{15}) \cdot \|a\|_2.$

(c) $\left\| a^\top B \prod_{\ell=\ell_2}^{\ell_1} \widetilde{D}_{i,\ell}\widetilde{W} - a^\top B \prod_{\ell=\ell_2}^{\ell_1} \widetilde{D}_{i,\ell} W \right\|_2 \leq O(\tau_0 L^7) \cdot \|a\|_2.$

(d) $\left\| a^\top B \prod_{\ell=\ell_2}^{\ell_1} \widetilde{D}_{i,\ell} W - a^\top B \prod_{\ell=\ell_2}^{\ell_1} D_{i,\ell} W \right\|_2 \leq O\big(\tau_0^{1/3} L^6 \log m \cdot m^{1/3}\big) \cdot \|a\|_2.$

We first use Lemma C.2 to derive that, with probability $\geq 1 - e^{-\Omega(L^6 \tau_0^{4/3} m^{1/3})}$, $\|D'_{i,\ell}\|_0 \leq s^2 m^{2/3}$, for parameter $s = O(L^{5/3}\tau_0^{1/3})$. We prove the five statements separately.

Proof of Lemma C.9a. Without loss of generality we assume $\|a\| = 1$. We define set $\mathcal{C}$ to be

$$\mathcal{C} = \left\{ \prod_{\ell=\ell_2}^{\ell_1} (D'_{i,\ell})^{c_\ell} (\widetilde{D}_{i,\ell})^{1-c_\ell} \widetilde{W} \;\middle|\; c_{\ell_1}, \ldots, c_{\ell_2} \in \{0,1\} \right\}.$$

For each $\ell \in [\ell_2 - \ell_1]$, we define set $\mathcal{C}_\ell$ to be

$$\mathcal{C}_\ell = \{ C \in \mathcal{C} \mid D' \text{ appears } \ell \text{ times in } C \}.$$

Ignoring subscripts for the ease of presentation, by Corollary B.18, we know

$$\|a^\top B \widetilde{D}\widetilde{W} \cdots \widetilde{D}\widetilde{W} D'\|_2 \leq (s \log m) m^{1/3} \ .$$

By Corollary B.15, we know

$$\left\| D'\widetilde{W}\widetilde{D} \cdots \widetilde{D}\widetilde{W}D' \right\|_2 \leq (s \log m) m^{-1/6} \ .$$

Thus, for each ℓ and $C \in \mathcal{C}_\ell$,

$$\begin{aligned} \|a^\top B C\|_2 &\leq \|a^\top B \widetilde{D}\widetilde{W} \cdots \widetilde{D}\widetilde{W} D'\|_2 \cdot \left\| D'\widetilde{W}\widetilde{D} \cdots \widetilde{D}\widetilde{W}D' \right\|_2^{\ell-1} \cdot \left\| D'\widetilde{W}\widetilde{D} \cdots \widetilde{D}\widetilde{W} \right\|_2 \\ &\leq (s \cdot m^{1/3} \log m) \times (s \cdot m^{-1/6} \log m)^{\ell-1} \times O(L^3) \ . \end{aligned}$$

Finally, the LHS of the Lemma C.9a becomes

$$\begin{aligned}
\text{LHS} &= \left\| \sum_{\ell=1}^{\ell_2-\ell_1} \sum_{C\in\mathcal{C}_\ell} a^\top BC \right\|_2 \leq \sum_{\ell=1}^{\ell_2-\ell_1} \sum_{C\in\mathcal{C}_\ell} \|a^\top BC\|_2 \\
&\leq \sum_{\ell=1}^{\ell_2-\ell_1} \binom{\ell_2-\ell_1}{\ell} (s\cdot m^{1/3}\log m) \times (s\cdot m^{-1/6}\log m)^{\ell-1} \times O(L^3) \leq O\big(L^4 \cdot s\log m \cdot m^{1/3}\big) \ .
\end{aligned}$$

□

Proof of Lemma C.9b. Again we assume $\|a\| = 1$ without loss of generality. We define set $\mathcal{C}$ to be

$$\mathcal{C} = \left\{ \prod_{\ell=\ell_2}^{\ell_1} D_{i,\ell}(W')^{c_\ell}\widetilde{W}^{1-c_\ell} \;\middle|\; c_{\ell_1},\dots,c_{\ell_2} \in \{0,1\} \right\}$$

For each $\ell \in [\ell_2-\ell_1]$, we define set $\mathcal{C}_\ell$

$$\mathcal{C}_\ell = \{C \in \mathcal{C} \mid W' \text{ appears } \ell \text{ times in } C\} \ .$$

Ignoring subscripts for the ease of presentation, we have for each $C \in \mathcal{C}_\ell$,

$$\|a^\top BC\|_2 \leq \|a^\top B\|_2 \cdot \|\widetilde{W}D\cdots\widetilde{W}D\|_2 \cdot \|W'\|_2 \cdot \left(\|D\widetilde{W}\cdots\widetilde{W}D\|_2\|W'\|_2\right)^{\ell-1} \cdot \|D\widetilde{W}\cdots D\widetilde{W}\|_2 \tag{C.7}$$

$$\leq O(\sqrt{m}L^7)\cdot\left(\frac{\tau_0}{\sqrt{m}}\right)\cdot\left(\frac{O(\tau_0L^7)}{\sqrt{m}}\right)^{\ell-1}\cdot O(L^7) \ . \tag{C.8}$$

Above, we have used $\|W'\|_2 \leq \frac{\tau_0}{\sqrt{m}}$, $\|\widetilde{W}D\cdots\widetilde{W}D\|_2 \leq O(L^7)$ from Lemma C.7, and $\|B\|_2 \leq O(\sqrt{m})$ with high probability. Finally, by triangle inequality, the LHS of Lemma C.9b becomes

$$\begin{aligned}
\text{LHS} &= \left\| \sum_{\ell=1}^{\ell_2-\ell_1} \sum_{C\in\mathcal{C}_\ell} a^\top BC \right\|_2 \leq \sum_{\ell=1}^{\ell_2-\ell_1} \sum_{C\in\mathcal{C}_\ell} \left\|a^\top BC\right\|_2 \\
&\leq \sum_{\ell=1}^{\ell_2-\ell_1} \binom{\ell_2-\ell_1}{\ell} O(\sqrt{m}L^7)\cdot\left(\frac{\tau_0}{\sqrt{m}}\right)\cdot\left(\frac{O(\tau_0L^7)}{\sqrt{m}}\right)^{\ell-1}\cdot O(L^7) \\
&\leq O(\tau_0 L^{15})
\end{aligned}$$

□

Proof of Lemma C.9c. It is similar to the proof of Lemma C.9b. □

Proof of Lemma C.9d. It is similar to the proof of Lemma C.9a. □

C.4 Special Rank-One Perturbation

We prove two lemmas regarding the special rank-one perturbation with respect to a coordinate $k \in [m]$. The first one talks about forward propagation.

Lemma C.10 (c.f. (6.4)). *Let $\rho = nLd\log m$, $\varrho = nLd\delta^{-1}\log(m/\varepsilon)$, $\tau_0 \in [\varrho^{-100}, \varrho^{50}]$, $N \in [1, \varrho^{100}]$. If $\widetilde{W}, A$ are at random initialization, then with probability at least $1 - e^{-\Omega(\rho^2)}$, for any rank-one perturbation $W' = yz^\top$ with $\|z\| = 1$, $\|y\|_0 = N$, $\|y\|_\infty \le \frac{\tau_0}{\sqrt{m}}$, it satisfies*

$$\forall i \in [n], \forall \ell \in [L], \forall k \in [m]: \quad |((\widetilde{W} + W')h'_{i,\ell})_k| \le O\Big(\frac{\rho^8 N^{2/3}\tau_0^{5/6}}{m^{2/3}}\Big) \ .$$

Proof. Without loss of generality we only prove the statement for fixed i, ℓ, k because one can take union bound at the end. We can rewrite $Wh'_{i,\ell}$ as follows

$$\begin{aligned}
(\widetilde{W} + W')h'_{i,\ell} &= W'h'_{i,\ell} + \widetilde{W}h'_{i,\ell} \\
&= W'h'_{i,\ell} + \widetilde{W}(h_{i,\ell} - \widetilde{h}_{i,\ell}) \\
&= W'h'_{i,\ell} + \widetilde{W}((\widetilde{D}_{i,\ell} + D'_{i,\ell})g_{i,\ell} - \widetilde{D}_{i,\ell}\widetilde{g}_{i,\ell}) \\
&= W'h'_{i,\ell} + \widetilde{W}\widetilde{D}_{i,\ell}g'_{i,\ell} + \widetilde{W}D'_{i,\ell}g_{i,\ell} \\
&= W'h'_{i,\ell} + \widetilde{W}\widetilde{D}_{i,\ell}((\widetilde{W} + W')h_{i,\ell-1} - \widetilde{W}\widetilde{h}_{i,\ell-1}) + \widetilde{W}D'_{i,\ell}g_{i,\ell} \\
&= W'h'_{i,\ell} + \widetilde{W}\widetilde{D}_{i,\ell}(\widetilde{W}h'_{i,\ell-1} + W'h_{i,\ell-1}) + \widetilde{W}D'_{i,\ell}g_{i,\ell} \\
&= W'h'_{i,\ell} + \widetilde{W}\widetilde{D}_{i,\ell}\underbrace{\widetilde{W}h'_{i,\ell-1}}_{\text{recurse}} + \widetilde{W}\widetilde{D}_{i,\ell}W'h_{i,\ell-1} + \widetilde{W}D'_{i,\ell}g_{i,\ell}
\end{aligned}$$

After recursively calculating as above, using $h'_{i,0} = 0$, and applying triangle inequality, we have

$$|((\widetilde{W} + W')h'_{i,\ell})_k| \le |(W'h'_{i,\ell})_k| + \sum_{a=0}^{\ell} |(\widetilde{W}\widetilde{D}_{i,\ell}\cdots\widetilde{W}\widetilde{D}_{i,a}W'h_{i,a})_k| + \sum_{a=0}^{\ell} |(\widetilde{W}\widetilde{D}_{i,\ell}\cdots\widetilde{W}\widetilde{D}_{i,a+1}\widetilde{W}D'_{i,a}g_{i,a})_k| \tag{C.9}$$

We bound the three types of terms on the RHS of (C.9) as follows.

- First, we can bound

$$|(W'h'_{i,\ell})_k| = |\mathbb{1}_k^\top yz^\top h'_{i,\ell}| = |\mathbb{1}_k^\top y| \cdot |z^\top h'_{i,\ell}| \le \|y\|_\infty \cdot \|h'_{i,\ell}\|_2 \le \frac{\tau_0}{\sqrt{m}} \cdot (\frac{O(L^6\tau_0\sqrt{N})}{\sqrt{m}})$$

  where the last step follows by $\|y\|_\infty \le \frac{\tau_0}{\sqrt{m}}$, $\|W'\|_2 \le \frac{\sqrt{N}\tau_0}{\sqrt{m}}$ and Lemma C.2a.

- Second, we can bound

$$\begin{aligned}
|(\widetilde{W}\widetilde{D}_{i,\ell}\cdots\widetilde{W}\widetilde{D}_{i,a}W'h_{i,a})_k| &= |\mathbb{1}_k^\top\widetilde{W}\widetilde{D}_{i,\ell}\cdots\widetilde{W}\widetilde{D}_{i,a}yz^\top h_{i,a}| \\
&= |\mathbb{1}_k^\top\widetilde{W}\widetilde{D}_{i,\ell}\cdots\widetilde{W}\widetilde{D}_{i,a}y| \cdot |z^\top h_{i,a}| \\
&\overset{①}{\le} |\mathbb{1}_k^\top\widetilde{W}\widetilde{D}_{i,\ell}\cdots\widetilde{W}\widetilde{D}_{i,a}y| \cdot \|h_{i,a}\|_2 \\
&\overset{②}{\le} |\mathbb{1}_k^\top\widetilde{W}\widetilde{D}_{i,\ell}\cdots\widetilde{W}\widetilde{D}_{i,a}y| \cdot O(L) \\
&\le N \cdot \|y\|_\infty \max_j |\mathbb{1}_k^\top\widetilde{W}\widetilde{D}_{i,\ell}\cdots\widetilde{W}\widetilde{D}_{i,a}\mathbb{1}_j| \cdot O(L) \\
&\overset{③}{\le} N \cdot \|y\|_\infty \cdot \frac{\rho}{\sqrt{m}} \cdot O(L) \\
&\overset{④}{\le} N \cdot (\tau_0\frac{1}{\sqrt{m}}) \cdot \frac{\rho}{\sqrt{m}} \cdot O(L) \le O(\frac{N\tau_0\rho L}{m}).
\end{aligned}$$

where ① follows by $\|z\|_2 \le 1$, ② follows by Lemma B.3, ③ follows by Corollary B.16, and ④ follows by $\|y\|_\infty \le \tau_0 \frac{1}{\sqrt{m}}$.

- Third, we can bound,

$$
\begin{aligned}
|(\widetilde{W}\widetilde{D}_{i,\ell}\cdots\widetilde{W}\widetilde{D}_{i,a+1}\widetilde{W}D'_{i,a}g_{i,a})_k| &\le \|\mathbb{1}_k\widetilde{W}\widetilde{D}_{i,\ell}\cdots\widetilde{W}\widetilde{D}_{i,a+1}\widetilde{W}D'_{i,a}\|_2\cdot\|D'_{i,a}g_{i,a}\|_2 \\
&\le O\left(\frac{L^{5/3}\tau_0^{1/3}N^{1/6}\log m}{m^{1/6}}\right)\cdot O\left(\frac{L^5\tau_0^{1/2}\sqrt{N}}{\sqrt{m}}\right)
\end{aligned}
$$

where the last step follows by Corollary B.15 (which relies on Lemma C.2b to give $s = O(L^{5/3}\tau_0^{1/3}N^{1/6})$) together with Lemma C.2c.

Putting the three bounds into (C.9) (where the term one is the dominating term), we have

$$
|((\widetilde{W}+W')h'_{i,\ell})_k| \le O\left(\frac{\varrho^8 N^{2/3}\tau_0^{5/6}}{m^{2/3}}\right) \ . \qquad \square
$$

The next one talks about backward propagation.

Lemma C.11 (c.f. (6.5))**.** *Let $\rho = nLd\log m$, $\varrho = nLd\delta^{-1}\log(m/\varepsilon)$, $\tau_0 \in [\varrho^{-100}, \varrho^{50}]$, $N \in [1, \varrho^{100}]$. If $\widetilde{W}, A, B$ are at random initialization, and given some perturbation $W' = yz^\top$ with $\|z\| = 1$, $\|y\|_0 = N$, $\|y\|_\infty \le \frac{\tau_0}{\sqrt{m}}$ where W' can only depend on the randomness of $\widetilde{W}$ and A (but not on B), then with probability at least $1 - e^{-\Omega(\rho^2)}$, for all $i \in [n]$, all $1 \le \ell_1 \le \ell_2 \le L$, all $k \in [m]$:*

(a) $\left|a^\top B\left(\prod_{\ell=\ell_2}^{\ell_1} D_{i,\ell}\widetilde{W}\right)\mathbb{1}_k - a^\top B\left(\prod_{\ell=\ell_2}^{\ell_1}\widetilde{D}_{i,\ell}\widetilde{W}\right)\mathbb{1}_k\right| \le \|a\|_2\cdot O\left(\frac{\rho^3\tau_0^{1/3}N^{1/6}}{m^{1/6}}\right)$,

(b) $\left|a^\top B\left(\prod_{\ell=\ell_2}^{\ell_1} D_{i,\ell}W\right)\mathbb{1}_k - a^\top B\left(\prod_{\ell=\ell_2}^{\ell_1} D_{i,\ell}\widetilde{W}\right)\mathbb{1}_k\right| \le \|a\|_2\cdot O\left(\frac{\rho^{12}N^{5/6}\tau_0^{5/3}}{m^{1/3}}\right)$.

Proof. We prove for a fixed $i \in [n]$ and drop the subscript in i for notational simplicity. Without loss of generality we assume $\|a\| = 1$.

(a) We define set $\mathcal{C}$ to be

$$
\mathcal{C} = \left\{\prod_{\ell=\ell_2}^{\ell_1}(D'_\ell)^{c_\ell}(\widetilde{D}_\ell)^{1-c_\ell}\widetilde{W} \;\middle|\; c_\ell \in \{0,1\}, \forall \ell \in [\ell_1,\ell_2]\right\}.
$$

For each $\ell \in [\ell_2 - \ell_1]$, we define set $\mathcal{C}_\ell$ to be

$$
\mathcal{C}_\ell = \{C \in \mathcal{C} \mid D' \text{ appears } \ell \text{ times in } C\}.
$$

Ignoring the subscripts for the ease of presentation, by Lemma B.11, we know

$$
\|\widetilde{D}\widetilde{W}\cdots\widetilde{D}\widetilde{W}\cdot D'\|_2 \le \|\widetilde{D}\widetilde{W}\cdots\widetilde{D}\widetilde{W}\|_2 \le O(L^3) \ .
$$

By Corollary B.15 (which relies on Lemma C.2b to give $s = O(L^{5/3}\tau_0^{1/3}N^{1/6})$), we know

$$
\begin{aligned}
\left\|D'\widetilde{W}\widetilde{D}\cdots\widetilde{D}\widetilde{W}D'\right\|_2 &\le O\left(\frac{L^{5/3}\tau_0^{1/3}N^{1/6}}{m^{1/6}}\log m\right) \ , \\
\forall k \in [m]\colon\ \left\|D'\widetilde{W}\widetilde{D}\cdots\widetilde{D}\widetilde{W}\mathbb{1}_k\right\|_2 &\le O\left(\frac{L^{5/3}\tau_0^{1/3}N^{1/6}}{m^{1/6}}\log m\right) \ .
\end{aligned}
$$

Thus, for each $k \in [m]$ and $C \in \mathcal{C}_\ell$, we can bound it

$$\|C\mathbb{1}_k\| \le L^3 \times \left(O(\frac{L^{5/3}\tau_0^{1/3}N^{1/6}}{m^{1/6}}\log m)\right)^\ell \quad .$$

As a consequence, letting

$$v = \left(\prod_{\ell=\ell_2}^{\ell_1} D_{i,\ell}\widetilde{W}\right)\mathbb{1}_k - \left(\prod_{\ell=\ell_2}^{\ell_1} \widetilde{D}_{i,\ell}\widetilde{W}\right)\mathbb{1}_k$$

we have

$$\begin{aligned}
\|v\| &= \left\|\sum_{\ell=1}^{\ell_2-\ell_1}\sum_{C\in\mathcal{C}_\ell} C\mathbb{1}_k\right\| \\
&\le \sum_{\ell=1}^{\ell_2-\ell_1}\sum_{C\in\mathcal{C}_\ell}\|C\mathbb{1}_k\| \le \sum_{\ell=1}^{\ell_2-\ell_1}\binom{L}{\ell}\cdot\left(O(\frac{L^{5/3}\tau_0^{1/3}N^{1/6}}{m^{1/6}}\log m)\right)^\ell \le O\left(\frac{\rho^3\tau_0^{1/3}N^{1/6}}{m^{1/6}}\right) \quad .
\end{aligned}$$

Finally, we use the randomness of B (recall that v does not depend on B), we conclude that with probability at least $1-e^{-\Omega(m)}$, it satisfies

$$\text{LHS} = \left\|a^\top Bv\right\| \le O\left(\frac{\rho^3\tau_0^{1/3}N^{1/6}}{m^{1/6}}\right) \quad .$$

(b) This time, we define set $\mathcal{C}$ to be

$$\mathcal{C} = \left\{\prod_{\ell=\ell_2}^{\ell_1} D_\ell (W')^{c_\ell}\widetilde{W}^{1-c_\ell} \;\middle|\; c_\ell \in \{0,1\}, \forall \ell \in [\ell_1,\ell_2]\right\}$$

For each $\ell \in [\ell_2-\ell_1]$, we define set $\mathcal{C}_\ell$

$$\mathcal{C}_\ell = \{C \in \mathcal{C} \mid W' \text{ appears } \ell \text{ times in } C\}.$$

One can carefully derive that[15]

$$|a^\top BD\widetilde{W}\cdots\widetilde{W}Dy| \leq O\left(\frac{\rho^4 N^{5/6}\tau_0^{5/3}}{m^{1/3}}\right) \ .$$

By Lemma C.7 we have

$$|z^\top D\widetilde{W}\cdots\widetilde{W}Dy| \leq \|z\|_2 \cdot O(L^7)\cdot\|y\|_2 \leq O\Big(\frac{L^7\sqrt{N}\tau_0}{\sqrt{m}}\Big) \ ,$$
$$|z^\top D\widetilde{W}\cdots\widetilde{W}D\mathbb{1}_k| \leq \|z\|_2 \cdot O(L^7)\cdot\|\mathbb{1}_k\|_2 \leq O(L^7) \ .$$

Therefore, for each $C\in\mathcal{C}_\ell$ we can upper bound $\|a^\top BC\|_2$ as follows

$$|a^\top BC\mathbb{1}_k| \leq O\left(\frac{\rho^4 N^{5/6}\tau_0^{5/3}}{m^{1/3}}\right)\cdot\left(O\Big(\frac{L^7\sqrt{N}\tau_0}{\sqrt{m}}\Big)\right)^{\ell-1}\cdot O(L^7)$$

and therefore

$$\begin{aligned}\text{LHS} &= \left|\sum_{\ell=1}^{\ell_2-\ell_1}\sum_{C\in\mathcal{C}_\ell} a^\top BC\mathbb{1}_k\right| \\ &\leq \sum_{\ell=1}^{\ell_2-\ell_1}\binom{\ell_2-\ell_1}{\ell}\cdot O\left(\frac{\rho^4 N^{5/6}\tau_0^{5/3}}{m^{1/3}}\right)\cdot\left(O\Big(\frac{L^7\sqrt{N}\tau_0}{\sqrt{m}}\Big)\right)^{\ell-1}\cdot O(L^7) \leq O\left(\frac{\rho^{12} N^{5/6}\tau_0^{5/3}}{m^{1/3}}\right) \ .\end{aligned}$$

□

[15] Since we have done this too many times, let us quickly point out the calculation. By Corollary B.19, we have with probability at least $1-e^{-\Omega(\rho^2)}$

$$|a^\top B\widetilde{D}\widetilde{W}\cdots\widetilde{W}\widetilde{D}y| \leq \|a\|_2\cdot\sqrt{|N|}\log m\cdot\|y\|_2 \leq \|y\|_\infty\|y\|_0\cdot\rho \leq \frac{\tau_0|N|}{\sqrt{m}} \ .$$

Using binomial expansion again, we have

$$\begin{aligned}|a^\top B\widetilde{D}\widetilde{W}\cdots\widetilde{W}\widetilde{D}y - a^\top BD\widetilde{W}\cdots\widetilde{W}Dy| &\leq \sum_{\ell=1}^{L}\binom{L}{\ell}\|a^\top B\widetilde{D}\widetilde{W}\cdots\widetilde{W}D'\|\cdot\|D'\widetilde{W}\cdots\widetilde{W}D'\|_2^{\ell-1}\cdot\|D'\widetilde{W}\cdots\widetilde{W}\widetilde{D}y\| \\ &\leq \sum_{\ell=1}^{L}\binom{L}{\ell}(sm^{1/3}\log m)\cdot\left(\frac{s\log m}{m^{1/6}}\right)^{\ell-1}\cdot\left(\frac{s\log m}{m^{1/6}}\right)\cdot\|y\| \\ &\leq O(L)(sm^{1/3}\log m)\left(\frac{s\log m}{m^{1/6}}\right)\cdot\frac{\sqrt{N}\tau_0}{\sqrt{m}} = O\left(\frac{\rho^4 N^{5/6}\tau_0^{5/3}}{m^{1/3}}\right)\end{aligned}$$

Here, we have used Corollary B.18 and Corollary B.15 with parameter $s = O(L^{5/3}\tau_0^{1/3}N^{1/6})$ chosen from Lemma C.2b, as well as $\|y\| \leq \frac{\sqrt{N}\tau_0}{\sqrt{m}}$.

D Indicator and Backward Coordinate Bounds

In this section, let $W \in \mathbb{R}^{m\times m}$, $A \in \mathbb{R}^{m\times d_x}$ and $B \in \mathbb{R}^{d\times m}$ denote three random matrices where each entry of W and A is i.i.d. sampled from $\mathcal{N}(0, \frac{2}{m})$ and each entry of B is i.i.d. sampled from $\mathcal{N}(0, \frac{1}{d})$. We let $\{\mathsf{loss}_{i,\ell} \in \mathbb{R}^d : i \in [n], \ell \in [L]\backslash\{1\}\}$ be arbitrary fixed vectors, and let $i^*, \ell^* = \arg\max_{i,\ell} \|\mathsf{loss}_{i,\ell}\|_2$.

- In Section D.1, we prove that there are sufficiently many coordinates $k \in [m]$ such that the quantity $|(g_{i,\ell+1})_k| = |(Wh_{i,\ell} + Ax_{i,\ell+1})_k|$ is small for $(i,\ell) = (i^*,\ell^*)$ but large for $(i,\ell) \in [n]\times\{\ell^*,\ell^*+1,L\}\backslash(i^*,\ell^*)$.
- In Section D.2, we prove that there are sufficiently many coordinates $k \in [m]$ such that $|\sum_a(\mathsf{Back}^\top_{i^*,\ell^*\to a}\cdot\mathsf{loss}_{i^*,a})_k|$ is large.

D.1 Indicator Coordinate Bound

For this section, we fix some choice of $i^* \in [n]$ and $\ell^* \in [L]\backslash\{1\}$, and define two parameters

Definition D.1.

$$\boxed{\beta_+ \stackrel{\text{def}}{=} \frac{\delta}{\rho^2} \quad \textit{and} \quad \beta_- \stackrel{\text{def}}{=} \frac{\delta}{\rho^{10}}}$$

Lemma D.2 (indicator coordinate bound, c.f. (7.3))**.** *Suppose W and A follow from random initialization. Given fixed set $N_1 \subseteq [m]$ and define*

$$N_4 \stackrel{\text{def}}{=} \left\{ k \in N_1 \;\middle|\; \left[\begin{array}{ll} |\big(Wh_{i,\ell} + Ax_{i,\ell+1}\big)_k| \le \beta_-/\sqrt{m}, & \textit{if } i = i^*, \ell = \ell^*; \\ |\big(Wh_{i,\ell} + Ax_{i,\ell+1}\big)_k| \ge \beta_+/\sqrt{m}, & \textit{if } i \ne i^* \textit{ and } \ell = \ell^*; \\ |\big(Wh_{i,\ell} + Ax_{i,\ell+1}\big)_k| \ge \beta_+/\sqrt{m}, & \textit{if } i \in [n] \textit{ and } \ell > \ell^*. \end{array}\right.\right\}$$

Then, as long as $|N_1| \ge n^2L^2/\beta_-$, we have

$$\Pr_{W,A}\left[|N_4| \ge \frac{\beta_-|N_1|}{64L}\right] \ge 1 - e^{-\Omega(\beta_-|N_1|/L)}.$$

Lemma D.3. *Suppose W and A follow from random initialization. Given fixed set $N_1 \subseteq [m]$ and define*

$$N_2 \stackrel{\text{def}}{=} \left\{ k \in N_1 \;\middle|\; |\langle W_k, h_{i^*,\ell^*}\rangle + \langle A_k, x_{i^*,\ell^*+1}\rangle| \le \frac{\beta_-}{\sqrt{m}} \right\}.$$

Then, we have

$$\Pr_{W,A}\left[|N_2| \ge \frac{\beta_-|N_1|}{16L}\right] \ge 1 - e^{-\Omega(\beta_-|N_1|/L)}.$$

Proof. Let $i = i^*$ and $\ell = \ell^*$. Similar to the proof of Lemma B.3, we can write

$$y = Wh_{i,\ell} + Ax_{i,\ell+1} = WU_\ell U_\ell^\top h_{i,\ell} + Ax_{i,\ell+1} = M_1z_1 + M_2z_2$$

where $M_1 \in \mathbb{R}^{m\times n\ell}$, $M_2 = A$, $z_1 = U_\ell^\top h_{i,\ell}$ and $z_2 = x_{i,\ell+1}$. Again, the entries of M_1 and M_2 are i.i.d. from $\mathcal{N}(0, \frac{2}{m})$.

Now, suppose z_1 is fixed (instead of depending on the randomness of W and A), and it satisfies $\|z_1\| \le 6L$. We have each entry of y is is distributed as $\mathcal{N}\big(0, \frac{2\|z_1\|^2+2\|z_2\|^2}{m}\big)$. By property of Gaussian (see Fact A.12), we have for each $y \in [m]$,

$$\mathbf{Pr}\left[|y_k| \le \frac{0.9\beta_-}{\sqrt{m}}\right] \ge \frac{2}{3}\frac{0.9\beta_-}{\sqrt{2\|z_1\|^2 + 2\|z_2\|^2}} \ge \frac{\beta_-}{15L} \enspace .$$

Since we have $|N_1|$ independent random variables, applying a Chernoff bound, we have

$$\mathbf{Pr}\left[\left|\left\{k \in N_1 \colon |y_k| \le \frac{0.9\beta_-}{\sqrt{m}}\right\}\right| \ge \frac{\beta_-}{16L}|N_1|\right] \ge 1 - \exp(-\Omega(\beta_-|N_1|/L)) \enspace .$$

Finally, by taking a union bound with respect to the ε-net over all possible choices of $z_1 \in \mathbb{R}^{n\ell}$, we have that

$$\mathbf{Pr}\left[\left|\left\{k \in N_1 \colon |y_k| \le \frac{\beta_-}{\sqrt{m}}\right\}\right| \ge \frac{\beta_-}{16L}|N_1|\right] \ge 1 - \exp(-\Omega(\beta_-|N_1|/L)) \enspace .$$

for all z_1 and a fixed z_2. Plugging in the random choice of $z_1 = U_\ell^\top h_{i,\ell}$ (and we have $\|z_1\| \le 6L$ by Lemma B.3), we have the desired result. □

Observe that the set N_2 produced by Lemma D.3 only depends on the randomness of A, WU_{ℓ^*-1} and Wh_{i^*,ℓ^*}. In other words, for any unit vector $z \in \mathbb{R}^m$ that is orthogonal to the columns of U_{ℓ^*-1} and orthogonal to h_{i^*,ℓ^*}, we have that Wz is independent of N_2.

We next proceed to refine N_2:

Lemma D.4 ($i \ne i^*, \ell = \ell^*$)**.** *Suppose W and A follow from random initialization, and $N_2 \subseteq [m]$ is a given set (that may only depend on the randomness of A, WU_{ℓ^*-1} and Wh_{i^*,ℓ^*}). Define*

$$N_3 = \left\{k \in N_2 \;\middle|\; |\langle W_k, h_{i,\ell^*}\rangle + \langle A_k, x_{i,\ell^*+1}\rangle| \ge \frac{\beta_-}{\sqrt{m}}, \forall i \in [n]\backslash i^*\right\}.$$

Then

$$\mathbf{Pr}\left[|N_3| \ge \frac{1}{2}|N_2| \;\middle|\; A, WU_{\ell^*-1}, Wh_{i^*,\ell^*}\right] \ge 1 - e^{-\Omega(|N_2|/n)}.$$

Proof. First fix some $i \in [n]\backslash i^*$ and $\ell = \ell^*$. Define set $N_{3,i}$,

$$N_{3,i} \stackrel{\text{def}}{=} \left\{k \in N_2 \;\middle|\; |\langle W_k, h_{i,\ell}\rangle + \langle A_k, x_{i,\ell+1}\rangle| \ge \frac{\beta_-}{\sqrt{m}}\right\} \enspace .$$

Let $v \stackrel{\text{def}}{=} \frac{(I-U_{\ell-1}U_{\ell-1}^\top)h_{i^*,\ell}}{\|(I-U_{\ell-1}U_{\ell-1}^\top)h_{i^*,\ell}\|}$, and let column orthonormal matrix

$$U \stackrel{\text{def}}{=} \mathsf{GS}(U_{\ell-1}, h_{i^*,\ell}) = [U_{\ell-1}, v] \enspace .$$

We have

$$\langle W_k, h_{i,\ell}\rangle + \langle A_k, x_{i,\ell+1}\rangle = \Big(WUUh_{i,\ell} + Ax_{i,\ell+1} + W(I - UU^\top)h_{i,\ell}\Big)_k \enspace .$$

The rest of the proof is focusing on having the WU and A being fixed (so the first two vectors $WUUh_{i,\ell} + Ax_{i,\ell+1}$ are fixed), and letting $W(I-UU^\top)h_{i,\ell}$ be the only random variable. We have

$$W(I - UU^\top)h_{i,\ell} \sim \mathcal{N}\left(0, \frac{2\|(I-UU^\top)h_{i,\ell}\|_2^2 \cdot I}{m}\right) \enspace .$$

and therefore for a fixed vector μ it satisfies

$$Wh_{i,\ell} + Ax_{i,\ell+1} = WUUh_{i,\ell} + Ax_{i,\ell+1} + W(I-UU^\top)h_{i,\ell} \sim \mathcal{N}\left(\mu, \frac{2\|(I-UU^\top)h_{i,\ell}\|_2^2 \cdot I}{m}\right) \enspace .$$

According to Lemma B.9, with probability at least $1-\exp(-\Omega(\sqrt{m}))$, it satisfies $\|(I-UU^\top)h_{i,\ell}\|_2 \geq \frac{\delta}{2}$. By property of Gaussian distribution (see Fact A.12), we have for each $k \in N_2$:

$$\mathbf{Pr}\left[|(Wh_{i,\ell} + Ax_{i,\ell+1})_k| \geq \frac{\beta_-}{\sqrt{m}} \,\Big|\, WU, A\right] \geq 1 - \frac{\beta_-}{\delta} \geq 1 - \frac{1}{4n} \enspace .$$

Applying Chernoff bound, with probability at least $1 - e^{-\Omega(|N_2|/n)}$, we have $|N_{3,i}| \geq (1 - \frac{1}{2n})|N_2|$. Applying union bound over all possible $i \in [n] \setminus \{i^*\}$, we have $N_3 = \bigcap_{i \neq i^*} N_{3,i}$ has cardinality at least $\frac{|N_2|}{2}$. $\square$

Lemma D.5 ($\ell > \ell^*$)**.** *Suppose W and A follow from random initialization, and $N_3 \subseteq [m]$ is a given set (that may only depend on the randomness of A, WU_{ℓ^*}). Define*

$$N_4 = \left\{ k \in N_3 \,\middle|\, |\langle W_k, h_{i,\ell}\rangle + \langle A_k, x_{i,\ell+1}\rangle| \geq \frac{\beta_-}{\sqrt{m}}, \forall i \in [n], \ell > \ell^* \right\} \enspace .$$

Then

$$\mathbf{Pr}\left[|N_4| \geq \frac{1}{2}|N_3| \,\middle|\, A, WU_{\ell^*}\right] \geq 1 - e^{-\Omega(|N_3|/nL)} \enspace .$$

Proof. Letting $N_{4,\ell^*} \stackrel{\text{def}}{=} N_3$. For any $i \in [n]$ and $\ell = \ell^*+1, \ell^*+2, \dots, L$, we define $N_{4,i,\ell}$ as

$$N_{4,i,\ell} \stackrel{\text{def}}{=} \left\{ k \in N_{4,\ell-1} \,\middle|\, |\langle W_k, h_{i,\ell}\rangle + \langle A_k, x_{i,\ell+1}\rangle| \geq \frac{\beta_-}{\sqrt{m}} \right\} \quad \text{and} \quad N_{4,\ell} \stackrel{\text{def}}{=} \bigcap_{i \in [n]} N_{4,i,\ell} \enspace .$$

We rewrite

$$\langle W_k, h_{i,\ell}\rangle + \langle A_k, x_{i,\ell+1}\rangle = (WU_{\ell-1}U_{\ell-1}^\top h_{i,\ell} + Ax_{i,\ell+1} + W(I - U_{\ell-1}U_{\ell-1}^\top)h_{i,\ell})_k \enspace .$$

The rest of the proof is focusing on having the $WU_{\ell-1}$ and A being fixed (so the first two vectors $WUUh_{i,\ell} + Ax_{i,\ell+1}$ are fixed and $N_{4,\ell-1}$ is fixed), and letting $W(I - U_{\ell-1}U_{\ell-1}^\top)h_{i,\ell}$ be the only random variable. We have

$$W(I - U_{\ell-1}U_{\ell-1}^\top)h_{i,\ell} \sim \mathcal{N}\left(0, \frac{2\|(I - U_{\ell-1}U_{\ell-1}^\top)h_{i,\ell}\|_2^2 \cdot I}{m}\right) \enspace .$$

and therefore for a fixed vector μ it satisfies

$$Wh_{i,\ell}+Ax_{i,\ell+1} = WU_{\ell-1}U_{\ell-1}h_{i,\ell}+Ax_{i,\ell+1}+W(I-U_{\ell-1}U_{\ell-1}^\top)h_{i,\ell} \sim \mathcal{N}\left(\mu, \frac{2\|(I - U_{\ell-1}U_{\ell-1}^\top)h_{i,\ell}\|_2^2 \cdot I}{m}\right) \enspace .$$

According to Lemma B.6, with probability at least $1-\exp(-\Omega(\sqrt{m}))$, it satisfies $\|(I-UU^\top)h_{i,\ell}\|_2 \geq \frac{\delta}{2}$. By property of Gaussian distribution (see Fact A.12), we have for each $k \in N_{4,\ell-1}$:

$$\mathbf{Pr}\left[|(\widetilde{W}h_{i,\ell} + Ax_{i,\ell+1})_k| \geq \frac{\beta_-}{\sqrt{m}} \,\Big|\, WU, A\right] \geq 1 - \frac{\beta_-}{\delta} \geq 1 - \frac{1}{4nL} \enspace .$$

Applying Chernoff bound, we have

$$\mathbf{Pr}\left[|N_{4,i,\ell}| \geq (1 - \frac{1}{3nL})|N_{4,\ell-1}| \,\Big|\, WU_{\ell-1}, A\right] \geq 1 - e^{-\Omega(|N_{4,\ell-1}|/nL)} \enspace .$$

Applying union bound over all possible $i \in [n]$, we have $N_{4,\ell} = \bigcap_{i \in [n]} N_{4,i,\ell}$ has cardinality at least $(1 - \frac{1}{3L})|N_{4,\ell-1}|$. After telescoping we have $|N_4| = |N_{4,L}| \geq \frac{1}{2}|N_3|$. $\square$

D.2 Backward Coordinate Bound

Recall the following notions

Definition D.6. *Given matrices W, A, B. For each i, for each ℓ, we define*

$$\mathsf{Back}_{i,\ell\to\ell} = B \in \mathbb{R}^{d\times m}.$$

For each i, for each $a \ge \ell+1$, we define

$$\mathsf{Back}_{i,\ell\to a} = B D_{i,a} W \cdots D_{i,\ell+1} W \in \mathbb{R}^{d\times m}.$$

The goal of this section is to prove Lemma D.7.

Lemma D.7 (backward coordinate bound, c.f. (7.2))**.** *Let $\rho = ndL\log m$ and $\varrho = nLd\delta^{-1}\log(m/\varepsilon)$. Let $\{\mathsf{loss}_{i,\ell} \in \mathbb{R}^d : i \in [n], \ell \in [L]\backslash\{1\}\}$ be fixed vectors and $i^*, \ell^* = \arg\max_{i,\ell} \| \mathsf{loss}_{i,\ell} \|_2$. If $W \in \mathbb{R}^{m\times m}, A \in \mathbb{R}^{m\times d_x}, B \in \mathbb{R}^{d\times m}$ are random, and $N_4 \subseteq [m]$ is a set with $|N_4| \in [\rho^4, \varrho^{100}]$ (N_4 can depend on the randomness of W and A, but not depend on B). Define*

$$N_5 \overset{\mathrm{def}}{=} \left\{ k \in N_4 \;\middle|\; \left| \sum_{a=\ell^*+1}^{L} (\mathsf{Back}_{i^*,\ell^*+1\to a}^\top \cdot \mathsf{loss}_{i^*,a})_k \right| \ge \frac{\| \mathsf{loss}_{i^*,\ell^*} \|_2}{6\sqrt{d}nL} \right\} .$$

Then with probability at least $1 - e^{-\Omega(\rho^2)}$, we have

$$|N_5| \ge \left(1 - \frac{1}{2nL}\right)|N_4|.$$

Proof. For notational simplicity, in the proof we use ℓ to denote ℓ^*, use N to denote N_4, and drop the subscript i^*. We denote $B \in \mathbb{R}^{d\times m}$ as $[b_1, b_2, \cdots, b_m]$ where each $b_i \in \mathbb{R}^d$. We denote by $C_{a,\ell+1} = D_a W \cdots D_{\ell+1} W$.

We define a function $v_k(b_1, b_2, \cdots, b_m) \in \mathbb{R}$ as follows

$$v_k(b_1, b_2, \cdots, b_m) \overset{\mathrm{def}}{=} \sum_{a=\ell+1}^{L} (\mathsf{Back}_{\ell+1\to a}^\top \cdot \mathsf{loss}_{i,a})_k.$$

We can rewrite

$$\begin{aligned}
v_k(b_1, b_2, \cdots, b_m) &= \left(B^\top \cdot \mathsf{loss}_{i,\ell+1} + \sum_{a=\ell+2}^{L} \mathsf{Back}_{\ell+1\to a}^\top \cdot \mathsf{loss}_{i,a} \right)_k \\
&= \left(B^\top \cdot \mathsf{loss}_{i,\ell+1} + \sum_{a=\ell+2}^{L} (C_{a,\ell+2})^\top B^\top \cdot \mathsf{loss}_{i,a} \right)_k \\
&= \langle b_k, \mathsf{loss}_{i,\ell+1} \rangle + \sum_{a=\ell+2}^{L} \left\langle \sum_{j=1}^{m} (C_{a,\ell+2})_{k,j} \cdot b_j, \mathsf{loss}_{i,a} \right\rangle \\
&= \left\langle b_k, \mathsf{loss}_{i,\ell+1} + \sum_{a=\ell+2}^{L} (C_{a,\ell+2})_{k,j} \cdot \mathsf{loss}_{i,a} \right\rangle + \sum_{j\in[m]\backslash k} \left\langle b_j, \sum_{a=\ell+2}^{L} (C_{a,\ell+2})_{k,j} \cdot \mathsf{loss}_{i,a} \right\rangle
\end{aligned}$$

where we have used $(C_{a,\ell+2})_{k,j}$ to denote the (k,j) entry of matrix $C_{a,\ell+2}$.

Choose parameter $\mathfrak{L}_h = \frac{3\sqrt{d}nL}{\| \mathsf{loss}_{i,\ell} \|_2}$ and we can define function $h : \mathbb{R} \to [0,1]$ by

$$h(t) \overset{\mathrm{def}}{=} \begin{cases} |t| \cdot \mathfrak{L}_h, & \text{if } |t| \le 1/\mathfrak{L}_h; \\ 1, & \text{otherwise.} \end{cases}$$

We have $h(t)$ is $\mathfrak{L}_h$-Lipschitz continuous.

We define two probabilistic events E_1, E_2.

- Event E_1 depends on the randomness of W and A:

$$E_1 \stackrel{\text{def}}{=} \left\{ |z^\top C_{a,\ell+2} y| \le \frac{\rho}{\sqrt{m}} \|z\| \|y\| \text{ for all } a, \ell \text{ and for all 1-sparse vectors } y, z \right\}$$

Corollary B.16 says $\mathbf{Pr}[E_1] \ge 1 - e^{-\Omega(\rho^2)}$.

- Event E_2 depends on the randomness of B:

$$E_2 \stackrel{\text{def}}{=} \left\{ |B_{i,j}| \le \frac{\rho}{\sqrt{d}} \text{ for all } i, j \right\}$$

By Gaussian tail bound, we have $\mathbf{Pr}[E_2] \ge 1 - e^{-\Omega(\rho^2)}$.

In the rest of the proof, we assume W and A are fixed and satisfy E_1. We let B be the only source of randomness but we condition on B satisfies E_2.

For simplicity we use subscript $-j$ to denote $[m] \setminus j$, and subscript $-N$ to denote $[m] \setminus N$. For instance, we shall write $b = (b_{-k}, b_k) = (b_N, b_{-N})$.

Step 1. Fixing b_{-N} and letting b_N be the only randomness, we claim for each $k \in N$:

$$\mathbb{E}_{b_N}[h(v_k)] \ge \mathbf{Pr}_{b_N}\left[|v_k| \ge \frac{1}{\mathfrak{L}_h} \right] \ge 1 - \frac{1}{4nL} \ .$$

We prove this inequality as follows. We can rewrite

$$\begin{aligned} v_k = &\left\langle b_k, \mathsf{loss}_{i,\ell+1} + \sum_{a=\ell+2}^{L} (C_{a,\ell+2})_{k,k} \cdot \mathsf{loss}_{i,a} \right\rangle + \sum_{j \in N \setminus \{k\}} \left\langle b_j, \sum_{a=\ell+2}^{L} (C_{a,\ell+2})_{k,j} \cdot \mathsf{loss}_{i,a} \right\rangle \\ &+ \sum_{j \in [m] \setminus N} \left\langle b_j, \sum_{a=\ell+2}^{L} (C_{a,\ell+2})_{k,j} \cdot \mathsf{loss}_{i,a} \right\rangle \end{aligned}$$

and v_k is distributed as Gaussian random variable $\mathcal{N}(\mu, \sigma^2)$, and μ and σ^2 are defined as follows

$$\begin{aligned} \mu &= \sum_{j \in [m] \setminus N} \left\langle b_j, \sum_{a=\ell+2}^{L} (C_{a,\ell+2})_{k,j} \cdot \mathsf{loss}_{i,a} \right\rangle \\ \sigma^2 &= \frac{1}{d} \left\| \mathsf{loss}_{i,\ell} + \sum_{a=\ell+2}^{L} (C_{a,\ell+2})_{k,k} \cdot \mathsf{loss}_{i,a} \right\|_2^2 + \frac{1}{d} \sum_{j \in N \setminus \{k\}} \left\| \sum_{a=\ell+2}^{L} (C_{a,\ell+2})_{k,j} \cdot \mathsf{loss}_{i,a} \right\|_2^2 \end{aligned}$$

Meanwhile, we calculate that

$$\begin{aligned} \left\| \sum_{a=\ell+2}^{L} (C_{a,\ell+2})_{k,j} \cdot \mathsf{loss}_{i,a} \right\|_2 &\overset{①}{\le} \sum_{a=\ell+2}^{L} |(C_{a,\ell+2})_{k,j}| \cdot \| \mathsf{loss}_{i,a} \|_2 \\ &\overset{②}{\le} \sum_{a=\ell+2}^{L} |(C_{a,\ell+2})_{k,j}| \cdot \| \mathsf{loss}_{i,\ell} \|_2 \overset{③}{\le} \frac{L\rho}{\sqrt{m}} \cdot \| \mathsf{loss}_{i,\ell} \|_2 \end{aligned} \tag{D.1}$$

where ① follows by triangle inequality, ② follows by $i, \ell = \arg\max_{i',\ell'} \| \mathsf{loss}_{i',\ell'} \|$, and ③ follows from event E_1. Therefore,

$$\sigma^2 \geq \frac{1}{d}\left(\|\,\mathsf{loss}_{i,\ell}\,\|_2 - \left\|\sum_{a=\ell+2}^{L}(C_{a,\ell+2})_{k,j}\cdot\mathsf{loss}_{i,a}\right\|_2\right)^2$$
$$\geq \frac{1}{d}\left(1-\frac{L\rho}{\sqrt{m}}\right)^2\|\,\mathsf{loss}_{i,\ell}\,\|_2^2 \geq \frac{1}{2d}\|\,\mathsf{loss}_{i,\ell}\,\|_2^2$$

where the first step follows by $\|a+b\|_2 \geq \|a\|_2 - \|b\|_2$, the second step follows by (D.1). In sum, we have that v_k is a random Gaussian with $\sigma^2 \geq \frac{1}{2d}\|\,\mathsf{loss}_{i,\ell}\,\|_2^2$. This means,

$$\Pr_{b_N}\left[|v_k| \geq \frac{1}{\mathfrak{L}_h}\right] \geq \Pr_{b_N}\left[|v_k| \geq \frac{\|\,\mathsf{loss}_{i,\ell}\,\|_2}{3\sqrt{d}nL}\right] \geq 1-\frac{1}{4nL}\ .$$

according to Fact A.12.

Step 2. For every $b_1, b_2, \cdots, b_m$ and $b'_k \in \mathbb{R}^d$ satisfying event E_2, for every $j \neq k$, we have

$$\begin{aligned}|v_j(b_{-k}, b_k) - v_j(b_{-k}, b'_k)| &= \left|\left\langle b_k - b'_k, \sum_{a=\ell+2}^{L}(C_{a,\ell+2})_{k,k}\cdot\mathsf{loss}_{i,a}\right\rangle\right| \\ &\leq \|b_k - b'_k\|_2 \cdot \left\|\sum_{a=\ell+2}^{L}(C_{a,\ell+2})_{k,k}\cdot\mathsf{loss}_{i,a}\right\|_2 \\ &\leq \rho\cdot\frac{L\rho}{\sqrt{m}}\|\,\mathsf{loss}_{i,\ell}\,\|_2 \leq \frac{\rho^3}{\sqrt{m}}\|\,\mathsf{loss}_{i,\ell}\,\|_2 \end{aligned} \tag{D.2}$$

where the second step follows by $|\langle a,b\rangle| \leq \|a\|_2 \cdot \|b\|_2$, the third step follows by (D.1).

Step 3. Fixing b_{-N} and letting function $g(b_1, b_2, \cdots, b_m) \stackrel{\text{def}}{=} \sum_{k\in N} h(v_k(b_1, b_2, \cdots, b_m))$. We have

$$\begin{aligned}|g(b_j, b_{-j}) - g(b'_j, b_{-j})| &= \left|\sum_{k\in N} h(v_k(b_j, b_{-j})) - \sum_{k\in N} h(v_k(b'_j, b_{-j}))\right| \\ &\leq \sum_{k\in N}\left|h(v_k(b_j, b_{-j})) - h(v_k(b'_j, b_{-j}))\right| \\ &= \left|h(v_j(b_j, b_{-j})) - h(v_j(b'_j, b_{-j}))\right| + \sum_{k\in N\setminus\{j\}}\left|h(v_k(b_j, b_{-j})) - h(v_k(b'_j, b_{-j}))\right| \\ &\stackrel{①}{\leq} 1 + \sum_{k\in N\setminus\{j\}}\left|h(v_k(b_j, b_{-j})) - h(v_k(b'_j, b_{-j}))\right| \\ &\leq 1 + \sum_{k\in N\setminus\{j\}}\mathfrak{L}_h\cdot|v_k(b_j, b_{-j}) - v_k(b'_j, b_{-j})| \\ &\stackrel{②}{\leq} 1 + |N|\cdot\frac{3\sqrt{d}nL}{\|\,\mathsf{loss}_{i,\ell}\,\|_2}\cdot\left((nLd\log m)^3\frac{\|\,\mathsf{loss}_{i,\ell}\,\|_2}{\sqrt{m}}\right) \\ &\leq 1 + \frac{|N|\cdot 3(nLd\log m)^4}{\sqrt{m}} \leq 2\ .\end{aligned}$$

Above, inequality ① follows by $h(t) \in [0,1]$, inequality ② follows from (D.2). Using McDiarmid inequality (see Lemma A.18), we have

$$\begin{aligned}
&\Pr_{b_N}\left[\left|g(b_1,\cdots,b_m)-\mathbb{E}_{b_N}[g]\right|\geq\varepsilon\right]\leq\exp\left(-\frac{\varepsilon^2}{8|N|}\right)\\
\Longrightarrow\quad&\Pr_{b_N}\left[\left|g(b_1,\cdots,b_m)-\mathbb{E}_{b_N}[g]\right|\geq\frac{|N|}{4nL}\right]<\exp\left(-\frac{|N|}{128(nL)^2}\right)
\end{aligned}$$

where we choose $\varepsilon=\frac{|N|}{4nL}$.

Finally, combining this with Step 1 —which gives $\mathbb{E}_{b_N}[g(b_1,\cdots,b_m)]\geq|N|(1-\frac{1}{4nL})$— we have with probability at least $1-e^{-\Omega(|N|/(nL)^2)}$ over randomness of b_N,

$$g(b_1,\cdots,b_m)\geq|N|\left(1-\frac{1}{2nL}\right)\ .$$

This implies at least $(1-\frac{1}{2nL})$-fraction of $k\in N$ will have $|v_k|\geq\frac{1}{2\mathfrak{L}_h}=\frac{\|\mathsf{loss}_{i,\ell}\|_2}{6nL\sqrt{d}}$. $\square$

E Gradient Bound at Random Initialization (Theorem 5)

The goal of this section is to understand the lower and upper bounds of gradients. Instead of analyzing the true gradient directly —where the forward and backward propagation have correlated randomness— we assume that the loss vectors $\{\mathsf{loss}_{i,a}\}_{i\in[n],a\in\{2,\dots,L\}}$ are fixed (no randomness) in this section, as opposed to being defined as $\mathsf{loss}_{i,a}=Bh_{i,a}-y^*_{i,a}$ which is random. We call this the "fake loss." We define a corresponding "fake gradient" with respect to this fixed loss.

Definition E.1 (fake gradient, c.f. (7.1))**.** *Given fixed vectors $\{\mathsf{loss}_{i,a}\}_{i\in[n],a\in\{2,\dots,L\}}$, we define*

$$\begin{aligned}
\widehat{\nabla}_k f(W)&\stackrel{\text{def}}{=}\sum_{i=1}^n\sum_{a=2}^L\sum_{\ell=1}^{a-1}(\mathsf{Back}^\top_{i,\ell+1\to a}\cdot\mathsf{loss}_{i,a})_k\cdot h_{i,\ell}\cdot\mathbf{1}_{\langle W_k,h_{i,\ell}\rangle+\langle A_k,x_{i,\ell+1}\rangle\geq0}\\
\widehat{\nabla} f(W)&\stackrel{\text{def}}{=}\sum_{i=1}^n\sum_{a=2}^L\sum_{\ell=1}^{a-1}D_{i,\ell+1}(\mathsf{Back}^\top_{i,\ell+1\to a}\cdot\mathsf{loss}_{i,a})\cdot h_{i,\ell}^\top\\
\widehat{\nabla} f_i(W)&\stackrel{\text{def}}{=}\sum_{a=2}^L\sum_{\ell=1}^{a-1}D_{i,\ell+1}(\mathsf{Back}^\top_{i,\ell+1\to a}\cdot\mathsf{loss}_{i,a})\cdot h_{i,\ell}^\top
\end{aligned}$$

In our analysis, we also write it as

$$\begin{aligned}
\widehat{\nabla}_k f(W)&=\sum_{i=1}^n\sum_{\ell=1}^{L-1}(u_{i,\ell})_k\cdot h_{i,\ell}\cdot\mathbf{1}_{(g_{i,\ell+1})_k\geq0}\\
\widehat{\nabla} f(W)&=\sum_{i=1}^n\sum_{\ell=1}^{L-1}D_{i,\ell+1}u_{i,\ell}\cdot h_{i,\ell}^\top\\
\widehat{\nabla} f_i(W)&=\sum_{\ell=1}^{L-1}D_{i,\ell+1}u_{i,\ell}\cdot h_{i,\ell}^\top
\end{aligned}$$

where vectors $u_{i,\ell},g_{i,\ell+1}\in\mathbb{R}^m$ are defined as

$$u_{i,\ell}\stackrel{\text{def}}{=}\sum_{a=\ell+1}^L\mathsf{Back}^\top_{i,\ell+1\to a}\cdot\mathsf{loss}_{i,a}\qquad\qquad g_{i,\ell+1}\stackrel{\text{def}}{=}Wh_{i,\ell}+Ax_{i,\ell+1}\ .$$

We first state a simple upper bound on the gradients.

Lemma E.2. *Letting $\rho = nLd\log m$, given fixed vectors $\{\mathsf{loss}_{i,\ell} \in \mathbb{R}^d\}_{i\in[n],\ell\in[L]\setminus\{1\}}$, with probability at least $1 - e^{-\Omega(m/L^2)}$ over W, A, B, we have*

$$\|\widehat{\nabla} f(W)\|_F \leq O(nL^6\sqrt{m}) \cdot \max_{i,\ell}\{\|\,\mathsf{loss}_{i,\ell}\,\|\} \quad \text{and}$$

$$\forall i \in [n] \colon \|\widehat{\nabla} f_i(W)\|_F \leq O(L^6\sqrt{m}) \cdot \max_{i,\ell}\{\|\,\mathsf{loss}_{i,\ell}\,\|\} \enspace .$$

With probability at least $1 - e^{-\Omega(\rho^2)}$, we have for every $k \in [m]$:

$$\|\widehat{\nabla}_k f(W)\|_2 \leq O(n\rho L^3) \cdot \max_{i,\ell}\{\|\,\mathsf{loss}_{i,\ell}\,\|\} \enspace .$$

Proof of Lemma E.2. For each i, a, ℓ, we can calculate that

$$\begin{aligned}
\left\|D_{i,\ell+1}(\mathsf{Back}_{i,\ell+1\to a}^\top \cdot \mathsf{loss}_{i,a}) \cdot h_{i,\ell}^\top\right\|_F &= \left\|D_{i,\ell+1}(\mathsf{Back}_{i,\ell+1\to a}^\top \cdot \mathsf{loss}_{i,a})\right\|_2 \cdot \|h_{i,\ell}^\top\|_2 \\
&\leq \|\mathsf{Back}_{i,\ell+1\to a}\|_2 \cdot \|\,\mathsf{loss}_{i,a}\,\|_2 \cdot \|h_{i,\ell}^\top\|_2 \\
&\overset{①}{\leq} O(\sqrt{m}) \cdot O(L^3) \cdot O(L) \cdot \|\,\mathsf{loss}_{i,a}\,\|_2 \enspace .
\end{aligned}$$

where inequality ① uses Lemma B.11, Lemma B.3, and $\|B\|_2 \leq O(\sqrt{m})$ with high probability. Applying triangle inequality and using the definition of fake gradient (see Definition E.1), we finish the proof of the first statement.

As for the second statement, we replace the use of Lemma B.11 with Corollary B.19. □

The rest of this section is devoted to proving a (much more involved) lower bound on this fake gradient.

Theorem 5 (gradient lower bound at random init, restated)**.** *Letting $\rho = nLd\log m$, given fixed vectors $\{\mathsf{loss}_{i,\ell} \in \mathbb{R}^d\}_{i\in[n],\ell\in[L]\setminus\{1\}}$, with probability at least $1 - e^{-\Omega(\rho^2)}$ over W, A, B, we have*

$$\|\widehat{\nabla} f(W)\|_F^2 \geq \Omega\left(\frac{\delta}{\rho^{14}}\right) \times m \times \max_{i,\ell}\{\|\,\mathsf{loss}_{i,\ell}\,\|^2\} \enspace .$$

In the rest of this section, we first present a elegant way to decompose the randomness in Section E.1 (motivated by smooth analysis [89]). We give a lower bound on the *expected* fake gradient in Section E.2. We then calculate the stability of fake gradient against rank-one perturbations in Section E.3. Finally, in Section E.4, we apply our extended McDiarmid's inequality to prove Theorem 5.

E.1 Randomness Decomposition

We introduce a parameter θ as follows:

Definition E.3. *Choose any*

$$\boxed{\theta \in \left[\rho^4 \cdot \beta_-, \rho^{-3} \cdot \beta_+\right]}$$

where the notions β_- and β_+ come from Definition D.1. We have $\theta \leq \frac{\delta}{\rho}$.

In the next lemma, we introduce a way to decompose the randomness of W into $W = W_2 + W'$ for two corollated random matrices W_2 and W'. We will make sure that the entries of W_2 are also i.i.d. from $N(0, \frac{2}{m})$. This definition requires us to choose a specific pair $(i^*, \ell^*) \in [n] \times [L]$.

Lemma E.4. *Suppose W and A follow from random initialization. Given any fixed $i^* \in [n]$, $\ell^* \in [L]$, and parameter $\theta \in \big(0, \frac{1}{2\rho}\big]$, letting random vector $v_{i^*,\ell^*} \in \mathbb{R}^m$ be defined as*

$$v_{i^*,\ell^*} \stackrel{\text{def}}{=} \frac{(I - U_{\ell^*-1}U_{\ell^*-1}^\top)h_{i^*,\ell^*}}{\|(I - U_{\ell^*-1}U_{\ell^*-1}^\top)h_{i^*,\ell^*}\|_2} \enspace ,$$

then we can write $W = W_2 + W'$ for two random matrices $W_2 \in \mathbb{R}^{m\times m}$ and $W' \in \mathbb{R}^{m\times m}$ where

(a) entries of W_2 are i.i.d. drawn from $\mathcal{N}(0, \frac{2}{m})$ (so W_2 is in the same distribution as W);

(b) W' is correlated with W_2 and can be written as $W' = uv_{i^*,\ell^*}^\top$ for some $u \in \mathbb{R}^m$;

(c) With probability at least $1 - e^{-\Omega(\rho^2)}$ over the randomness of W_2:

$$\forall k\colon \quad \Pr_{W'}\left[u_k = (W'v_{i^*,\ell^*})_k \geq \frac{\theta}{2\sqrt{m}} \;\middle|\; W_2\right] \geq \frac{1}{4} \quad \textit{and}$$
$$\Pr_{W'}\left[u_k = (W'v_{i^*,\ell^*})_k \leq -\frac{\theta}{2\sqrt{m}} \;\middle|\; W_2\right] \geq \frac{1}{4}$$

(d) With probability at least $1 - e^{-\Omega(\rho^2)}$ over the randomness of W_2:

$$\Pr_{W'}\left[\|u\|_\infty = \|W'v_{i^*,\ell^*}\|_\infty \leq \frac{3\theta\rho}{2\sqrt{m}} \;\middle|\; W_2\right] \geq 1 - e^{-\Omega(\rho^2)} \enspace .$$

Proof. Since v_{i^*,ℓ^*} only depends on the randomness of WU_{ℓ^*-1}, we have $Wv_{i^*,\ell^*} \sim \mathcal{N}(0, \frac{2I}{m})$. As a result, letting $g_1, g_2 \in \mathbb{R}^m$ be two independent random Gaussian vectors sampled from $\mathcal{N}(0, \frac{2I}{m})$, we can couple them to make sure

$$Wv_{i^*,\ell^*} = \sqrt{1-\theta^2}g_1 + \theta g_2 \enspace .$$

We define $W_2, W' \in \mathbb{R}^{n\times n}$ as follows

$$\begin{aligned}
W_2 &\stackrel{\text{def}}{=} WU_{\ell^*-1}U_{\ell^*-1}^\top + g_1 v_{i^*,\ell^*}^\top + W(I - U_{\ell^*-1}U_{\ell^*-1}^\top)(I - v_{i^*,\ell^*}v_{i^*,\ell^*}^\top) \\
W' &\stackrel{\text{def}}{=} uv_{i^*,\ell^*}^\top \quad \text{where} \quad u \stackrel{\text{def}}{=} \theta g_2 - (1 - \sqrt{1-\theta^2})g_1
\end{aligned}$$

- It is easy to verify that

$$\begin{aligned}
W_2 + W' &= W(I - v_{i^*,\ell^*}v_{i^*,\ell^*}^\top) + g_1 v_{i^*,\ell^*}^\top + (\theta g_2 - (1 - \sqrt{1-\theta^2})g_1)v_{i^*,\ell^*}^\top \\
&= W(I - v_{i^*,\ell^*}v_{i^*,\ell^*}^\top) + (\theta g_2 + \sqrt{1-\theta^2}g_1)v_{i^*,\ell^*}^\top \\
&= W(I - v_{i^*,\ell^*}v_{i^*,\ell^*}^\top) + Wv_{i^*,\ell^*}v_{i^*,\ell^*}^\top \\
&= W.
\end{aligned}$$

- We verify that the entries of W_2 are i.i.d. Gaussian in two steps.

  On one hand, let $U = [\widehat{h}_1, \dots, \widehat{h}_m] \in \mathbb{R}^{m\times m}$ be an arbitrary orthonormal matrix where its first nL columns are identical to $U_L \in \mathbb{R}^{m\times nL}$. For each $k \in [nL]$, although $\widehat{h}_k$ may depend on the randomness of $W = W_2 + W'$, it can only depend on $W[\widehat{h}_1, \dots, \widehat{h}_{k-1}]$. Therefore, conditioning on the randomness of $W_2[\widehat{h}_1, \dots, \widehat{h}_{k-1}]$, we have that $W_2\widehat{h}_k$ is still $\mathcal{N}(0, \frac{2I}{m})$. This shows that all entries of W_2U still follow from $N(0, \frac{2}{m})$.

  On the other hand, let $V = W_2U = [v_1, \dots, v_m] \in \mathbb{R}^{m\times m}$, then we have $v_k = W_2\widehat{h}_k$ is independent of $\widehat{h}_k$ as argued above. Since each $v_k \sim N(0, \frac{2I}{m})$, we have that all entries of $VU^\top = \sum_k v_k\widehat{k}_k^\top$ are i.i.d. from $N(0, \frac{2}{m})$. Since $W_2 = VU^\top$ this concludes the proof.

- By standard Gaussian tail bound, we have with probability at least $1 - e^{-\Omega(\rho^2)}$, it satisfies $\|g_1\|_\infty \le \frac{\rho}{\sqrt{m}}$ and therefore $\|(1-\sqrt{1-\theta^2})g_1\|_\infty \le \frac{\theta^2\rho}{\sqrt{m}}$. If this happens, for every $k \in [m]$, we have (see for instance Fact A.12) that $\mathbf{Pr}_{g_2}[(\theta g_2)_k > \frac{\theta}{\sqrt{m}}] \ge \frac{1}{4}$ and $\mathbf{Pr}_{g_2}[(\theta g_2)_k < -\frac{\theta}{\sqrt{m}}] \ge \frac{1}{4}$. Recalling
$$u = (\theta g_2 - (1-\sqrt{1-\theta^2})g_1 \enspace .$$
and using the fact that $\theta \le \frac{1}{2\rho}$, we conclude that
$$\mathop{\mathbf{Pr}}_{W'}\left[u_k > \frac{\theta}{2\sqrt{m}} \,\Big|\, W_2, A\right] \ge \frac{1}{4} \quad\text{and}\quad \mathop{\mathbf{Pr}}_{W'}\left[u_k < -\frac{\theta}{2\sqrt{m}} \,\Big|\, W_2, A\right] \ge \frac{1}{4} \enspace .$$
- Again by Gaussian tail bound, we have with probability at least $1 - e^{-\Omega(\rho^2)}$, it satisfies $\|g_1\|_\infty, \|g_2\|_\infty \le \frac{\rho}{\sqrt{m}}$. Therefore, $\|(1-\sqrt{1-\theta^2})g_1\|_\infty \le \frac{\theta^2\rho}{\sqrt{m}}$ and using our assumption on θ, we have
$$\|u\|_\infty = \left\|(\theta g_2 - (1-\sqrt{1-\theta^2})g_1\right\|_\infty \le \frac{3\theta\rho}{2\sqrt{m}} \enspace .$$

□

Lemma E.5. *Given any fixed $i^* \in [n]$, $\ell^* \in [L]$, and parameter $\theta \in \left(0, \frac{1}{2\rho}\right]$, and suppose suppose W and A are at random initialization and $W = W_2 + W'$ where $W' = uv_{i^*,\ell^*}^\top$ (following Lemma E.4).*
Now, we introduce another two ways of decomposing randomness based on this definition.

(a) Given fixed $k \in [m]$, we can write $W = W_0 + W'_k$ where $W'_k = u_k v_{i^*,\ell^*}^\top$ for $u_k = (0,\dots,0,u_k,0,\dots,0)$ that is non-zero only at the k-th coordinate. We again have the entries of W_0 are i.i.d. from $N(0,\frac{2}{m})$.

(b) Given fixed $N \subseteq [m]$, we can write $W = W_1 + W'_N$ where $W'_N = u_N v_{i^*,\ell^*}^\top$ for u_N being the projection of u onto coordinates in N (so $\|u\|_0 = |N|$). We again have the entries of W_0 are i.i.d. from $N(0,\frac{2}{m})$.

(c) Given fixed $N \subseteq [m]$, we can write $W = W_2 + W'_N + W'_{-N}$ where $W'_N = u_N v_{i^*,\ell^*}^\top$ and $W'_{-N} = u_{-N} v_{i^*,\ell^*}^\top$. Here, u_{-N} is the projection of u onto coordinates in $[m]\setminus N$.

Proof. The proofs are analogous to Lemma E.4. □

E.2 Gradient Lower Bound in Expectation

The goal of this subsection is to prove Lemma E.6 and then translate it into our Core Lemma A (see Lemma E.7).

Lemma E.6 (c.f. (7.6))**.** *Let $\rho = nLd\log m$ and $\varrho = nLd\delta^{-1}\log(m/\varepsilon)$. Fix integer $N \in [\rho^6/\beta_-, \varrho^{100}]$, θ from Definition E.3, fix vectors $\{\mathsf{loss}_{i,\ell} \in \mathbb{R}^d\}_{i\in[n],\ell\in[L]\setminus\{1\}}$, let $(i^*,\ell^*) = \arg\min_{i,\ell}\{\|\mathsf{loss}_{i,\ell}\|\}$. Suppose W_1, A, B are at random initialization, and suppose W'_N is defined according to Lemma E.5b so that $W = W_1 + W'_N$ is also at random initialization. With probability at least $1-e^{-\Omega(\rho^2)}$ over the randomness of W_1, A, B, we have*
$$\left|\left\{k \in N \;\middle|\; \mathop{\mathbb{E}}_{W'_N}\left[\|\widehat{\nabla}_k f(W_1 + W'_N)\|_2\right] \ge \Omega\Big(\frac{\|\mathsf{loss}_{i^*,\ell^*}\|_2}{nL\sqrt{d}}\Big)\right\}\right| \ge \frac{\beta_-|N|}{100L}$$

E.2.1 Proof of Lemma E.6

Throughout this proof we denote by $\widetilde{W} = W_1$. We introduce some (old and new) notations:

$$\widetilde{\mathsf{Back}}_{i,\ell\to\ell} = B, \qquad \mathsf{Back}_{i,\ell\to\ell} = B$$
$$\widetilde{\mathsf{Back}}_{i,\ell\to a} = B\widetilde{D}_{i,a}\widetilde{W}\cdots\widetilde{D}_{i,\ell+1}\widetilde{W}, \qquad \mathsf{Back}_{i,\ell\to a} = BD_{i,a}W\cdots D_{i,\ell+1}W,$$

and we let $\mathsf{Back}'_{i,\ell\to a} = \mathsf{Back}_{i,\ell\to a} - \widetilde{\mathsf{Back}}_{i,\ell\to a}$. For each $i \in [n]$, $\ell \in [L]$, we define

$$\widetilde{g}_{i,\ell} = Ax_{i,\ell} + \widetilde{W}\widetilde{D}_{i,\ell-1}\widetilde{g}_{i,\ell-1} \qquad g_{i,\ell} = Ax_{i,\ell} + WD_{i,\ell-1}g_{i,\ell-1} \qquad g'_{i,\ell} = g_{i,\ell} - \widetilde{g}_{i,\ell}$$
$$\widetilde{h}_{i,\ell} = \widetilde{D}_{i,\ell}(Ax_{i,\ell} + \widetilde{W}\widetilde{h}_{i,\ell-1}) \qquad h_{i,\ell} = D_{i,\ell}(Ax_{i,\ell} + Wh_{i,\ell-1}) \qquad h'_{i,\ell} = h_{i,\ell} - \widetilde{h}_{i,\ell}$$
$$\widetilde{u}_{i,\ell} = \sum_{a=\ell}^{L} \widetilde{\mathsf{Back}}^\top_{\ell\to a}\,\mathsf{loss}_{i,a} \qquad u_{i,\ell} = \sum_{a=\ell}^{L} \mathsf{Back}^\top_{\ell\to a}\,\mathsf{loss}_{i,a} \qquad u'_{i,\ell} = u_{i,\ell} - \widetilde{u}_{i,\ell}$$

We also let

$$v_{i^*,\ell^*} = \frac{(I - U_{\ell^*-1}U^\top_{\ell^*-1})\widetilde{h}_{i^*,\ell^*}}{\|(I - U_{\ell^*-1}U^\top_{\ell^*-1})\widetilde{h}_{i^*,\ell^*}\|_2}$$

and recall from Lemma E.4d, we have

$$W'_N = yv^\top_{i^*,\ell^*} \text{ for some } y \in \mathbb{R}^m \text{ with } \|y\|_0 = |N| \text{ and } \|y\|_\infty \le \frac{\tau_0}{\sqrt{m}} \stackrel{\text{def}}{=} \frac{3\theta\rho}{2\sqrt{m}} \in \left[\frac{\varrho^{-20}}{\sqrt{m}}, \frac{\varrho^{20}}{\sqrt{m}}\right]$$

Proof of Lemma E.6. We let $N_1 = N$, apply Lemma D.2 to obtain $N_4 \subseteq N_1$ with $|N_4| \ge \frac{\beta_-|N_1|}{64L}$, and apply Lemma D.7 to obtain $N_5 \subseteq N_4$. According to the statements of these lemmas, we know that with probability at least $1 - e^{-\Omega(\rho^2)}$, the random choice of W_1, A, B will satisfy $|N_5| \ge \frac{\beta_-|N_1|}{100L}$.

Fixing such $k \in N_5$, we use $u_{i,\ell}$ to denote $(u_{i,\ell})_k$ and $g_{i,\ell}$ to denote $(g_{i,\ell})_k$ for notational simplicity. We can rewrite $\widehat{\nabla}_k f(W)$ as follows and apply triangle inequality:

$$\begin{aligned}\|\widehat{\nabla}_k f(W)\| &= \left\|u_{i^*,\ell^*}\cdot h_{i^*,\ell^*}\cdot \mathbf{1}_{g_{i^*,\ell^*+1}\ge 0} + \textstyle\sum_{(i,\ell)\ne(i^*,\ell^*)} u_{i,\ell}\cdot h_{i,\ell}\cdot \mathbf{1}_{g_{i,\ell+1}\ge 0}\right\| \\ &\ge \underbrace{\left\|\widetilde{u}_{i^*,\ell^*}\cdot \widetilde{h}_{i^*,\ell^*}\cdot \mathbf{1}_{g_{i^*,\ell^*+1}\ge 0} + \textstyle\sum_{(i,\ell)\ne(i^*,\ell^*)} \widetilde{u}_{i,\ell}\cdot \widetilde{h}_{i,\ell}\cdot \mathbf{1}_{g_{i,\ell+1}\ge 0}\right\|_2}_{\clubsuit} \\ &\quad - \underbrace{\textstyle\sum_{i,\ell}\left\|u'_{i,\ell}\cdot\widetilde{h}_{i,\ell} + \widetilde{u}_{i,\ell}\cdot h'_{i,\ell} + u'_{i,\ell}\cdot h'_{i,\ell}\right\|_2}_{\spadesuit}\end{aligned} \tag{E.1}$$

where recall $g_{i,\ell+1} = \langle W_k, h_{i,\ell}\rangle + \langle A_k, x_{i,\ell+1}\rangle$.

Step 1. To bound the $\spadesuit$ term on the RHS of (E.1), we consider the following bounds:

1. Using Lemma C.11, we have with probability at least $1 - e^{-\Omega(\rho^2)}$, for all i, ℓ, a

$$|(\widetilde{\mathsf{Back}}_{i,\ell\to a}\cdot \mathsf{loss}_{i,a})_k - (\mathsf{Back}_{i,\ell\to a}\cdot \mathsf{loss}_{i,a})_k| \le O\big(\frac{\rho^3\tau_0^{1/3}N^{1/6}}{m^{1/6}}\big)\cdot\|\,\mathsf{loss}_{i,a}\,\| \enspace .$$

   This implies, by triangle inequality and the fact that $\|\,\mathsf{loss}_{i,a}\,\| \le \|\,\mathsf{loss}_{i^*,a^*}\,\|$,

$$|u'_{i,\ell}| = |u_{i,\ell} - \widetilde{u}_{i,\ell}| \le O\big(\frac{\rho^4\tau_0^{1/3}N^{1/6}}{m^{1/6}}\big)\cdot\|\,\mathsf{loss}_{i^*,a^*}\,\| \enspace .$$

2. Using Corollary B.19, we have with probability at least $1 - e^{-\Omega(\rho^2)}$, for all i, ℓ, a

$$|(\widetilde{\mathsf{Back}}_{i,\ell\to a}\cdot \mathsf{loss}_{i,a})_k| \le \rho\|\,\mathsf{loss}_{i,a}\,\| \enspace .$$

This implies, by triangle inequality and the fact that $\| \mathsf{loss}_{i,a} \| \leq \| \mathsf{loss}_{i^*,a^*} \|$,

$$|\widetilde{u}_{i,\ell}| \leq \rho^2 \| \mathsf{loss}_{i^*,a^*} \| \enspace .$$

3. Using Lemma B.3 and Lemma C.2a we have

$$\|\widetilde{h}_{i,\ell}\|_2 \leq O(L) \quad \text{and} \quad \|h'_{i,\ell}\|_2 \leq O(L^6 \tau_0 \sqrt{N})/m^{1/2}$$

All together, they imply

$$\textstyle\sum_{i,\ell} \left\| u'_{i,\ell} \cdot \widetilde{h}_{i,\ell} + \widetilde{u}_{i,\ell} \cdot h'_{i,\ell} + u'_{i,\ell} \cdot h'_{i,\ell} \right\|_2 \leq O\big(\frac{\rho^6 \tau_0^{1/3} N^{1/6}}{m^{1/6}}\big) \cdot \| \mathsf{loss}_{i^*,a^*} \| \enspace . \tag{E.2}$$

Step 2. We next turn to the ♣ term on the RHS of (E.1). We divide into three cases to analyze the sign change of $g_{i,\ell+1}$ for all possible i and ℓ in the above formula (under the randomness of W'_N). Before we do so, first note that Lemma C.10 implies

$$|(W h'_{i,\ell})_k| = |(W_1 + W'_N) h'_{i,\ell}| \leq O\Big(\frac{\rho^8 N^{2/3} \tau_0^{5/6}}{m^{2/3}}\Big) \tag{E.3}$$

1. Consider $g_{i,\ell+1}$ for the case of $i = i^*, \ell = \ell^*$. We want to show that the sign of $g_{i,\ell+1}$ changes (at least with constant probability) with respect to the randomness of W'_N. One can easily check that:[16]

$$g_{i,\ell+1} = \langle A_k, x_{i,\ell+1} \rangle + (W_1 \widetilde{h}_{i,\ell})_k + (W h'_{i,\ell})_k + (W'_N v_{i,\ell})_k \cdot \|(I - U_{\ell-1} U_{\ell-1}^\top) h_{i,\ell}\|_2. \tag{E.4}$$

By Lemma D.2, we have $|\langle A_k, x_{i,\ell} \rangle + (W_1 \widetilde{h}_{i,\ell})_k| \leq \frac{\beta_-}{\sqrt{m}}$ Thus, putting this and (E.3) into (E.4), we can write

$$g_{i,\ell+1} = \Xi + (W'_N v_{i,\ell})_k \cdot \|(I - U_{\ell-1} U_{\ell-1}^\top) \widetilde{h}_{i,\ell}\|_2$$

for some value Ξ with $|\Xi| \leq \frac{2\beta_-}{\sqrt{m}}$. By Lemma B.6, we have $\|(I - U_{\ell-1} U_{\ell-1}^\top) \widetilde{h}_{i,\ell}\|_2 \geq \Omega\big(\frac{1}{L^2 \log^3 m}\big)$. By Lemma E.4c, we have $(W'_N v_{i,\ell})_k$ shall be larger than $\frac{\theta}{2\sqrt{m}}$ and smaller than $-\frac{\theta}{2\sqrt{m}}$ each with probability at least 1/4. Since by our choice of θ (see Definition E.3),

$$\frac{\theta}{2\sqrt{m}} \cdot \Omega\big(\frac{1}{L^2 \log^3 m}\big) \gg \frac{2\beta_-}{\sqrt{m}} \geq |\Xi|$$

This concludes that, with probability at least $1 - e^{-\Omega(\rho^2)}$ over W_1 and A, it satisfies

$$\mathop{\mathbf{Pr}}_{W'_N}[g_{i,\ell+1} > 0 \mid W_1, A] \geq \frac{1}{5} \quad \text{and} \quad \mathop{\mathbf{Pr}}_{W'_N}[g_{i,\ell+1} < 0 \mid W_1, A] \geq \frac{1}{5}$$

[16] We can write $(g_{i,\ell+1})_k = \langle A_k, x_{i,\ell+1} \rangle + (W(\widetilde{h}_{i,\ell} + h'_{i,\ell}))_k = \langle A_k, x_{i,\ell+1} \rangle + (W\widetilde{h}_{i,\ell})_k + (W h'_{i,\ell})_k$. For the second term in the above equation, we have

$$\begin{aligned}
(W\widetilde{h}_{i,\ell})_k &= \left(W U_{\ell-1} U_{\ell-1}^\top \widetilde{h}_{i,\ell} + \frac{W(I - U_{\ell-1} U_{\ell-1}^\top)\widetilde{h}_{i,\ell}}{\|(I - U_{\ell-1} U_{\ell-1}^\top)\widetilde{h}_{i,\ell}\|_2} \|(I - U_{\ell-1} U_{\ell-1}^\top)\widetilde{h}_{i,\ell}\|_2 \right)_k \\
&\overset{①}{=} \Big(W U_{\ell-1} U_{\ell-1}^\top \widetilde{h}_{i,\ell} + W v_{i,\ell} \|(I - U_{\ell-1} U_{\ell-1}^\top \widetilde{h}_{i,\ell})\|_2 \Big)_k \\
&\overset{②}{=} \Big(W_1 U_{\ell-1} U_{\ell-1}^\top \widetilde{h}_{i,\ell} + W v_{i,\ell} \|(I - U_{\ell-1} U_{\ell-1}^\top \widetilde{h}_{i,\ell})\|_2 \Big)_k \\
&\overset{③}{=} \Big(W_1 U_{\ell-1} U_{\ell-1}^\top \widetilde{h}_{i,\ell} + W_1 v_{i,\ell} \|(I - U_{\ell-1} U_{\ell-1}^\top \widetilde{h}_{i,\ell})\|_2 \Big)_k + \Big(W'_N v_{i,\ell} \cdot \|(I - U_{\ell-1} U_{\ell-1}^\top \widetilde{h}_{i,\ell})\|_2 \Big)_k \\
&= (W_1 \widetilde{h}_{i,\ell})_k + (W'_N v_{i,\ell})_k \cdot \|(I - U_{\ell-1} U_{\ell-1}^\top \widetilde{h}_{i,\ell})\|_2
\end{aligned}$$

where ① follows by definition of $v_{i,\ell}$, ② follows by definition of W_1, ③ follows by $W = W_1 + W'_N$.

2. Consider $g_{i,\ell+1}$ for the case of $i \in [n] \backslash \{i^*\}, \ell = \ell^*$ or $i \in [n], \ell > \ell^*$. We want to show the sign of $g_{i,\ell+1}$ is fixed, meaning with high probability independent of the choice of W'_N. One can easily check that:

$$g_{i,\ell+1} = \langle A_k, x_{i,\ell+1} \rangle + (W_1 \widetilde{h}_{i,\ell})_k + (W h'_{i,\ell})_k + (W'_N \widetilde{h}_{i,\ell})_k \ . \tag{E.5}$$

By Lemma D.2, we have $|\langle A_k, x_{i,\ell} \rangle + (W_1 \widetilde{h}_{i,\ell})_k| \geq \frac{\beta_+}{\sqrt{m}}$. Thus, putting this and (E.3) into (E.6), we can write

$$g_{i,\ell+1} = \Xi + (W'_N \widetilde{h}_{i,\ell})_k = \Xi + (W'_N v_{i^*,\ell^*})_k \cdot \langle v_{i^*,\ell^*}, \widetilde{h}_{i,\ell} \rangle$$

for some value Ξ with $|\Xi| \geq \frac{\beta_+}{2\sqrt{m}}$. By Lemma B.3, we have $|\langle v_{i^*,\ell^*}, \widetilde{h}_{i,\ell} \rangle| \leq \|\widetilde{h}_{i,\ell}\|_2 \leq O(L)$. By Lemma E.4d, we have $|(W'_N v_{i^*,\ell^*})_k| \leq \frac{2\theta\rho}{\sqrt{m}}$ and by our choice of θ (see Definition E.3),

$$\frac{2\theta\rho}{\sqrt{m}} \cdot O(L) \ll \frac{2\beta_+}{\sqrt{m}} \leq |\Xi|$$

This concludes that, with probability at least $1 - e^{-\Omega(\rho^2)}$ over W_1 and A, it satisfies

$$\exists s \in \{-1, 1\} \colon \quad \Pr_{W'_N} [\mathrm{sgn}(g_{i,\ell+1}) = s \mid W_1, A] \geq 1 - e^{-\Omega(\rho^2)} \ .$$

3. Consider $g_{i,\ell+1}$ for the case of $i \in [n]$ and $\ell < \ell^*$. We want to show such $g_{i,\ell+1}$ is fixed, meaning independent of W'_N. One can easily check that:

$$g_{i,\ell+1} = \langle A_k, x_{i,\ell+1} \rangle + (W_1 \widetilde{h}_{i,\ell})_k + (W h'_{i,\ell})_k + (W'_N \widetilde{h}_{i,\ell})_k \ . \tag{E.6}$$

Now, since $\ell < \ell^*$, we know $h_{i,\ell} = \widetilde{h}_{i,\ell}$ because both of which only depend on the randomness of $W U_{\ell^*-1} = W_1 U_{\ell^*-1}$ and A; this implies $h'_{i,\ell} = 0$. Also, since $W'_N = (W'_N v_{i^*,\ell^*}) v_{i^*,\ell^*}^\top$ but $\langle v_{i^*,\ell^*}, \widetilde{h}_{i,\ell} \rangle = 0$, we know that

$$g_{i,\ell+1} = \langle A_k, x_{i,\ell+1} \rangle + (W_1 \widetilde{h}_{i,\ell})_k \ .$$

This is a fixed value, independent of W'_N.

Therefore, combining the three cases above, we conclude that with at least constant probability, the sign of g_{i^*,ℓ^*+1} can possibly change, but the sign of $g_{i,\ell+1}$ will not change for any other $(i, \ell) \neq (i^*, \ell^*)$. This means,

$$\begin{aligned} \mathbb{E}[\clubsuit] = \ & \mathbb{E}\left[\left\| \widetilde{u}_{i^*,\ell^*} \cdot \widetilde{h}_{i^*,\ell^*} \cdot \mathbf{1}_{g_{i^*,\ell^*+1} \geq 0} + \textstyle\sum_{(i,\ell) \neq (i^*,\ell^*)} \widetilde{u}_{i,\ell} \cdot \widetilde{h}_{i,\ell} \cdot \mathbf{1}_{g_{i,\ell+1} \geq 0} \right\|_2 \right] \\ & \geq \Omega(\|\widetilde{u}_{i^*,\ell^*} \widetilde{h}_{i^*,\ell^*}\|_2) = \Omega\big(|\widetilde{u}_{i^*,\ell^*}| \cdot \|\widetilde{h}_{i^*,\ell^*}\|_2\big) \end{aligned}$$

Finally, for each $k \in N_5$, owing to Lemma D.7 and Lemma B.3, we have

$$|\widetilde{u}_{i^*,\ell^*}|_2 = \left| \sum_{a=\ell^*}^{L} (\widetilde{\mathsf{Back}}_{i^*,\ell^* \to a}^\top \mathsf{loss}_{i^*,a})_k \right| \geq \frac{\| \mathsf{loss}_{i^*,\ell^*} \|_2}{6\sqrt{d}nL} \quad \text{and} \quad \|\widetilde{h}_{i^*,\ell^*}\|_2 \geq \frac{1}{2} \ .$$

This means,

$$\mathbb{E}[\clubsuit] \geq \Omega\left(\frac{\| \mathsf{loss}_{i^*,\ell^*} \|_2}{\sqrt{d}nL} \right) \ . \tag{E.7}$$

Putting (E.2) and (E.7) back to (E.1), we conclude that

$$\forall k \in N_5 \colon \quad \mathbb{E}[\|\widehat{\nabla}_k f(W)\|] \geq \Omega\left(\frac{\| \mathsf{loss}_{i^*,\ell^*} \|_2}{\sqrt{d}nL} \right) \ .$$

$\square$

E.2.2 Core Lemma A

In Lemma E.6, we split W into $W = W_1 + W'_N$ for a fixed $N \subseteq [m]$. In this subsection, we split W into three parts $W = W_2 + W'_N + W'_{-N}$ following Lemma E.5c. Our purpose is to rewrite Lemma E.6 into the following variant:

Lemma E.7 (Core Lemma A)**.** *Let $\rho = nLd\log m$ and $\varrho = nLd\delta^{-1}\log(m/\varepsilon)$. Fix vectors $\{\mathsf{loss}_{i,\ell} \in \mathbb{R}^d\}_{i\in[n],\ell\in[L]\setminus\{1\}}$, integer $|N| \in [\rho^6/\beta_-, \varrho^{100}]$, and parameter $t = m/|N|$. Let $N_1, \dots, N_t$ be i.i.d. random subsets of $[m]$ with cardinality $|N_i| = |N|$. Then,*

- *with probability $\geq 1 - e^{-\Omega(\rho^2)}$ over the randomness of W_2, A, B and $N_1, \cdots N_t$, the following holds:*
  - *for every $N \in \{N_1, N_2, \cdots, N_t\}$, the following holds:*
    - *with probability at $1 - e^{-\Omega(\rho^2)}$ over the randomness of W'_{-N}, the following holds:*

$$\sum_{k\in N} \mathbb{E}_{W'_N}\left[\|\widehat{\nabla}_k f(W_2 + W'_N + W'_{-N})\|_2^2\right] \geq \Omega\left(\frac{\beta_-|N|}{\rho^2}\right) \cdot \max_{i,\ell} \|\,\mathsf{loss}_{i,\ell}\,\|_2^2.$$

Proof of Lemma E.7. We denote by $(i^*, \ell^*) = \arg\max_{i,\ell} \|\,\mathsf{loss}_{i,\ell}\,\|_2$. We first note that Lemma E.6 directly implies

- with probability $\geq 1 - e^{-\Omega(\rho^2)}$ over the randomness of W_2, W'_{-N}, A, B, N, the following holds:

$$\left|\left\{k \in N \;\middle|\; \mathbb{E}_{W'_N}\left[\|\widehat{\nabla}_k f(W_2 + W'_{-N} + W'_N)\|_2\right] \geq \Omega\Big(\frac{\|\,\mathsf{loss}_{i^*,\ell^*}\,\|_2}{nL\sqrt{d}}\Big)\right\}\right| \geq \frac{\beta_-}{100L}|N|$$

or putting it in another way, using our choice of q,

- with probability $\geq 1 - e^{-\Omega(\rho^2)}$ over the randomness of W_2, W'_{-N}, A, B, N, the following holds ($q = \frac{\beta_-|N|}{c\rho^2}$ for some sufficiently large constant):

$$\sum_{k\in N} \mathbb{E}_{W'_N}\left[\|\widehat{\nabla}_k f(W_2 + W'_{-N} + W_N)\|_2^2\right] \geq q \cdot \|\,\mathsf{loss}_{i^*,\ell^*}\,\|_2^2.$$

Applying simple tricks to switch the ordering of randomness (see Fact A.2), we have

- with probability $\geq 1 - e^{-\Omega(\rho^2)}$ over the randomness of W_2, A, B, the following holds:
  - with probability $\geq 1 - e^{-\Omega(\rho^2)}$ over the randomness of N, the following holds:
    - with probability $\geq 1 - e^{-\Omega(\rho^2)}$ over randomness of W'_{-N}, the following holds:

$$\sum_{k\in N} \mathbb{E}_{W'_N}[\|\widehat{\nabla}_k f(W_2 + W'_N + W'_{-N})\|_2^2] \geq q \cdot \|\,\mathsf{loss}_{i^*,\ell^*}\,\|_2^2.$$

Repeating the choice of N for t times, and using union bound, we have

- with probability $\geq 1 - e^{-\Omega(\rho^2)}$ over the randomness of W_2, A, B, the following holds:
  - with probability $\geq 1 - e^{-\Omega(\rho^2)}$ over the randomness of $N_1, N_2, \cdots, N_t$, the following holds:
    - for all $N \in \{N_1, N_2, \cdots, N_t\}$, the following holds:
      - with probability $\geq 1 - e^{-\Omega(\rho^2)}$ over randomness of W'_{-N}, the following holds:

$$\sum_{k\in N} \mathbb{E}_{W'_N}[\|\widehat{\nabla}_k f(W_2 + W'_N + W'_{-N})\|_2^2] \geq q \cdot \|\,\mathsf{loss}_{i^*,\ell^*}\,\|_2^2.$$

Combining the above statement with Fact A.1 (to merge randomness), we conclude the proof. $\square$

E.3 Gradient Stability

The goal of this subsection is to prove Lemma E.8 and then translate it into our Core Lemma B (see Lemma E.10).

Lemma E.8 (c.f. (7.7))**.** *Fix parameter $\rho = nLd\log m$, vectors $\{\mathsf{loss}_{i,\ell} \in \mathbb{R}^d\}_{i\in[n],\ell\in[L]\setminus\{1\}}$, $(i^*,\ell^*) = \arg\min_{i,\ell}\ \{\|\mathsf{loss}_{i,\ell}\|\}$, index $j \in [m]$. Let $N \subseteq [m]$ be a random subset containing j of fixed cardinality $|N|$, let W, A, B be at random initialization, and W''_j be defined as Definition E.5a so that $W + W''_j$ is also at random initialization. Then, with probability at least $1 - e^{-\Omega(\rho^2)}$*

$$\sum_{k\in[N]} \left| \|\widehat{\nabla}_k(W)\|_2^2 - \|\widehat{\nabla}_k(W + W''_j)\|_2^2 \right| \leq O\left(\rho^8 + \frac{\rho^{11}\theta^{1/3}}{m^{1/6}}|N|\right) \ . \tag{E.8}$$

E.3.1 Proof of Lemma E.8

Throughout this proof we denote by $\widetilde{W} = W$ to emphasize that W is at random initialization. We first show some properties before proving Lemma E.8.

Claim E.9. *Fix parameter $\rho = nLd\log m$, vectors $\{\mathsf{loss}_{i,\ell} \in \mathbb{R}^d\}_{i\in[n],\ell\in[L]\setminus\{1\}}$, $(i^*,\ell^*) = \arg\min_{i,\ell}$ $\{\|\mathsf{loss}_{i,\ell}\|\}$, index $j \in [m]$. Let $\widetilde{W}, A, B$ be at random initialization, and W''_j be defined as Lemma E.5a so that $W = \widetilde{W} + W''_j$ is also at random initialization. Then, then with probability $\geq 1 - e^{-\Omega(\rho^2)}$ over $\widetilde{W}, A, B, W''_j$,*

- $\left|\left\{k \in [m] \ \middle|\ \left|\|\widehat{\nabla}_k f(\widetilde{W} + W''_j)\|_2^2 - \|\widehat{\nabla}_k f(\widetilde{W})\|_2^2\right| \leq O\left(\frac{\rho^{11}\theta^{1/3}}{m^{1/6}}\right)\right\}\right| \geq \left(1 - \frac{\rho^5\theta^{2/3}}{m^{1/3}}\right) m$.
- $\left|\|\widehat{\nabla}_k f(\widetilde{W} + W''_j)\|_2^2 - \|\widehat{\nabla}_k f(\widetilde{W})\|_2^2\right| \leq O(\rho^6)$ *for every $k \in [m]$.*

Proof of Claim E.9. As before, let $\widetilde{D}_{i,\ell}$ denote the diagonal matrix consisting of the indicator function $\mathbf{1}_{(\widetilde{g}_{i,\ell+1})_k\geq 0}$ for its k-th diagonal entry. Let $D_{i,\ell}$ be the same thing but with respect to weight matrix $\widetilde{W} + W''_j$. Let $D''_{i,\ell} = D_{i,\ell} - \widetilde{D}_{i,\ell}$ be the diagonal matrix of sign change.

- By Lemma C.2 (with $\tau_0 = \frac{3\theta\rho}{2}$ owing to Lemma E.4d), we have $\|D''_\ell\|_0 \leq \rho^4\theta^{2/3}m^{2/3}$. Therefore, letting $J \stackrel{\text{def}}{=} [m]\setminus \cup_{i,\ell}\operatorname{supp}(D_{i,\ell})$ be the set of all indices of sign changes, we have

$$|J| \geq m - nL\rho^4\theta^{2/3}m^{2/3} \geq \left(1 - \frac{\rho^5\theta^{2/3}}{m^{1/3}}\right)\cdot m.$$

  Next, for every index $k \in J$, we have

$$\begin{aligned}
&\|\widehat{\nabla}_k f(\widetilde{W} + W''_j) - \widehat{\nabla}_k f(\widetilde{W})\|_2 \\
&\overset{①}{=} \left\| \sum_{i=1}^n \sum_{\ell=1}^L (\widetilde{u}_{i,\ell})_k \cdot \widetilde{h}_{i,\ell} \cdot \mathbf{1}_{(\widetilde{g}_{i,\ell+1})_k\geq 0} - \sum_{i=1}^n \sum_{\ell=1}^L (u_{i,\ell})_k \cdot h_{i,\ell} \cdot \mathbf{1}_{(g_{i,\ell+1})_k\geq 0} \right\|_2 \\
&\overset{②}{=} \left\| \sum_{i=1}^n \sum_{\ell=1}^L (\widetilde{u}_{i,\ell})_k \cdot \widetilde{h}_{i,\ell} \cdot \mathbf{1}_{(g_{i,\ell+1})_k\geq 0} - \sum_{i=1}^n \sum_{\ell=1}^L (u_{i,\ell})_k \cdot h_{i,\ell} \cdot \mathbf{1}_{(g_{i,\ell+1})_k\geq 0} \right\|_2 \\
&\overset{③}{\leq} \sum_{i=1}^n \sum_{\ell=1}^L \|((\widetilde{u}_{i,\ell})_k \cdot \widetilde{h}_{i,\ell} - (u_{i,\ell})_k \cdot h_{i,\ell}) \cdot \mathbf{1}_{(g_{i,\ell+1})_k\geq 0}\|_2 \ .
\end{aligned}$$

where ① follows by definition, ② follows by our choice of $k \in J$, and ③ follows by triangle inequality. For each i, ℓ, using the same proof as (E.2) (so invoking Lemma C.11, Corollary B.19, Lemma B.3 and Lemma C.2a), we have with probability at least $1 - e^{-\Omega(\rho^2)}$:

$$\|(\widetilde{u}_{i,\ell})_k \cdot \widetilde{h}_{i,\ell} - (u_{i,\ell})_k \cdot h_{i,\ell}\|_2 \le O\big(\frac{\rho^6 \tau_0^{1/3}}{m^{1/6}}\big) \cdot \| \mathsf{loss}_{i^*,a^*} \| \ . \tag{E.9}$$

which implies

$$\|\widehat{\nabla}_k f(\widetilde{W} + W''_j) - \widehat{\nabla}_k f(\widetilde{W})\|_2 \le O\big(\frac{\rho^6 \tau_0^{1/3}}{m^{1/6}}\big) \cdot \| \mathsf{loss}_{i^*,a^*} \| \ .$$

Combining this with $\|\widehat{\nabla}_k f(\widetilde{W})\|_2 \le O(\rho^4)$ from Lemma E.2, we finish the proof of the first item.

- For every index $k \in [m]$, we have

$$\begin{aligned}
&\|\widehat{\nabla}_k f(\widetilde{W} + W''_j) - \widehat{\nabla}_k f(\widetilde{W})\|_2 \\
&= \left\| \sum_{i=1}^n \sum_{\ell=1}^L (\widetilde{u}_{i,\ell})_k \cdot \widetilde{h}_{i,\ell} \cdot \mathbf{1}_{(\widetilde{g}_{i,\ell+1})_k \ge 0} - \sum_{i=1}^n \sum_{\ell=1}^L (u_{i,\ell})_k \cdot h_{i,\ell} \cdot \mathbf{1}_{(g_{i,\ell+1})_k \ge 0} \right\|_2 \\
&\le \sum_{i=1}^n \sum_{\ell=1}^L \|(\widetilde{u}_{i,\ell})_k \cdot \widetilde{h}_{i,\ell}\|_2 + \|(u_{i,\ell})_k \cdot h_{i,\ell})\|_2 \ .
\end{aligned}$$

Using (E.9) and $\|(\widetilde{u}_{i,\ell})_k \cdot \widetilde{h}_{i,\ell}\|_2 \le O(\rho^2 L)$ (see the proof of Lemma E.2), we immediately have the desired bound.

□

Proof of Lemma E.8. Letting $J \stackrel{\text{def}}{=} \left\{ k \in [m] \;\middle|\; \left| \|\widehat{\nabla}_k f(\widetilde{W} + W''_j)\|_2^2 - \|\widehat{\nabla}_k f(\widetilde{W})\|_2^2 \right| \le O\left(\frac{\rho^{11}\theta^{1/3}}{m^{1/6}}\right) \right\}$, we have $|J| \ge \left(1 - \frac{\rho^5 \theta^{2/3}}{m^{1/3}}\right) m$ according to Claim E.9.

Now, for the indices in N that are randomly sampled from $[m]$, if $|N \cap J| \ge |N| - S$ for a parameter $S = \rho^2$, then we have

$$\sum_{k \in N} \left| \|\widehat{\nabla}_k f(\widetilde{W})\|_2^2 - \|\widehat{\nabla}_k f(\widetilde{W} + W''_j)\|_2^2 \right| \le |S| \cdot O(\rho^6) + |N| \cdot O\left(\frac{\rho^{11}\theta^{1/3}}{m^{1/6}}\right) \le O\left(\rho^8 + \frac{\rho^{11}\theta^{1/3}}{m^{1/6}} |N|\right) \ .$$

Otherwise, if $|N \cap J| \le |N| - S$, this means at least S indices that are chosen from N are outside $J \subseteq [m]$. This happens with probability at most $\left(\frac{\rho^5 \theta^{2/3}}{m^{1/3}}\right)^S \le e^{-\Omega(\rho^2)}$. □

E.3.2 Core Lemma B

In this subsection, we split W into three parts $W = W_2 + W'_N + W'_{-N}$ following def:random-decomp:N-N. Our purpose is to rewrite Lemma E.8 into the following variant:

Lemma E.10 (Core Lemma B). *Let $\rho = nLd\log m$ and $\varrho = nLd\delta^{-1}\log(m/\varepsilon)$. Fix parameter θ from Definition E.3, vectors $\{\mathsf{loss}_{i,\ell} \in \mathbb{R}^d\}_{i\in[n],\ell\in[L]\setminus\{1\}}$, integer $|N| \in [1, \varrho^{100}]$, parameter $t = m/|N|$. Let $N_1, \ldots, N_t$ be i.i.d. random subsets of $[m]$ with cardinality $|N_i| = |N|$. Then, with probability $\geq 1 - e^{-\Omega(\rho^2)}$ over the randomness of W_2, A, B and $N_1, \cdots N_t$, the following holds:*

- *for every $N \in \{N_1, N_2, \cdots, N_t\}$, the following holds:*
  - *with probability at $1 - e^{-\Omega(\rho^2)}$ over the randomness of W'_{-N}, the following holds:*
    - *with probability at least $1 - e^{-\Omega(\rho^2)}$ over the randomness of W'_N, we have :*
      - *for all $j \in N$, and for all $W''_j = u_j v^\top_{i^*,\ell^*}$ satisfying*

$$\|u_j\|_0 = 1, \quad \|u_j\|_\infty \leq \tfrac{3\theta\rho}{\sqrt{m}}, \quad v_{i^*,\ell^*} = \frac{(I - U_{\ell^*-1}U^\top_{\ell-1})h_{i^*,\ell}h_{i^*,\ell^*}}{\|(I - U_{\ell^*-1}U^\top_{\ell^*-1})h_{i^*,\ell^*}\|_2}\ , \tag{E.10}$$

we have

$$\textstyle\sum_{k\in N}\left|\|\widehat{\nabla}_k f(W_2 + W'_N + W'_{-N})\|_2^2 - \|\widehat{\nabla}_k f(W_2 + W'_N + W'_{-N} + W''_j)\|_2^2\right|$$
$$\leq O(\rho^8)\cdot \max_{i,\ell}\|\mathsf{loss}_{i,\ell}\|_2^2\ .$$

Proof of Lemma E.10. Without loss of generality we assume $\max_{i,\ell}\|\mathsf{loss}_{i,\ell}\|_2^2 = 1$ in the proof. We now first rewrite Lemma E.8 as follows (recalling m is sufficiently large so we only need to keep the $O(\rho^8)$ term):

- with probability $\geq 1 - e^{-\Omega(\rho^2)}$ over random $N \subset [m]$, random $j \in N$, random W, A, B, and random W'_j, the following holds:

$$\sum_{k\in[N]}\left|\|\widehat{\nabla}_k(W)\|_2^2 - \|\widehat{\nabla}_k(W + W''_j)\|_2^2\right| \leq O\left(\rho^8 + \frac{\rho^{11}\theta^{1/3}}{m^{1/6}}|N|\right) \leq O(\rho^8)\ . \tag{E.11}$$

We can split the randomness (see Fact A.1) and derive that

- With probability $\geq 1 - e^{-\Omega(\rho^2)}$ over $N \subseteq [m]$, W, A, B, the following holds:
  - With probability $\geq 1 - e^{-\Omega(\rho^2)}$ over $j \in N$ and W''_j, Eq. (E.11) holds.

Applying standard ε-net argument, we derive that

- With probability $\geq 1 - e^{-\Omega(\rho^2)}$ over $N \subseteq [m]$, W, A, B, the following holds:
  - for all $j \in N$, and for all $W''_j = u_j v^\top_{i^*,\ell^*}$ satisfying (E.10), we have Eq. (E.11) holds.

Finally, letting $W = W_2 + W'_N + W'_{-N}$, we have

- With probability $\geq 1 - e^{-\Omega(\rho^2)}$ over the randomness of $N \subseteq [m], W_2, W_N, W'_{-N}, A, B$, the following holds:
  - for all $j \in N$, and for all $W''_j = u_j v^\top_{i^*,\ell^*}$ satisfying (E.10), we have

$$\sum_{k\in N}\left|\|\widehat{\nabla}_k f(W_2 + W'_N + W'_{-N})\|_2^2 - \|\widehat{\nabla}_k f(W_2 + W'_N + W_{-N'} + W''_j)\|_2^2\right| \leq O\left(\rho^8\right)\ .$$

Finally, splitting the randomness (using Fact A.1), and applying a union bound over multiple samples $N_1, \ldots, N_t$, we finish the proof. □

E.4 Main Theorem: Proof of Theorem 5

Proof of Theorem 5. Without loss of generality, we assume that $\max_{i,\ell}\|\mathsf{loss}_{i,\ell}\|_2^2 = 1$.

Combining Core Lemma A and B (i.e., Lemma E.7 and Lemma E.10), we know that if $|N|$ is appropriately chosen, with probability $\geq 1 - e^{-\Omega(\rho^2)}$ over the randomness of W_2, A, B and $N_1, \cdots N_t$, the following holds:

- for every $N \in \{N_1, N_2, \cdots, N_t\}$, the following holds:

– with probability at $1 - e^{-\Omega(\rho^2)}$ over the randomness of W'_{-N}, the following boxed statement holds:

the boxed statement

- (core A)

$$\mathbb{E}_{W'_N}\left[\sum_{k\in N}\|\widehat{\nabla}_k f(W_2 + W'_N + W'_{-N})\|_2^2\right] \geq q \stackrel{\text{def}}{=} \Omega\left(\frac{\beta_-|N|}{\rho^2}\right)$$

- (core B) with probability at least $1 - e^{-\Omega(\rho^2)}$ over the randomness of W'_N, the following holds:
  – for all $j \in N$, and for all $W''_j = u_j v_{i^*,\ell^*}^\top$ satisfying

$$\|u_j\|_0 = 1, \quad \|u_j\|_\infty \leq \frac{3\theta\rho}{\sqrt{m}}, \quad v_{i^*,\ell^*} = \frac{(I - U_{\ell^*-1}U_{\ell-1}^\top)h_{i^*,\ell}h_{i^*,\ell^*}}{\|(I - U_{\ell^*-1}U_{\ell^*-1}^\top)h_{i^*,\ell^*}\|_2} \ ,$$

  we have

$$\sum_{k\in N}\left|\|\widehat{\nabla}_k f(W_2 + W'_N + W'_{-N})\|_2^2 - \|\widehat{\nabla}_k f(W_2 + W'_N + W'_{-N} + W''_j)\|_2^2\right| \leq p \stackrel{\text{def}}{=} O(\rho^8) \ .$$

We now wish to apply our extended McDiarmid's inequality (see Lemma A.19) to the boxed statement. For this goal, recalling that $W'_N = u_N v_{i^*,\ell^*}^\top$ for a vector $u_N \in \mathbb{R}^m$ that is only supported on indices in N. That is, $(u_N)_k = 0$ for all $k \notin N$. Therefore, we can define function $\mathcal{F}(u_N) \stackrel{\text{def}}{=} \sum_{k\in N}\|\widehat{\nabla}_k f(W_2 + W'_{-N} + W'_N)\|^2$ where $W'_N = u_N v_{i^*,\ell^*}^\top$. We emphasize here that, inside the boxed statement, W_2, W'_{-N} and v_{i^*,ℓ^*} are already fixed, and u_N is the only source of randomness.

Below, we condition on the high probability event (see Lemma E.4d) that $\|u_N\|_\infty \leq \frac{3\theta\rho}{2\sqrt{m}}$. The boxed statement tells us:

- $\mathbb{E}_{u_N}[\mathcal{F}(u_N)] \geq q$, and
- With probability $\geq 1 - e^{-\Omega(\rho^2)}$ over u_N, it satisfies

$$\forall j \in [N], \forall u''_j \colon \ \left|\mathcal{F}(u_{-j}, u_j) - \mathcal{F}(u_{-j}, u''_j)\right| \leq p \ .$$

At the same time, we have $0 \leq \mathcal{F}(u_N) \leq O(N\rho^8)$ owing to Lemma E.2. Therefore, scaling down $\mathcal{F}(u_N)$ by $\Theta(N\rho^8)$ (to make sure the function value stays in $[0,1]$), we can applying extended McDiarmid's inequality (see Lemma A.19), we have

$$\Pr_{u_N}\left[\mathcal{F}(u_N) \geq \frac{q}{2}\right] \geq 1 - N^2 \cdot e^{-\Omega(\rho^2)} - \exp\left(-\Omega\left(\frac{(q/\rho^8)^2}{N(p/\rho^8)^2 + Ne^{-\Omega(\rho^2)}}\right)\right) \ . \tag{E.12}$$

As long as $N \geq \frac{\rho^{22}}{\beta_-^2}$, we have that the above probability is at least $1 - e^{-\Omega(\rho^2)}$.

Finally, we choose

$$\boxed{N = \frac{\rho^{22}}{\beta_-^2}} \tag{E.13}$$

in order to satisfy Lemma E.7. We replace the boxed statement with (E.12). This tells us

- with probability $\geq 1 - e^{-\Omega(\rho^2)}$ over the randomness of W_2, A, B and $N_1, \cdots N_t$, the following holds:
  – for every $N \in \{N_1, N_2, \cdots, N_t\}$, the following holds:
    * with probability at $1 - e^{-\Omega(\rho^2)}$ over the randomness of W'_{-N} and W'_N, the following holds:

$$\sum_{k\in N}\|\widehat{\nabla}_k f(W_2 + W'_N + W'_{-N})\|_2^2 \geq \Omega\left(\frac{\beta_-|N|}{\rho^2}\right)$$

After rearranging randomness, and using $W = W_2 + W'_N + W'_{-N}$, we have

- with probability $\geq 1 - e^{-\Omega(\rho^2)}$ over the randomness of W, A, B, the following holds:
  - with probability $\geq 1 - e^{-\Omega(\rho^2)}$ over the randomness of $N_1, \cdots N_t$, the following holds:
    - for every $N \in \{N_1, N_2, \cdots, N_t\}$, the following holds:

$$\sum_{k \in N} \|\widehat{\nabla}_k f(W)\|_2^2 \geq \Omega\left(\frac{\beta_- |N|}{\rho^2}\right)$$

Finally, since $N_1, \ldots, N_t$ are t random subsets of $[m]$, we know that with probability at least $1 - e^{-\Omega(\rho^2)}$, for each index $k \in [m]$, it is covered by at most ρ^2 random subsets. Therefore,

$$\sum_{k \in N} \|\widehat{\nabla}_k f(W)\|_2^2 \geq \frac{1}{\rho^2} \times t \times \Omega\left(\frac{\beta_- |N|}{\rho^2}\right) = \Omega\left(\frac{\delta}{\rho^{14}}\right) \times m \ .$$

□

F Gradient Bound After Perturbation (Theorem 3)

We use the same fake gradient notion (see Definition E.1) and first derive the following result based on Lemma E.2 and Theorem 5.

Lemma F.1. *Let $\rho = nLd\log m$ and $\varrho = nLd\delta^{-1}\log(m/\varepsilon)$. Given fixed vectors $\{\mathsf{loss}_{i,\ell} \in \mathbb{R}^d\}_{i\in[n],\ell\in[L]\setminus\{1\}}$, with probability at least $1-e^{-\Omega(\rho^2)}$ over $\widetilde{W}, A, B$, it satisfies for* all $W' \in \mathbb{R}^{m\times m}$ *with $\|W'\|_2 \le \frac{\tau_0}{\sqrt{m}}$ with $\tau_0 \le \varrho^{100}$,*

$$\|\widehat{\nabla} f(\widetilde{W}+W')\|_F^2 \ge \Omega\left(\frac{\delta}{\rho^{14}}\right) \times m \times \max_{i,\ell}\{\|\mathsf{loss}_{i,\ell}\|^2\}$$
$$\|\widehat{\nabla} f(\widetilde{W}+W')\|_F^2 \le O(\rho^{12}m) \times \max_{i,\ell}\{\|\mathsf{loss}_{i,\ell}\|^2\}$$
$$\|\widehat{\nabla} f_i(\widetilde{W}+W')\|_F^2 \le \frac{1}{n^2}O(\rho^{12}m) \times \max_{i,\ell}\{\|\mathsf{loss}_{i,\ell}\|^2\} \ .$$

Proof of Lemma F.1. Like before, we denote by $\widetilde{D}_{i,\ell}, \widetilde{u}_{i,\ell}, \widetilde{h}_{i,\ell}$ the corresponding matrices or vectors when the weight matrix is $\widetilde{W}$, and by $D_{i,\ell}, u_{i,\ell}, h_{i,\ell}$ the corresponding terms when when the weight matrix is $W = \widetilde{W} + W'$.

Lower bound on $\|\widehat{\nabla} f(\widetilde{W}+W')\|_F$. Recall

$$\widehat{\nabla} f(\widetilde{W}+W') = \sum_{i=1}^n \sum_{\ell=1}^L D_{i,\ell+1}\left(\sum_{a=\ell}^L \mathsf{Back}_{i,\ell\to a}^\top \mathsf{loss}_{i,a}\right) h_{i,\ell}^\top = \sum_{i=1}^n\sum_{\ell=1}^L D_{i,\ell+1}u_{i,\ell}h_{i,\ell}^\top$$

In Theorem 5 of the previous section, we already have a lower bound on $\|\widehat{\nabla} f(\widetilde{W})\|_F$. We just need to upper bound $\|\widehat{\nabla} f(\widetilde{W}+W') - \widehat{\nabla} f(\widetilde{W})\|_F$.

$$\begin{aligned}
\|\widehat{\nabla} f(\widetilde{W}+W') - \widehat{\nabla} f(\widetilde{W})\|_F &\overset{①}{=} \left\|\sum_{i=1}^n\sum_{\ell=1}^L D_{i,\ell+1}u_{i,\ell}h_{i,\ell}^\top - \widetilde{D}_{i,\ell+1}\widetilde{u}_{i,\ell}\widetilde{h}_{i,\ell}^\top\right\|_F \\
&\overset{②}{\le} \sum_{i=1}^n\sum_{\ell=1}^L \left\|D_{i,\ell+1}u_{i,\ell}h_{i,\ell}^\top - \widetilde{D}_{i,\ell+1}\widetilde{u}_{i,\ell}\widetilde{h}_{i,\ell}^\top\right\|_F \\
&\overset{③}{\le} \sum_{i=1}^n\sum_{\ell=1}^L \|D'_{i,\ell+1}\widehat{u}_{i,\ell}\widehat{h}_{i,\ell}^\top\|_F + \|\widetilde{D}_{i,\ell+1}u'_{i,\ell}\widehat{h}_{i,\ell}^\top\|_F + \|\widetilde{D}_{i,\ell+1}\widetilde{u}_{i,\ell}{h'_{i,\ell}}^\top\|_F + o(m^{1/3}).
\end{aligned}$$

where ① follows by definition and ② and ③ follow by the triangle inequality. Note that in inequality ③ we have hidden four more higher order terms in $o(m^{1/3})$. We ignore the details for how to bound them, for the ease of presentation. We bound the three terms separately:

- Using Corollary B.18 (with $s^2 = O(L^{10/3}\tau_0^{2/3})$ from Lemma C.2b) and Lemma B.3 we have
$$\begin{aligned}
\|D'_{i,\ell+1}\widehat{u}_{i,\ell}\widehat{h}_{i,\ell}^\top\|_F &\le \|D'_{i,\ell+1}\widehat{u}_{i,\ell}\|_2 \cdot \|\widehat{h}_{i,\ell}\|_2 \\
&\le O(L\cdot L^4\tau_0^{1/3}m^{1/3}\log m \cdot \|\mathsf{loss}_{i^*,\ell^*}\|_2)\cdot L
\end{aligned}$$
- Using $\|\widetilde{D}_{i,\ell+1}\|_2 \le 1$, Lemma C.9 and Lemma B.3 we have
$$\begin{aligned}
\|\widetilde{D}_{i,\ell+1}u'_{i,\ell}\widehat{h}_{i,\ell}^\top\|_F &\le \|\widetilde{D}_{i,\ell+1}\|\cdot\|u'_{i,\ell}\|_2\cdot\|\widehat{h}_{i,\ell}\|_2 \\
&\le O(L\cdot L^6\tau_0^{1/3}m^{1/3}\log m\cdot\|\mathsf{loss}_{i^*,\ell^*}\|_2)\cdot L
\end{aligned}$$

- Using $\|\widetilde{D}_{i,\ell+1}\|_2 \leq 1$, $\|\widetilde{u}_{i,\ell}\|_2 \leq O(L^4)\sqrt{m}\|\operatorname{\mathsf{loss}}_{i^*,\ell^*}\|_2$ (implied by Lemma B.11) and Lemma C.2a, we have

$$\begin{aligned}\|\widetilde{D}_{i,\ell+1}\widetilde{u}_{i,\ell}{h'_{i,\ell}}^\top\|_F &\leq \|\widetilde{D}_{i,\ell+1}\| \cdot \|\widetilde{u}_{i,\ell}\|_2 \cdot \|h'_{i,\ell}\|_2 \\ &\leq O(L^4\sqrt{m} \cdot \|\operatorname{\mathsf{loss}}_{i^*,\ell^*}\|_2) \cdot (L^6\tau_0^{1/2}\frac{1}{\sqrt{m}})\end{aligned}$$

Putting it all together, we have

$$\|\widehat{\nabla} f(\widetilde{W}+W') - \widehat{\nabla} f(\widetilde{W})\|_F \leq O(\rho^8\tau_0^{1/3}m^{1/3}) \cdot \|\operatorname{\mathsf{loss}}_{i^*,\ell^*}\|_F \ . \tag{F.1}$$

Finally, using $(a-b)^2 \geq \frac{1}{2}a^2 - b^2$, we have

$$\begin{aligned}\|\widehat{\nabla} f(\widetilde{W}+W')\|_F^2 &\geq (\|\widehat{\nabla} f(\widetilde{W})\|_F - \|\widehat{\nabla} f(\widetilde{W}+W') - \widehat{\nabla} f(\widetilde{W})\|_F)^2 \\ &\geq \frac{1}{2}\|\widehat{\nabla} f(\widetilde{W})\|_F^2 - \|\widehat{\nabla} f(\widetilde{W}+W') - \widehat{\nabla} f(\widetilde{W})\|_F^2 \\ &\geq \Omega\left(\frac{\delta}{\rho^{14}}\right) \times m \times \|\operatorname{\mathsf{loss}}_{i^*,\ell^*}\|^2 \ .\end{aligned}$$

where the last step follows by (F.1) (with our sufficiently large choice of m) and Theorem 5.

Upper bound on $\|\widehat{\nabla} f(\widetilde{W}+W')\|_F$. Using $(a+b)^2 \leq 2a^2 + 2b^2$, we have

$$\begin{aligned}\|\widehat{\nabla} f(\widetilde{W}+W')\|_F^2 &\leq 2\|\widehat{\nabla} f(\widetilde{W})\|_F^2 + 2\|\widehat{\nabla} f(\widetilde{W}+W') - \widehat{\nabla} f(\widetilde{W})\|_F^2 \\ &\overset{①}{\leq} O(\rho^{12}m) \cdot \|\operatorname{\mathsf{loss}}_{i^*,\ell^*}\|_F^2 + 2\|\widehat{\nabla} f(\widetilde{W}+W') - \widehat{\nabla} f(\widetilde{W})\|_F^2 \overset{②}{\leq} O(\rho^{12}m) \cdot \|\operatorname{\mathsf{loss}}_{i^*,\ell^*}\|_F^2\end{aligned}$$

where ① follows by Lemma E.2 and ② follows by (F.1).

Upper bound on $\|\widehat{\nabla} f_i(\widetilde{W}+W')\|_F$. This is completely analogous so we do not replicate the proofs here. □

By applying an ε-net argument over all possible $\operatorname{\mathsf{loss}}_{i,\ell} \in \mathbb{R}^d$, we can modify Lemma F.1 from "for fixed loss" to "for all loss." This allows us to plug in the true loss $\operatorname{\mathsf{loss}}_{i,\ell} = Bh_{i,\ell} - y^*_{i,\ell}$ and derive that:

Theorem 3 (restated)**.** *Let $\rho = nLd\log m$ and $\varrho = nLd\delta^{-1}\log(m/\varepsilon)$. With probability at least $1 - e^{-\Omega(\rho^2)}$ over W, A, B, it satisfies for* all *$W' \in \mathbb{R}^{m\times m}$ with $\|W'\|_2 \leq \frac{\varrho^{100}}{\sqrt{m}}$,*

$$\begin{aligned}\|\nabla f(\widetilde{W}+W')\|_F^2 &\geq \Omega\left(\frac{\delta}{\rho^{14}}\right) \times m \times \max_{i,\ell}\{\|\operatorname{\mathsf{loss}}_{i,\ell}\|^2\} \\ \|\nabla f(\widetilde{W}+W')\|_F^2 &\leq O(\rho^{12}m) \times \max_{i,\ell}\{\|\operatorname{\mathsf{loss}}_{i,\ell}\|^2\} \\ \|\nabla f_i(\widetilde{W}+W')\|_F^2 &\leq \frac{1}{n^2}O(\rho^{12}m) \times \max_{i,\ell}\{\|\operatorname{\mathsf{loss}}_{i,\ell}\|^2\} \ .\end{aligned}$$

G Objective Smoothness (Theorem 4)

The purpose of this section is to prove

Theorem 4 (objective smoothness, restated). *Let $\rho = nLd\log m$, $\varrho = nLd\delta^{-1}\log(m/\varepsilon)$, $\tau_0 \in [\varrho^{-100}, \varrho^{100}]$, and $\widetilde{W}, A, B$ be at random initialization. With probability at least $1 - e^{-\Omega(\rho^2)}$ over the randomness of $\widetilde{W}, A, B$, we have for every $\breve{W} \in \mathbb{R}^{m\times m}$ with $\|\breve{W} - \widetilde{W}\| \le \frac{\tau_0}{\sqrt{m}}$, and for every $W' \in \mathbb{R}^{m\times m}$ with $\|W'\| \le \frac{\tau_0}{\sqrt{m}}$, and letting $(i^*, \ell^*) = \arg\max_{i,\ell}\{\|\breve{\mathsf{loss}}_{i,\ell}\|_2\}$,*

$$f(\breve{W} + W') \le f(\breve{W}) + \langle \nabla f(\breve{W}), W'\rangle + O(\rho^{11}\tau_0^{1/3}m^{1/3}) \cdot \|\breve{\mathsf{loss}}_{i^*,\ell^*}\|_2 \cdot \|W'\|_2 + O(L^{18}nm)\|W'\|_2^2$$

We introduce the following notations before we go to proofs.

Definition G.1. *For $i \in [n]$ and $\ell \in [L]$:*

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

$$\breve{\mathsf{loss}}_{i,\ell} = B\breve{h}_{i,\ell} - y^*_{i,\ell}$$

Define diagonal matrices $\widetilde{D}_{i,\ell}$ and $\breve{D}_{i,\ell}$ respectively by letting

$$(\widetilde{D}_{i,\ell})_{k,k} = \mathbf{1}_{(\widetilde{g}_{i,\ell})_k \ge 0} \text{ and } (\breve{D}_{i,\ell})_{k,k} = \mathbf{1}_{(\breve{g}_{i,\ell})_k \ge 0}.$$

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

Claim G.2 (c.f. (8.1)). *There exist diagonal matrices $D''_{i,\ell} \in \mathbb{R}^{m\times m}$ with entries in $[-1,1]$ such that,*

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

Furthermore, we have $\|h_{i,\ell} - \breve{h}_{i,\ell}\| \le O(L^9)\|W'\|_2$ and $\|D''_{i,\ell}\|_0 \le O(L^{10/3}\tau_0^{2/3}m^{2/3})$.

Proof of Claim G.2. We ignore the subscript in i for cleanness, and calculate that

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

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

- We have $\|h_\ell - \breve{h}_\ell\| \le O(L^9)\|W'\|_2$. This is because we have
  - $\big\|(\breve{D}_\ell + D''_\ell)\breve{W}\cdots\breve{W}(\breve{D}_{a+1} + D''_{a+1})\big\|_2 \le O(L^7)$ by Lemma C.7;
  - $\|h_a\| \le O(L)$ (by $\|\widetilde{h}_a\| \le O(L)$ from Lemma B.3 and $\|\widetilde{h}_a - h_a\| \le o(1)$ from Lemma C.2a); and

- $\|W'h_a\| \le \|W'\|_2 \|h_a\|$.

- We have $\|D''_\ell\|_0 \le O(L^{10/3}\tau_0^{2/3} m^{2/3})$.

  This is because, $(D''_\ell)_{k,k}$ is non-zero only at the coordinates $k \in [m]$ where the signs of $\breve{g}_\ell$ and g_ℓ are opposite (by Proposition G.3). Such a coordinate k must satisfy either $(\widetilde{D}_\ell)_{k,k} \neq (\breve{D}_\ell)_{k,k}$ or $(\widetilde{D}_\ell)_{k,k} \neq (D_\ell)_{k,k}$, and therefore by Lemma C.2 — with probability $\ge 1 - e^{-\Omega(L^6 \tau_0^{4/3} m^{1/3})}$ — there are at most $O(L^{10/3}\tau_0^{2/3} m^{2/3})$ such coordinates k.

$\square$

Proof of Theorem 4. First of all, since

$$\frac{1}{2}\|Bh_{i,\ell} - y^*_{i,\ell}\|^2 = \frac{1}{2}\|\breve{\mathsf{loss}}_{i,\ell} + B(h_{i,\ell} - \breve{h}_{i,\ell})\|^2 = \frac{1}{2}\|\breve{\mathsf{loss}}_{i,\ell}\|^2 + \breve{\mathsf{loss}}_{i,\ell}^\top B(h_{i,\ell} - \breve{h}_{i,\ell}) + \frac{1}{2}\|B(h_{i,\ell} - \breve{h}_{i,\ell})\|^2 \tag{G.2}$$

we can write

$$\begin{aligned}
& f(\breve{W} + W') - f(W) - \langle \nabla f(W), W' \rangle \\
& \overset{①}{=} -\langle \nabla f(\breve{W}), W' \rangle + \frac{1}{2}\sum_{i=1}^n \sum_{\ell=2}^L \|Bh_{i,\ell} - y^*_{i,\ell}\|^2 - \|B\breve{h}_{i,\ell} - y^*_{i,\ell}\|^2 \\
& \overset{②}{=} -\langle \nabla f(\breve{W}), W' \rangle + \sum_{i=1}^n \sum_{\ell=2}^L \breve{\mathsf{loss}}_{i,\ell}^\top B(h_{i,\ell} - \breve{h}_{i,\ell}) + \frac{1}{2}\|B(h_{i,\ell} - \breve{h}_{i,\ell})\|^2 \\
& \overset{③}{=} \sum_{i=1}^n \sum_{\ell=2}^L \breve{\mathsf{loss}}_{i,\ell}^\top B\left((h_{i,\ell} - \breve{h}_{i,\ell}) - \sum_{a=1}^{\ell-1} \breve{D}_{i,\ell}\breve{W}\cdots\breve{W}\breve{D}_{i,a+1}W'\breve{h}_{i,a}\right) + \frac{1}{2}\|B(h_{i,\ell} - \breve{h}_{i,\ell})\|^2 \\
& \overset{④}{=} \sum_{i=1}^n \sum_{\ell=2}^L \breve{\mathsf{loss}}_{i,\ell}^\top B\left(\sum_{a=1}^{\ell-1} (\breve{D}_{i,\ell} + D''_{i,\ell})\breve{W}\cdots\breve{W}(\breve{D}_{i,a+1} + D''_{i,a+1})W'h_{i,a} - \breve{D}_{i,\ell}\breve{W}\cdots\breve{W}\breve{D}_{i,a+1}W'\breve{h}_{i,a}\right) \\
& \qquad + \frac{1}{2}\|B(h_{i,\ell} - \breve{h}_{i,\ell})\|^2
\end{aligned} \tag{G.3}$$

Above, ① is by the definition of $f(\cdot)$; ② is by (G.2); ③ is by the definition of $\nabla f(\cdot)$ (see Fact 2.5 for an explicit form of the gradient); ④ is by Claim G.2.

We next bound the RHS of (G.3). We drop subscripts in i for notational simplicity. We first use (G.1) in Claim G.2 to calculate

$$\begin{aligned}
\|B(h_\ell - \breve{h}_\ell)\| &\le \sum_{a=1}^{\ell-1} \|B(\breve{D}_\ell + D''_\ell)\breve{W}\cdots\breve{W}(\breve{D}_{a+1} + D''_{a+1})W'h_a\| \\
&\le \sum_{a=1}^{\ell-1} \|B\|_2 \|(\breve{D}_\ell + D''_\ell)\breve{W}\cdots\breve{W}(\breve{D}_{a+1} + D''_{a+1})\|_2 \|W'\|_2 \|h_a\| \\
&\overset{①}{\le} \sum_{a=1}^{\ell-1} O(\sqrt{m}) \cdot O(L^7) \cdot O(L) \cdot \|W'\|_2 \le O(L^9\sqrt{m}) \cdot \|W'\|_2 \ .
\end{aligned} \tag{G.4}$$

In the last inequality ① above, we have used $\left\|(\breve{D}_\ell + D''_\ell)\breve{W}\cdots\breve{W}(\breve{D}_{a+1} + D''_{a+1})\right\|_2 \le O(L^7)$ from Lemma C.7; we have used $\|B\|_2 \le O(\sqrt{m})$ with high probability; and we have used $\|h_a\| \le O(L)$ (by $\|\widetilde{h}_a\| \le O(L)$ from Lemma B.3 and $\|\widetilde{h}_a - h_a\| \le o(1)$ from Lemma C.2a).

Since for each D''_ℓ, we can write it as $D''_\ell = D^{0/1}_\ell D''_\ell D^{0/1}_\ell$, where each $D^{0/1}_\ell$ is a diagonal matrix satisfying

$$(D^{0/1}_\ell)_{k,k} = \begin{cases} 1, & (D''_\ell)_{k,k} \neq 0; \\ 0, & (D''_\ell)_{k,k} = 0. \end{cases} \quad \text{and} \quad \|D^{0/1}_\ell\|_0 \leq O(L^{10/3}\tau_0^{2/3}m^{2/3})$$

Therefore,

$$\begin{aligned}
&\left|\breve{\mathsf{loss}}_\ell^\top B(\breve{D}_\ell + D''_\ell)\breve{W}\cdots\breve{W}(\breve{D}_{a+1} + D''_{a+1})W'h_a - \breve{\mathsf{loss}}_\ell^\top B\breve{D}_\ell\breve{W}\cdots\breve{D}_{a+1}W'h_a\right| \\
&\overset{①}{\leq} \|\breve{\mathsf{loss}}_\ell\|_2 \cdot \sum_{b=1}^{\ell-a}\binom{\ell-a}{b}\|B\breve{D}\breve{W}\cdots\breve{D}\breve{W}D^{0/1}\|_2 \cdot \|D^{0/1}\breve{W}\breve{D}\cdots\breve{D}\breve{W}D^{0/1}\|_2^{b-1} \cdot \|D^{0/1}\breve{W}\cdots\breve{W}\breve{D}W'h_a\|_2 \\
&\overset{②}{\leq} \|\breve{\mathsf{loss}}_\ell\|_2 \cdot \sum_{b=1}^{\ell-a}\binom{\ell-a}{b} \cdot O(\rho^2\tau_0^{1/3}m^{1/3}) \cdot \left(O(\frac{\rho^2}{m^{1/6}})\right)^{b-1} \cdot O(L^7) \cdot \|W'\|_2 \cdot O(L) \\
&\overset{②}{\leq} \|\breve{\mathsf{loss}}_\ell\|_2 \cdot O(\rho^{11}\tau_0^{1/3}m^{1/3}) \cdot \|W'\|_2 \qquad \text{(G.5)}
\end{aligned}$$

Above, ① is an abbreviation and we have dropped the subscripts for the easy of presentation; ② uses Corollary B.18, Corollary B.15, Lemma C.7 and Lemma B.3.

Finally, we also have

$$\begin{aligned}
\left|\breve{\mathsf{loss}}_\ell^\top B\breve{D}_\ell\breve{W}\cdots\breve{D}_{a+1}W'(h_a - \breve{h}_a)\right| &\overset{①}{\leq} \|\breve{\mathsf{loss}}_\ell\|_2 \cdot O(\sqrt{m}) \cdot O(L^7) \cdot \|W'\|_2 \cdot \|h_a - \breve{h}_a\|_2 \\
&\overset{②}{\leq} O(\rho^{16}\sqrt{m}) \cdot \|\breve{\mathsf{loss}}_\ell\|_2 \cdot \|W'\|_2^2 \qquad \text{(G.6)}
\end{aligned}$$

where ① uses Lemma C.7 and ② uses Claim G.2 to bound $\|h_a - \breve{h}_a\|_2$.

Putting (G.4), (G.5) and (G.6) back to (G.3), and using triangle inequality, we have the desired result. □

G.1 Tool

Proposition G.3. *Given vectors $a, b \in \mathbb{R}^m$ and $D \in \mathbb{R}^{m\times m}$ the diagonal matrix where $D_{k,k} = \mathbf{1}_{a_k\geq 0}$. Then, then there exists a diagonal matrix $D'' \in \mathbb{R}^{m\times m}$ with*

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

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

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

H Convergence Rate of Gradient Descent (Theorem 1)

Theorem 1 (gradient descent, restated)**.** *There exists some absolute constant $C > 1$ such that the following holds. Let $\rho = nLd\log m$, $\varepsilon \in (0,1]$, $\varrho = nLd\delta^{-1}\log(m/\varepsilon)$, and $\eta \stackrel{\text{def}}{=} \frac{\delta}{\rho^{44}m}$. Suppose $m \geq C\varrho^C$, and $W^{(0)}, A, B$ be at random initialization. Then, with probability at least $1 - e^{-\Omega(\rho^2)}$ over the randomness of $W^{(0)}, A, B$, suppose we start at $W^{(0)}$ and for each $t = 0, 1, \dots, T-1$,*

$$W^{(t+1)} = W^{(t)} - \eta \nabla f(W^{(t)}) \ .$$

Then, it satisfies

$$f(W^{(T)}) \leq \varepsilon \quad \textit{for all} \quad T \in \Big[\frac{\rho^{59}}{\delta^2}\log\frac{1}{\varepsilon}, \frac{\rho^{59}\varrho^{28}}{\delta^2}\log\frac{1}{\varepsilon}\Big] \ .$$

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

Proof of Theorem 1. Using Lemma B.3 and the randomness of B, it is easy to show that $\|Bh_{i,\ell} - y^*_{i,\ell}\|^2 \leq O(\rho^2 L^2)$ with at least $1 - e^{-\Omega(\rho^2)}$ (where $h_{i,\ell}$ is defined with respect to $W^{(0)}$), and therefore

$$f(W^{(0)}) \leq O(n\rho^2 L^3) \ .$$

In the rest of the proof, we first assume that for every $t = 0, 1, \dots, T-1$, the following holds

$$\|W^{(t)} - W^{(0)}\|_F \leq \frac{\tau_0}{\sqrt{m}} \stackrel{\text{def}}{=} \frac{\varrho^{50}}{\sqrt{m}} \ . \tag{H.1}$$

We shall prove the convergence of gradient descent assuming (H.1), so that previous statements such as Theorem 4 and Theorem 3 can be applied. At the end of the proof, we shall verify that (H.1) is satisfied throughout the gradient descent process.

For each $t = 0, 1, \dots, T-1$, we let $\{\mathsf{loss}^{(t)}_{i,\ell}\}_{i,\ell}$ denote the loss vectors with respect to the current point $W^{(t)}$. We denote by $(i^*, \ell^*) = \arg\max_{i,\ell}\{\|\mathsf{loss}^{(t)}_{i,\ell}\|_2\}$ and $\nabla_t = \nabla f(W^{(t)})$.

We calculate that

$$\begin{aligned}
f(W^{(t+1)}) &\overset{①}{\leq} f(W^{(t)}) - \eta\|\nabla_t\|_F^2 + O(\rho^{11}\tau_0^{1/3}m^{1/3}) \cdot \|\mathsf{loss}^{(t)}_{i^*,\ell^*}\|_2 \cdot \eta\|\nabla_t\|_2 + O(L^{18}nm\eta^2)\|\nabla_t\|_2^2 \\
&\overset{②}{\leq} f(W^{(t)}) - \eta\|\nabla_t\|_F^2 + O\left(\rho^{30}\eta^2 m^2\right) \cdot \|\mathsf{loss}^{(t)}_{i^*,\ell^*}\|_2^2 \\
&\overset{③}{\leq} f(W^{(t)}) - \left(\Omega\Big(\frac{\eta\delta}{\rho^{14}}m\Big) - O(\rho^{30}\eta^2 m^2)\right) \cdot \|\mathsf{loss}^{(t)}_{i^*,\ell^*}\|_2^2 \\
&\overset{④}{\leq} f(W^{(t)}) - \Omega\Big(\frac{\eta\delta}{\rho^{14}}m\Big) \cdot \|\mathsf{loss}^{(t)}_{i^*,\ell^*}\|_2^2 \\
&\overset{⑤}{\leq} \left(1 - \Omega\Big(\frac{\eta\delta}{\rho^{15}}m\Big)\right) f(W^{(t)})
\end{aligned}$$

Above, ① uses Theorem 4; ② uses Theorem 3 (which gives $\|\nabla_t\|_2 \leq \|\nabla_t\|_F \leq O(\rho^6\sqrt{m}) \times \|\mathsf{loss}^{(t)}_{i^*,\ell^*}\|$), and our choices of τ_0 and m; ③ uses Theorem 3; ④ uses our choice of η; and ⑤ uses $f(W^{(t)}) \leq nL\|\mathsf{loss}^{(t)}_{i^*,\ell^*}\|^2$. In other words, after $T = \Omega(\frac{\rho^{15}}{\eta\delta m})\log\frac{nL^2}{\varepsilon}$ iterations we have $f(W^{(T)}) \leq \varepsilon$.

We need to verify for each t, $\|W^{(t)} - W^{(0)}\|_F$ is small so that (H.1) holds. By Theorem 3,

$$\|W^{(t)} - W^{(0)}\|_F \le \sum_{i=0}^{t-1} \|\eta \nabla f(W^{(i)})\|_F \le O(\eta\rho^6\sqrt{m}) \cdot \sum_{i=0}^{t-1} \sqrt{f(W^{(i)})} \le O(\eta\rho^6\sqrt{m}) \cdot O(T \cdot \sqrt{n\rho^2 L^3})$$
$$\le \eta T \cdot O(\rho^{8.5}\sqrt{m}) \le \frac{\varrho^{50}}{\sqrt{m}} \enspace .$$

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

I Convergence Rate of Stochastic Gradient Descent (Theorem 2)

Theorem 2 (stochastic gradient descent, stated)**.** *There exists some absolute constant $C > 1$ such that the following holds. Let $\rho = nLd\log m$, $\varepsilon \in (0,1]$, $\varrho = nLd\delta^{-1}\log(m/\varepsilon)$, and $\eta \stackrel{\text{def}}{=} \frac{\delta}{\rho^{42}m}$. Suppose $m \ge C\varrho^C$, and $W^{(0)}, A, B$ be at random initialization. Then, with probability at least $1 - e^{-\Omega(\rho^2)}$ over the randomness of $W^{(0)}, A, B$, suppose we start at $W^{(0)}$ and for each $t = 0, 1, \dots, T-1$,*

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

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

$$f(W^{(T)}) \le \varepsilon \quad \textit{for all} \quad T \in \left[\frac{\rho^{57}}{\delta^2}\log\frac{1}{\varepsilon}, \frac{\rho^{57}\varrho^{28}}{\delta^2}\log\frac{1}{\varepsilon}\right] \enspace .$$

Proof of Theorem 2. The proof is almost identical to that of Theorem 1. We again have with probability at least $1 - e^{-\Omega(\rho^2)}$

$$f(W^{(0)}) \le O(n\rho^2 L^3) \enspace .$$

Again, we first assume for every $t = 0, 1, \dots, T-1$, the following holds

$$\|W^{(t)} - W^{(0)}\|_F \le \frac{\tau_0}{\sqrt{m}} \stackrel{\text{def}}{=} \frac{\varrho^{50}}{\sqrt{m}} \enspace . \tag{I.1}$$

We shall prove the convergence of SGD assuming (I.1), so that previous statements such as Theorem 4 and Theorem 3 can be applied. At the end of the proof, we shall verify that (I.1) is satisfied throughout the SGD with high probability.

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

$$\begin{aligned}
&\mathop{\mathbb{E}}_{S_t}[f(W^{(t+1)})] \\
&\stackrel{①}{\le} f(W^{(t)}) - \eta\|\nabla f(W^{(t)})\|_F^2 + O(\rho^{11}\tau_0^{1/3}m^{1/3}) \cdot \|\,\mathsf{loss}_{i^*,\ell^*}^{(t)}\,\|_2 \cdot \eta \mathop{\mathbb{E}}_{S_t}[\|\nabla_t\|_2] + O(L^{18}nm\eta^2)\mathop{\mathbb{E}}_{S_t}[\|\nabla_t\|_2^2] \\
&\stackrel{②}{\le} f(W^{(t)}) - \eta\|\nabla_t\|_F^2 + O\left(\rho^{30}\eta^2 m^2\right) \cdot \|\,\mathsf{loss}_{i^*,\ell^*}^{(t)}\,\|_2^2 \\
&\stackrel{③}{\le} \left(1 - \Omega\Big(\frac{\eta\delta}{\rho^{15}}m\Big)\right) f(W^{(t)}) \enspace .
\end{aligned} \tag{I.2}$$

Above, ① uses Theorem 4 and $\mathbb{E}_{S_t}[\nabla_t] = \nabla f(W^{(t)})$; ② uses Theorem 3 (which gives $\|\nabla_t\|_2^2 \le \frac{n^2}{|S_t|^2}\sum_{i\in S_t}\|\nabla f_i(W^{(t)})\|_F^2 \le O(\rho^{12}m) \times \|\,\mathsf{loss}_{i^*,\ell^*}^{(t)}\,\|^2$); ③ is identical to the proof of Theorem 1.

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

$$
\begin{aligned}
& f(W^{(t+1)}) \\
\overset{①}{\leq}\ & f(W^{(t)}) + \eta \|\nabla f(W^{(t)})\|_F \cdot \|\nabla_t\|_F + O(\rho^{11}\tau_0^{1/3}m^{1/3}) \cdot \| \mathsf{loss}_{i^*,\ell^*}^{(t)} \|_2 \cdot \eta \|\nabla_t\|_2 + O(L^{18}nm\eta^2)\|\nabla_t\|_2^2 \\
\overset{②}{\leq}\ & f(W^{(t)}) + O\left(\rho^{12}\eta m + \rho^{30}\eta^2 m^2\right) \cdot \| \mathsf{loss}_{i^*,\ell^*}^{(t)} \|_2^2 \\
\overset{③}{\leq}\ & \left(1 + O\big(\rho^{12}\eta m\big)\right) f(W^{(t)}) \ . \qquad \text{(I.3)}
\end{aligned}
$$

Above, ① uses Theorem 4 and Cauchy-Shwartz $\langle A, B\rangle \leq \|A\|_F \|B\|_F$, and ② uses Theorem 3 and the derivation from (I.2).

Next, taking logarithm on both sides of (I.2) and (I.3), and using Jensen's inequality $\mathbb{E}[\log X] \leq \log \mathbb{E}[X]$, we have

$$
\mathbb{E}[\log f(W^{(t+1)})] \leq \log f(W^{(t)}) - \Omega\Big(\frac{\eta\delta}{\rho^{15}}m\Big) \quad \text{and} \quad \log f(W^{(t+1)}) \leq \log f(W^{(t)}) + O\big(\rho^{12}\eta m\big)
$$

By one-sided Azuma's inequality (a.k.a. martingale concentration), we have with probability at least $1 - e^{-\Omega(\rho^2)}$, for every $t = 1, 2, \dots, T$:

$$
\log f(W^{(t)}) - \mathbb{E}[\log f(W^{(t)})] \leq \sqrt{t} \cdot O\big(\rho^{12}\eta m\big) \cdot \rho \ .
$$

This implies two things.

- On one hand, after $T = \Omega(\frac{\rho^{15}\log(nL^2/\varepsilon)}{\eta\delta m})$ iterations we have

  $$
  \begin{aligned}
  \log f(W^{(T)}) &\leq \sqrt{T} \cdot O\big(\rho^{12}\eta m\big) \cdot \rho + \log f(W^{(0)}) - \Omega\Big(\frac{\eta\delta}{\rho^{15}}m\Big)T \\
  &\leq \log f(W^{(0)}) - \Omega\Big(\frac{\eta\delta}{\rho^{15}}m\Big)T \leq \log O(n\rho^2L^3) - \Omega(\log \frac{\rho^5}{\varepsilon}) \leq \log \varepsilon \ .
  \end{aligned}
  $$

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

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

  $$
  \begin{aligned}
  \log f(W^{(t)}) &\leq \sqrt{t} \cdot O\big(\rho^{12}\eta m\big) \cdot \rho + \log f(W^{(0)}) - \Omega\Big(\frac{\eta\delta}{\rho^{15}}m\Big)t \\
  &\overset{①}{=} \log f(W^{(0)}) - \left(\sqrt{\frac{\eta\delta m}{\rho^{15}}} \cdot \Omega(\sqrt{t}) - \sqrt{\frac{\rho^{15}}{\eta\delta m}} \cdot O(\rho^{13}\eta m)\right)^2 + O\Big(\frac{\rho^{41}\eta m}{\delta}\Big) \\
  &\overset{②}{\leq} \log f(W^{(0)}) + 1
  \end{aligned}
  $$

  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 $\eta \leq O\big(\frac{\delta}{\rho^{42}m}\big)$. This implies $f(W^{(t)}) \leq O(n\rho^2L^3)$. We can now verify for each t, $\|W^{(t)} - W^{(0)}\|_F$ is small so that (I.1) holds. By Theorem 3,

  $$
  \begin{aligned}
  \|W^{(t)} - W^{(0)}\|_F \leq \sum_{i=0}^{t-1} \|\eta\nabla_t\|_F \leq O(\eta\rho^6\sqrt{m}) \cdot \sum_{i=0}^{t-1} \sqrt{f(W^{(i)})} &\leq O(\eta\rho^6\sqrt{m}) \cdot O(T\sqrt{n\rho^2L^3}) \\
  &\leq \eta T \cdot O(\rho^{8.5}\sqrt{m}) \leq \frac{\varrho^{50}}{\sqrt{m}} \ .
  \end{aligned}
  $$

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

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] 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.

[3] Alexandr Andoni, Rina Panigrahy, Gregory Valiant, and Li Zhang. Learning polynomials with neural networks. In *International Conference on Machine Learning (ICML)*, pages 1908–1916, 2014.

[4] Sanjeev Arora, Aditya Bhaskara, Rong Ge, and Tengyu Ma. Provable bounds for learning some deep representations. In *International Conference on Machine Learning (ICML)*, pages 584–592, 2014.

[5] Sanjeev Arora, Rong Ge, Tengyu Ma, and Andrej Risteski. Provable learning of noisy-or networks. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pages 1057–1066. ACM, 2017.

[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] Andrew R Barron. Approximation and estimation bounds for artificial neural networks. *Machine learning*, 14(1):115–133, 1994.

[9] 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.

[10] 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.

[11] 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.

[12] Alon Brutzkus, Amir Globerson, Eran Malach, and Shai Shalev-Shwartz. SGD learns over-parameterized networks that provably generalize on linearly separable data. In *ICLR*, 2018. URL `https://arxiv.org/abs/1710.10174`.

[13] Minmin Chen, Jeffrey Pennington, and Samuel S. Schoenholz. Dynamical isometry and a mean field theory of RNNs: Gating enables signal propagation in recurrent neural networks. *arXiv:1806.05394*, 2018. URL `http://arxiv.org/abs/1806.05394`.

[14] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

[15] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*, pages 1724–1734, 2014.

[16] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[17] Nadav Cohen and Amnon Shashua. Convolutional rectifier networks as generalized tensor decompositions. In *International Conference on Machine Learning (ICML)*, pages 955–963, 2016.

[18] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[19] 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.

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

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

[22] Amit Daniely, Roy Frostig, and Yoram Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2253–2261, 2016.

[23] 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.

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

[25] 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.

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

[27] Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

[28] C. Daniel Freeman and Joan Bruna. Topology and geometry of half-rectified network optimization. In *ICLR*, 2017. URL `https://arxiv.org/abs/1611.01540`.

[29] Ken-Ichi Funahashi. On the approximate realization of continuous mappings by neural networks. *Neural networks*, 2(3):183–192, 1989.

[30] 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`.

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

[32] 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.

[33] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5-6):602–610, 2005.

[34] 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.

[35] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *CoRR*, abs/1410.5401, 2014. URL `http://arxiv.org/abs/1410.5401`.

[36] Benjamin D Haeffele and René Vidal. Global optimality in tensor factorization, deep learning, and beyond. *arXiv preprint arXiv:1506.07540*, 2015.

[37] Richard LT Hahnloser. On the piecewise analysis of networks of linear threshold neurons. *Neural Networks*, 11(4):691–697, 1998.

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

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

[40] 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.

[41] 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.

[42] 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.

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

[44] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[45] Majid Janzamin, Hanie Sedghi, and Anima Anandkumar. Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods. *arXiv preprint arXiv:1506.08473*, 2015.

[46] Nal Kalchbrenner and Phil Blunsom. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, 2013.

[47] Guy Katz, Clark Barrett, David L Dill, Kyle Julian, and Mykel J Kochenderfer. Reluplex: An efficient smt solver for verifying deep neural networks. In *International Conference on Computer Aided Verification (CAV)*, pages 97–117. Springer, 2017.

[48] Valentin Khrulkov, Alexander Novikov, and Ivan Oseledets. Expressive power of recurrent neural networks. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=S1WRibb0Z`.

[49] 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.

[50] 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.

[51] Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

[52] Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as Gaussian processes. *arXiv:1711.00165*, 2017. URL `http://arxiv.org/abs/1711.00165`.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] S Lojasiewicz. A topological property of real analytic subsets. *Coll. du CNRS, Les équations aux dérivées partielles*, 117:87–89, 1963.

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

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

[60] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33): E7665–E7671, 2018.

[61] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[62] Tomáš Mikolov, Stefan Kombrink, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. Extensions of recurrent neural network language model. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5528–5531. IEEE, 2011.

[63] Marco Mondelli and Andrea Montanari. On the connection between learning two-layers neural networks and tensor decomposition. *arXiv preprint arXiv:1802.07301*, 2018.

[64] Quynh Nguyen and Matthias Hein. The loss surface of deep and wide neural networks. *arXiv preprint arXiv:1704.08045*, 2017.

[65] Quynh Nguyen and Matthias Hein. Optimization landscape and expressivity of deep cnns. In *International Conference on Machine Learning (ICML)*, pages 3727–3736, 2018.

[66] Junier B Oliva, Barnabás Póczos, and Jeff Schneider. The statistical recurrent unit. In *International Conference on Machine Learning (ICML)*, pages 2671–2680, 2017.

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

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

[69] Jeffrey Pennington and Yasaman Bahri. Geometry of neural network loss surfaces via random matrix theory. In *International Conference on Machine Learning (ICML)*, International Convention Centre, Sydney, Australia, 2017.

[70] Jeffrey Pennington and Pratik Worah. Nonlinear random matrix theory for deep learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[71] Jeffrey Pennington, Samuel S. Schoenholz, and Surya Ganguli. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. *arXiv:1711.04735*, November 2017. URL `http://arxiv.org/abs/1711.04735`.

[72] Allan Pinkus. Approximation theory of the mlp model in neural networks. *Acta Numerica*, 8:143–195, 1999.

[73] Boris Teodorovich Polyak. Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653, 1963.

[74] Ben Poole, Subhaneil Lahiri, Maithreyi Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. In *Advances In Neural Information Processing Systems*, pages 3360–3368, 2016. 00047.

[75] 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.

[76] Haşim Sak, Andrew Senior, and Françoise Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Fifteenth annual conference of the international speech communication association*, 2014.

[77] Hojjat Salehinejad, Julianne Baarbe, Sharan Sankar, Joseph Barfett, Errol Colak, and Shahrokh Valaee. Recent advances in recurrent neural networks. *arXiv preprint arXiv:1801.01078*, 2017.

[78] Emilio Salinas and Laurence F. Abbott. A model of multiplicative neural responses in parietal cortex. *Proceedings of the National Academy of Sciences*, 93(21):11956–11961, 1996.

[79] Samuel S. Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation. In *ICLR*, 2017. URL `https://openreview.net/pdf?id=H1W1UN9gg`.

[80] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[81] Hanie Sedghi and Anima Anandkumar. Provable methods for training neural networks with sparse connectivity. In *ICLR*. arXiv preprint arXiv:1412.2693, 2015.

[82] Hava T Siegelmann and Eduardo D Sontag. Turing computability with neural nets. *Applied Mathematics Letters*, 4(6):77–80, 1991.

[83] 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.

[84] 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.

[85] 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.

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

[87] Mahdi Soltanolkotabi, Adel Javanmard, and Jason D Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 2018.

[88] 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.

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

[90] Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. Lstm neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*, 2012.

[91] Ilya Sutskever, James Martens, and Geoffrey Hinton. Generating text with recurrent neural networks. In *International Conference on Machine Learning (ICML)*, pages 1017–1024, 2011.

[92] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems (NIPS)*, pages 3104–3112, 2014.

[93] Matus Telgarsky. Benefits of depth in neural networks. In *Conference on Learning Theory (COLT)*. arXiv preprint arXiv:1602.04485, 2016.

[94] 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.

[95] Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. Tensor decomposition for compressing recurrent neural network. *CoRR*, abs/1802.10410, 2018. URL `http://arxiv.org/abs/1802.10410`.

[96] Santosh Vempala and John Wilmes. Polynomial convergence of gradient descent for training one-hidden-layer neural networks. *arXiv preprint arXiv:1805.02677*, 2018.

[97] Tsui-Wei Weng, Huan Zhang, Hongge Chen, Zhao Song, Cho-Jui Hsieh, Duane Boning, Inderjit S Dhillon, and Luca Daniel. Towards fast computation of certified robustness for ReLU networks. In *International Conference on Machine Learning (ICML)*. arXiv preprint arXiv:1804.09699, 2018.

[98] Paul J Werbos. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, 1990.

[99] Lechao Xiao, Yasaman Bahri, Jascha Sohl-Dickstein, Samuel S. Schoenholz, and Jeffrey Pennington. Dynamical isometry and a mean field theory of CNNs: How to train 10,000-layer vanilla convolutional neural networks. In *International Conference on Machine Learning (ICML)*, 2018.

[100] Greg Yang and Sam S. Schoenholz. Deep mean field theory: Layerwise variance and width variation as methods to control gradient explosion. *ICLR open review*, 2018. URL `https://openreview.net/forum?id=rJGY8GbR-`.

[101] Greg Yang and Samuel Schoenholz. Mean field residual networks: On the edge of chaos. In *Advances in Neural Information Processing Systems (NIPS)*, pages 7103–7114, 2017.

[102] 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.

[103] 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.

[104] Yuchen Zhang, Jason D Lee, and Michael I Jordan. ℓ_1-regularized neural networks are improperly learnable in polynomial time. In *International Conference on Machine Learning (ICML)*, pages 993–1001, 2016.

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

[106] 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.

[107] Yi Zhou and Yingbin Liang. Critical points of neural networks: Analytical forms and landscape properties. *arXiv preprint arXiv:1710.11205*, 2017.