# Gradient Descent Finds Global Minima of Deep Neural Networks

Simon S. Du*[1], Jason D. Lee*[2], Haochuan Li*†[3,5], Liwei Wang*[4,5], and Xiyu Zhai*[6]

[1]*Machine Learning Department, Carnegie Mellon University*

[2]*Data Science and Operations Department, University of Southern California*

[3]*School of Physics, Peking University*

[4]*Key Laboratory of Machine Perception, MOE, School of EECS, Peking University*

[5]*Center for Data Science, Peking University, Beijing Institute of Big Data Research*

[6]*Department of EECS, Massachusetts Institute of Technology*

December 3, 2018

**Abstract**

Gradient descent finds a global minimum in training deep neural networks despite the objective function being non-convex. The current paper proves gradient descent achieves zero training loss in polynomial time for a deep over-parameterized neural network with residual connections (ResNet). Our analysis relies on the particular structure of the Gram matrix induced by the neural network architecture. This structure allows us to show the Gram matrix is stable throughout the training process and this stability implies the global optimality of the gradient descent algorithm. We further extend our analysis to deep residual convolutional neural networks and obtain a similar convergence result.

# 1 Introduction

One of the mysteries in deep learning is random initialized first order methods like gradient descent achieve zero training loss, even if the labels are arbitrary [Zhang et al., 2016]. Over-parameterization is widely believed to be the main reason for this phenomenon as only if the neural network has a sufficiently large capacity, it is possible for this neural network to fit all the training data. In practice, many neural network architectures are highly over-parameterized. For example,

*Alphabetical ordering.

†This work was completed while Haochuan Li was visiting University of Southern California.

Wide Residual Networks have 100x parameters than the number of training data [Zagoruyko and Komodakis, 2016].

The second mysterious phenomenon in training deep neural networks is "deeper networks are harder to train." To solve this problem, He et al. [2016] proposed the deep residual network (ResNet) architecture which enables randomly initialized first order method to train neural networks with an order of magnitude more layers. Theoretically, Hardt and Ma [2016] showed that residual links in linear networks prevent gradient vanishing in a large neighborhood of zero, but for neural networks with non-linear activations, the advantages of using residual connections are not well understood.

In this paper, we demystify these two mysterious phenomena. We consider the setting where there are $n$ data points, and the neural network has $H$ layers with width $m$. We focus on the least-squares loss and assume the activation function is Lipschitz and smooth. This assumption holds for many activation functions including the soft-plus. Our contributions are summarized below.

- As a warm-up, we first consider a fully-connected feedforward network. We show if $m = \Omega\left(\mathrm{poly}(n)2^{O(H)}\right)$[1], then randomly initialized gradient descent converges to zero training loss at a linear rate.
- Next, we consider the ResNet architecture. We show as long as $m = \Omega\left(\mathrm{poly}(n, H)\right)$, then randomly initialized gradient descent converges to zero training loss at a linear rate. Comparing with the first result, the dependence on the number of layers improves exponentially for ResNet. This theory demonstrates the advantage of using residual connections.
- Lastly, we apply the same technique to analyze convolutional ResNet. We show if $m = \mathrm{poly}(n, p, H)$ where $p$ is the number of patches, then randomly initialized gradient descent achieves zero training loss.

Our proof builds on two ideas from previous work on gradient descent for two-layer neural networks. First, we use the observation by Li and Liang [2018] that if the neural network is over-parameterized, every weight matrix is close to its initialization. Second, following Du et al. [2018b], we analyze the dynamics of the predictions whose convergence is determined by the least eigenvalue of the Gram matrix induced by the neural network architecture and to lower bound the least eigenvalue, it is sufficient to bound the distance of each weight matrix from its initialization.

Different from these two works, in analyzing deep neural networks, we need to exploit more structural properties of deep neural networks and develop new techniques. See Section 6 and the Appendix for more details.

## 1.1 Organization

This paper is organized as follows. In Section 2, we formally state the problem setup. In Section 3, we give our main result for the deep fully-connected neural network. In Section 4, we give our main result for the ResNet. In Section 5, we give our main result for the convolutional ResNet.

[1] The precise polynomials and data-dependent parameters are stated in Section 3, 4, 5.

In Section 6, we present a unified proof strategy for these three architectures. We conclude in Section 7 and defer all proofs to the appendix.

## 1.2 Related Works

Recently, many works try to study the optimization problem in deep learning. Since optimizing a neural network is a non-convex problem, one approach is first to develop a general theory for a class of non-convex problems which satisfy desired geometric properties and then identify that the neural network optimization problem belongs to this class. One promising candidate class is the set of functions that satisfy all local minima are global and there exists a negative curvature for every saddle point. For this function class, researchers have shown gradient descent [Jin et al., 2017, Ge et al., 2015, Lee et al., 2016, Du et al., 2017a] can find a global minimum. Many previous works thus try to study the optimization landscape of neural networks with different activation functions [Soudry and Hoffer, 2017, Safran and Shamir, 2018, 2016, Zhou and Liang, 2017, Freeman and Bruna, 2016, Hardt and Ma, 2016, Nguyen and Hein, 2017, Kawaguchi, 2016, Venturi et al., 2018, Soudry and Carmon, 2016, Du and Lee, 2018, Soltanolkotabi et al., 2018, Haeffele and Vidal, 2015]. However, even for a deep linear network, there exists a saddle point that does not have a negative curvature [Kawaguchi, 2016], so it is unclear whether this approach can be used to obtain the global convergence guarantee of first-order methods.

Another way to attack this problem is to study the dynamics of a specific algorithm for a specific neural network architecture. Our paper also belongs to this category. Many previous works put assumptions on the input distribution and assume the label is generated according to a planted neural network. Based on these assumptions, one can obtain global convergence of gradient descent for some shallow neural networks [Tian, 2017, Soltanolkotabi, 2017, Brutzkus and Globerson, 2017, Du et al., 2018a, Li and Yuan, 2017, Du et al., 2017b]. Some local convergence results have also been proved [Zhong et al., 2017a,b, Zhang et al., 2018]. In comparison, our paper does not try to recover the underlying neural network. Instead, we focus the empirical loss minimization problem and rigorously prove that randomly initialized gradient descent can achieve zero training loss.

The most related papers are Li and Liang [2018], Du et al. [2018b] who observed that when training a two-layer full connected neural network, most of the patterns do not change over iterations, which we also use to show the stability of the Gram matrix. They used this observation to obtain the convergence rate of gradient descent on a two-layer over-parameterized neural network for the cross-entropy and least-squares loss. More recently, Allen-Zhu et al. [2018a] generalizes ideas from Li and Liang [2018] to derive convergence rates of training recurrent neural networks. Our work extends these previous results in several ways: a) we consider deep networks, b) we generalize to ResNet architectures, and c) we generalize to convolutional networks. To improve the width dependence $m$ on sample size $n$, we utilize a smooth activation (e.g. smooth ReLU). For example, our results specialized to depth $H = 2$ improve upon Du et al. [2018b] in the required amount of overparametrization from $m = \Omega\left(\frac{n^6}{\lambda_0^4}\right)$ to $m = \Omega\left(\frac{n^4}{\lambda_0^4}\right)$. See Theorem 3.1 for the precise statement.

Chizat and Bach [2018], Wei et al. [2018], Mei et al. [2018] used optimal transport theory to analyze gradient descent on over-parameterized models. However, their results are limited to two-layer neural networks and may require an exponential amount of over-parametrization.

Daniely [2017] developed the connection between deep neural networks with kernel methods and showed stochastic gradient descent can learn a function that is competitive with the best function in the conjugate kernel space of the network. Andoni et al. [2014] showed that gradient descent can learn networks that are competitive with polynomial classifiers. However, these results do not imply gradient descent can find a global minimum for the empirical loss minimization problem.

Finally in concurrent work, Allen-Zhu et al. [2018b] also analyze gradient descent on deep neural networks. The primary difference between the two papers are that we analyze general smooth activations, and Allen-Zhu et al. [2018b] develop specific analysis for ReLU activation. The two papers also differ significantly on their data assumptions. *We wish to emphasize a fair comparison is not possible due to the difference in setting and data assumptions. We view the two papers as complementary since they address different neural net architectures.*

For ResNet, the primary focus of this manuscript, the required width per layer[2] for Allen-Zhu et al. [2018b] is $m \gtrsim n^{30}H^{30}\log^2\frac{1}{\epsilon}$ and for Theorem 4.1 is $m \gtrsim n^4H^2$ [3]. Our paper requires a width $m$ that does not depend on the desired accuracy $\epsilon$. As a consequence, Theorem 4.1 guarantees the convergence of gradient descent to a global minimizer. The iteration complexity of Allen-Zhu et al. [2018b] is $T \gtrsim n^6H^2\log\frac{1}{\epsilon}$ and of Theorem 4.1 is $T \gtrsim n^2\log\frac{1}{\epsilon}$.

For fully-connected networks, Allen-Zhu et al. [2018b] requires width $m \gtrsim n^{30}H^{30}\log^2\frac{1}{\epsilon}$ and iteration complexity $T \geq n^6H^2\log\frac{1}{\epsilon}$. Theorem 3.1 requires width $m \geq n^4(\frac{1}{\lambda})^{O(H)}$ and iteration complexity $T \geq n^2 2^{O(H)}\log\frac{1}{\epsilon}$. The primary difference is for very deep fully-connected networks, Allen-Zhu et al. [2018b] requires less width and iterations, but we have much milder dependence on sample size $n$. Commonly used fully-connected networks such as VGG are not deep ($H = 16$), yet the dataset size such as ImageNet ($n \sim 10^6$) is very large.

In a second concurrent work, Zou et al. [2018] also analyzed the convergence of gradient descent on fully-connected networks with ReLU activation. The emphasis is on different loss functions (e.g. hinge loss), so the results are not directly comparable. Both Zou et al. [2018] and Allen-Zhu et al. [2018b] also analyze stochastic gradient descent.

# 2 Preliminaries

## 2.1 Notations

We Let $[n] = \{1, 2, \ldots, n\}$. Given a set $S$, we use $\mathrm{unif}\left\{S\right\}$ to denote the uniform distribution over $S$. We use $N(\mathbf{0}, \mathbf{I})$ to denote the standard Gaussian distribution. For a matrix $\mathbf{A}$, we use $\mathbf{A}_{ij}$ to denote its $(i, j)$-th entry. For a vector $\mathbf{v}$, we use $\left\|\mathbf{v}\right\|_2$ to denote the Euclidean norm. For a matrix $\mathbf{A}$ we use $\left\|\mathbf{A}\right\|_F$ to denote the Frobenius norm and $\left\|\mathbf{A}\right\|_2$ to denote the operator norm. If a matrix $\mathbf{A}$ is positive semi-definite, we use $\lambda_{\min}(\mathbf{A})$ to denote its smallest eigenvalue. We

[2] In all comparisons, we ignore the polynomial dependency on data-dependent parameters which only depends on the input data and the activation function. The two papers use different measures $\lambda$ and $\lambda_0$ in our manuscript and $\delta$ in Allen-Zhu et al. [2018b], so they are not directly comparable.

[3] $\lambda_0 = \frac{c_{\mathcal{D}}}{H^2}$ and $c_{\mathcal{D}}$ does not depend on the depth $H$. We substituted this into Theorem 4.1

use $\langle \cdot, \cdot \rangle$ to denote the standard Euclidean inner product between two vectors or matrices. We use $\sigma(\cdot)$ to denote the activation function, which in this paper we assume is Lipschitz continuous and smooth (Lipschitz gradient). The guiding example is softplus: $\sigma(z) = \log(1 + \exp(z))$ whose Lipschitz and smoothness constants are bounded by 1. We will use $c$ (e.g. $c_x, c_\lambda$) to denote universal constants that only depend on the activation. Lastly, let $O(\cdot)$ and $\Omega(\cdot)$ denote standard Big-O and Big-Omega notations, only hiding absolute constants.

## 2.2 Problem Setup

In this paper, we focus on the empirical risk minimization problem with the quadratic loss function

$$\min_{\mathbf{w}} L(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n}(f(\mathbf{w}, \mathbf{x}_i) - y_i)^2 \tag{1}$$

where $\{\mathbf{x}_i\}_{i=1}^n$ are the training inputs, $\{y_i\}_{i=1}^n$ are the labels, $\mathbf{w}$ is the parameter we optimize over and $f$ is the prediction function, which in our case is a neural network. We consider the following architectures.

- **Multilayer fully-connected neural networks:** Let $\mathbf{x} \in \mathbb{R}^d$ be the input, $\mathbf{W}^{(1)} \in \mathbb{R}^{m \times d}$ is the first weight matrix, $\mathbf{W}^{(h)} \in \mathbb{R}^{m \times m}$ is the weight at the $h$-th layer for $2 \le h \le H$, $\mathbf{a} \in \mathbb{R}^m$ is the output layer and $\sigma(\cdot)$ is the activation function.[4] The prediction function is defined as

$$f(\mathbf{x}, \mathbf{w}) = \mathbf{a}^\top \sqrt{\frac{c_\sigma}{m}}\sigma\left(\mathbf{W}^{(H)}\sqrt{\frac{c_\sigma}{m}}\sigma\left(\mathbf{W}^{(H-1)}\cdots\sqrt{\frac{c_\sigma}{m}}\sigma\left(\mathbf{W}^{(1)}\mathbf{x}\right)\right)\right) \tag{2}$$

  where $c_\sigma = \left(\mathbb{E}_{x \sim N(0,1)}\left[\sigma(x)^2\right]\right)^{-1}$ is a scaling factor to normalize the input in the initialization phase.

- **ResNet**[5]: We use the same notations as the multilayer fully connected neural networks. We define the prediction recursively.

$$\begin{aligned}
\mathbf{x}^{(1)} &= \sqrt{\frac{c_\sigma}{m}}\sigma\left(\mathbf{W}^{(1)}\mathbf{x}\right), \\
\mathbf{x}^{(h)} &= \mathbf{x}^{(h-1)} + \frac{c_\lambda \lambda^{3/2}}{H\sqrt{m}}\sigma\left(\mathbf{W}^{(h)}\mathbf{x}^{(h-1)}\right), 2 \le h \le H, \\
f_{res}(\mathbf{x}, \mathbf{w}) &= \mathbf{a}^\top \mathbf{x}^{(H)}.
\end{aligned} \tag{3}$$

  where $0 < c_\lambda < 1$ is a constant that only depends on the activation, $c_\sigma$, and $\lambda$ are constants specified in Section 4. Note here we use a $\frac{c_\lambda \lambda^{3/2}}{H\sqrt{m}}$ scaling. This scaling plays an important role

[4] We assume intermediate layers are square matrices for simplicity. It is not difficult to generalize our analysis to rectangular weight matrices.

[5] We will refer to this architecture as ResNet, although this differs by the standard ResNet architecture since the skip-connections at every layer, instead of every two layers. This architecture was previously studied in Hardt and Ma [2016]. We study this architecture for the ease of presentation and analysis. It is not hard to generalize our analysis to architectures with skip-connections are every two or more layers.

in guaranteeing the width per layer only needs to scale polynomially with $H$. In practice, the small scaling is enforced by a small initialization of the residual connection [Hardt and Ma, 2016, Anonymous, 2018], which obtains state-of-art performance for deep residual network. We choose to use an explicit scaling, instead of altering the initialization scheme for notational convenience.

- **Convolutional ResNet**: Lastly, we consider the convolutional ResNet architecture. Again we define the prediction function in a recursive way. Let $\mathbf{x}^{(0)} \in \mathbb{R}^{d_0 \times p}$ be the input, where $d_0$ is the number of input channels and $p$ is the number of pixels. For $h \in [H]$, we let the number of channels be $d_h = m$ and number of pixels be $p$. Given $\mathbf{x}^{(h-1)} \in \mathbb{R}^{d_{h-1} \times p}$ for $h \in [H]$, we first use an operator $\phi_h(\cdot)$ to divide $\mathbf{x}^{(h-1)}$ into $p$ patches. Each patch has size $qd_{h-1}$ and this implies $\phi_h(\mathbf{x}^{(h-1)}) \in \mathbb{R}^{qd_{h-1} \times p}$. For example, when the stride is $1$ and $q = 3$

$$\phi_h(\mathbf{x}^{(h-1)}) = \begin{pmatrix} \left(\mathbf{x}^{(h-1)}_{1,0:2}\right)^\top, & \left(\mathbf{x}^{(h-1)}_{1,1:3}\right)^\top, & \ldots & ,\left(\mathbf{x}^{(h-1)}_{1,p-1:p+1}\right)^\top \\ \ldots, & \ldots, & \ldots, & \ldots \\ \left(\mathbf{x}^{(h-1)}_{d_{h-1},0:2}\right)^\top, & \left(\mathbf{x}^{(h-1)}_{d_{h-1},1:3}\right)^\top, & \ldots, & \left(\mathbf{x}^{(h-1)}_{d_{h-1},p-1:p+1}\right)^\top \end{pmatrix},$$

  where we defined $\mathbf{x}^{(h-1)}_{:,0} = \mathbf{x}^{(h-1)}_{:,p+1} = \mathbf{0}$, i.e., zero-padding. Note this operator has the property

$$\left\|\mathbf{x}^{(h-1)}\right\|_F \le \left\|\phi_h(\mathbf{x}^{(h-1)})\right\|_F \le \sqrt{q}\left\|\mathbf{x}^{(h-1)}\right\|_F.$$

  because each element from $\mathbf{x}^{(h-1)}$ at least appears once and at most appears $q$ times. In practice, $q$ is often small like $3 \times 3$, so throughout the paper we treat $q$ as a constant in our theoretical analysis. To proceed, let $\mathbf{W}^{(h)} \in \mathbb{R}^{d_h \times qd_{h-1}}$, we have

$$\begin{aligned} \mathbf{x}^{(1)} =& \sqrt{\frac{c_\sigma}{m}}\sigma\left(\mathbf{W}^{(h)}\phi(\mathbf{x})\right) \in \mathbb{R}^{m \times p}, \\ \mathbf{x}^{(h)} =& \mathbf{x}^{(h-1)} + \frac{c_\lambda \lambda^{3/2}}{H\sqrt{m}}\sigma\left(\mathbf{W}^{(h)}\phi(\mathbf{x}^{(h-1))}\right) \in \mathbb{R}^{m \times p}, 2 \le h \le H \end{aligned}$$

  where $0 < c_\lambda <$ is a small constant depending only on $\sigma$, $c_\sigma$, and $\lambda$ are constants we will specify in Section 5. Finally, for $\mathbf{a} \in \mathbb{R}^{m \times p}$, the output is defined as

$$f_{cnn}(\mathbf{x}, \mathbf{w}) = \langle \mathbf{a}, \mathbf{x}^{(H)} \rangle.$$

  Note here we use the similar scaling $O(\frac{1}{H\sqrt{m}})$ as ResNet.

To learn the deep neural network, we consider the randomly initialized gradient descent algorithm to find the global minimizer of the empirical loss (1). Specifically, we use the following random initialization scheme. For every level $h \in [H]$, each entry is sampled from a standard Gaussian distribution, $\mathbf{W}^{(h)}_{ij} \sim N(0,1)$. For multilayer fully-connected neural networks and ResNet, each entry of the output layer is sampled from a Rademacher distribution,

$\mathbf{a}_i \sim \text{unif}\left[\{-1,+1\}\right]$. For convolutional ResNet, for $r \in [m]$, we sample $\mathbf{a}_{r,1} \sim \text{unif}\left[\{-1,+1\}\right]$ and set $\mathbf{a}_{r,2} = \mathbf{a}_{r,3} = \cdots = \mathbf{a}_{r,p} = \mathbf{a}_{r,1}$, i.e., we make each channel have the same corresponding output weight. In this paper, we fix the output layer and train lower layers by gradient descent, for $k = 1, 2, \ldots$, and $h \in [H]$

$$\mathbf{W}^{(h)}(k) = \mathbf{W}^{(h)}(k-1) - \eta \frac{\partial L(\mathbf{w}(k-1))}{\partial \mathbf{W}^{(h)}(k-1)},$$

where $\eta > 0$ is the step size. Hoffer et al. [2018] show that fixing the last layer results in little to no loss in test accuracy.

# 3 Warm Up: Deep Fully-connected Neural Networks

In this section, we show gradient descent with a constant positive step size converges to the global minimum with a linear rate. To formally state our assumption, we first define the following population Gram matrices in a recursive way. For $(i,j) \in [n] \times [n]$, let

$$\begin{aligned}
\mathbf{K}_{ij}^{(0)} =& \langle \mathbf{x}_i, \mathbf{x}_j \rangle, \\
\mathbf{K}_{ij}^{(h)} =& c_\sigma \mathbb{E}_{(u,v)^\top \sim N\left(\mathbf{0}, \begin{pmatrix} \mathbf{K}_{ii}^{(h-1)} & \mathbf{K}_{ij}^{(h-1)} \\ \mathbf{K}_{ji}^{(h-1)} & \mathbf{K}_{jj}^{(h-1)} \end{pmatrix}\right)} \left[\sigma(u)\sigma(v)\right], \text{ and for } h \in [H-1], \\
\mathbf{K}_{ij}^{(H)} =& c_\sigma \mathbf{K}_{ij}^{(H-1)} \cdot \mathbb{E}_{(u,v)^\top \sim N\left(\mathbf{0}, \begin{pmatrix} \mathbf{K}_{ii}^{(H-1)} & \mathbf{K}_{ij}^{(H-1)} \\ \mathbf{K}_{ji}^{(H-1)} & \mathbf{K}_{jj}^{(H-1)} \end{pmatrix}\right)} \left[\sigma'(u)\sigma'(v)\right].
\end{aligned} \tag{4}$$

Intuitively, $\mathbf{K}^{(h)}$ represents the Gram matrix after compositing $h$ times kernel induced by the activation function. As will be apparent in the proof, this Gram matrix natural comes up as $m$ goes to infinity. Based on these definitions, we make the following assumption.

**Assumption 3.1.** $\lambda_{\min}(\mathbf{K}^{(H)}) \triangleq \lambda_0 > 0$ *and* $\lambda \triangleq \min_{h \in [H], (i,j) \in [n] \times [n]} \lambda_{\min} \begin{pmatrix} \mathbf{K}_{ii}^{(h)} & \mathbf{K}_{ij}^{(h)} \\ \mathbf{K}_{ji}^{(h)} & \mathbf{K}_{jj}^{(h)} \end{pmatrix} > 0.$

The first part of the assumption states that the Gram matrix at the last layer is strictly positive definite and this least eigenvalue determines the convergence rate of the gradient descent algorithm. Note this assumption is a generalization of the non-degeneracy assumption used in Du et al. [2018b] in which they considered a two-layer neural network and assumed the Gram matrix from the second layer is strictly positive definite. The second part of the assumption states every two by two sub-matrix of every layer has a lower bounded eigenvalue. The inverse of this quantity can be viewed as a measure of the stability of the population Gram matrix. As will be apparent in the proof, this condition guarantees that if $m$ is large, at the initialization phase, our Gram matrix is close to the population Gram matrix.

Now we are ready to state our main theorem for deep multilayer fully-connected neural networks.

**Theorem 3.1** (Convergence Rate of Gradient Descent for Deep Fully Connected Neural Networks)**.** *Assume for all* $i \in [n]$*,* $\|\mathbf{x}_i\|_2 = 1$*,* $|y_i| \leq C$ *for some constant* $C$ *and the number of hidden nodes per layer* $m = \Omega\left(\max\{\frac{n^4 2^{O(H)}}{\lambda_0^4}, \frac{n 2^{O(H)}}{\delta}, \frac{n^2 log(Hn^2/\delta)}{\lambda_0^2 \lambda^{3H/2}}\}\right)$*. If we set the step size* $\eta = O\left(\frac{\lambda_0}{n^2 2^{O(H)}}\right)$*, then with probability at least* $1-\delta$ *over the random initialization we have for* $k = 1, 2, \ldots$

$$L(\mathbf{w}(k)) \leq \left(1 - \frac{\eta\lambda_0}{2}\right)^k L(\mathbf{w}(0)).$$

This theorem states that if the width $m$ is large enough and we set step size appropriately then gradient descent converges to the global minimum with zero loss at linear rate. The width depends on $n$, $H$, $\lambda_0$ and $\lambda$. The dependency on $n$ and the least eigenvalue $\lambda_0$ is only polynomial, which is the same as previous work on shallow neural networks [Du et al., 2018b, Li and Liang, 2018, Allen-Zhu et al., 2018a]. However, the dependency on the number of layers $H$ and stability parameter of the Gram matrices $\frac{1}{\lambda}$ is exponential. As will be clear in Section 6 and proofs, this exponential dependency results from the amplification factor of multilayer fully-connected neural network architecture.

# 4 ResNet

In this section we consider the convergence of gradient descent for training a ResNet. We focus on how much over-parameterization is needed to ensure the global convergence of gradient descent. Similar to the previous section, we define the population Gram matrices, for $(i, j) \in [n] \times [n]$

$$\begin{aligned}
&\mathbf{K}_{ij}^{(0)} = \langle \mathbf{x}_i, \mathbf{x}_j \rangle, \\
&\mathbf{K}_{ij}^{(1)} = \mathbb{E}_{(u,v)^\top \sim N\left(\mathbf{0}, \begin{pmatrix} \mathbf{K}_{ii}^{(0)} & \mathbf{K}_{ij}^{(0)} \\ \mathbf{K}_{ji}^{(0)} & \mathbf{K}_{jj}^{(0)} \end{pmatrix}\right)} c_\sigma \sigma(u)\sigma(v), \mathbf{b}_i^{(1)} = \sqrt{c_\sigma}\mathbb{E}_{u\sim N(0,\mathbf{K}_{ii}^{(0)})}\left[\sigma(u)\right], \\
&\mathbf{K}_{ij}^{(h)} = \mathbf{K}_{ij}^{(h-1)} + \\
&\mathbb{E}_{(u,v)^\top \sim N\left(\mathbf{0}, \begin{pmatrix} \mathbf{K}_{ii}^{(h-1)} & \mathbf{K}_{ij}^{(h-1)} \\ \mathbf{K}_{ji}^{(h-1)} & \mathbf{K}_{jj}^{(h-1)} \end{pmatrix}\right)} \left[\frac{c_\lambda \lambda^{3/2} \mathbf{b}_i^{(h-1)} \sigma(u)}{H} + \frac{c_\lambda \lambda^{3/2} \mathbf{b}_j^{(h-1)} \sigma(v)}{H} + \frac{c_\lambda^2 \lambda^3 \sigma(u)\sigma(v)}{H^2}\right], \\
&\mathbf{b}_i^{(h)} = \mathbf{b}_i^{(h-1)} + \frac{c_\lambda \lambda^{3/2}}{H}\mathbb{E}_{u\sim N(0,\mathbf{K}_{ii}^{(h-1)})}\left[\sigma(u)\right] \text{ for } 2 \leq h \leq H-1, \\
&\mathbf{K}_{ij}^{(H)} = \frac{c_\lambda^2 \lambda^3}{H^2} \mathbf{K}_{ij}^{(H-1)} \cdot \mathbb{E}_{(u,v)^\top \sim N\left(\mathbf{0}, \begin{pmatrix} \mathbf{K}_{ii}^{(H-1)} & \mathbf{K}_{ij}^{(H-1)} \\ \mathbf{K}_{ji}^{(H-1)} & \mathbf{K}_{jj}^{(H-1)} \end{pmatrix}\right)} \left[\sigma'(u)\sigma'(v)\right].
\end{aligned} \tag{5}$$

These are the asymptotic Gram matrices as $m$ goes to infinity. We make the following assumption which determines the convergence rate and the amount of over-parameterization.

**Assumption 4.1.** $\lambda_{\min}(\mathbf{K}^{(H)}) \triangleq \lambda_0 > 0$ *and* $\lambda \triangleq \min_{(i,j)\in[n]\times[n]} \lambda_{\min}\begin{pmatrix} \mathbf{K}_{ii}^{(0)} & \mathbf{K}_{ij}^{(0)} \\ \mathbf{K}_{ji}^{(0)} & \mathbf{K}_{jj}^{(0)} \end{pmatrix} > 0.$

Note $\lambda$ defined here is different from that of the deep fully-connected neural network because here $\lambda$ only depends on $\mathbf{K}^{(0)}$. In general, unless there are two data points that are parallel, $\lambda$ is always positive here. Now we are ready to state our main theorem for ResNet.

**Theorem 4.1** (Convergence Rate of Gradient Descent for ResNet)**.** *Assume for all* $i \in [n]$, $\|\mathbf{x}_i\|_2 = 1$, $|y_i| \le C$ *for some constant* $C$ *and the number of hidden nodes per layer* $m = \Omega\left(\max\{\frac{n^4}{\lambda_0^4 H^6}, \frac{n^2}{\lambda_0^2 H^2}, \frac{n}{\delta}, \frac{n^2 log(Hn^2/\delta)}{\lambda_0^2}\}\right)$. *If we set the step size* $\eta = O\left(\frac{\lambda_0 H^2}{n^2}\right)$, *then with probability at least* $1-\delta$ *over the random initialization we have for* $k = 1, 2, \ldots$

$$L(\mathbf{w}(k)) \le \left(1 - \frac{\eta\lambda_0}{2}\right)^k L(\mathbf{w}(0)).$$

In sharp contrast to Theorem 3.1, this theorem is fully polynomial in the sense that both the number of neurons and the convergence rate is polynomially in $n$ and $H$. Note the amount of over-parameterization depends on $\lambda_0$ which is the smallest eigenvalue of the $H$-th layer's Gram matrix. In Section D.1, we show $\lambda_0 \sim \frac{1}{H^2}$, and does not depend inverse exponentially on $H$. The main reason that we do not have any exponential factor here is that the skip connection block makes the overall architecture more stable in both the initialization phase and the training phase. See Section B for details.

# 5 Convolutional ResNet

In this section we present convergence result of gradient descent for convolutional ResNet. Again, we focus on how much over-parameterization is needed to ensure the global convergence of gradient descent. Similar to previous sections, we first define the population Gram matrix in order to formally state our assumptions. These are the asymptotic Gram matrices as $m$ goes to infinity.

$$\begin{aligned} \mathbf{K}_{ij}^{(0)} =& \phi_1\left(\mathbf{x}_i\right)^\top \phi_1\left(\mathbf{x}_j\right) \in \mathbb{R}^{p\times p}, \\ \mathbf{K}_{ij}^{(1)} =& \mathbb{E}_{(\mathbf{U},\mathbf{V})\sim N\left(\mathbf{0}, \begin{pmatrix} \mathbf{K}_{ii}^{(0)} & \mathbf{K}_{ij}^{(0)} \\ \mathbf{K}_{ji}^{(0)} & \mathbf{K}_{jj}^{(0)} \end{pmatrix}\right)} c_\sigma \sigma\left(\mathbf{U}\right)\sigma\left(\mathbf{V}\right), \mathbf{b}_i^{(1)} = \sqrt{c_\sigma}\mathbb{E}_{\mathbf{U}\sim N\left(\mathbf{0},\mathbf{K}_{ii}^{(0)}\right)}\left[\sigma\left(\mathbf{U}\right)\right], \end{aligned}$$

$$\begin{aligned} &\mathbf{H}_{ij}^{(h)} = \mathbf{K}_{ij}^{(h-1)} + \\ &\mathbb{E}_{(\mathbf{U},\mathbf{V})\sim N\left(\mathbf{0}, \begin{pmatrix} \mathbf{K}_{ii}^{(h-1)} & \mathbf{K}_{ij}^{(h-1)} \\ \mathbf{K}_{ji}^{(h-1)} & \mathbf{K}_{jj}^{(h-1)} \end{pmatrix}\right)} \left[\frac{c_\lambda \lambda^{3/2}\mathbf{b}_i^{(h-1)}\sigma\left(\mathbf{U}\right)}{H} + \frac{c_\lambda \lambda^{3/2}\mathbf{b}_j^{(h-1)}\sigma\left(\mathbf{V}\right)}{H} + \frac{c_\lambda^2 \lambda^3 \sigma\left(\mathbf{U}\right)\sigma\left(\mathbf{V}\right)}{H^2}\right], \end{aligned}$$

$$\mathbf{K}_{ij,lr}^{(h)} = \sum_{s\in D_l^{(h)}} \sum_{t\in D_r^{(h)}} \mathbf{H}_{ij,st}^{(h)},$$

$$\mathbf{b}_i^{(h)} = \mathbf{b}_i^{(h-1)} + \frac{c_\lambda \lambda^{3/2}}{H}\mathbb{E}_{\mathbf{U}\sim N\left(\mathbf{0},\mathbf{K}_{ii}^{(h-1)}\right)}\left[\sigma\left(\mathbf{U}\right)\right] \text{ for } 2 \leq h \leq H-1,$$

$$\mathbf{K}_{ij}^{(H)} = \frac{c_\lambda^2 \lambda^3}{H^2}\langle \mathbf{K}_{ij}^{(H-1)}, \mathbb{E}_{(\mathbf{U},\mathbf{V})\sim N\left(\mathbf{0},\begin{pmatrix}\mathbf{K}_{ii}^{(h-1)} & \mathbf{K}_{ij}^{(h-1)} \\ \mathbf{K}_{ji}^{(h-1)} & \mathbf{K}_{jj}^{(h-1)}\end{pmatrix}\right)}\left[\sigma'(\mathbf{U})\sigma'(\mathbf{V})\right]\rangle \in \mathbb{R}, \tag{6}$$

where $D_l^{(h)} \triangleq \{s : \mathbf{x}_{:,s}^{(h-1)} \in \text{the } l^{th} \text{ patch}\}$. We make the following assumption which determines the convergence rate and the amount of over-parameterization.

**Assumption 5.1.** *$\lambda_{\min}(\mathbf{K}^{(H)}) \triangleq \lambda_0 > 0$ and $\lambda \triangleq \min_{(i,j)\in[n]\times[n]} \lambda_{\min}\begin{pmatrix}\mathbf{K}_{ii}^{(0)} & \mathbf{K}_{ij}^{(0)} \\ \mathbf{K}_{ji}^{(0)} & \mathbf{K}_{jj}^{(0)}\end{pmatrix} > 0.$*

Note this assumption is basically the same as Assumption 4.1. Now we state our main convergence theorem for the convolutional ResNet.

**Theorem 5.1** (Convergence Rate of Gradient Descent for Convolutional ResNet)**.** *Assume for all $i \in [n]$, $\|\mathbf{x}_i\|_F = 1$, $|y_i| \leq C$ for some constant $C$ and the number of hidden nodes per layer $m = \Omega\left(\max\{\frac{n^4}{\lambda_0^4 H^6}, \frac{n^4}{\lambda_0^4 H^2}, \frac{n}{\delta}, \frac{n^2 log(Hn^2/\delta)}{\lambda_0^2}\}\text{poly}(p)\right)$. If we set the step size $\eta = O\left(\frac{\lambda_0 H^2}{n^2 \text{poly}(p)}\right)$, then with probability at least $1-\delta$ over the random initialization we have for $k = 1, 2, \ldots$*

$$L(\mathbf{w}(k)) \leq \left(1 - \frac{\eta\lambda_0}{2}\right)^k L(\mathbf{w}(0)).$$

This theorem is similar to that of ResNet. The number of neurons required per layer is only polynomial in the depth and the number of data points and step size is only polynomially small. The analysis is similar to ResNet and we refer readers to Section C for details.

# 6 Proof Sketch

In this section we develop a unified proof strategy for proving Theorem 3.1, 4.1 and 5.1. First, following Du et al. [2018b], we define the individual prediction at the $k$-th iteration

$$u_i(k) = f(\mathbf{w}(k), \mathbf{x}_i)$$

and we denote $\mathbf{u}(k) = (u_1(k), \ldots, u_n(k))^\top$. Note with this notation, we can write the loss as

$$L(\mathbf{w}(k)) = \frac{1}{2}\left\|\mathbf{y} - \mathbf{u}(k)\right\|_2^2.$$

Our induction hypothesis is just the following convergence rate of empirical loss.

**Condition 6.1.** *At the $k$-th iteration, we have*

$$\left\|\mathbf{y} - \mathbf{u}(k)\right\|_2^2 \leq (1 - \frac{\eta\lambda_0}{2})^k \left\|\mathbf{y} - \mathbf{u}(0)\right\|_2^2.$$

Note this condition implies the conclusions we want to prove. To prove Condition 6.1, we consider one iteration on the loss function.

$$\begin{aligned}
&\left\|\mathbf{y}-\mathbf{u}(k+1)\right\|_2^2 \\
=&\left\|\mathbf{y}-\mathbf{u}(k)-\left(\mathbf{u}(k+1)-\mathbf{u}(k)\right)\right\|_2^2 \\
=&\left\|\mathbf{y}-\mathbf{u}(k)\right\|_2^2-2\left(\mathbf{y}-\mathbf{u}(k)\right)^\top\left(\mathbf{u}(k+1)-\mathbf{u}(k)\right)+\left\|\mathbf{u}(k+1)-\mathbf{u}(k)\right\|_2^2. \qquad (7)
\end{aligned}$$

This equation shows if $2\left(\mathbf{y}-\mathbf{u}(k)\right)^\top\left(\mathbf{u}(k+1)-\mathbf{u}(k)\right) > \left\|\mathbf{u}(k+1)-\mathbf{u}(k)\right\|_2^2$, the loss decreases. Note both terms involves $\mathbf{u}(k+1)-\mathbf{u}(k)$, which we now more carefully analyze. To simplify notations, we define

$$u_i'(\mathbf{w}) \triangleq \frac{\partial u_i}{\partial \mathbf{w}}, \qquad u_i'^{(h)}(\mathbf{w}) \triangleq \frac{\partial u_i}{\partial \mathbf{W}^{(h)}} \quad \text{and} \quad L'(\mathbf{w}) = \frac{\partial L(\mathbf{w})}{\partial \mathbf{w}}, \quad L'^{(h)}(\mathbf{W}^{(h)}) = \frac{\partial L(\mathbf{w})}{\partial \mathbf{W}^{(h)}}.$$

We look one coordinate of $\mathbf{u}(k+1)-\mathbf{u}(k)$.

Using Taylor expansion, we have

$$\begin{aligned}
&u_i(k+1)-u_i(k) \\
=&u_i\left(\mathbf{w}(k)-\eta L'(\mathbf{w}(k))\right)-u_i\left(\mathbf{w}(k)\right) \\
=&-\int_{s=0}^{\eta}\langle L'(\mathbf{w}(k)), u_i'\left(\mathbf{w}(k)-sL'(\mathbf{w}(k))\right)\rangle ds \\
=&-\int_{s=0}^{\eta}\langle L'(\mathbf{w}(k)), u_i'\left(\mathbf{w}(k)\right)\rangle ds+\int_{s=0}^{\eta}\langle L'(\mathbf{w}(k)), u_i'\left(\mathbf{w}(k)\right)-u_i'\left(\mathbf{w}(k)-sL'(\mathbf{w}(k))\right)\rangle ds \\
\triangleq& I_1^i(k)+I_2^i(k).
\end{aligned}$$

Denote $\mathbf{I}_1(k)=\left(I_1^1(k),\ldots,I_1^n(k)\right)^\top$ and $\mathbf{I}_2(k)=\left(I_2^1(k),\ldots,I_2^n(k)\right)^\top$ and so $\mathbf{u}(k+1)-\mathbf{u}(k)=\mathbf{I}_1(k)+\mathbf{I}_2(k)$. We will show the $\mathbf{I}_1(k)$ term, which is proportional to $\eta$, drives the loss function to decrease and the $\mathbf{I}_2(k)$ term, which is a perturbation term but it is proportional to $\eta^2$ so it is small. We further unpack the $I_1^i(k)$ term,

$$\begin{aligned}
I_1^i =& -\eta\langle L'(\mathbf{w}(k)), u_i'\left(\mathbf{w}(k)\right)\rangle \\
=& -\eta\sum_{j=1}^{n}(u_j-y_j)\langle u_j'(\mathbf{w}(k)), u_i'\left(\mathbf{w}(k)\right)\rangle \\
\triangleq& -\eta\sum_{j=1}^{n}(u_j-y_j)\mathbf{G}_{ij}(k)
\end{aligned}$$

where $\mathbf{G}(k)\in\mathbb{R}^{n\times n}$ with $(i,j)$-th entry being $\langle u_j'(\mathbf{w}(k)), u_i'\left(\mathbf{w}(k)\right)\rangle$. Now we unpack this $\mathbf{G}(k)$ matrix. Recall

$$\langle u_j'(\mathbf{w}(k)), u_i'\left(\mathbf{w}(k)\right)\rangle = \sum_{h=1}^{H}\langle u_j'^{(h)}(\mathbf{w}(k)), u_i'^{(h)}(\mathbf{w}(k))\rangle.$$

Define $\mathbf{G}^{(h)}(k) \in \mathbb{R}^{n\times n}$ with $\mathbf{G}_{ij}^{(h)}(k) = \langle u_j'^{(h)}(\mathbf{w}(k)), u_i'^{(h)}(\mathbf{w}(k))\rangle$. Note by definition $\mathbf{G}^{(h)}(k)$ is a Gram matrix and thus it is positive semi-definite. Therefore we have $\mathbf{G}(k) = \sum_{h=1}^{H} \mathbf{G}^{(h)}(k) \succcurlyeq \mathbf{G}^{(H)}(k)$

$$\mathbf{G}_{i,j}^{(H)}(k) = (\mathbf{x}_i^{(H-1)}(k))^\top \mathbf{x}_j^{(H-1)}(k) \cdot \frac{c_\sigma}{m} \sum_{r=1}^{m} a_r^2 \sigma'((\mathbf{w}_r^{(H)}(k))^\top \mathbf{x}_i^{(H-1)}(k)) \sigma'((\mathbf{w}_r^{(H)}(k))^\top \mathbf{x}_j^{(H-1)}(k)).$$

Now we analyze $\mathbf{I}_1(k)$. We can write $\mathbf{I}_1$ in a more compact form with $\mathbf{G}(k)$.

$$\mathbf{I}_1(k) = -\eta \left(\mathbf{u}(k) - \mathbf{y}\right) \sum_{h=1}^{H} \mathbf{G}^{(h)}(k).$$

Now observe that

$$\begin{aligned}(\mathbf{y} - \mathbf{u}(k))^\top \mathbf{I}_1(k) =& \eta \left(\mathbf{y} - \mathbf{u}(k)\right)^\top \mathbf{G}(k)(\mathbf{y} - \mathbf{u}(k)) \\ \geq& \lambda_{\min}\left(\mathbf{G}(k)\right) \left\|\mathbf{y} - \mathbf{u}(k)\right\|_2^2 \\ \geq& \lambda_{\min}\left(\mathbf{G}^{(H)}(k)\right) \left\|\mathbf{y} - \mathbf{u}(k)\right\|_2^2\end{aligned}$$

Now recall the progress of loss function in Equation (7):

$$\begin{aligned}&\left\|\mathbf{y} - \mathbf{u}(k+1)\right\|_2^2 \\ =& \left\|\mathbf{y} - \mathbf{u}(k)\right\|_2^2 - 2\left(\mathbf{y} - \mathbf{u}(k)\right)^\top \mathbf{I}_1(k) - 2\left(\mathbf{y} - \mathbf{u}(k)\right)^\top \mathbf{I}_2(k) + \left\|\mathbf{u}(k+1) - \mathbf{u}(k)\right\|_2^2 \\ \leq& \left(1 - \eta\lambda_{\min}\left(\mathbf{G}^{(H)}(k)\right)\right) \left\|\mathbf{y} - \mathbf{u}(k)\right\|_2^2 - 2\left(\mathbf{y} - \mathbf{u}(k)\right)^\top \mathbf{I}_2(k) + \left\|\mathbf{u}(k+1) - \mathbf{u}(k)\right\|_2^2.\end{aligned}$$

For the perturbation terms, through standard calculations, we can show both $-2\left(\mathbf{y} - \mathbf{u}(k)\right)^\top \mathbf{I}_2(k)$ and $\left\|\mathbf{u}(k+1) - \mathbf{u}(k)\right\|_2$ are proportional to $\eta^2 \left\|\mathbf{y} - \mathbf{u}(k)\right\|_2^2$ so if we set $\eta$ sufficiently small, this term is smaller than $\eta\lambda_{\min}\left(\mathbf{G}^{(H)}(k)\right) \left\|\mathbf{y} - \mathbf{u}(k)\right\|_2^2$ and thus the loss function decreases with a linear rate. See Lemmas A.6 and A.7.

Therefore, to prove the induction hypothesis, it suffices to prove $\lambda_{\min}\left(\mathbf{G}^{(H)}(k)\right) \geq \frac{\lambda_0}{2}$ for $k' = 0, \ldots, k$, where $\lambda_0$ is independent of $m$. To analyze the least eigenvalue, we first look at the initialization. Using assumptions of the population kernel matrix and concentration inequalities, we can show at the beginning $\left\|\mathbf{G}^{(H)}(0) - \mathbf{K}^{(H)}(0)\right\|_2 \leq \frac{1}{4}\lambda_0$, which implies

$$\lambda_{\min}\left(\mathbf{G}^{(H)}(0)\right) \geq \frac{3}{4}\lambda_0.$$

Now for the $k$-th iteration, by matrix perturbation analysis, we know it is sufficient to show $\left\|\mathbf{G}^{(H)}(k) - \mathbf{G}^{(H)}(0)\right\|_2 \leq \frac{1}{4}\lambda_0$. To do this, we use a similar approach as in Du et al. [2018b]. We show as long as $m$ is large enough, every weight matrix is close its initialization in a relative error sense. Ignoring all other parameters except $m$, $\left\|\mathbf{W}^{(h)}(k) - \mathbf{W}^{(h)}(0)\right\|_F \lesssim 1$, and thus the average per-neuron distance from initialization is $\frac{\left\|\mathbf{W}^{(h)}(k) - \mathbf{W}^{(h)}(0)\right\|_F}{\sqrt{m}} \lesssim \frac{1}{\sqrt{m}}$ which tends to zero as $m$ increases. See Lemma A.5 for precise statements with all the dependencies.

This fact in turn shows $\left\|\mathbf{G}^{(h)}(k) - \mathbf{G}^{(h)}(0)\right\|_2$ is small (see Lemma A.4). The main difference from Du et al. [2018b] is that we are considering deep neural networks, and when translating the small deviation, $\left\|\mathbf{W}^{(h)}(k) - \mathbf{W}^{(h)}(0)\right\|_F$ to $\left\|\mathbf{G}^{(h)}(k) - \mathbf{G}^{(h)}(0)\right\|_2$, there is an amplification factor which depends on the neural network architecture.

For deep fully connected neural networks, we show this amplification factor is exponential in $H$. On the other hand, for ResNet and convolutional ResNet we show this amplification factor is only polynomial in $H$. We further show the width $m$ required is proportional to this amplification factor.

# 7 Conclusion

In this paper, we show that gradient descent on deep overparametrized networks can obtain zero training loss. The key technique is to show that the Gram matrix is increasingly stable under overparametrization, and so every step of gradient descent decreases the loss at a geometric rate.

We list some directions for future research:

1. The current paper focuses on the train loss, but does not address the test loss. It would be an important problem to show that gradient descent can also find solutions of low test loss. In particular, existing work only demonstrate that gradient descent works under the same situations as kernel methods and random feature methods [Daniely, 2017, Li and Liang, 2018].

2. The width of the layers $m$ is polynomial in all the parameters for the ResNet architecture, but still very large. Realistic networks have number of parameters, not width, a large constant multiple of $n$. We consider improving the analysis to cover commonly utilized networks an important open problem.

3. The current analysis is for gradient descent, instead of stochastic gradient descent. We believe the analysis can be extended to stochastic gradient, while maintaining the linear convergence rate.

# 8 Acknowledgment

We thank Lijie Chen and Ruosong Wang for useful discussions.

# References

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

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. *arXiv preprint arXiv:1811.03962*, 2018b.

Alexandr Andoni, Rina Panigrahy, Gregory Valiant, and Li Zhang. Learning polynomials with neural networks. In *International Conference on Machine Learning*, pages 1908–1916, 2014.

Anonymous. The unreasonable effectiveness of (zero) initialization in deep residual learning. *Openreview* `https://openreview.net/pdf?id=H1gsz30cKX`, 2018.

Alon Brutzkus and Amir Globerson. Globally optimal gradient descent for a ConvNet with gaussian inputs. In *International Conference on Machine Learning*, pages 605–614, 2017.

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

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

Simon S Du and Jason D Lee. On the power of over-parametrization in neural networks with quadratic activation. *Proceedings of the 35th International Conference on Machine Learning*, pages 1329–1338, 2018.

Simon S Du, Chi Jin, Jason D Lee, Michael I Jordan, Aarti Singh, and Barnabas Poczos. Gradient descent can take exponential time to escape saddle points. In *Advances in Neural Information Processing Systems*, pages 1067–1077, 2017a.

Simon S Du, Jason D Lee, and Yuandong Tian. When is a convolutional filter easy to learn? *arXiv preprint arXiv:1709.06129*, 2017b.

Simon S Du, Jason D Lee, Yuandong Tian, Barnabas Poczos, and Aarti Singh. Gradient descent learns one-hidden-layer cnn: Don't be afraid of spurious local minima. *Proceedings of the 35th International Conference on Machine Learning*, pages 1339–1348, 2018a.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018b.

C Daniel Freeman and Joan Bruna. Topology and geometry of half-rectified network optimization. *arXiv preprint arXiv:1611.01540*, 2016.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points – online stochastic gradient for tensor decomposition. In *Proceedings of The 28th Conference on Learning Theory*, pages 797–842, 2015.

Benjamin D Haeffele and René Vidal. Global optimality in tensor factorization, deep learning, and beyond. *arXiv preprint arXiv:1506.07540*, 2015.

Moritz Hardt and Tengyu Ma. Identity matters in deep learning. *arXiv preprint arXiv:1611.04231*, 2016.

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.

Elad Hoffer, Itay Hubara, and Daniel Soudry. Fix your classifier: the marginal value of training the last weight layer. *arXiv preprint arXiv:1801.04540*, 2018.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, and Michael I. Jordan. How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1724–1732, 2017.

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

Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on Learning Theory*, pages 1246–1257, 2016.

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

Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with relu activation. In *Advances in Neural Information Processing Systems*, pages 597–607, 2017.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layers neural networks. *Proceedings of the National Academy of Sciences*, pages E7665–E7671, 2018.

Quynh Nguyen and Matthias Hein. The loss surface of deep and wide neural networks. In *International Conference on Machine Learning*, pages 2603–2612, 2017.

Itay Safran and Ohad Shamir. On the quality of the initial basin in overspecified neural networks. In *International Conference on Machine Learning*, pages 774–782, 2016.

Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer ReLU neural networks. *In International Conference on Machine Learning*, pages 4433–4441, 2018.

Mahdi Soltanolkotabi. Learning ReLus via gradient descent. In *Advances in Neural Information Processing Systems*, pages 2007–2017, 2017.

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.

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

Daniel Soudry and Elad Hoffer. Exponentially vanishing sub-optimal local minima in multilayer neural networks. *arXiv preprint arXiv:1702.05777*, 2017.

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*, pages 3404–3413, 2017.

Luca Venturi, Afonso Bandeira, and Joan Bruna. Neural networks with finite intrinsic dimension have no spurious valleys. *arXiv preprint arXiv:1802.06384*, 2018.

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

Colin Wei, Jason D Lee, Qiang Liu, and Tengyu Ma. On the margin theory of feedforward neural networks. *arXiv preprint arXiv:1810.05369*, 2018.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *NIN*, 8:35–67, 2016.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Xiao Zhang, Yaodong Yu, Lingxiao Wang, and Quanquan Gu. Learning one-hidden-layer relu networks via gradient descent. *arXiv preprint arXiv:1806.07808*, 2018.

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

Kai Zhong, Zhao Song, Prateek Jain, Peter L Bartlett, and Inderjit S Dhillon. Recovery guarantees for one-hidden-layer neural networks. *arXiv preprint arXiv:1706.03175*, 2017b.

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

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Stochastic gradient descent optimizes over-parameterized deep relu networks. *arXiv preprint arXiv:1811.08888*, 2018.

# Appendix

In the proof we will use the geometric series function $g_\alpha(n) = \sum_{i=0}^{n-1} \alpha^i$ extensively. Some constants we will define below may be different for different network structures, such as $c_x$, $c_{w,0}$ and $c_{x,0}$. We will also use $c$ to denote a small enough constant, which may be different in different lemmas.

# A Proofs for Section 3

We first derive the formula of the gradient for the multilayer fully connected neural network is

$$\frac{\partial L(\mathbf{w})}{\partial \mathbf{W}^{(h)}} = \sum_{i=1}^{n} \left(f(\mathbf{x}_i, \mathbf{w}, \mathbf{a}) - y_i\right) \mathbf{x}_i^{(h-1)} \mathbf{a}^\top \mathbf{J}_i^{(H)} \mathbf{W}^{(H)} \cdots \mathbf{W}^{(h+1)} \mathbf{J}_i^{(h)}$$

where

$$\mathbf{J}^{(h')} \triangleq \mathbf{diag}\left(\sigma'\left((\mathbf{w}_1^{(h')})^\top \mathbf{x}^{(h'-1)}\right), \ldots, \sigma'\left((\mathbf{w}_m^{(h')})^\top \mathbf{x}^{(h'-1)}\right)\right) \in \mathbb{R}^{m \times m}.$$

are the derivative matrices induced by the activation function and

$$\mathbf{x}^{(h')} = \sqrt{\frac{c_\sigma}{m}} \sigma\left(\mathbf{W}^{(h')} \mathbf{x}^{(h'-1)}\right).$$

is the output of the $h'$-th layer.

Through standard calculation, we can get the expression of $\mathbf{G}_{i,j}^{(H)}$ of the following form

$$\mathbf{G}_{i,j}^{(H)} = (\mathbf{x}_i^{(H-1)})^\top \mathbf{x}_j^{(H-1)} \cdot \frac{c_\sigma}{m} \sum_{r=1}^{m} a_r^2 \sigma'((\mathbf{w}_r^{(H)})^\top \mathbf{x}_i^{(H-1)}) \sigma'((\mathbf{w}_r^{(H)})^\top \mathbf{x}_j^{(H-1)}).$$

We first present a lemma shows with high probability the feature of each layer is approximately normalized.

**Lemma A.1** (Lemma on Initialization Norms)**.** *If $\sigma(\cdot)$ is $L-$Lipschitz and $m = \Omega\left(\frac{nHg_C(H)^2}{\delta}\right)$, where $C \triangleq c_\sigma L\left(2\left|\sigma(0)\right|\sqrt{\frac{2}{\pi}} + 2L\right)$, then with probability at least $1-\delta$ over random initialization, for every $h \in [H]$ and $i \in [n]$, we have*

$$\frac{1}{c_{x,0}} \le \left\|\mathbf{x}_i^{(h)}(0)\right\|_2 \le c_{x,0}$$

*where $c_{x,0} = 2$.*

We follow the proof sketch described in Section 6. We first analyze the spectral property of $\mathbf{G}^{(H)}(0)$ at the initialization phase. The following lemma lower bounds its least eigenvalue. This lemma is a direct consequence of Theorem D.1 and Remark D.4.

**Lemma A.2** (Least Eigenvalue at the Initialization)**.** *If* $m = \Omega\left(\text{poly}(n, \frac{1}{\lambda_0}) \cdot (\frac{1}{\lambda})^{O(H)}\right)$*, we have*

$$\lambda_{\min}(\mathbf{G}^{(H)}(0)) \geq \frac{3}{4}\lambda_0.$$

Now we proceed to analyze the training process. We prove the following lemma which characterizes how the perturbation from weight matrices propagates to the input of each layer. This Lemma is used to prove the subsequent lemmas.

**Lemma A.3.** *Suppose for every* $h \in [H]$*,* $\left\|\mathbf{W}^{(h)}(0)\right\|_2 \leq c_{w,0}\sqrt{m}$*,* $\left\|\mathbf{x}^{(h)}(0)\right\|_2 \leq c_{x,0}$ *and* $\left\|\mathbf{W}^{(h)}(k) - \mathbf{W}^{(h)}(0)\right\|_F \leq \sqrt{m}R$ *for some constant* $c_{w,0}, c_{x,0} > 0$ *and* $R \leq c_{w,0}$*. If* $\sigma(\cdot)$ *is* $L-$*Lipschitz, we have*

$$\left\|\mathbf{x}^{(h)}(k) - \mathbf{x}^{(h)}(0)\right\|_2 \leq \sqrt{c_\sigma} L c_{x,0} g_{c_x}(h) R$$

*where* $c_x = 2\sqrt{c_\sigma} L c_{w,0}$*.*

Here the assumption of $\left\|\mathbf{W}^{(h)}(0)\right\|_2 \leq c_{w,0}\sqrt{m}$ can be shown using Lemma E.4 and taking union bound over $h \in [H]$. Next, we show with high probability over random initialization, perturbation in weight matrices leads to small perturbation in the Gram matrix.

**Lemma A.4.** *Suppose* $\sigma(\cdot)$ *is* $L-$*Lipschitz and* $\beta-$*smooth. Suppose for* $h \in [H]$*,* $\left\|\mathbf{W}^{(h)}(0)\right\|_2 \leq c_{w,0}\sqrt{m}$*,* $\frac{1}{c_{x,0}} \leq \left\|\mathbf{x}^{(h)}(0)\right\|_2 \leq c_{x,0}$*, if* $\left\|\mathbf{W}^{(h)}(k) - \mathbf{W}^{(h)}(0)\right\|_F \leq \sqrt{m}R$ *where* $R \leq c g_{c_x}(H)^{-1}\lambda_0 n^{-1}$ *and* $R \leq c g_{c_x}(H)^{-1}$ *for some small constant* $c$ *and* $c_x = 2\sqrt{c_\sigma}Lc_{w,0}$*, we have*

$$\left\|\mathbf{G}^{(H)}(k) - \mathbf{G}^{(H)}(0)\right\|_2 \leq \frac{\lambda_0}{4}.$$

The following lemma shows if the induction holds, we have every weight matrix close to its initialization.

**Lemma A.5.** *If Condition 6.1 holds for* $k' = 1, \ldots, k$*, we have for any* $s = 1, \ldots, k+1$

$$\begin{aligned}\left\|\mathbf{W}^{(h)}(s) - \mathbf{W}^{(h)}(0)\right\|_F &\leq R'\sqrt{m}\\ \left\|\mathbf{W}^{(h)}(s) - \mathbf{W}^{(h)}(s-1)\right\|_F &\leq \eta Q'(s-1)\end{aligned}$$

*where* $R' = \frac{8c_{x,0}(c_x)^H\sqrt{n}\|\mathbf{y}-\mathbf{u}(0)\|_2}{\lambda_0\sqrt{m}} \leq c g_{c_x}(H)^{-1}$ *for some small constant* $c$ *with* $c_x = \max\{2\sqrt{c_\sigma}Lc_{w,0}, 1\}$ *and* $Q'(s) = 2c_{x,0}\left(c_x\right)^H\sqrt{n}\left\|\mathbf{y} - \mathbf{u}(s)\right\|_2$

Now we proceed to analyze the perturbation terms.

**Lemma A.6.** *If Condition 6.1 holds for* $k' = 1, \ldots, k$*, suppose* $\eta \leq c\lambda_0\left(n^2H^2(c_x)^{3H}g_{2c_x}(H)\right)^{-1}$ *for some small constant* $c$*, we have*

$$\left\|\mathbf{I}_2(k)\right\|_2 \leq \frac{1}{8}\eta\lambda_0\left\|\mathbf{y} - \mathbf{u}(k)\right\|_2.$$

**Lemma A.7.** *If Condition 6.1 holds for $k' = 1, \ldots, k$, suppose $\eta \le c\lambda_0 \left(n^2 H^2 (c_x)^{2H} g_{2c_x}(H)\right)^{-1}$ for some small constant c, then we have $\left\|\mathbf{u}(k+1) - \mathbf{u}(k)\right\|_2^2 \le \frac{1}{8}\eta\lambda_0 \left\|\mathbf{y} - \mathbf{u}(k)\right\|_2^2$.*

We now proceed with the proof of Theorem 3.1. By induction, we assume Condition 6.1 for all $k' < k$. Using Lemma A.5, this establishes

$$\begin{aligned}\left\|\mathbf{W}^{(h)}(k) - \mathbf{W}^{(h)}(0)\right\|_F &\le R'\sqrt{m} \\ &\le R\sqrt{m} && \text{( using the choice of } m \text{ in the theorem.)}\end{aligned}$$

By Lemma A.4, this establishes $\lambda_{\min}(\mathbf{G}^{(H)}(k)) \ge \frac{\lambda_0}{2}$.

With these estimates in hand, we are ready to prove the induction hypothesis of Condition 6.1.

$$\begin{aligned}&\left\|\mathbf{y} - \mathbf{u}(k+1)\right\|_2^2 \\ &= \left\|\mathbf{y} - \mathbf{u}(k)\right\|_2^2 - 2\eta\left(\mathbf{y} - \mathbf{u}(k)\right)^\top \sum_{h=1}^{H} \mathbf{G}^{(h)}(k)\left(\mathbf{y} - \mathbf{u}(k)\right) - 2\left(\mathbf{y} - \mathbf{u}(k)\right)^\top \mathbf{I}_2 + \left\|\mathbf{u}(k+1) - \mathbf{u}(k)\right\|_2^2 \quad \le \|\mathbf{y} - \\ &\le (1 - \eta\lambda_0)\left\|\mathbf{y} - \mathbf{u}(k)\right\|_2^2 - 2\left(\mathbf{y} - \mathbf{u}(k)\right)^\top \mathbf{I}_2 + \left\|\mathbf{u}(k+1) - \mathbf{u}(k)\right\|_2^2 \\ &\le (1 - \frac{\eta\lambda_0}{2})\left\|\mathbf{y} - \mathbf{u}(k)\right\|_2^2.\end{aligned}$$

The first inequality drops the positive terms $\left(\mathbf{y} - \mathbf{u}(k)\right)^\top \sum_{h=1}^{H-1} \mathbf{G}^{(h)}(k)\left(\mathbf{y} - \mathbf{u}(k)\right)$. The second inequality uses the argument above that establishes $\lambda_{\min}(\mathbf{G}^{(H)}(k)) \ge \frac{\lambda_0}{2}$. The third inequality uses Lemmas A.6 and A.7.

## A.1 Proofs of Lemmas

*Proof of Lemma A.1.* We will bound $\left\|\mathbf{x}_i^{(h)}(0)\right\|_2$ by induction on layers. The induction hypothesis is that with probability at least $1 - (h-1)\frac{\delta}{nH}$ over $\mathbf{W}^{(1)}(0), \ldots, \mathbf{W}^{(h-1)}(0)$, for every $1 \le h' \le h-1$, $\frac{1}{2} \le 1 - \frac{g_C(h')}{2g_C(H)} \le \left\|\mathbf{x}_i^{(h')}(0)\right\|_2 \le 1 + \frac{g_C(h')}{2g_C(H)} \le 2$. Note that it is true for $h = 1$. We calculate the expectation of $\left\|\mathbf{x}_i^{(h)}(0)\right\|_2^2$ over the randomness from $\mathbf{W}^h(0)$. Recall

$$\left\|\mathbf{x}_i^{(h)}(0)\right\|_2^2 = \frac{c_\sigma}{m}\sum_{r=1}^{m} \sigma\left(\mathbf{w}_r^{(h)}(0)^\top \mathbf{x}_i^{(h-1)}(0)\right)^2.$$

Therefore we have

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{x}_i^{(h)}(0)\right\|_2^2\right] =& c_\sigma \mathbb{E}\left[\sigma\left(\mathbf{w}_r^{(h)}(0)^\top \mathbf{x}_i^{(h-1)}(0)\right)^2\right] \\ =& c_\sigma \mathbb{E}_{X\sim N(0,1)}\sigma(\left\|\mathbf{x}_i^{(h-1)}(0)\right\|_2 X)^2.\end{aligned}$$

Note that $\sigma(\cdot)$ is $L-$Lipschitz, for any $\frac{1}{2} \le \alpha \le 2$, we have

$$\left|\mathbb{E}_{X\sim N(0,1)}\sigma(\alpha X)^2 - \mathbb{E}_{X\sim N(0,1)}\sigma(X)^2\right|$$

$$
\begin{aligned}
&\leq \mathbb{E}_{X\sim N(0,1)} \left|\sigma(\alpha X)^2 - \sigma(X)^2\right| \\
&\leq L\left|\alpha-1\right| \mathbb{E}_{X\sim N(0,1)} \left|X\left(\sigma(\alpha X)+\sigma(X)\right)\right| \\
&\leq L\left|\alpha-1\right| \mathbb{E}_{X\sim N(0,1)} \left|X\right| \left(\left|2\sigma(0)\right| + L\left|(\alpha+1)X\right|\right) \\
&\leq L\left|\alpha-1\right| \left(2\left|\sigma(0)\right| \mathbb{E}_{X\sim N(0,1)}\left|X\right| + L\left|\alpha+1\right| \mathbb{E}_{X\sim N(0,1)} X^2\right) \\
&= L\left|\alpha-1\right| \left(2\left|\sigma(0)\right| \sqrt{\frac{2}{\pi}} + L\left|\alpha+1\right|\right) \\
&\leq \frac{C}{c_\sigma}\left|\alpha-1\right|,
\end{aligned}
$$

where $C \triangleq c_\sigma L\left(2\left|\sigma(0)\right|\sqrt{\frac{2}{\pi}} + 2L\right)$, which implies

$$
1 - \frac{C g_C(h-1)}{2g_C(H)} \leq \mathbb{E}\left[\left\|\mathbf{x}_i^{(h)}(0)\right\|_2^2\right] \leq 1 + \frac{C g_C(h-1)}{2g_C(H)}.
$$

For the variance we have

$$
\begin{aligned}
\mathrm{Var}\left[\left\|\mathbf{x}_i^{(h)}(0)\right\|_2^2\right] &= \frac{c_\sigma^2}{m}\mathrm{Var}\left[\sigma\left(\mathbf{w}_r^{(h)}(0)^\top \mathbf{x}_i^{(h-1)}(0)\right)^2\right] \\
&\leq \frac{c_\sigma^2}{m}\mathbb{E}\left[\sigma\left(\mathbf{w}_r^{(h)}(0)^\top \mathbf{x}_i^{(h-1)}(0)\right)^4\right] \\
&\leq \frac{c_\sigma^2}{m}\mathbb{E}\left[\left(\left|\sigma(0)\right| + L\left|\mathbf{w}_r^{(h)}(0)^\top \mathbf{x}_i^{(h-1)}(0)\right|\right)^4\right] \\
&\leq \frac{C_2}{m}.
\end{aligned}
$$

where $C_2 \triangleq \sigma(0)^4 + 8\left|\sigma(0)\right|^3 L\sqrt{2/\pi} + 24\sigma(0)^2 L^2 + 64\sigma(0)L^3\sqrt{2/\pi} + 512L^4$ and the last inequality we used the formula for the first four absolute moments of Gaussian.

Applying Chebyshev's inequality and plugging in our assumption on $m$, we have with probability $1-\frac{\delta}{nH}$ over $\mathbf{W}^{(h)}$,

$$
\left|\left\|\mathbf{x}_i^{(h)}(0)\right\|_2^2 - \mathbb{E}\left\|\mathbf{x}_i^{(h)}(0)\right\|_2^2\right| \leq \frac{1}{2g_C(H)}.
$$

Thus with probability $1 - h\frac{\delta}{nH}$ over $\mathbf{W}^{(1)},\ldots,\mathbf{W}^{(h)}$,

$$
\left|\left\|\mathbf{x}_i^{(h)}(0)\right\|_2 - 1\right| \leq \left|\left\|\mathbf{x}_i^{(h)}(0)\right\|_2^2 - 1\right| \leq \frac{C g_C(h-1)}{2g_C(H)} + \frac{1}{2g(H)} = \frac{g_C(h)}{2g_C(H)}.
$$

Using union bounds over $[n]$, we prove the lemma. □

*Proof of Lemma A.3.* We prove this lemma by induction. Our induction hypothesis is

$$
\left\|\mathbf{x}^{(h)}(k) - \mathbf{x}^{(h)}(0)\right\|_2 \leq \sqrt{c_\sigma} L R c_{x,0} g_{c_x}(h),
$$

where

$$c_x = 2\sqrt{c_\sigma} L c_{w,0}.$$

For $h = 0$, since the input data is fixed, we know the induction hypothesis holds. Now suppose the induction hypothesis holds for $h' = 0, \ldots, h-1$, we consider $h' = h$.

$$\begin{aligned}
\left\|\mathbf{x}^{(h)}(k) - \mathbf{x}^{(h)}(0)\right\|_2 =& \sqrt{\frac{c_\sigma}{m}} \left\|\sigma\left(\mathbf{W}^{(h)}(k)\mathbf{x}^{(h-1)}(k)\right) - \sigma\left(\mathbf{W}^{(h)}(0)\mathbf{x}^{(h-1)}(0)\right)\right\|_2 \\
\leq& \sqrt{\frac{c_\sigma}{m}} \left\|\sigma\left(\mathbf{W}^{(h)}(k)\mathbf{x}^{(h-1)}(k)\right) - \sigma\left(\mathbf{W}^{(h)}(k)\mathbf{x}^{(h-1)}(0)\right)\right\|_2 \\
&+ \sqrt{\frac{c_\sigma}{m}} \left\|\sigma\left(\mathbf{W}^{(h)}(k)\mathbf{x}^{(h-1)}(0)\right) - \sigma\left(\mathbf{W}^{(h)}(0)\mathbf{x}^{(h-1)}(0)\right)\right\|_2 \\
\leq& \sqrt{\frac{c_\sigma}{m}} L \left(\left\|\mathbf{W}^{(h)}(0)\right\|_2 + \left\|\mathbf{W}^{(h)}(k) - \mathbf{W}^{(h)}(0)\right\|_F\right) \cdot \left\|\mathbf{x}^{(h-1)}(k) - \mathbf{x}^{(h-1)}(0)\right\|_2 \\
&+ \sqrt{\frac{c_\sigma}{m}} L \left\|\mathbf{W}^{(h)}(k) - \mathbf{W}^{(h)}(0)\right\|_F \left\|\mathbf{x}^{h-1}(0)\right\|_2 \\
\leq& \sqrt{\frac{c_\sigma}{m}} L \left(c_{w,0}\sqrt{m} + R\sqrt{m}\right) \sqrt{c_\sigma} L R c_{x,0} g_{c_x}(h-1) + \sqrt{\frac{c_\sigma}{m}} L \sqrt{m} R c_{x,0} \\
\leq& \sqrt{c_\sigma} L R c_{x,0} \left(c_x g_{c_x}(h-1) + 1\right) \\
\leq& \sqrt{c_\sigma} L R c_{x,0} g_{c_x}(h).
\end{aligned}$$

□

*Proof of Lemma A.4.* Because Frobenius-norm of a matrix is bigger than the operator norm, it is sufficient to bound $\left\|\mathbf{G}^{(H)}(k) - \mathbf{G}^{(H)}(0)\right\|_F$. For simplicity define $z_{i,r}(k) = \mathbf{w}_r^{(H)}(k)^\top \mathbf{x}_i^{(H-1)}(k)$, we have

$$\begin{aligned}
&\left|\mathbf{G}_{i,j}^{(H)}(k) - \mathbf{G}_{i,j}^{(H)}(0)\right| \\
=& \Big|\mathbf{x}_i^{(H-1)}(k)^\top \mathbf{x}_j^{(H-1)}(k) \frac{c_\sigma}{m} \sum_{r=1}^{m} a_r^2 \sigma'\left(z_{i,r}(k)\right) \sigma'\left(z_{j,r}(k)\right) \\
&- \mathbf{x}_i^{(H-1)}(0)^\top \mathbf{x}_j^{(H-1)}(0) \frac{c_\sigma}{m} \sum_{r=1}^{m} a_r^2 \sigma'\left(z_{i,r}(0)\right) \sigma'\left(z_{j,r}(0)\right) \Big| \\
\leq& \left|\mathbf{x}_i^{(H-1)}(k)^\top \mathbf{x}_j^{(H-1)}(k) - \mathbf{x}_i^{(H-1)}(0)^\top \mathbf{x}_j^{(H-1)}(0)\right| \frac{c_\sigma}{m} \sum_{r=1}^{m} \left|\sigma'\left(z_{i,r}(k)\right) \sigma'\left(z_{j,r}(k)\right)\right| \\
&+ \left|\mathbf{x}_i^{(H-1)}(0)^\top \mathbf{x}_j^{(H-1)}(0)\right| \frac{c_\sigma}{m} \left|\sum_{r=1}^{m} \sigma'\left(z_{i,r}(k)\right) \sigma'\left(z_{j,r}(k)\right) - \sigma'\left(z_{i,r}(0)\right) \sigma'\left(z_{j,r}(0)\right)\right| \\
\leq& L^2 c_\sigma \left|\mathbf{x}_i^{(H-1)}(k)^\top \mathbf{x}_j^{(H-1)}(k) - \mathbf{x}_i^{(H-1)}(0)^\top \mathbf{x}_j^{(H-1)}(0)\right|
\end{aligned}$$

$$
\begin{aligned}
&+c_{x,0}^2\frac{c_\sigma}{m}\left|\sum_{r=1}^{m}\sigma'\left(z_{i,r}(k)\right)\sigma'\left(z_{j,r}(k)\right)-\sigma'\left(z_{i,r}(0)\right)\sigma'\left(z_{j,r}(0)\right)\right|\\
\triangleq& I_1^{i,j}+I_2^{i,j}.
\end{aligned}
$$

For $I_1^{i,j}$, using Lemma A.3, we have

$$
\begin{aligned}
I_1^{i,j}=&L^2c_\sigma\left|\mathbf{x}_i^{(H-1)}(k)^\top\mathbf{x}_j^{(H-1)}(k)-\mathbf{x}_i^{(H-1)}(0)^\top\mathbf{x}_j^{(H-1)}(0)\right|\\
\leq&L^2c_\sigma\left|(\mathbf{x}_i^{(H-1)}(k)-\mathbf{x}_i^{(H-1)}(0))^\top\mathbf{x}_j^{(H-1)}(k)\right|+L^2c_\sigma\left|\mathbf{x}_i^{(H-1)}(0)^\top(\mathbf{x}_j^{(H-1)}(k)-\mathbf{x}_j^{(H-1)}(0))\right|\\
\leq&c_\sigma\sqrt{c_\sigma}L^3c_{x,0}g_{c_x}(H)R\cdot(c_{x,0}+\sqrt{c_\sigma}Lc_{x,0}g_{c_x}(H)R)+c_\sigma\sqrt{c_\sigma}L^3c_{x,0}g_{c_x}(H)Rc_{x,0}\\
\leq&3c_\sigma c_{x,0}^2\sqrt{c_\sigma}L^3g_{c_x}(H)R.
\end{aligned}
$$

For $I_2^{i,j}$, we have

$$
\begin{aligned}
I_2^{i,j}=&c_{x,0}^2\frac{c_\sigma}{m}\left|\sum_{r=1}^{m}\sigma'\left(z_{i,r}(k)\right)\sigma'\left(z_{j,r}(k)\right)-\sigma'\left(z_{i,r}(0)\right)\sigma'\left(z_{j,r}(0)\right)\right|\\
\leq&c_{x,0}^2\frac{c_\sigma}{m}\sum_{r=1}^{m}\left|\left(\sigma'\left(z_{i,r}(k)\right)-\sigma'\left(z_{i,r}(0)\right)\right)\sigma'\left(z_{j,r}(k)\right)\right|+\left|\left(\sigma'\left(z_{j,r}(k)\right)-\sigma'\left(z_{j,r}(0)\right)\right)\sigma'\left(z_{i,r}(0)\right)\right|\\
\leq&\frac{\beta Lc_\sigma c_{x,0}^2}{m}\left(\sum_{r=1}^{m}\left|z_{i,r}(k)-z_{i,r}(0)\right|+\left|z_{j,r}(k)-z_{j,r}(0)\right|\right)\\
\leq&\frac{\beta Lc_\sigma c_{x,0}^2}{\sqrt{m}}\left(\sqrt{\sum_{r=1}^{m}\left|z_{i,r}(k)-z_{i,r}(0)\right|^2}+\sqrt{\sum_{r=1}^{m}\left|z_{j,r}(k)-z_{j,r}(0)\right|^2}\right).
\end{aligned}
$$

Using the same proof for Lemma A.3, it is easy to see

$$
\sum_{r=1}^{m}\left|z_{i,r}(t)-z_{i,r}(0)\right|^2\leq c_{x,0}^2g_{c_x}(H)^2mR^2.
$$

Thus

$$
I_2^{i,j}\leq 2\beta c_\sigma c_{x,0}^3Lg_{c_x}(H)R.
$$

Therefore we can bound the perturbation

$$
\begin{aligned}
\left\|\mathbf{G}^{(H)}(t)-\mathbf{G}^{(H)}(0)\right\|_F=&\sqrt{\sum_{(i,j)}^{n,n}\left|\mathbf{G}_{i,j}^{(H)}(t)-\mathbf{G}_{i,j}^{(H)}(0)\right|^2}\\
&\leq\left(2\beta c_{x,0}+3\sqrt{c_\sigma}L^2\right)Lc_\sigma c_{x,0}^2g_{c_x}(H)nR.
\end{aligned}
$$

Plugging in the bound on $R$, we have the desired result.

□

*Proof of Lemma A.5.* We will prove this corollary by induction. The induction hypothesis is

$$\left\|\mathbf{W}^{(h)}(s) - \mathbf{W}^{(h)}(0)\right\|_F \leq \sum_{s'=0}^{s-1}(1 - \frac{\eta\lambda_0}{2})^{s'/2}\frac{1}{4}\eta\lambda_0 R'\sqrt{m} \leq R'\sqrt{m}, s \in [k+1].$$

First it is easy to see it holds for $s' = 0$. Now suppose it holds for $s' = 0, \ldots, s$, we consider $s' = s+1$. We have

$$\begin{aligned}
&\left\|\mathbf{W}^{(h)}(s+1) - \mathbf{W}^{(h)}(s)\right\|_F \\
=&\eta\left\|\left(\frac{c_\sigma}{m}\right)^{\frac{H-h+1}{2}}\sum_{i=1}^{n}(y_i - u_i(s))\mathbf{x}_i^{(h-1)}(s)\left(\mathbf{a}^\top\mathbf{J}_i^{(H)}(s)\mathbf{W}^{(H)}(s)\cdots\mathbf{W}^{(h+1)}(s)\mathbf{J}_i^{(h)}(s)\right)\right\|_F \\
\leq&\eta\left(\frac{c_\sigma}{m}\right)^{\frac{H-h+1}{2}}\left\|\mathbf{a}\right\|_2\sum_{i=1}^{n}\left|y_i - u_i(s)\right|\left\|\mathbf{x}_i^{(h-1)}(s)\right\|_2\prod_{k=h+1}^{H}\left\|\mathbf{W}^{(k)}(s)\right\|_2\prod_{k=h}^{H}\left\|\mathbf{J}^{(k)}(s)\right\|_2.
\end{aligned}$$

To bound $\left\|\mathbf{x}_i^{(h-1)}(s)\right\|_2$, we can just apply Lemma A.3 and get

$$\left\|\mathbf{x}_i^{(h-1)}(s)\right\|_2 \leq \sqrt{c_\sigma}Lc_{x,0}g_{c_x}(h)R' + c_{x,0} \leq 2c_{x,0}.$$

To bound $\left\|\mathbf{W}^{(k)}(s)\right\|_2$, we use our assumption

$$\begin{aligned}
\prod_{k=h+1}^{H}\left\|\mathbf{W}^{(k)}(s)\right\|_2 \leq &\prod_{k=h+1}^{H}\left(\left\|\mathbf{W}^{(k)}(0)\right\|_2 + \left\|\mathbf{W}^{(k)}(s) - \mathbf{W}^{(k)}(0)\right\|_2\right) \\
\leq &\prod_{k=h+1}^{H}\left(c_{w,0}\sqrt{m} + R'\sqrt{m}\right) \\
= &\left(c_{w,0} + R'\right)^{H-h} m^{\frac{H-h}{2}} \\
\leq &\left(2c_{w,0}\right)^{H-h} m^{\frac{H-h}{2}}.
\end{aligned}$$

Note that $\left\|\mathbf{J}^{(k)}(s)\right\|_2 \leq L$. Plugging in these two bounds back, we obtain

$$\begin{aligned}
\left\|\mathbf{W}^{(h)}(s+1) - \mathbf{W}^{(h)}(s)\right\|_F \leq &2\eta c_{x,0}c_x^H\sum_{i=1}^{n}\left|y_i - u(s)\right| \\
\leq &2\eta c_{x,0}c_x^H\sqrt{n}\left\|\mathbf{y} - \mathbf{u}(s)\right\|_2 \\
= &\eta Q'(s) \\
\leq &(1 - \frac{\eta\lambda_0}{2})^{s/2}\frac{1}{4}\eta\lambda_0 R'\sqrt{m}.
\end{aligned}$$

Thus

$$\left\|\mathbf{W}^{(h)}(s+1) - \mathbf{W}^{(h)}(0)\right\|_F$$

$$
\begin{aligned}
&\le \left\|\mathbf{W}^{(h)}(s+1)-\mathbf{W}^{(h)}(s)\right\|_F+\left\|\mathbf{W}^{(h)}(s)-\mathbf{W}^{(h)}(0)\right\|_F\\
&\le \sum_{s'=0}^{s}\eta(1-\frac{\eta\lambda_0}{2})^{s'/2}\frac{1}{4}\eta\lambda_0 R'\sqrt{m}.
\end{aligned}
$$

□

*Proof of Lemma A.6.* Fix $i\in[n]$, we bound

$$
\left|I_2^i(k)\right| \le \eta \max_{0\le s\le \eta}\sum_{h=1}^{H}\left\|L'^{(h)}(\mathbf{w}(k))\right\|_F\left\|u_i'^{(h)}\left(\mathbf{w}(k)\right)-u_i'^{(h)}\left(\mathbf{w}(k)-sL'^{(h)}(\mathbf{w}(k))\right)\right\|_F.
$$

For the gradient norm, we have

$$
\begin{aligned}
&\left\|L'^{(h)}(\mathbf{w}(k))\right\|_F\\
=&\left\|\left(\frac{c_\sigma}{m}\right)^{\frac{H-h+1}{2}}\sum_{i=1}^{n}(y_i-u_i(k))\mathbf{x}_i^{(h-1)}(k)\left(\mathbf{a}^\top\mathbf{J}_i^{(H)}(k)\mathbf{W}^{(H)}(k)\cdots\mathbf{W}^{(h+1)}(k)\mathbf{J}_i^{(h)}(k)\right)\right\|_F.
\end{aligned}
$$

Similar to the proof for Lemma A.5, we have

$$
\left\|L'^{(h)}(\mathbf{w}(k))\right\|_F\le Q'(k).
$$

Let $\mathbf{w}(k,s)=\mathbf{w}(k)-sL'(\mathbf{w}(k))$,

$$
\begin{aligned}
&\left\|u_i'^{(h)}\left(\mathbf{w}(k)\right)-u_i'^{(h)}\left(\mathbf{w}(k,s)\right)\right\|_F\\
=&\left(\frac{c_\sigma}{m}\right)^{\frac{H-h+1}{2}}\left\|\mathbf{x}_i^{(h-1)}(k)\left(\mathbf{a}^\top\mathbf{J}_i^{(H)}(k)\mathbf{W}^{(H)}(k)\cdots\mathbf{W}^{(h+1)}(k)\mathbf{J}_i^{(h)}(k)\right)\right.\\
&\left.-\mathbf{x}_i^{(h-1)}(k,s)\left(\mathbf{a}^\top\mathbf{J}_i^{(H)}(k,s)\mathbf{W}^{(H)}(k,s)\cdots\mathbf{W}^{(h+1)}(k,s)\mathbf{J}_i^{(h)}(k,s)\right)\right\|_F
\end{aligned}
$$

Through standard calculations, we have

$$
\begin{aligned}
\left\|\mathbf{W}^{(l)}(k)-\mathbf{W}^{(l)}(k,s)\right\|_F\le&\eta Q'(k),\\
\left\|\mathbf{x}_i^{(h-1)}(k)-\mathbf{x}_i^{(h-1)}(k,s)\right\|_F\le&2\eta\sqrt{c_\sigma}Lc_{x,0}g_{2c_x}(H)\frac{Q'(k)}{\sqrt{m}},\\
\left\|\mathbf{J}_i^{(l)}(k)-\mathbf{J}_i^{(l)}(k,s)\right\|_F\le&2\eta\beta\sqrt{c_\sigma}Lc_{x,0}g_{2c_x}(H)Q'(k).
\end{aligned}
$$

According to Lemma E.1, we have

$$
\begin{aligned}
&\left\|u_i'^{(h)}\left(\mathbf{w}(k)\right)-u_i'^{(h)}\left(\mathbf{w}(k,s)\right)\right\|_F\\
\le&2c_{x,0}c_x^H\eta\frac{Q'(k)}{\sqrt{m}}\left(\frac{H}{2c_{w,0}}+\left[\frac{1}{2c_{x,0}}+\frac{H\beta\sqrt{m}}{L}\right]2\sqrt{c_\sigma}Lc_{x,0}g_{2c_x}(H)\right)
\end{aligned}
$$

$$\leq 8H\sqrt{c_\sigma}c_{x,0}^2 c_x^H g_{2c_x}(H)\beta\eta Q'(k).$$

Thus we have

$$\left|I_2^i\right| \leq 8H^2\sqrt{c_\sigma}c_{x,0}^2 c_x^H g_{2c_x}(H)\beta\eta^2 Q'(k)^2.$$

Since this holds for all $i \in [n]$, plugging in $\eta$ and noting that $\|\mathbf{y}-\mathbf{u}(0)\|_2 = O(\sqrt{n})$, we have

$$\|\mathbf{I}_2(k)\|_2 \leq \frac{1}{8}\eta\lambda_0 \|\mathbf{y}-\mathbf{u}(k)\|_2 .$$

□

*Proof of Lemma A.7.*

$$\begin{aligned}
\|\mathbf{u}(k+1)-\mathbf{u}(k)\|_2^2 =& \sum_{i=1}^n \left(\mathbf{a}^\top \mathbf{x}_i^{(H)}(k+1) - \mathbf{a}^\top \mathbf{x}_i^{(H)}(k)\right)^2 \\
\leq& m\sum_{i=1}^n \left\|\mathbf{x}_i^{(H)}(k+1) - \mathbf{x}_i^{(H)}(k)\right\|_2^2 \\
\leq& n\left(2\eta\sqrt{c_\sigma}Lc_{x,0}g_{2c_x}(H)Q'(k)\right)^2 \\
\leq& \frac{1}{8}\eta\lambda_0\|\mathbf{y}-\mathbf{u}(k)\|_2^2 .
\end{aligned}$$

□

# B Proofs for Section 4

The gradient for ResNet is

$$\begin{aligned}
\frac{\partial L}{\partial \mathbf{W}^{(h)}} =& \frac{c_\lambda \lambda^{3/2}}{H\sqrt{m}}\sum_{i=1}^n (y_i - u_i)\mathbf{x}_i^{(h-1)} \cdot \\
& \left[\mathbf{a}^\top \left(\mathbf{I} + \frac{1}{H\sqrt{m}}\mathbf{J}_i^{(H)}\mathbf{W}^{(H)}\right)\cdots\left(\mathbf{I} + \frac{1}{H\sqrt{m}}\mathbf{J}_i^{(h+1)}\mathbf{W}^{(h+1)}\right)\mathbf{J}_i^{(h)}\right]
\end{aligned}$$

For ResNets, $\mathbf{G}^{(H)}$ has the following form:

$$\mathbf{G}_{ij}^{(H)} = \frac{c_\lambda^2\lambda^3}{H^2 m}(\mathbf{x}_i^{(H-1)})^\top \mathbf{x}_j^{(H-1)} \sum_{r=1}^m a_r^2 \sigma'((\mathbf{w}_r^{(H)})^\top \mathbf{x}_i^{(H-1)})\sigma'((\mathbf{w}_r^{(H)})^\top \mathbf{x}_j^{(H-1)}).$$

Similar to Lemma A.1, we can show with high probability the feature of each layer is approximately normalized.

**Lemma B.1** (Lemma on Initialization Norms)**.** *If $\sigma(\cdot)$ is $L-$Lipschitz and $m = \Omega\left(\frac{n}{\delta}\right)$, assuming $\left\|\mathbf{W}^{(h)}(0)\right\|_2 \leq c_{w,0}\sqrt{m}$ for $h \in [2, H]$ and $c_{w,0} \approx 2$ for Gaussian initialization. We have with probability at least $1-\delta$ over random initialization, for every $h \in [H]$ and $i \in [n]$,*

$$\frac{1}{c_{x,0}} \leq \left\|\mathbf{x}_i^{(h)}(0)\right\|_2 \leq c_{x,0}$$

*for some universal constant $c_{x,0} > 1$ (only depends on $\sigma$).*

The following lemma lower bounds $\mathbf{G}^{(H)}(0)$'s least eigenvalue. This lemma is a direct consequence of Theorem D.1, Theorem D.1 and Remark D.4.

**Lemma B.2** (Least Eigenvalue at the Initialization)**.** *If $m = \Omega\left(\text{poly}(n, \frac{1}{\lambda_0}), \frac{1}{\lambda}\right)$, we have*

$$\lambda_{\min}(\mathbf{G}^{(H)}(0)) \geq \frac{3}{4}\lambda_0.$$

Next, we characterize how the perturbation on the weight matrices affects the input of each layer.

**Lemma B.3.** *Suppose $\sigma(\cdot)$ is L-Lipschitz and for $h \in [H]$, $\left\|\mathbf{W}^{(h)}(0)\right\|_2 \leq c_{w,0}\sqrt{m}$, $\left\|\mathbf{x}^{(h)}(0)\right\|_2 \leq c_{x,0}$ and $\left\|\mathbf{W}^{(h)}(k) - \mathbf{W}^{(h)}(0)\right\|_F \leq \sqrt{m}R$ for some constant $c_{w,0}, c_{x,0} > 0$ and $R \leq c_{w,0}$ . Then we have*

$$\left\|\mathbf{x}^{(h)}(k) - \mathbf{x}^{(h)}(0)\right\|_2 \leq \left(\sqrt{c_\sigma}L + \frac{c_{x,0}}{c_{w,0}}\right) e^{2c_\lambda c_{w,0}L}R.$$

Next, we characterize how the perturbation on the weight matrices affect $\mathbf{G}^{(H)}$.

**Lemma B.4.** *Suppose $\sigma(\cdot)$ is differentiable, $L-$Lipschitz and $\beta-$smooth. Using the same notations in Lemma A.4, if $\left\|\mathbf{W}^{(h)}(k) - \mathbf{W}^{(h)}(0)\right\|_F \leq \sqrt{m}R$ where $R \leq c\lambda_0 H^2 n^{-1}$ and $R \leq c$ for some small constant $c$, we have*

$$\left\|\mathbf{G}^{(H)}(k) - \mathbf{G}^{(H)}(0)\right\|_2 \leq \frac{\lambda_0}{2}.$$

We prove Theorem 4.1 by induction. Our induction hypothesis is just the following convergence rate of empirical loss.

A directly corollary of this condition is the following bound of deviation from the initialization. The proof only involves standard calculations so we defer it to appendix.

**Lemma B.5.** *If Condition 6.1 holds for $k' = 1, \ldots, k$, we have for any $s \in [k+1]$*

$$\begin{aligned}\left\|\mathbf{W}^{(h)}(s) - \mathbf{W}^{(h)}(0)\right\|_F &\leq R'\sqrt{m},\\ \left\|\mathbf{W}^{(h)}(s) - \mathbf{W}^{(h)}(s-1)\right\|_F &\leq \eta Q'(s-1),\end{aligned}$$

*where $R' = \frac{8c_\lambda c_{x,0}Le^{2c_\lambda c_{w,0}L}\sqrt{n}\|\mathbf{y}-\mathbf{u}(0)\|_2}{H\lambda_0\sqrt{m}} < c$ for some small constant $c$ and*
*$Q'(s) = 2c_\lambda c_{x,0}Le^{2c_\lambda c_{w,0}L}\sqrt{n}\left\|\mathbf{y}-\mathbf{u}(s)\right\|_2/H$.*

The next lemma bounds the $\mathbf{I}_2$ term.

**Lemma B.6.** *If Condition 6.1 holds for* $k' = 1, \ldots, k$ *and* $\eta \le c\lambda_0 H^2 n^{-2}$ *for some small constant* $c$*, we have*

$$\left\|\mathbf{I}_2(k)\right\|_2 \le \frac{1}{8}\eta\lambda_0 \left\|\mathbf{y} - \mathbf{u}(k)\right\|_2 .$$

Next we bound the quadratic term.

**Lemma B.7.** *If Condition 6.1 holds for* $k' = 1, \ldots, k$ *and* $\eta \le c\lambda_0 H^2 n^{-2}$ *for some small constant* $c$*, we have* $\left\|\mathbf{u}(k+1) - \mathbf{u}(k)\right\|_2^2 \le \frac{1}{8}\eta\lambda_0 \left\|\mathbf{y} - \mathbf{u}(k)\right\|_2^2$.

Now using the same argument as in the proof for multilayer fully connected neural network, we finish our proof for ResNet.

## B.1 Proofs of Lemmas

*Proof of Lemma B.1.* We will bound $\left\|\mathbf{x}_i^{(h)}(0)\right\|_2$ layer by layer. For the first layer, we can calculate

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{x}_i^{(1)}(0)\right\|_2^2\right] =& c_\sigma \mathbb{E}\left[\sigma\left(\mathbf{w}_r^{(1)}(0)^\top \mathbf{x}_i\right)^2\right] \\
=& c_\sigma \mathbb{E}_{X\sim N(0,1)} \sigma(X)^2 \\
=& 1.
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}\left[\left\|\mathbf{x}_i^{(1)}(0)\right\|_2^2\right] =& \frac{c_\sigma^2}{m} \mathrm{Var}\left[\sigma\left(\mathbf{w}_r^{(1)}(0)^\top \mathbf{x}_i(0)\right)^2\right] \\
\le& \frac{c_\sigma^2}{m} \mathbb{E}_{X\sim N(0,1)} \sigma(X)^4 \\
\le& \frac{c_\sigma^2}{m} \mathbb{E}\left[\left(\left|\sigma(0)\right| + L\left|\mathbf{w}_r^{(1)}(0)^\top \mathbf{x}_i\right|\right)^4\right] \\
\le& \frac{C_2}{m},
\end{aligned}$$

where $C_2 \triangleq \sigma(0)^4 + 4\left|\sigma(0)\right|^3 L\sqrt{2/\pi} + 6\sigma(0)^2 L^2 + 8\left|\sigma(0)\right| L^3\sqrt{2/\pi} + 32L^4$. We have with probability at least $1 - \frac{\delta}{n}$,

$$\frac{1}{2} \le \left\|\mathbf{x}_i^{(1)}(0)\right\|_2 \le 2.$$

By definition we have for $2 \le h \le H$,

$$\left\|\mathbf{x}_i^{(h-1)}(0)\right\|_2 - \left\|\frac{c_\lambda \lambda^{3/2}}{H\sqrt{m}} \sigma\left(\mathbf{W}^{(h)}(0)\mathbf{x}_i^{(h-1)}(0)\right)\right\|_2 \le \left\|\mathbf{x}^{(h)}(0)\right\|_2$$

$$\leq \left\|\mathbf{x}_i^{(h-1)}(0)\right\|_2 + \left\|\frac{c_\lambda \lambda^{3/2}}{H\sqrt{m}}\sigma\left(\mathbf{W}^{(h)}(0)\mathbf{x}^{(h-1)}(0)\right)\right\|_2,$$

where

$$\left\|\frac{c_\lambda \lambda^{3/2}}{H\sqrt{m}}\sigma\left(\mathbf{W}^{(h)}(0)\mathbf{x}_i^{(h-1)}(0)\right)\right\|_2 \leq \frac{c_\lambda c_{w,0} L}{H}\left\|\mathbf{x}_i^{(h-1)}(0)\right\|_2.$$

Thus

$$\left\|\mathbf{x}_i^{(h-1)}(0)\right\|_2\left(1-\frac{c_\lambda c_{w,0}L}{H}\right) \leq \left\|\mathbf{x}^{(h)}(0)\right\|_2 \leq \left\|\mathbf{x}_i^{(h-1)}(0)\right\|_2\left(1+\frac{c_\lambda c_{w,0}L}{H}\right),$$

which implies

$$\frac{1}{2}e^{-c_\lambda c_{w,0}L} \leq \left\|\mathbf{x}^{(h)}(0)\right\|_2 \leq 2e^{c_\lambda c_{w,0}L}.$$

Choosing $c_{x,0} = 2e^{c_\lambda c_{w,0}L}$ and using union bounds over $[n]$, we prove the lemma.

□

*Proof of Lemma B.3.* We prove this lemma by induction. Our induction hypothesis is

$$\left\|\mathbf{x}^{(h)}(k)-\mathbf{x}^{(h)}(0)\right\|_2 \leq g(h),$$

where

$$g(h) = g(h-1)\left[1+\frac{2c_\lambda c_{w,0}L}{H}\right] + \frac{L}{H}Rc_{x,0}.$$

For $h=1$, we have

$$\begin{aligned}\left\|\mathbf{x}^{(1)}(k)-\mathbf{x}^{(1)}(0)\right\|_2 &\leq \sqrt{\frac{c_\sigma}{m}}\left\|\sigma\left(\mathbf{W}^{(1)}(k)\mathbf{x}\right)-\sigma\left(\mathbf{W}^{(1)}(0)\mathbf{x}\right)\right\|_2\\ &\leq \sqrt{\frac{c_\sigma}{m}}\left\|\mathbf{W}^{(1)}(k)-\mathbf{W}^{(1)}(0)\right\|_F \leq \sqrt{c_\sigma}LR,\end{aligned}$$

which implies $g(1)=\sqrt{c_\sigma}LR$, for $2 \leq h \leq H$, we have

$$\begin{aligned}&\left\|\mathbf{x}^{(h)}(k)-\mathbf{x}^{(h)}(0)\right\|_2 \leq \frac{c_\lambda}{H\sqrt{m}}\left\|\sigma\left(\mathbf{W}^{(h)}(k)\mathbf{x}^{(h-1)}(k)\right)-\sigma\left(\mathbf{W}^{(h)}(0)\mathbf{x}^{(h-1)}(0)\right)\right\|_2\\ &+\left\|\mathbf{x}^{(h-1)}(k)-\mathbf{x}^{(h-1)}(0)\right\|_2\\ &\leq \frac{c_\lambda}{H\sqrt{m}}\left\|\sigma\left(\mathbf{W}^{(h)}(k)\mathbf{x}^{(h-1)}(k)\right)-\sigma\left(\mathbf{W}^{(h)}(k)\mathbf{x}^{(h-1)}(0)\right)\right\|_2\\ &+\frac{c_\lambda}{H\sqrt{m}}\left\|\sigma\left(\mathbf{W}^{(h)}(k)\mathbf{x}^{(h-1)}(0)\right)-\sigma\left(\mathbf{W}^{(h)}(0)\mathbf{x}^{(h-1)}(0)\right)\right\|_2\\ &+\left\|\mathbf{x}^{(h-1)}(k)-\mathbf{x}^{(h-1)}(0)\right\|_2\\ &\leq \frac{c_\lambda L}{H\sqrt{m}}\left(\left\|\mathbf{W}^{(h)}(0)\right\|_2+\left\|\mathbf{W}^{(h)}(k)-\mathbf{W}^{(h)}(0)\right\|_F\right)\cdot\left\|\mathbf{x}^{(h-1)}(k)-\mathbf{x}^{(h-1)}(0)\right\|_2\end{aligned}$$

$$
\begin{aligned}
&+\frac{c_\lambda L}{H\sqrt{m}}\left\|\mathbf{W}^{(h)}(k)-\mathbf{W}^{(h)}(0)\right\|_F\left\|\mathbf{x}^{h-1}(0)\right\|_2+\left\|\mathbf{x}^{(h-1)}(k)-\mathbf{x}^{(h-1)}(0)\right\|_2\\
&\leq\left[1+\frac{c_\lambda L}{H\sqrt{m}}\left(c_{w,0}\sqrt{m}+R\sqrt{m}\right)\right]g(h-1)+\frac{c_\lambda L}{H\sqrt{m}}\sqrt{m}Rc_{x,0}\\
&\leq\left(1+\frac{2c_\lambda c_{w,0}L}{H}\right)g(h-1)+\frac{c_\lambda}{H}Lc_{x,0}R.
\end{aligned}
$$

Lastly, simple calculations show $g(h)\leq\left(\sqrt{c_\sigma}L+\frac{c_{x,0}}{c_{w,0}}\right)e^{2c_\lambda c_{w,0}L}R$.

□

*Proof of Lemma B.4.* Similar to the proof of Lemma A.4, we can obtain

$$
\left|\mathbf{G}_{i,j}^{(H)}(k)-\mathbf{G}_{i,j}^{(H)}(0)\right|\leq\frac{c_\lambda^2}{H^2}\left(I_1^{i,j}+I_2^{i,j}\right).
$$

For $I_1^{i,j}$, using Lemma B.3, we have

$$
\begin{aligned}
I_1^{i,j}=&L^2\left|\mathbf{x}_i^{(H-1)}(k)^\top\mathbf{x}_j^{(H-1)}(k)-\mathbf{x}_i^{(H-1)}(0)^\top\mathbf{x}_j^{(H-1)}(0)\right|\\
\leq&L^2\left|(\mathbf{x}_i^{(H-1)}(k)-\mathbf{x}_i^{(H-1)}(0))^\top\mathbf{x}_j^{(H-1)}(k)\right|+L^2\left|\mathbf{x}_i^{(H-1)}(0)^\top(\mathbf{x}_i^{(H-1)}(k)-\mathbf{x}_i^{(H-1)}(0))\right|\\
\leq&c_xL^2R\cdot(c_{x,0}+c_xR)+c_{x,0}c_xL^2R\\
\leq&3c_{x,0}c_xL^2R,
\end{aligned}
$$

where $c_x\triangleq\left(\sqrt{c_\sigma}L+\frac{c_{x,0}}{c_{w,0}}\right)e^{2c_\lambda c_{w,0}L}$. To bound $I_2^{i,j}$, we have

$$
\begin{aligned}
I_2^{i,j}=&c_{x,0}^2\frac{1}{m}\left|\sum_{r=1}^m\sigma'\left(z_{i,r}(k)\right)\sigma'\left(z_{j,r}(k)\right)-\sigma'\left(z_{i,r}(0)\right)\sigma'\left(z_{j,r}(0)\right)\right|\\
\leq&c_{x,0}^2\frac{1}{m}\sum_{r=1}^m\left|\left(\sigma'\left(z_{i,r}(k)\right)-\sigma'\left(z_{i,r}(0)\right)\right)\sigma'\left(z_{j,r}(k)\right)\right|+\left|\left(\sigma'\left(z_{j,r}(k)\right)-\sigma'\left(z_{j,r}(0)\right)\right)\sigma'\left(z_{i,r}(0)\right)\right|\\
\leq&\frac{\beta Lc_{x,0}^2}{m}\left(\sum_{r=1}^m\left|z_{i,r}(k)-z_{i,r}(0)\right|+\left|z_{j,r}(k)-z_{j,r}(0)\right|\right)\\
\leq&\frac{\beta Lc_{x,0}^2}{\sqrt{m}}\left(\sqrt{\sum_{r=1}^m\left|z_{i,r}(k)-z_{i,r}(0)\right|^2}+\sqrt{\sum_{r=1}^m\left|z_{j,r}(k)-z_{j,r}(0)\right|^2}\right).
\end{aligned}
$$

Using the same proof for Lemma B.3, it is easy to see

$$
\sum_{r=1}^m\left|z_{i,r}(k)-z_{i,r}(0)\right|^2\leq\left(2c_xc_{w,0}+c_{x,0}\right)^2L^2mR^2.
$$

Thus

$$I_2^{i,j} \leq 2\beta c_{x,0}^2 \left(2c_x c_{w,0} + c_{x,0}\right) L^2 R.$$

Therefore we can bound the perturbation

$$\begin{aligned}\left\|\mathbf{G}^{(H)}(k) - \mathbf{G}^{(H)}(0)\right\|_F =& \sqrt{\sum_{(i,j)}^{n,n} \left|\mathbf{G}_{i,j}^{(H)}(k) - \mathbf{G}_{i,j}^{(H)}(0)\right|^2} \\ \leq& \frac{c_\lambda^2}{H^2}\left[3c_{x,0}c_x + 2\beta c_{x,0}^2\left(2c_x c_{w,0} + c_{x,0}\right)\right] L^2 nR.\end{aligned}$$

Plugging in the bound on $R$, we have the desired result.

□

*Proof of Lemma B.5.* We will prove this corollary by induction. The induction hypothesis is

$$\left\|\mathbf{W}^{(h)}(s) - \mathbf{W}^{(h)}(0)\right\|_F \leq \sum_{s'=0}^{s-1}(1 - \frac{\eta\lambda_0}{2})^{s'/2}\frac{1}{4}\eta\lambda_0 R'\sqrt{m} \leq R'\sqrt{m}, s \in [k+1].$$

First it is easy to see it holds for $s' = 0$. Now suppose it holds for $s' = 0, \ldots, s$, we consider $s' = s + 1$. Similar to Lemma A.5, we have

$$\begin{aligned}&\left\|\mathbf{W}^{(h)}(s+1) - \mathbf{W}^{(h)}(s)\right\|_F \\ \leq& \eta\frac{Lc_\lambda}{H\sqrt{m}}\left\|\mathbf{a}\right\|_2 \sum_{i=1}^n \left|y_i - u_i(s)\right| \left\|\mathbf{x}_i^{(h-1)}(s)\right\|_2 \prod_{k=h+1}^H \left\|\mathbf{I} + \frac{c_\lambda \lambda^{3/2}}{H\sqrt{m}}\mathbf{J}_i^{(k)}(s)\mathbf{W}^{(k)}(s)\right\|_2 \\ \leq& 2\eta c_\lambda c_{x,0} L e^{2c_\lambda c_{w,0} L}\sqrt{n}\left\|\mathbf{y} - \mathbf{u}(s)\right\|_2 / H \\ =& \eta Q'(s) \\ \leq& (1 - \frac{\eta\lambda_0}{2})^{s/2}\frac{1}{4}\eta\lambda_0 R'\sqrt{m}.\end{aligned}$$

Thus

$$\begin{aligned}&\left\|\mathbf{W}^{(h)}(s+1) - \mathbf{W}^{(h)}(0)\right\|_F \\ \leq& \left\|\mathbf{W}^{(h)}(s+1) - \mathbf{W}^{(h)}(s)\right\|_F + \left\|\mathbf{W}^{(h)}(s) - \mathbf{W}^{(h)}(0)\right\|_F \\ \leq& \sum_{s'=0}^{s} \eta(1 - \frac{\eta\lambda_0}{2})^{s'/2}\frac{1}{4}\eta\lambda_0 R'\sqrt{m}.\end{aligned}$$

□

*Proof of Lemma B.6.* Similar to Lemma A.6, we first bound the gradient norm.

$$
\begin{aligned}
&\left\|L'^{(h)}(\mathbf{w}(k))\right\|_F \\
=&\left\| \frac{c_\lambda}{H\sqrt{m}} \sum_{i=1}^{n} (y_i - u_i(k)) \mathbf{x}_i^{(h-1)}(k) \cdot \right. \\
&\left. \left[\mathbf{a}^\top \left(\mathbf{I} + \frac{c_\lambda \lambda^{3/2}}{H\sqrt{m}} \mathbf{J}_i^{(H)}(k) \mathbf{W}^{(H)}(k)\right) \cdots \left(\mathbf{I} + \frac{c_\lambda \lambda^{3/2}}{H\sqrt{m}} \mathbf{J}_i^{(h+1)}(k) \mathbf{W}^{(h+1)}(k)\right) \mathbf{J}_i^{(h)}(k)\right] \right\|_F \\
\leq& \frac{c_\lambda L}{H\sqrt{m}} \left\|\mathbf{a}\right\|_2 \sum_{i=1}^{n} \left|y_i - u_i(k)\right| \left\|\mathbf{x}^{(h-1)}(k)\right\|_2 \prod_{k=h+1}^{H} \left\|\mathbf{I} + \frac{c_\lambda \lambda^{3/2}}{H\sqrt{m}} \mathbf{J}_i^{(k)}(k) \mathbf{W}^{(k)}(k)\right\|_2 .
\end{aligned}
$$

We have bounded the RHS in the proof for Lemma B.5, thus

$$
\left\|L'^{(h)}(\mathbf{w}(k))\right\|_F \leq \lambda_0 Q'(k).
$$

Let $\mathbf{w}(k,s) = \mathbf{w}(k) - sL'(\mathbf{w}(k))$, we have

$$
\begin{aligned}
&\left\|u_i'^{(h)}(\mathbf{w}(k)) - u_i'^{(h)}(\mathbf{w}(k,s))\right\|_F = \\
&\frac{c_\lambda \lambda^{3/2}}{H\sqrt{m}} \left\| \mathbf{x}_i^{(h-1)}(k) \mathbf{a}^\top \left(\mathbf{I} + \frac{c_\lambda \lambda^{3/2}}{H\sqrt{m}} \mathbf{J}_i^{(H)}(k) \mathbf{W}^{(H)}(k)\right) \cdots \left(\mathbf{I} + \frac{c_\lambda \lambda^{3/2}}{H\sqrt{m}} \mathbf{J}_i^{(h+1)}(k) \mathbf{W}^{(h+1)}(k)\right) \mathbf{J}_i^{(h)}(k) \right. \\
&\left. -\mathbf{x}_i^{(h-1)}(k,s) \mathbf{a}^\top \left(\mathbf{I} + \frac{c_\lambda \lambda^{3/2}}{H\sqrt{m}} \mathbf{J}_i^{(H)}(k,s) \mathbf{W}^{(H)}(k,s)\right) \cdots \left(\mathbf{I} + \frac{c_\lambda \lambda^{3/2}}{H\sqrt{m}} \mathbf{J}_i^{(h+1)}(k,s) \mathbf{W}^{(h+1)}(k,s)\right) \mathbf{J}_i^{(h)}(k,s) \right\|_F .
\end{aligned}
$$

Through standard calculations, we have

$$
\begin{aligned}
\left\|\mathbf{W}^{(l)}(k) - \mathbf{W}^{(l)}(k,s)\right\|_F \leq& \eta Q'(k), \\
\left\|\mathbf{x}_i^{(h-1)}(k) - \mathbf{x}_i^{(h-1)}(k,s)\right\|_F \leq& \eta c_x \frac{Q'(k)}{\sqrt{m}}, \\
\left\|\mathbf{J}^{(l)}(k) - \mathbf{J}^{(l)}(k,s)\right\|_F \leq& 2\left(c_{x,0} + c_{w,0} c_x\right) \eta \beta Q'(k),
\end{aligned}
$$

where $c_x \triangleq \left(\sqrt{c_\sigma} L + \frac{c_{x,0}}{c_{w,0}}\right) e^{3c_\lambda c_{w,0} L}$. According to Lemma E.1, we have

$$
\begin{aligned}
&\left\|u_i'^{(h)}(\mathbf{w}(k)) - u_i'^{(h)}(\mathbf{w}(k,s))\right\|_F \\
\leq& \frac{2}{H} c_\lambda c_{x,0} L e^{2Lc_{w,0}} \eta \frac{Q'(k)}{\sqrt{m}} \left(\frac{c_x}{c_{x,0}} + \frac{2}{L}\left(c_{x,0} + c_{w,0} c_x\right) \beta \sqrt{m} + 4c_{w,0}\left(c_{x,0} + c_{w,0} c_x\right) \beta + L\right) \\
\leq& \frac{16}{H} c_\lambda c_{x,0} e^{2Lc_{w,0}} \left(c_{x,0} + c_{w,0} c_x\right) \beta \eta Q'(k).
\end{aligned}
$$

Thus we have

$$
\left|I_2^i\right| \leq 16 c_\lambda c_{x,0} e^{2Lc_{w,0}} \left(c_{x,0} + c_{w,0} c_x\right) \beta \eta^2 Q'(k)^2 \leq \frac{1}{8} \eta \lambda_0 \left\|\mathbf{y} - \mathbf{u}(k)\right\|_2 ,
$$

where we used the bound of $\eta$ and that $\left\|\mathbf{y} - \mathbf{u}(0)\right\|_2 = O(\sqrt{n})$,. □

*Proof of Lemma B.7.*

$$
\begin{aligned}
\left\|\mathbf{u}(k+1)-\mathbf{u}(k)\right\|_2^2 =& \sum_{i=1}^{n}\left(\mathbf{a}^\top \mathbf{x}_i^{(H)}(k+1)-\mathbf{a}^\top \mathbf{x}_i^{(H)}(k)\right)^2\\
\leq & m\sum_{i=1}^{n}\left\|\mathbf{x}_i^{(H)}(k+1)-\mathbf{x}_i^{(H)}(k)\right\|_2^2\\
\leq & n\left(\eta c_x Q'(k)\right)^2\\
\leq & \frac{1}{8}\eta\lambda_0\left\|\mathbf{y}-\mathbf{u}(k)\right\|_2^2.
\end{aligned}
$$

□

# C Proofs for Section 5

For CNN, denote $\mathbf{x}_{i,l} = \phi\left(\mathbf{x}_{i,l}\right)_{:,l}$, $\mathbf{G}^{(H)}$ has the following form:

$$
\mathbf{G}_{ij}^{(H)} = \frac{c_\lambda^2 \lambda^3}{H^2 m}\sum_{r=1}^{m}\left[\sum_{l=1}^{p} a_{l,r}\mathbf{x}_{i,l}^{(H-1)}\sigma'\left(\left(\mathbf{w}_r^{(H)}\right)^\top \mathbf{x}_{i,l}^{(H-1)}\right)\right]^\top\left[\sum_{k=1}^{p} a_{k,r}\mathbf{x}_{j,k}^{(H-1)}\sigma'\left(\left(\mathbf{w}_r^{(H)}\right)^\top \mathbf{x}_{j,k}^{(H-1)}\right)\right].
$$

We define a constant $c_{\sigma,c_0} = \left(\min_{c_0\leq\alpha\leq 1}\mathbb{E}_{X\sim N(0,1)}\sigma(\alpha X)^2\right)^{-1} > 0$, where $0 < c_0 \leq 1$. In particular, it is easy to see for smooth ReLU, $c_{\sigma,\frac{1}{\sqrt{p}}} = \text{poly}(p)$.

Similar to Lemma A.1, we can show with high probability the feature of each layer is approximately normalized.

**Lemma C.1** (Lemma on Initialization Norms)**.** *If $\sigma(\cdot)$ is $L-$Lipschitz and $m = \Omega\left(\frac{p^2 n}{c_{\sigma,\frac{1}{\sqrt{p}}}^2\delta}\right)$, assuming $\left\|\mathbf{W}^{(h)}(0)\right\|_2 \leq c_{w,0}\sqrt{m}$ for $h\in[H]$, we have with probability at least $1-\delta$ over random initialization, for every $h\in[H]$ and $i\in[n]$,*

$$
\frac{1}{c_{x,0}} \leq \left\|\mathbf{x}_i^{(h)}(0)\right\|_F \leq c_{x,0}
$$

*for some constant $c_{x,0} = poly(p) > 1$.*

The following lemma lower bounds $\mathbf{G}^{(H)}(0)$'s least eigenvalue. This lemma is a direct consequence of Theorem D.1, Theorem D.1 and Remark D.4.

**Lemma C.2** (Least Eigenvalue at the Initialization)**.** *If $m = \Omega\left(\text{poly}(n,\frac{1}{\lambda_0}),\frac{1}{\lambda}\right)$, we have*

$$
\lambda_{\min}(\mathbf{G}^{(H)}(0)) \geq \frac{3}{4}\lambda_0.
$$

Next, we prove the following lemma which characterizes how the perturbation from weight matrices propagates to the input of each layer.

**Lemma C.3.** *Suppose $\sigma(\cdot)$ is $L-$Lipschitz and for $h \in [H]$, $\left\|\mathbf{W}^{(h)}(0)\right\|_2 \le c_{w,0}\sqrt{m}$, $\left\|\mathbf{x}^{(h)}(0)\right\|_F \le c_{x,0}$ and $\left\|\mathbf{W}^{(h)}(k) - \mathbf{W}^{(h)}(0)\right\|_F \le \sqrt{m}R$ for some constant $c_{w,0}, c_{x,0} > 1$ and $R \le c_{w,0}$ . Then we have*

$$\left\|\mathbf{x}^{(h)}(k) - \mathbf{x}^{(h)}(0)\right\|_F \le \left(\sqrt{c_\sigma}L\sqrt{q} + \frac{c_{x,0}}{c_{w,0}}\right) e^{2c_{w,0}L\sqrt{q}c_\lambda} R.$$

Next, we show with high probability over random initialization, perturbation in weight matrices leads to small perturbation in the Gram matrix.

**Lemma C.4.** *Suppose $\sigma(\cdot)$ is differentaible, $L-$Lipschitz and $\beta-$smooth. Using the same notations in Lemma A.4, if $\left\|\mathbf{W}^{(h)}(k) - \mathbf{W}^{(h)}(0)\right\|_F \le \sqrt{m}R$ where $R \le c\lambda_0 H^2 \left(n\right)^{-1} poly(p)^{-1}$ for some small constant $c$, we have*

$$\left\|\mathbf{G}^{(H)}(k) - \mathbf{G}^{(H)}(0)\right\|_2 \le \frac{\lambda_0}{2}.$$

**Lemma C.5.** *If Condition 6.1 holds for $k' = 1, \ldots, k$, we have for any $s \in [k+1]$*

$$\begin{aligned}\left\|\mathbf{W}^{(h)}(s) - \mathbf{W}^{(h)}(0)\right\|_F &\le R'\sqrt{m},\\ \left\|\mathbf{W}^{(h)}(s) - \mathbf{W}^{(h)}(s-1)\right\|_F &\le \eta Q'(s-1),\end{aligned}$$

*where $R' = \frac{8c_\lambda c_{x,0}L\sqrt{pq}e^{2c_\lambda c_{w,0}L\sqrt{q}}\sqrt{n}\|\mathbf{y}-\mathbf{u}(0)\|_2}{H\lambda_0\sqrt{m}} < c$ for some small constant $c$ and*

$$Q'(s) = 2c_\lambda c_{x,0}L\sqrt{pq}e^{2c_\lambda c_{w,0}L\sqrt{q}}\sqrt{n}\left\|\mathbf{y} - \mathbf{u}(s)\right\|_2 / H.$$

The follow lemma bounds the norm of $\mathbf{I}_2$.

**Lemma C.6.** *If Condition 6.1 holds for $k' = 1, \ldots, k$ and $\eta \le c\lambda_0 H^2 n^{-2} poly(1/p)$ for some small constant $c$, we have*

$$\left\|\mathbf{I}_2(k)\right\|_2 \le \frac{1}{8}\eta\lambda_0\left\|\mathbf{y} - \mathbf{u}(k)\right\|_2.$$

Next we also bound the quadratic term.

**Lemma C.7.** *If Condition 6.1 holds for $k' = 1, \ldots, k$ and $\eta \le c\lambda_0 H^2 n^{-2} poly(1/p)$ for some small constant $c$, we have $\left\|\mathbf{u}(k+1) - \mathbf{u}(k)\right\|_2^2 \le \frac{1}{8}\eta\lambda_0\left\|\mathbf{y} - \mathbf{u}(k)\right\|_2^2$.*

Now using the same argument as in the proof for multilayer fully connected neural network, we finish our proof for CNN.

## C.1 Proofs of Lemmas

*Proof of Lemma C.1.* We will bound $\left\|\mathbf{x}_i^{(h)}(0)\right\|_F$ layer by layer. For the first layer, we can calculate

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{x}_i^{(1)}(0)\right\|_F^2\right] =& c_\sigma \sum_{l=1}^{p_1} \mathbb{E}\left[\sigma\left(\mathbf{w}_r^{(1)}(0)^\top \mathbf{x}_{i,l}\right)^2\right] \\ \geq& \frac{c_\sigma}{c_{\sigma,\frac{1}{\sqrt{p}}}},\end{aligned}$$

where the inequality we use the definition of $c_{\sigma,\frac{1}{\sqrt{p}}}$ and the fact that there must exist $l' \in [p]$ such that $\|\mathbf{x}_{i,l'}\|_2^2 \geq \frac{1}{p_1} \geq \frac{1}{p}$. For the variance,

$$\begin{aligned}\mathrm{Var}\left[\left\|\mathbf{x}_i^{(1)}(0)\right\|_F^2\right] =& \frac{c_\sigma^2}{m}\mathrm{Var}\left[\sum_{l=1}^{p_1}\sigma\left(\mathbf{w}_r^{(1)}(0)^\top \mathbf{x}_{i,l}\right)^2\right] \\ \leq& \frac{c_\sigma^2}{m}\mathbb{E}\left[\left(\sum_{l=1}^{p_1}\left(|\sigma(0)| + L\left|\mathbf{w}_r^{(1)}(0)^\top \mathbf{x}_{i,l}\right|\right)^2\right)^2\right] \\ \leq& \frac{p^2 C_2}{m},\end{aligned}$$

where $C_2 \triangleq \sigma(0)^4 + 4|\sigma(0)|^3 L\sqrt{2/\pi} + 6\sigma(0)^2L^2 + 8|\sigma(0)|L^3\sqrt{2/\pi} + 32L^4$. We have with probability at least $1 - \frac{\delta}{n}$,

$$\left\|\mathbf{x}_i^{(1)}(0)\right\|_F^2 \geq \frac{c_\sigma}{2c_{\sigma,\frac{1}{\sqrt{p}}}}.$$

It is easy to get its upper bound

$$\left\|\mathbf{x}_i^{(1)}(0)\right\|_F^2 = \frac{c_\sigma}{m}\left\|\sigma\left(\mathbf{W}^{(1)}\phi(\mathbf{x}_i)\right)\right\|_F^2 \leq qL^2 c_\sigma c_{w,0}^2.$$

By defination we have for $2 \leq h \leq H$

$$\begin{aligned}\left\|\mathbf{x}_i^{(h-1)}(0)\right\|_F - \left\|\frac{c_\lambda \lambda^{3/2}}{H\sqrt{m}}\sigma\left(\mathbf{W}^{(h)}(0)\phi\left(\mathbf{x}_i^{(h-1)}(0)\right)\right)\right\|_F \leq \left\|\mathbf{x}_i^{(h)}(0)\right\|_F \\ \leq \left\|\mathbf{x}_i^{(h-1)}(0)\right\|_F + \left\|\frac{c_\lambda \lambda^{3/2}}{H\sqrt{m}}\sigma\left(\mathbf{W}^{(h)}(0)\phi\left(\mathbf{x}_i^{(h-1)}(0)\right)\right)\right\|_F,\end{aligned}$$

where

$$\left\|\frac{c_\lambda \lambda^{3/2}}{H\sqrt{m}}\sigma\left(\mathbf{W}^{(h)}(0)\phi\left(\mathbf{x}_i^{(h-1)}(0)\right)\right)\right\|_F \leq \frac{\sqrt{q}c_\lambda c_{w,0}L}{H}\left\|\mathbf{x}_i^{(h-1)}(0)\right\|_F.$$

Thus

$$\left\|\mathbf{x}_i^{(h-1)}(0)\right\|_F\left(1 - \frac{\sqrt{q}c_\lambda c_{w,0}L}{H}\right) \leq \left\|\mathbf{x}^{(h)}(0)\right\|_F \leq \left\|\mathbf{x}_i^{(h-1)}(0)\right\|_F\left(1 + \frac{\sqrt{q}c_\lambda c_{w,0}L}{H}\right),$$

which implies

$$\sqrt{\frac{c_\sigma}{2c_{\sigma,\frac{1}{\sqrt{p}}}}e^{-\sqrt{q}c_\lambda c_{w,0}L}} \leq \left\|\mathbf{x}^{(h)}(0)\right\|_F \leq \sqrt{qL^2c_\sigma c_{w,0}^2}e^{\sqrt{q}c_\lambda c_{w,0}L}.$$

Choosing $c_{x,0} = \max\{\sqrt{qL^2c_\sigma c_{w,0}^2}, \sqrt{\frac{2c_{\sigma,\frac{1}{\sqrt{p}}}}{c_\sigma}}\}e^{\sqrt{q}c_\lambda c_{w,0}L}$ and using union bounds over $[n]$, we prove the lemma.

□

*Proof of Lemma C.3.* We prove this lemma by induction. Our induction hypothesis is

$$\left\|\mathbf{x}^{(h)}(k) - \mathbf{x}^{(h)}(0)\right\|_F \leq g(h),$$

where

$$g(h) = g(h-1)\left[1 + \frac{2c_\lambda c_{w,0}L\sqrt{q}}{H}\right] + \frac{c_\lambda L\sqrt{q}}{H}Rc_{x,0}.$$

For $h=1$, we have

$$\begin{aligned}\left\|\mathbf{x}^{(1)}(k) - \mathbf{x}^{(1)}(0)\right\|_F \leq& \sqrt{\frac{c_\sigma}{m}}\left\|\sigma\left(\mathbf{W}^{(1)}(k)\phi_1(\mathbf{x})\right) - \sigma\left(\mathbf{W}^{(1)}(0)\phi_1(\mathbf{x})\right)\right\|_F \\ \leq& \sqrt{\frac{c_\sigma}{m}}L\sqrt{q}\left\|\mathbf{W}^{(1)}(k) - \mathbf{W}^{(1)}(0)\right\|_F \leq \sqrt{c_\sigma}L\sqrt{q}R,\end{aligned}$$

which implies $g(1) = \sqrt{c_\sigma}L\sqrt{q}R$, for $2 \leq h \leq H$, we have

$$\begin{aligned}&\left\|\mathbf{x}^{(h)}(k) - \mathbf{x}^{(h)}(0)\right\|_F \\ \leq& \frac{c_\lambda}{H\sqrt{m}}\left\|\sigma\left(\mathbf{W}^{(h)}(k)\phi_h\left(\mathbf{x}^{(h-1)}(k)\right)\right) - \sigma\left(\mathbf{W}^{(h)}(0)\phi_h\left(\mathbf{x}^{(h-1)}(0)\right)\right)\right\|_F + \left\|\mathbf{x}^{(h-1)}(k) - \mathbf{x}^{(h-1)}(0)\right\|_F \\ \leq& \frac{c_\lambda}{H\sqrt{m}}\left\|\sigma\left(\mathbf{W}^{(h)}(k)\phi_h\left(\mathbf{x}^{(h-1)}(k)\right)\right) - \sigma\left(\mathbf{W}^{(h)}(k)\phi_h\left(\mathbf{x}^{(h-1)}(0)\right)\right)\right\|_F \\ &+ \frac{c_\lambda}{H\sqrt{m}}\left\|\sigma\left(\mathbf{W}^{(h)}(k)\phi_h\left(\mathbf{x}^{(h-1)}(0)\right)\right) - \sigma\left(\mathbf{W}^{(h)}(0)\phi_h\left(\mathbf{x}^{(h-1)}(0)\right)\right)\right\|_F + \left\|\mathbf{x}^{(h-1)}(k) - \mathbf{x}^{(h-1)}(0)\right\|_F \\ \leq& \frac{L\sqrt{q}c_\lambda}{H\sqrt{m}}\left(\left\|\mathbf{W}^{(h)}(0)\right\|_2 + \left\|\mathbf{W}^{(h)}(k) - \mathbf{W}^{(h)}(0)\right\|_F\right)\cdot\left\|\mathbf{x}^{(h-1)}(k) - \mathbf{x}^{(h-1)}(0)\right\|_F \\ &+ \frac{L\sqrt{q}c_\lambda}{H\sqrt{m}}\left\|\mathbf{W}^{(h)}(k) - \mathbf{W}^{(h)}(0)\right\|_F\left\|\mathbf{x}^{h-1}(0)\right\|_F + \left\|\mathbf{x}^{(h-1)}(k) - \mathbf{x}^{(h-1)}(0)\right\|_F \\ \leq& \left[1 + \frac{L\sqrt{q}c_\lambda}{H\sqrt{m}}\left(c_{w,0}\sqrt{m} + R\sqrt{m}\right)\right]g(h-1) + \frac{L\sqrt{q}c_\lambda}{H\sqrt{m}}\sqrt{m}Rc_{x,0} \\ \leq& \left(1 + \frac{2c_{w,0}L\sqrt{q}c_\lambda}{H}\right)g(h-1) + \frac{1}{H}L\sqrt{q}c_\lambda c_{x,0}R.\end{aligned}$$

Lastly, simple calculations show $g(h) \leq \left(\sqrt{c_\sigma}L\sqrt{q} + \frac{c_{x,0}}{c_{w,0}}\right)e^{2c_{w,0}L\sqrt{q}c_\lambda}R$.

□

*Proof of Lemma C.4.* Similar to Lemma B.4, define $z_{i,l,r} = \left(\mathbf{w}_r^{(H)}\right)^\top \mathbf{x}_{i,l}^{(H-1)}$, we have

$$
\begin{aligned}
&\left|\mathbf{G}_{i,j}^{(H)}(k) - \mathbf{G}_{i,j}^{(H)}(0)\right| \\
=&\frac{c_\lambda^2\lambda^2}{H^2} | \sum_{l=1}^{p}\sum_{k=1}^{p} \mathbf{x}_{i,l}^{(H-1)}(k)^\top \mathbf{x}_{j,k}^{(H-1)}(k) \frac{1}{m}\sum_{r=1}^{m} a_{r,l}a_{r,k}\sigma'\left(z_{i,l,r}(k)\right)\sigma'\left(z_{j,k,r}(k)\right) \\
&- \sum_{l=1}^{p}\sum_{k=1}^{p} \mathbf{x}_{i,l}^{(H-1)}(0)^\top \mathbf{x}_{j,k}^{(H-1)}(0) \frac{1}{m}\sum_{r=1}^{m} a_{r,l}a_{r,k}\sigma'\left(z_{i,l,r}(0)\right)\sigma'\left(z_{j,k,r}(0)\right) | \\
\leq&\frac{c_\lambda^2 L^2}{H^2}\left|\sum_{l=1}^{p}\sum_{k=1}^{p} \mathbf{x}_{i,l}^{(H-1)}(k)^\top \mathbf{x}_{j,k}^{(H-1)}(k) - \mathbf{x}_{i,l}^{(H-1)}(0)^\top \mathbf{x}_{j,k}^{(H-1)}(0)\right| \\
&+ \frac{c_\lambda^2}{H^2}\left|\sum_{l=1}^{p}\sum_{k=1}^{p} \mathbf{x}_{i,l}^{(H-1)}(0)^\top \mathbf{x}_{j,k}^{(H-1)}(0)\right| \frac{1}{m}\sum_{r=1}^{m}\left|\sigma'\left(z_{i,l,r}(k)\right)\sigma'\left(z_{j,k,r}(k)\right) - \sigma'\left(z_{i,l,r}(0)\right)\sigma'\left(z_{j,k,r}(0)\right)\right| \\
\triangleq&\frac{c_\lambda^2}{H^2}\left(I_1^{i,j} + I_2^{i,j}\right).
\end{aligned}
$$

For $I_1^{i,j}$, using Lemma C.3, we have

$$
\begin{aligned}
I_1^{i,j} =&L^2\left|\sum_{l=1}^{p}\sum_{k=1}^{p} \mathbf{x}_{i,l}^{(H-1)}(k)^\top \mathbf{x}_{j,k}^{(H-1)}(k) - \mathbf{x}_{i,l}^{(H-1)}(0)^\top \mathbf{x}_{j,k}^{(H-1)}(0)\right| \\
\leq&L^2\sum_{l=1}^{p}\sum_{k=1}^{p}\left|(\mathbf{x}_{i,l}^{(H-1)}(k) - \mathbf{x}_{i,l}^{(H-1)}(0))^\top \mathbf{x}_{j,k}^{(H-1)}(k)\right| + L^2\sum_{l=1}^{p}\sum_{k=1}^{p}\left|\mathbf{x}_{i,l}^{(H-1)}(0)^\top(\mathbf{x}_{j,k}^{(H-1)}(k) - \mathbf{x}_{j,k}^{(H-1)}(0))\right| \\
\leq&L^2\sqrt{\sum_{l=1}^{p}\sum_{k=1}^{p}\left\|\mathbf{x}_{i,l}^{(H-1)}(k) - \mathbf{x}_{i,l}^{(H-1)}(0)\right\|_2^2}\sqrt{\sum_{l=1}^{p}\sum_{k=1}^{p}\left\|\mathbf{x}_{j,k}^{(H-1)}(k)\right\|_2^2} \\
&+ L^2\sqrt{\sum_{l=1}^{p}\sum_{k=1}^{p}\left\|\mathbf{x}_{i,l}^{(H-1)}(0)\right\|_2^2}\sqrt{\sum_{l=1}^{p}\sum_{k=1}^{p}\left\|\mathbf{x}_{j,k}^{(H-1)}(k) - \mathbf{x}_{j,k}^{(H-1)}(0)\right\|_2^2} \\
\leq&L^2 p\left\|\mathbf{x}_i^{(H-1)}(k) - \mathbf{x}_i^{(H-1)}(0)\right\|_F\left\|\mathbf{x}_j^{(H-1)}(k)\right\|_F + L^2 p\left\|\mathbf{x}_i^{(H-1)}(0)\right\|_F\left\|\mathbf{x}_j^{(H-1)}(k) - \mathbf{x}_j^{(H-1)}(0)\right\|_F \\
\leq&3c_{x,0}c_x L^2 pR,
\end{aligned}
$$

where $c_x \triangleq \left(\sqrt{c_\sigma}L\sqrt{q} + \frac{c_{x,0}}{c_{w,0}}\right)e^{2c_\lambda c_{w,0}L\sqrt{q}}$. To bound $I_2^{i,j}$, we have

$$
\begin{aligned}
I_2^{i,j} =&\sum_{l=1}^{p}\sum_{k=1}^{p}\left|\mathbf{x}_{i,l}^{(H-1)}(0)^\top \mathbf{x}_{j,k}^{(H-1)}(0)\right| \frac{1}{m}\left|\sum_{r=1}^{m}\sigma'\left(z_{i,l,r}(k)\right)\sigma'\left(z_{j,k,r}(k)\right) - \sigma'\left(z_{i,l,r}(0)\right)\sigma'\left(z_{j,k,r}(0)\right)\right| \\
\leq&\sum_{l=1}^{p}\sum_{k=1}^{p}\left|\mathbf{x}_{i,l}^{(H-1)}(0)^\top \mathbf{x}_{j,k}^{(H-1)}(0)\right| \frac{\beta L}{m}\left(\sum_{r=1}^{m}\left|z_{i,l,r}(k) - z_{i,l,r}(0)\right| + \left|z_{j,k,r}(k) - z_{j,k,r}(0)\right|\right)
\end{aligned}
$$

$$
\begin{aligned}
&\leq \frac{\beta L}{m} \sqrt{\sum_{l=1}^{p} \sum_{k=1}^{p} \left\| \mathbf{x}_{i,l}^{(H-1)}(0) \right\|_2^2 \left\| \mathbf{x}_{j,k}^{(H-1)}(0) \right\|_2^2} \\
&\quad \left( \sqrt{\sum_{l=1}^{p} \sum_{k=1}^{p} \left( \sum_{r=1}^{m} \left| z_{i,l,r}(k) - z_{i,l,r}(0) \right| \right)^2} + \sqrt{\sum_{l=1}^{p} \sum_{k=1}^{p} \left( \sum_{r=1}^{m} \left| z_{j,k,r}(k) - z_{j,k,r}(0) \right| \right)^2} \right) \\
&\leq \frac{\beta L c_{x,0}^2}{m} \left( \sqrt{m \sum_{l=1}^{p} \sum_{k=1}^{p} \sum_{r=1}^{m} \left| z_{i,l,r}(k) - z_{i,l,r}(0) \right|^2} + \sqrt{m \sum_{l=1}^{p} \sum_{k=1}^{p} \sum_{r=1}^{m} \left| z_{j,k,r}(k) - z_{j,k,r}(0) \right|^2} \right) \\
&\leq \frac{\beta L \sqrt{p} c_{x,0}^2}{\sqrt{m}} \left( \left\| \mathbf{z}_i \right\|_F + \left\| \mathbf{z}_j \right\|_F \right).
\end{aligned}
$$

Using the same proof for Lemma C.3, it is easy to see

$$
\left\| \mathbf{z}_i \right\|_F \leq \left( 2 c_x c_{w,0} \sqrt{q} + c_{x,0} \right) R \sqrt{m}.
$$

Thus

$$
I_2^{i,j} \leq 2 \beta L \sqrt{p} c_{x,0}^2 \left( 2 c_x c_{w,0} \sqrt{q} + c_{x,0} \right) R.
$$

Therefore we can bound the perturbation

$$
\begin{aligned}
\left\| \mathbf{G}^{(H)}(k) - \mathbf{G}^{(H)}(0) \right\|_2 \leq & \left\| \mathbf{G}^{(H)}(k) - \mathbf{G}^{(H)}(0) \right\|_F \\
= & \sqrt{\sum_{(i,j)}^{n,n} \left| \mathbf{G}_{i,j}^{(H)}(k) - \mathbf{G}_{i,j}^{(H)}(0) \right|^2} \\
\leq & \frac{c_\lambda^2}{H^2} \left[ 3 c_{x,0} c_x L p + 2 \beta c_{x,0}^2 \sqrt{p} \left( 2 c_x c_{w,0} \sqrt{q} + c_{x,0} \right) \right] L n R.
\end{aligned}
$$

Plugging in the bound on $R$, we have the desired result.

□

*Proof of Lemma C.5.* We will prove this corollary by induction. The induction hypothesis is

$$
\left\| \mathbf{W}^{(h)}(s) - \mathbf{W}^{(h)}(0) \right\|_F \leq \sum_{s'=0}^{s-1} (1 - \frac{\eta \lambda_0}{2})^{s'/2} \frac{1}{4} \eta \lambda_0 R' \sqrt{m} \leq R' \sqrt{m}, s \in [k+1].
$$

First it is easy to see it holds for $s' = 0$. Now suppose it holds for $s' = 0, \ldots, s$, we consider $s' = s + 1$. Similar to Lemma A.5, we have

$$
\begin{aligned}
& \left\| \mathbf{W}^{(h)}(s+1) - \mathbf{W}^{(h)}(s) \right\|_F \\
\leq & \eta \frac{c_\lambda L}{H \sqrt{m}} \left\| \mathbf{a} \right\|_F \sum_{i=1}^{n} \left| y_i - u(s) \right| \left\| \phi_h \left( \mathbf{x}^{(h-1)}(s) \right) \right\|_F \prod_{k=h+1}^{H} \left\| \mathbf{I} + \frac{c_\lambda \lambda^{3/2}}{H \sqrt{m}} \mathbf{W}^{(k)}(s) \phi_k \right\|_{op}
\end{aligned}
$$

$$
\begin{aligned}
&\leq 2\eta c_\lambda c_{x,0} L\sqrt{pq} e^{2c_\lambda c_{w,0} L\sqrt{q}} \sqrt{n} \left\|\mathbf{y}-\mathbf{u}(s)\right\|_2 / H \\
&= \eta Q'(s) \\
&\leq (1-\frac{\eta\lambda_0}{2})^{s/2} \frac{1}{4}\eta\lambda_0 R'\sqrt{m},
\end{aligned}
$$

where $\left\|\cdot\right\|_{op}$ denotes the operator norm. Thus

$$
\begin{aligned}
&\left\|\mathbf{W}^{(h)}(s+1)-\mathbf{W}^{(h)}(0)\right\|_F \\
\leq &\left\|\mathbf{W}^{(h)}(s+1)-\mathbf{W}^{(h)}(s)\right\|_F + \left\|\mathbf{W}^{(h)}(s)-\mathbf{W}^{(h)}(0)\right\|_F \\
\leq &\sum_{s'=0}^{s} \eta (1-\frac{\eta\lambda_0}{2})^{s'/2} \frac{1}{4}\eta\lambda_0 R'\sqrt{m}.
\end{aligned}
$$

□

*Proof of Lemma C.6.*

$$
\left|I_2^i\right| \leq \eta \max_{0\leq s\leq\eta} \sum_{h=1}^{H} \left\|L'^{(h)}(\mathbf{w}(k))\right\|_F \left\|u_i'^{(h)}\left(\mathbf{w}(k)\right) - u_i'^{(h)}\left(\mathbf{w}(k) - sL'^{(h)}(\mathbf{w}(k))\right)\right\|_F .
$$

For the gradient norm, we have

$$
\begin{aligned}
&\left\|L'^{(h)}(\mathbf{w}(k))\right\|_F \\
\leq &\frac{Lc_\lambda}{H\sqrt{m}} \left\|\mathbf{a}\right\|_F \sum_{i=1}^{n} \left|y_i - u_i(k)\right| \left\|\phi_h\left(\mathbf{x}_i^{(h-1)}(k)\right)\right\|_F \prod_{k=h+1}^{H} \left\|\mathbf{I} + \frac{c_\lambda \lambda^{3/2}}{H\sqrt{m}} \mathbf{J}_i^{(k)}(k)\mathbf{W}^{(k)}(k)\phi_k\right\|_{op},
\end{aligned}
$$

which we have bounded in Lemma C.5, thus

$$
\left\|L'^{(h)}(\mathbf{w}(k))\right\|_F \leq Q'(k).
$$

Let $\mathbf{w}(k,s) = \mathbf{w}(k) - sL'(\mathbf{w}(k))$ ,similar to the proof of Lemma A.6, we have

$$
\begin{aligned}
&\left\|u_i'^{(h)}\left(\mathbf{w}(k)\right) - u_i'^{(h)}\left(\mathbf{w}(k,s)\right)\right\|_F \\
\leq &\frac{2}{H} c_\lambda c_{x,0} L\sqrt{q} e^{2c_\lambda L c_{w,0}\sqrt{q}} \eta \frac{Q'(k)}{\sqrt{m}} \left(\frac{c_x}{c_{x,0}} + \frac{2}{L}\left(c_{x,0}+c_{w,0}c_x\right)\beta\sqrt{m} + 4\sqrt{q} c_{w,0}\left(c_{x,0}+c_{w,0}c_x\right)\beta\sqrt{m} + L\sqrt{q}\right) \\
\leq &\frac{24}{H} c_\lambda c_{x,0} L\sqrt{q} c_{w,0} e^{2c_\lambda L c_{w,0}\sqrt{q}} \left(c_{x,0}+c_{w,0}c_x\right)\beta\eta Q'(k).
\end{aligned}
$$

Thus

$$
\left|I_2^i\right| \leq 24 c_\lambda c_{x,0} L\sqrt{q} c_{w,0} e^{2c_\lambda L c_{w,0}} \left(c_{x,0}+c_{w,0}c_x\right)\beta\eta^2\lambda_0\sqrt{m} Q'(k) R' \leq \frac{1}{8}\eta\lambda_0 \left\|\mathbf{y}-\mathbf{u}(k)\right\|_2 .
$$

where we used the bound of $\eta$ and that $\left\|\mathbf{y}-\mathbf{u}(0)\right\|_2 = O(\sqrt{n})$. □

*Proof of Lemma C.7.*

$$\begin{aligned}\|\mathbf{u}(k+1)-\mathbf{u}(k)\|_2^2 =&\sum_{i=1}^{n}\left(\langle\mathbf{a},\mathbf{x}_i^{(H)}(k+1)\rangle-\langle\mathbf{a},\mathbf{x}_i^{(H)}(k+1)\rangle\right)^2\\ \leq& mp\sum_{i=1}^{n}\left\|\mathbf{x}_i^{(H)}(k+1)-\mathbf{x}_i^{(H)}(k)\right\|_F^2\\ \leq& np\left(\eta c_x Q'(k)\right)^2\\ \leq&\frac{1}{8}\eta\lambda_0\left\|\mathbf{y}-\mathbf{u}(k)\right\|_2^2.\end{aligned}$$

□

# D A General Framework for Analyzing Initialization

In this Section, we provide a general framework to study how much over-parameterization is needed to ensure the Gram matrices at each layer is close to its population ($m\to\infty$) counter part. We begin with some notations. Suppose that we have a sequence of real vector space

$$\mathbb{R}^{n^{(0)}}\to\mathbb{R}^{n^{(1)}}\to\cdots\to\mathbb{R}^{n^{(H)}}.$$

For each pair $(\mathbb{R}^{n^{(h)}},\mathbb{R}^{n^{(l+1)}})$, let $\mathcal{W}^{(h)}\subset\mathcal{L}(\mathbb{R}^{n^{(h-1)}},\mathbb{R}^{n^{(h)}})=\mathbb{R}^{n^{(h)}\times n^{(h-1)}}$ be a subspace, paired with a Gaussian distribution $\mathcal{P}^{(h)}$ over $\mathcal{W}^{(h)}$. We also consider a deterministic linear mapping $a^{(h)}:\mathbb{R}^{n^{(h-1)}}\to\mathbb{R}^{n^{(h)}}$ (with $a^{(1)}=0$). Let $\rho^{(1)},\cdots,\rho^{(H)}$ be a sequence of Lipschitz functions over $\mathbb{R}$.

**Remark D.1.** *For full-connected neural networks, we take $a^{(h)}$ to be zero. For ResNet, we take $a^{(h)}$ to be the identity mapping.*

**Remark D.2.** *In the case of CNN for image recognition, $d,m,\cdots,m$ corresponds to the number of channels and $n^{(h)}$ is the number of grid points at layer $l$ and and $\mathcal{W}^{(h)}$ is the space of all convolutions of a certain kernel size.*

Now we recursively define the output of each layer in this setup. In the following, we use $h\in[H]$ to index layers, $i\in[n]$ to index data points, $\alpha,\beta\in[m]$ or $[d]$ to index channels. We denote $X_i^{(h),[\alpha]}$ an $n^{(h)}$-dimensional vector which is the output at $(h-1)$-the layer. We have the following recursive formula

$$X_i^{(h),[\alpha]}=a^{(h)}(X_i^{(h-1),[\alpha]})+\rho^{(h)}\left(\frac{\sum_\beta W_{[\beta]}^{(h),[\alpha]}X_i^{(h-1),[\beta]}}{\mathbb{I}\left\{h\geq 2\right\}\sqrt{m}}\right)\tag{8}$$

where $W_{[\beta]}^{(h),[\alpha]}$ is $n^{(h)}\times n^{(h-1)}$ matrix generated according to the following rule

- for $l = 1$, $W^{(h),[\alpha]}_{[\beta]}$ is defined for $1 \leq \alpha \leq m$ and $1 \leq \beta \leq d$; for $l > 1$, $W^{(h),[\alpha]}_{[\beta]}$ is defined for $1 \leq \alpha \leq m$ and $1 \leq \beta \leq m$;
- the set of random variables $\{W^{(h),[\alpha]}_{[\beta]}\}_{l,\alpha,\beta}$ are independently generated;
- for fixed $l, \alpha, \beta$, $W^{(h),[\alpha]}_{[\beta]} \sim \mathcal{P}^{(h)}$.

**Remark D.3.** *Note here $X_i^{(h)} = \mathbf{x}_i^{(h)}\sqrt{m}$ for $h \geq 1$ and $X_i^{(h)} = \mathbf{x}_i^{(h)}$ for $h = 0$ in the main text. We change the scaling here to simplify the calculation of expectation and the covariance.*

Our main result for this setup is the following concentration inequality.

**Theorem D.1.** *Then there are $n^{(h)} \times n^{(h)}$-dimensional matrices $K_{ij}^{(h)}$ with $1 \leq l \leq H, 1 \leq i, j \leq n$ and $n^{(h)}$-dimensional vectors $b_i^{(h)}$ with $1 \leq i \leq n$,$1 \leq l \leq H$ such that with probability $1 - \delta$, for any $1 \leq l \leq H, 1 \leq i, j \leq n$,*

$$\left\| \frac{1}{m} \sum_{\alpha=1}^{m} (X_i^{(h),[\alpha]})^\top X_j^{(h),[\alpha]} - K_{ij}^{(h)} \right\|_\infty \leq \mathcal{E}_K \sqrt{\frac{\log(Hn^2(\max_l n^{(h)})^2/\delta)}{m}} \tag{9}$$

*and any $l \in [H], \forall 1 \leq i \leq n$,*

$$\| \frac{1}{m} \sum_{\alpha=1}^{m} X_i^{(h),[\alpha]} - b_i^{(h)} \|_\infty \leq \mathcal{E}_b \sqrt{\frac{\log(Hn^2(\max_l n^{(h)})^2/\delta)}{m}} \tag{10}$$

*where the matrices $K_{ij}^{(h)}$ and vectors $b_i^{(h)}$ can be defined recursively by*

$$\begin{aligned}
K_{ij}^{(0)} =& \sum_{\gamma} (X_i^{(0),[\gamma]})^\top X_j^{(0),[\gamma]}, b_i^{(0)} = 0, \\
K_{ij}^{(h)} =& a^{(h)} K_{ij}^{(h-1)} a^{(h)\top} + \mathop{\mathbb{E}}_{(U,V)} \left( \rho(U) a^{(h)} (b_j^{(h-1)})^\top + (a^{(h)}(b_i^{(h-1)}))\rho(V)^\top + \rho(U)\rho(V)^\top) \right), \\
b_i^{(h)} =& a^{(h)}(b_i^{(h-1)}) + \mathbb{E}_U \rho^{(h)}(U),
\end{aligned}$$

*where*

$$(U, V) \sim \mathcal{N}\left(0, \begin{pmatrix} \mathcal{W}^{(h)}(K_{ii}^{(h-1)}) & \mathcal{W}^{(h)}(K_{ij}^{(h-1)}) \\ \mathcal{W}^{(h)}(K_{ji}^{(h-1)}) & \mathcal{W}^{(h)}(K_{jj}^{(h-1)}) \end{pmatrix}\right)$$

*with*

$$\mathcal{W}^{(h)}(K_{ij}^{(h-1)}) = \mathbb{E}_{W \sim \mathcal{P}^{(h)}} W K_{ij}^{(h-1)} W^\top.$$

*Here $\mathcal{E}_K$ and $\mathcal{E}_b$ satisfy*

$$\begin{aligned}
\mathfrak{W}\mathcal{E}_K = \sqrt{M}\mathcal{E}_b =& \prod_{i=2}^{H} \max\left\{ \left( A_{(h)}^2 + \mathfrak{W}\left( 2A_{(h)}C_{(h)}(1 + \frac{BM}{\lambda^{3/2}}) + C_{(h)}^2 \sqrt{\frac{M}{\lambda}} \right) \right), (A_{(h)} + C_{(h)}\frac{M^{3/2}}{\lambda^{3/2}}) \right\} \times \\
& \max\{\mathfrak{W}\sqrt{16(1 + 2C_{(1)}^2/\sqrt{\pi})M^2}, \sqrt{2MC_{(1)}^2 M}\}
\end{aligned}$$

*where $M, B, \lambda, C_{(h)}, A_{(h)}, \mathfrak{W}_{(h)}$ are defined by:*

- $M = 100 \max_{h\in[H], i\in[n], j\in[n], p\in[n^{(h)}], q\in[n^{(h)}]} |\mathcal{W}^{(h)}(K_{ij}^{(h-1)})_{pq}|$
- $A_{(h)} = \|a^{(h)}\|_{L^\infty \to L^\infty}$
- $B = 100 \max_{h\in[H], i\in[n], p\in[n^{(h)}]} |b_{ip}^{(h)}|$
- $C_{(h)} = |\rho^{(h)}(0)| + \sup_{x\in\mathbb{R}} |(\rho^{(h)})'(x)|$
- $\mathfrak{W}_{(h)} = \|\mathcal{W}^{(h)}\|_{L^\infty \to L^\infty}$*, where* $\mathcal{W}^{(h)}$ *is viewed as a linear operator acting on matrices.*
- $\mathfrak{W} = \max_h \mathfrak{W}_{(h)}$.
- $\lambda = \min_{h\in[H],(i,j)\in[n]\times[n]}$ *least eigenvalue of* $\begin{pmatrix} \mathcal{W}^{(h)}(K_{ii}^{(h-1)}) & \mathcal{W}^{(h)}(K_{ij}^{(h-1)}) \\ \mathcal{W}^{(h)}(K_{ji}^{(h-1)}) & \mathcal{W}^{(h)}(K_{jj}^{(h-1)}) \end{pmatrix}$

This theorem shows if $m$ is large enough, the Gram matrices will concentrate around $\mathbf{K}^{(h)}$s.

*Proof of Theorem D.1.* The proof is by induction. For the base case, $h = 1$, we consider

$$X_i^{(1),[\alpha]} = \rho^{(1)}(\sum_\beta W_{[\beta]}^{(1),[\alpha]} X_i^{(0),[\beta]})$$

Denote

$$U_i^{(1),[\alpha]} = \sum_\beta W_{[\beta]}^{(1),[\alpha]} X_i^{(0),[\beta]}$$

The collection $\{U_i^{(1),[\beta]}\}_{1\le i\le n, 1\le\beta\le m}$ is a mean-zero Gaussian variable with covariance matrix:

$$\begin{aligned}
&\mathbb{E} U_i^{(1),[\alpha]} U_j^{(1),[\beta]\top} \\
=&\mathbb{E} \sum_{\gamma,\gamma'} W_{[\gamma]}^{(1),[\alpha]} X_i^{(0),[\gamma]} X_i^{(0),[\gamma']\top} W_{[\gamma']}^{(1),[\beta]\top} \\
=&\delta_{\alpha\beta} \mathcal{W}^{(1)}(\sum_\gamma X_i^{(0),[\gamma]} X_j^{(0),[\gamma]\top}) \\
=&\delta_{\alpha\beta} \mathcal{W}^{(1)}(K_{ij}^{(0)})
\end{aligned}$$

As a result, we have

$$\mathbb{E}\frac{1}{m}\sum_{i=1}^m X_i^{(1),[\alpha]} X_j^{(1),[\alpha]\top} = K_{ij}^{(1)}$$

and

$$\mathbb{E}\frac{1}{m}\sum_{i=1}^m X_i^{(1),[\alpha]} = b_i^{(1)}$$

Now applying standard Bernstein bound and Hoeffding bound, With probability at least $1-\frac{\delta}{H}$, we have

$$\max_{i,j}\left\|\frac{1}{m}\sum_{i=1}^{m} X_i^{(1),[\alpha]}X_j^{(1),[\alpha]\top}-K_{ij}^{(1)}\right\|_\infty<\sqrt{\frac{16(1+2C_{(1)}^2/\sqrt{\pi})M^2\log(4Hn^2(n^{(1)})^2/\delta)}{m}}$$

and

$$\max_{i,p}\left|\frac{1}{m}\sum_{\alpha=1}^{m} X_{ip}^{(1),[\alpha]}-b_{ip}^{(1)}\right|\le\sqrt{\frac{2C_{(1)}^2M\log(2nn^{(1)}H/\delta)}{m}}$$

Now we prove the induction step. For $h\in[H]$, define

$$\hat{K}_{ij}^{(h)}=\frac{1}{m}\sum_{\gamma}X_i^{(h),[\gamma]}X_j^{(h),[\gamma]\top},\ \text{and}\ \hat{b}_i^{(h)}=\frac{1}{m}\sum_{\gamma}X_i^{(1),[\gamma]}.$$

We use $\mathbb{E}^{(h)}$ to denote expectation conditioned on first $h-1$ layers. Suppose that Equation (9) and (10) hold for $1\le l\le h$. We want to show these equations holds for $l=h+1$ with probability at least $1-\delta/H$. Let $l=h+1$, then

$$X_i^{(h),[\alpha]}=a^{(h)}(X^{(h-1)})+\rho^{(h)}(\sum_{\beta}W_{[\beta]}^{(h),[\alpha]}X_i^{(h-1),[\beta]}/\sqrt{m})$$

Denote

$$U_i^{(h),[\alpha]}=\sum_{\beta}W_{[\beta]}^{(h),[\alpha]}X_i^{(h-1),[\beta]}/\sqrt{m}$$

The collection $\{U_i^{(h),[\beta]}\}_{1\le i\le n,1\le\beta\le m}$ is a mean-zero Gaussian variable with covariance matrix:

$$\begin{aligned}
&\mathbb{E}U_i^{(1),[\alpha]}U_j^{(1),[\beta]\top}\\
&=\mathbb{E}\frac{1}{m}\sum_{\gamma,\gamma'}W_{[\gamma]}^{(h),[\alpha]}X_i^{(0),[\gamma]}X_i^{(h-1),[\gamma']\top}W_{[\gamma']}^{(h),[\beta]\top}\\
&=\delta_{\alpha\beta}\mathcal{W}^{(h)}(\frac{1}{m}\sum_{\gamma}X_i^{(h-1),[\gamma]}X_j^{(h-1),[\gamma]\top})\\
&=\delta_{\alpha\beta}\mathcal{W}^{(h)}(\hat{K}_{ij}^{(h-1)}).
\end{aligned}$$

As a result, we have

$$\begin{aligned}
\mathbb{E}^{(h)}[\hat{K}_{ij}^{(h)}]=&a^{(h)}\hat{K}_{ij}^{(h-1)}a^{(h)\top}+\mathop{\mathbb{E}}_{(U,V)}\Big(\rho(U)^\top a^{(h)}(\hat{b}_j^{(h-1)})+(a^{(h)}(\hat{b}_i^{(h-1)}))^\top\rho(V)+\rho(U)^\top\rho(V))\Big)\\
\mathbb{E}^{(h)}\hat{b}_i^{(h)}=&a^{(h)}(\hat{b}_i^{(h-1)})+\mathbb{E}_U\rho^{(h)}(U)
\end{aligned}$$

with

$$(U,V)\sim\mathcal{N}\left(0,\begin{pmatrix}\mathcal{W}^{(h)}(\hat{K}_{ii}^{(h-1)}) & \mathcal{W}^{(h)}(\hat{K}_{ij}^{(h-1)})\\ \mathcal{W}^{(h)}(\hat{K}_{ji}^{(h-1)}) & \mathcal{W}^{(h)}(\hat{K}_{jj}^{(h-1)})\end{pmatrix}\right)\tag{11}$$

Again by concentration inequality, we can show that with probability at least $1-\delta/H$,

$$\max_{ij}\|\mathbb{E}^{(h)}\hat{K}_{ij}^{(h)}-\hat{K}_{ij}^{(h)}\|_\infty \le \sqrt{\frac{16(1+2C_{(1)}^2/\sqrt{\pi})M^2\log(4Hn^2(n^{(1)})^2/\delta)}{m}}$$

and

$$\max_{i}\|\mathbb{E}^{(h)}\hat{b}_{i}^{(h)}-\hat{b}_{i}^{(h)}\|_\infty \le \sqrt{\frac{2C_{(1)}^2M\log(2nn^{(1)}H/\delta)}{m}}$$

Now it remains to bound the how the perturbation propagates, i.e., we need bound

$$\max_{ij}\|\mathbb{E}^{(h)}\hat{K}_{ij}^{(h)}-K_{ij}^{(h)}\|_2 \text{ and } \max_{i}\|\mathbb{E}^{(h)}\hat{b}_{ij}^{(h)}-b_{ij}^{(h)}\|_2.$$

Through standard calculations, we can show

$$\max_{ij}\|\mathbb{E}^{(h)}\hat{K}_{ij}^{(h)}-K_{ij}^{(h)}\|_\infty \le \left(A_{(h)}^2+\mathfrak{W}\left(2A_{(h)}BC_{(h)}\frac{M}{\lambda^{3/2}}+C_{(h)}^2\sqrt{\frac{M}{\lambda}}\right)\right)\max_{ij}\|\hat{K}_{ij}^{(h-1)}-K_{ij}^{(h-1)}\|_\infty$$
$$+2C_{(h)}\sqrt{M}A_{(h)}\max_{i}\|\hat{b}_{ij}^{(h-1)}-b_{ij}^{(h-1)}\|_2$$

and

$$\max_{i}\|\mathbb{E}^{(h)}\hat{b}_{ij}^{(h)}-b_{ij}^{(h)}\|_2 \le \frac{C_{(h)}M\mathfrak{W}}{\lambda^{3/2}}\max_{ij}\|\hat{K}_{ij}^{(h-1)}-K_{ij}^{(h-1)}\|_\infty+A_{(h)}\max_{i}\|\hat{b}_{ij}^{(h-1)}-b_{ij}^{(h-1)}\|_2$$

Then we have

$$\mathfrak{W}\max_{ij}\|\mathbb{E}^{(h)}\hat{K}_{ij}^{(h)}-K_{ij}^{(h)}\|_\infty \le \left(A_{(h)}^2+\mathfrak{W}\left(2A_{(h)}C_{(h)}(1+\frac{BM}{\lambda^{3/2}})+C_{(h)}^2\sqrt{\frac{M}{\lambda}}\right)\right)$$
$$\max\{\mathfrak{W}\max_{ij}\|\hat{K}_{ij}^{(h-1)}-K_{ij}^{(h-1)}\|_\infty,\sqrt{M}\max_{i}\|\hat{b}_{ij}^{(h-1)}-b_{ij}^{(h-1)}\|_2\}$$

and

$$\sqrt{M}\max_{i}\|\mathbb{E}^{(h)}\hat{b}_{ij}^{(h)}-b_{ij}^{(h)}\|_2 \le (A_{(h)}+C_{(h)}\frac{M^{3/2}}{\lambda^{3/2}})\times$$
$$\max\{\mathfrak{W}\max_{ij}\|\hat{K}_{ij}^{(h-1)}-K_{ij}^{(h-1)}\|_\infty,\sqrt{M}\max_{i}\|\hat{b}_{ij}^{(h-1)}-b_{ij}^{(h-1)}\|_2\}$$

As a result, we have

$$\begin{aligned}&\max\{\mathfrak{W}\max_{ij}\|\mathbb{E}^{(h)}\hat{K}_{ij}^{(h)}-K_{ij}^{(h)}\|_\infty,\sqrt{M}\max_{i}\|\mathbb{E}^{(h)}\hat{b}_{ij}^{(h)}-b_{ij}^{(h)}\|_2\}\\ &\le\max\left\{\left(A_{(h)}^2+\mathfrak{W}\left(2A_{(h)}C_{(h)}(1+\frac{BM}{\lambda^{3/2}})+C_{(h)}^2\sqrt{\frac{M}{\lambda}}\right)\right),(A_{(h)}+C_{(h)}\frac{M^{3/2}}{\lambda^{3/2}})\right\}\times\\ &\max\{\mathfrak{W}\max_{ij}\|\hat{K}_{ij}^{(h-1)}-K_{ij}^{(h-1)}\|_\infty,\sqrt{M}\max_{i}\|\hat{b}_{ij}^{(h-1)}-b_{ij}^{(h-1)}\|_2\}\end{aligned}$$

Now we have proved the theorem. □

**Remark D.4.** *For deep fully-connected neural network, $M$, $C_{(h)}$ and $\mathfrak{W}$ are constants and $A_{(h)}$ and $B = 0$. For ResNet and convolutional ResNet, $M, A_(h), B$ and $\mathfrak{W}$ are constants and $C_{(h)}$ is proportional $\frac{c_\lambda \lambda^{3/2}}{H}$.*

Note in the bound, $\lambda$ depends on the specific choice of $a^{(h)}$s and architecture. Here we show for the ResNet structure, $\lambda$ has a positive lower bound.

**Lemma D.1.** *If $a^{(h)}$ is the identity mapping then $\lambda \geq \min_{(i,j)\in[n]\times[n]} \lambda_{\min} \left(\begin{array}{cc} K_{ii}^{(0)} & K_{ij}^{(0)} \\ K_{ji}^{(0)} & K_{jj}^{(0)} \end{array}\right)$.*

*Proof of Lemma D.1.* First recall

$$(U,V) \sim \mathcal{N}\left(0, \left(\begin{array}{cc} \mathcal{W}^{(h)}(K_{ii}^{(h-1)}) & \mathcal{W}^{(h)}(K_{ij}^{(h-1)}) \\ \mathcal{W}^{(h)}(K_{ji}^{(h-1)}) & \mathcal{W}^{(h)}(K_{jj}^{(h-1)}) \end{array}\right)\right)$$

Then we compute

$$\begin{aligned} & K_{ij}^{(h)} - b_i^{(h)} b_j^{(h)\top} \\ & = a^{(h)} K_{ij}^{(h-1)} a^{(h)\top} + \mathop{\mathbb{E}}_{(U,V)} \left(\rho(U) a^{(h)} (b_j^{(h-1)})^\top + (a^{(h)}(b_i^{(h-1)}))\rho(V)^\top + \rho(U)\rho(V)^\top)\right) \\ & - \left(a^{(h)}(b_i^{(h-1)}) + \mathbb{E}_U \rho^{(h)}(U)\right) \left(a^{(h)}(b_j^{(h-1)}) + \mathbb{E}_V \rho^{(h)}(V))\right)^\top \\ & = a^{(h)} \left(K_{ij}^{(h-1)} - b_i^{(h-1)} b_j^{(h-1)\top}\right) a^{(h)\top} + \mathop{\mathbb{E}}_{(U,V)} \left(\rho(U)\rho(V)^\top)\right) - \left(\mathbb{E}_U \rho^{(h)}(U)\right) \left(\mathbb{E}_V \rho^{(h)}(V))\right)^\top \end{aligned}$$

For ResNet, $a^{(h)}$ is the identity mapping so we have

$$\begin{aligned} & K_{ij}^{(h)} - b_i^{(h)} b_j^{(h)\top} \\ & = K_{ij}^{(h-1)} - b_i^{(h-1)} b_j^{(h-1)\top} + \mathop{\mathbb{E}}_{(U,V)} \left(\rho(U)\rho(V)^\top)\right) - \left(\mathbb{E}_U \rho^{(h)}(U)\right) \left(\mathbb{E}_V \rho^{(h)}(V))\right)^\top. \end{aligned}$$

To proceed, we calculate

$$\begin{aligned} & \left(\begin{array}{cc} K_{ii}^{(h)} & K_{ij}^{(h)} \\ K_{ji}^{(h)} & K_{jj}^{(h)} \end{array}\right) - \left(\begin{array}{c} b_i^{(h)} \\ b_j^{(h)} \end{array}\right) \left(\begin{array}{c} b_i^{(h)\top}, b_j^{(h)\top} \end{array}\right) \\ = & \left(\begin{array}{cc} K_{ii}^{(h-1)} & K_{ij}^{(h-1)} \\ K_{ji}^{(h-1)} & K_{jj}^{(h-1)} \end{array}\right) - \left(\begin{array}{c} b_i^{(h-1)} \\ b_j^{(h-1)} \end{array}\right) \left(\begin{array}{c} b_i^{(h-1)\top}, b_j^{(h-1)\top} \end{array}\right) \\ & + \left(\mathbb{E}_{U,V} \left(\begin{array}{cc} \rho^{(h)}(U)\rho^{(h)}(U)^\top & \rho^{(h)}(U)\rho^{(h)}(V)^\top \\ \rho^{(h)}(V)\rho^{(h)}(U)^\top & \rho^{(h)}(V)\rho^{(h)}(V)^\top \end{array}\right) - \mathbb{E}_{U,V} \left(\begin{array}{c} \rho(U) \\ \rho(V) \end{array}\right) \mathbb{E}_{U,V} \left(\begin{array}{c} \rho(U)^\top, \rho(V)^\top \end{array}\right)\right) \\ \geq & \left(\begin{array}{cc} K_{ii}^{(h-1)} & K_{ij}^{(h-1)} \\ K_{ji}^{(h-1)} & K_{jj}^{(h-1)} \end{array}\right) - \left(\begin{array}{c} b_i^{(h-1)} \\ b_j^{(h-1)} \end{array}\right) \left(\begin{array}{c} b_i^{(h-1)\top}, b_j^{(h-1)\top} \end{array}\right) \end{aligned}$$

As a result, we have

$$
\begin{aligned}
&\lambda_{\min}\begin{pmatrix} K_{ii}^{(h)} & K_{ij}^{(h)} \\ K_{ji}^{(h)} & K_{jj}^{(h)} \end{pmatrix} \\
\geq&\lambda_{\min}\begin{pmatrix} K_{ii}^{(h)} & K_{ij}^{(h)} \\ K_{ji}^{(h)} & K_{jj}^{(h)} \end{pmatrix} - \begin{pmatrix} b_i^{(h)} \\ b_j^{(h)} \end{pmatrix}\begin{pmatrix} b_i^{(h)\top}, b_j^{(h)\top} \end{pmatrix} \\
\geq& \min\lambda_{\min}\begin{pmatrix} K_{ii}^{(h-1)} & K_{ij}^{(h-1)} \\ K_{ji}^{(h-1)} & K_{jj}^{(h-1)} \end{pmatrix} - \begin{pmatrix} b_i^{(h-1)} \\ b_j^{(h-1)} \end{pmatrix}\begin{pmatrix} b_i^{(h-1)\top}, b_j^{(h-1)\top} \end{pmatrix} \\
\geq&\cdots \\
\geq&\lambda_{\min}\begin{pmatrix} K_{ii}^{(0)} & K_{ij}^{(0)} \\ K_{ji}^{(0)} & K_{jj}^{(0)} \end{pmatrix} - \begin{pmatrix} b_i^{(0)} \\ b_j^{(0)} \end{pmatrix}\begin{pmatrix} b_i^{(0)\top}, b_j^{(0)\top} \end{pmatrix} \\
=&\lambda_{\min}\begin{pmatrix} K_{ii}^{(0)} & K_{ij}^{(0)} \\ K_{ji}^{(0)} & K_{jj}^{(0)} \end{pmatrix}
\end{aligned} \tag{12}
$$

We now prove the theorem. □

## D.1 Dependence of $\lambda_0$ on Layers for ResNet.

The following proposition shows that for ResNet the dependence on $H$ is only via the constant $c_H := \frac{c_\lambda \lambda^{3/2}}{H^2} \sim \frac{1}{H^2}$.

**Proposition D.1.** *Assume $\sigma(\cdot)$ is analytic and not a polynomial function. Recall that*

$$
\mathbf{K}_{ij}^{(H)} = c_H \mathbf{K}_{ij}^{(H-1)} \cdot \mathbb{E}_{(u,v)^\top \sim N\left(\mathbf{0}, \begin{pmatrix} \mathbf{K}_{ii}^{(H-1)} & \mathbf{K}_{ij}^{(H-1)} \\ \mathbf{K}_{ji}^{(H-1)} & \mathbf{K}_{jj}^{(H-1)} \end{pmatrix}\right)} \left[\sigma'(u)\sigma'(v)\right],
$$

*then $\lambda_0 := \lambda_{\min}(\mathbf{K}^{(H)}) \geq c_H \kappa$, where $\kappa$ is a constant that only depends on the activation $\sigma$ and the input data $\{x_i\}$. In particular, $\kappa$ does not depend on the depth.*

*Proof of Proposition D.1.* First note $\mathbf{K}_{ii}^{(H-1)} \in [1/c_{x,0}^2, c_{x,0}^2]$ for all $H$, so it is in a bounded range that does not depend on the depth (c.f. Lemma B.1). Define a function

$$
\mathbf{G} : \mathbb{R}^{n \times n} \to \mathbb{R}^{n \times n}
$$

such that $\mathbf{G}(\mathbf{K})_{ij} = \mathbf{K}_{ij} \mathbb{E}_{(u,v)^\top \sim N\left(\mathbf{0}, \begin{pmatrix} \mathbf{K}_{ii} & \mathbf{K}_{ij} \\ \mathbf{K}_{ji} & \mathbf{K}_{jj} \end{pmatrix}\right)} [\sigma'(u)\sigma'(v)]$. Now define a scalar function

$$
g(\lambda) = \min_{\mathbf{K}: \mathbf{K} \succ 0, \frac{1}{c_{x,0}^2} \leq \mathbf{K}_{ii} \leq c_{x,0}, \lambda(\mathbf{K}) \geq \lambda} \lambda_{\min}(\mathbf{G}(\mathbf{K}))
$$

with

$$\lambda(\mathbf{K}) = \min_{ij} \begin{pmatrix} \mathbf{K}_{ii} & \mathbf{K}_{ij} \\ \mathbf{K}_{ji} & \mathbf{K}_{jj} \end{pmatrix}.$$

By Lemma D.1. $\lambda(\mathbf{K}^{(H-1)}) \geq c_H g(\lambda^{(H-1)}) \geq c_H g(\lambda^{(0)})$.

Next, let $\mathbf{U}\mathbf{D}\mathbf{U}^\top = \mathbf{K}^{(H-1)}$ be the eigen-ecomposition of $\mathbf{K}$, and $\mathbf{Z} = \mathbf{D}^{1/2}\mathbf{U}^\top$ be the feature embedding into $\mathbb{R}^n$. Since $\begin{pmatrix} \mathbf{z}_i^\top \mathbf{z}_i & \mathbf{z}_i^\top \mathbf{z}_j \\ \mathbf{z}_j^\top \mathbf{z}_i & \mathbf{z}_j^\top \mathbf{z}_j \end{pmatrix}$ is full rank, then $\mathbf{z}_i \notin \text{span}(\mathbf{z}_j)$. Then using Lemma D.2 , $g(\lambda^{(0)}) > 0$. Thus we have established that $\lambda_{\min}(\mathbf{K}^{(H)}) \geq c_H g(\lambda^{(0)})$ , where $g(\lambda^{(0)})$ only depends on the input data and activation $\sigma$. In particular, it is independent of the depth. □

**Lemma D.2.** *Assume $\sigma(\cdot)$ is analytic and not a polynomial function. Consider data $Z = \{\mathbf{z}_i\}_{i \in [n]}$ of $n$ non-parallel points (meaning $\mathbf{z}_i \notin$ span$(\mathbf{z}_j)$ for all $i \neq j$). Define*

$$\mathbf{G}_{ij}(Z) = \mathbb{E}_{\mathbf{w} \sim N(\mathbf{0}, \mathbf{I})}[\sigma'(\mathbf{w}^\top \mathbf{z}_i)\sigma'(\mathbf{w}^\top \mathbf{z}_j)(\mathbf{z}_i^\top \mathbf{z}_j)].$$

*Then $\lambda_{\min}(\mathbf{G}(Z)) > 0$.*

*Proof of Lemma D.2.* The feature map induced by the kernel $\mathbf{G}$ is given by $\phi_{\mathbf{z}}(\mathbf{w}) = \sigma'(\mathbf{w}^\top \mathbf{z})\mathbf{z}$. To show that $\mathbf{G}(Z)$ is strictly positive definite, we need to show $\phi_{\mathbf{z}_1}(\mathbf{w}), \ldots, \phi_{\mathbf{z}_n}(\mathbf{w})$ are linearly independent functions. Assume that there are $a_i$ such that

$$0 = \sum_i a_i \phi_{\mathbf{z}_i} = \sum_i a_i \sigma'(\mathbf{w}^\top \mathbf{z}_i)\mathbf{z}_i.$$

We wish to show that $a_i = 0$. Differentiating the above equation $(n-2)$ times with respect to $w$, we have

$$0 = \sum_i \left(a_i \sigma^n(\mathbf{w}^\top \mathbf{z}_i)\right) \mathbf{z}_i^{\otimes(n-1)}.$$

Using Lemma D.3, we know $\left\{\mathbf{z}_i^{\otimes(n-1)}\right\}_{i=1}^n$ are linearly independent. Therefore, we must have $a_i \sigma^n(\mathbf{w}^\top \mathbf{z}_i) = 0$ for all $i$. Now choosing a $\mathbf{w}$ such that $\sigma^{(n)}\left(\mathbf{w}^\top \mathbf{z}_i\right) \neq 0$ for all $i \in [n]$ (such $\mathbf{w}$ exists because of our assumption on $\sigma$), we have $a_i = 0$ for all $i \in [n]$. □

**Lemma D.3.** *If $\mathbf{v}_1, \ldots, \mathbf{v}_n \in \mathbb{R}^d$ satisfy that $\|\mathbf{v}_i\|_2 = 1$ and non-parallel (meaning $\mathbf{v}_i \notin$ span$(\mathbf{v}_j)$ for $i \neq j$), then the matrix $\left[vec\left(\mathbf{v}_1^{\otimes n}\right), \ldots, vec\left(\mathbf{v}_n^{\otimes n}\right)\right] \in \mathbb{R}^{d^n \times n}$ has rank-$n$.*

*Proof of Lemma D.3.* We prove by induction. For $n = 2$, $v_1 v_1^\top, v_2 v_2^\top$ are linearly independent under the non-parallel assumption. By induction suppose $\{\text{vec}\left(\mathbf{v}_1^{\otimes n-1}\right), \ldots, \text{vec}\left(\mathbf{v}_{n-1}^{\otimes n-1}\right)\}$ are linearly independent. Suppose the conclusion does not hold, then there exists $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ not identically 0, such that

$$\sum_{i=1}^n \alpha_i \text{vec}\left(\mathbf{v}_i^{\otimes n}\right) = 0,$$

which implies for $p = 1, \ldots, d$

$$\sum_{i=1}^{n} (\alpha_i \mathbf{v}_{i,p}) \text{vec}\left(\mathbf{v}_i^{\otimes(n-1)}\right) = 0.$$

Note by induction hypothesis any size $(n-1)$ subset of $\left\{\text{vec}\left(\mathbf{v}_1^{\otimes(n-1)}\right), \ldots, \text{vec}\left(\mathbf{v}_n^{\otimes(n-1)}\right)\right\}$ is linearly independent. This implies if $\alpha_i \mathbf{v}_{i,p} = 0$ for some $i \in [n]$ and $p \in [d]$, then we must have $\alpha_j \mathbf{v}_{j,p} = 0$ for all $j \in [n]$. Combining this observation with the assumption that every $\mathbf{v}_i$ is non-zero, there must exist $p \in [d]$ such that $\mathbf{v}_{i,p} \neq 0$ for all $i \in [n]$. Without loss of generality, we assume $\mathbf{v}_{i,1} \neq 0$ for all $i \in [n]$.

Next, note if there exists $\alpha_i = 0$, then we have $\alpha_j = 0$ for all $j \in [n]$ because $\mathbf{v}_{j,p} \neq 0$ for all $j \in [n]$ and the linear independence induction hypothesis. Therefore from now on we assume $\alpha_i \neq 0$ for all $i \in [n]$.

For any $p \in [d]$ , we have

$$\sum_{i=1}^{n} (\alpha_i \mathbf{v}_{i,p}) \text{vec}\left(\mathbf{v}_i^{\otimes(n-1)}\right) = 0 \text{ and } \sum_{i=1}^{n} (\alpha_i \mathbf{v}_{i,1}) \text{vec}\left(\mathbf{v}_i^{\otimes(n-1)}\right) = 0.$$

By multiplying the second equation by $\frac{\mathbf{v}_{1,p}}{\mathbf{v}_{1,1}}$ and subtracting,

$$\sum_{i=2}^{n} (\alpha_i \mathbf{v}_{i,p} - \alpha_i \frac{\mathbf{v}_{1,p}}{\mathbf{v}_{1,1}} \mathbf{v}_{i,1}) \text{vec}\left(\mathbf{v}_i^{\otimes(n-1)}\right) = 0.$$

Using the linear independence induction hypothesis, we know for $i = 2, \ldots, n$:

$$\frac{\mathbf{v}_{i,p}}{\mathbf{v}_{1,1}} = \frac{\mathbf{v}_{1,p}}{\mathbf{v}_{1,1}}.$$

Therefore we know

$$\frac{\mathbf{v}_{1,p}}{\mathbf{v}_{1,1}} = \cdots = \frac{\mathbf{v}_{n,p}}{\mathbf{v}_{n,1}}.$$

Thus there exists $c_2, \ldots, c_d \in \mathbb{R}^d$ such that

$$\mathbf{v}_{i,p} = c_p \mathbf{v}_{i,1} \text{ for all } i \in [n].$$

Note this implies all $\mathbf{v}_i$, $i \in [n]$ are on the same line. This contradicts with the non-parallel assumption. □

# E Useful Technical Lemmas

**Lemma E.1.** *Given a set of matrices $\{\mathbf{A}_i, \mathbf{B}_i : i \in [n]\}$, if $\|\mathbf{A}_i\|_2 \leq M_i$, $\|\mathbf{B}_i\|_2 \leq M_i$ and $\|\mathbf{A}_i - \mathbf{B}_i\|_F \leq \alpha_i M_i$, we have*

$$\left\|\prod_{i=1}^{n} \mathbf{A}_i - \prod_{i=1}^{n} \mathbf{B}_i\right\|_F \leq \left(\sum_{i=1}^{n} \alpha_i\right) \prod_{i=1}^{n} M_i.$$

*Proof of Lemma E.1.*

$$\begin{aligned}
&\left\|\prod_{i=1}^{n}\mathbf{A}_i - \prod_{i=1}^{n}\mathbf{B}_i\right\|_F \\
=&\left\|\sum_{i=1}^{n}\left(\prod_{j=1}^{i-1}\mathbf{A}_j\right)(\mathbf{A}_i-\mathbf{B}_i)\left(\prod_{k=i+1}^{n}\mathbf{B}_k\right)\right\|_F \\
\le&\sum_{i=1}^{n}\left\|\left(\prod_{j=1}^{i-1}\mathbf{A}_j\right)(\mathbf{A}_i-\mathbf{B}_i)\left(\prod_{k=i+1}^{n}\mathbf{B}_k\right)\right\|_F \\
\le&\left(\sum_{i=1}^{n}\alpha_i\right)\prod_{i=1}^{n}M_i.
\end{aligned}$$

□

Here fix a differentiable activation function $\rho$ with $C_\rho = 10\sqrt{|\rho(0)|^2 + \sup|\rho'(x)|^2}$

**Lemma E.2.** *Given an function $\rho : \mathbb{R} \to \mathbb{R}$ whose Lipschitz and smoothness constants are bounded by $L$, we define $F : \mathbb{R}^{2\times 2} \to \mathbb{R}$, $F(\mathbf{A}) = \mathbb{E}_{(u,v)^\top\sim\mathcal{N}(0,\mathbf{A})}\rho(u)\rho(v)$. Then we have for $\mathbf{A},\mathbf{B} \succ 0$,*

$$|F(\mathbf{A}^2) - F(\mathbf{B}^2)| \le 10L\max\{\|\mathbf{A}\|_2, \|\mathbf{B}\|_2\}\|\mathbf{A}-\mathbf{B}\|_2$$

*Proof of Lemma E.2.*

$$\begin{aligned}
&F(\mathbf{A}^2) - F(\mathbf{B}^2) \\
=&\mathbb{E}_{(u,v)\sim\mathcal{N}(0,\mathbf{A}^2)}\rho(u)\rho(v) - \mathbb{E}_{(u,v)\sim\mathcal{N}(0,\mathbf{B}^2)}\rho(u)\rho(v) \\
=&\mathbb{E}_{\mathbf{u}\sim N(0,\mathbf{I})}\rho((\mathbf{Au})_1)\rho((\mathbf{Au})_2) - \mathbb{E}_{\mathbf{u}\sim N(0,\mathbf{I})}\rho((\mathbf{Bu})_1)\rho((\mathbf{Bu})_2) \\
\le&10L\max\{\|\mathbf{A}\|_2, \|\mathbf{B}\|_2\}\|\mathbf{A}-\mathbf{B}\|_2
\end{aligned}$$

□

**Lemma E.3.** *Given an function $\rho : \mathbb{R} \to \mathbb{R}$ whose Lipschitz and smoothness constants are bounded by $L$, we define $F : \mathbb{R}^{2\times 2} \to \mathbb{R}$, $F(\mathbf{A}) = \mathbb{E}_{(u,v)^\top\sim\mathcal{N}(0,\mathbf{A})}\rho(u)\rho(v)$. Then we any for $\mathbf{A},\mathbf{B} \succcurlyeq \lambda\mathbf{I}$,*

$$|F(\mathbf{A}) - F(\mathbf{B})| \le \frac{40L}{\sqrt{\lambda}}\sqrt{\max\{\|\mathbf{A}\|_2, \|\mathbf{B}\|_2\}}\|\mathbf{A}-\mathbf{B}\|_\infty$$

*Proof of Lemma E.3.*

$$2(\mathbf{A}-\mathbf{B}) = (\sqrt{\mathbf{A}}+\sqrt{\mathbf{B}})(\sqrt{\mathbf{A}}-\sqrt{\mathbf{B}}) + (\sqrt{\mathbf{A}}-\sqrt{\mathbf{B}})(\sqrt{\mathbf{A}}+\sqrt{\mathbf{B}})$$

When $\mathbf{B}-\mathbf{A}$ is infinitesimal, we have

$$d\mathbf{A} = \sqrt{\mathbf{A}}d\sqrt{\mathbf{A}} + (d\sqrt{\mathbf{A}})\sqrt{\mathbf{A}}$$

Choose a coordinate system where $\sqrt{\mathbf{A}}$ is diagnized to be $\text{diag}(\lambda_1, \lambda_2)$, then we have

$$d\mathbf{A} = \begin{pmatrix} 2\lambda_1(d\sqrt{\mathbf{A}})_{11} & (\lambda_1+\lambda_2)(d\sqrt{\mathbf{A}})_{12} \\ (\lambda_1+\lambda_2)(d\sqrt{\mathbf{A}})_{21} & 2\lambda_2(d\sqrt{\mathbf{A}})_{22} \end{pmatrix}$$

Under this coordinate system, we have

$$\|d\sqrt{\mathbf{A}}\|_2 \le \|d\sqrt{\mathbf{A}}\|_F \le 2\|d\sqrt{\mathbf{A}}\|_\infty \le \frac{1}{\sqrt{\lambda}}\|d\mathbf{A}\|_\infty \le \frac{1}{\sqrt{\lambda}}\|d\mathbf{A}\|_2$$

Integrating this, we have

$$\|\sqrt{\mathbf{A}} - \sqrt{\mathbf{B}}\|_2 \le \frac{2}{\sqrt{\lambda}}\|\mathbf{A} - \mathbf{B}\|_2.$$

Now applying the previous Lemma E.2 we obtain the desired result.

□

**Lemma E.4.** *Given a matrix* $\mathbf{W} \in \mathbb{R}^{m \times cm}$ *with* $\mathbf{W}_{i,j} \sim N(0,1)$*, where* $c$ *is a constant. We have with probability at least* $1 - \exp\left(-\frac{\left(c_{w,0}-\sqrt{c}-1\right)^2 m}{2}\right)$

$$\|\mathbf{W}\|_2 \le c_{w,0}\sqrt{m},$$

*where* $c_{w,0} > \sqrt{c} + 1$ *is a constant.*

*Proof of Lemma E.4.* The lemma is a consequence of well-known deviations bounds concerning the singular values of Gaussian random matrices Vershynin [2010]

$$P\left(\lambda_{\max}(\mathbf{W}) > \sqrt{m} + \sqrt{cm} + t\right) \le e^{t^2/2}.$$

Choosing $t = (c_{w,0} - \sqrt{c} - 1)\sqrt{m}$, we prove the lemma. □