TOWARDS AN UNDERSTANDING OF RESIDUAL NETWORKS USING NEURAL TANGENT HIERARCHY

by

Yuqing Li

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of Mathematics

West Lafayette, Indiana

May 2021

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Nung Kwan Yip, Chair

Department of Mathematics

Dr. Vinayak Rao

Department of Statistics Department of Statistics and Computer Science (by courtesy)

Dr. Samy Tindel

Department of Mathematics

Dr. Guang Lin

Department of Mathematics, Statistics and School of Mechanical Engineering

Approved by:

Dr. Plamen Stefanov

To my parents.

To my beloved meemaw and deceased papaw.

To my cutest stuffed dog, who proudly and shamelessly named himself Dr. Doge.

To my undergraduate friend, William Chengfei Wu, for his great jokes on Dr. Doge.

To my teddy bears, especially the largest one whom I found abandoned by someone else near a trash can.

ACKNOWLEDGMENTS

First of all, most importantly, I shall give my sincere appreciation to my supervisor, Prof. Aaron Nung Kwan Yip, who has been giving me tremendous support on my study, research and daily life.

Next, I gratefully acknowledge Prof. Vinayak Rao, Prof. Samy Tindel, and Prof. Guang Lin for their kindness and readiness to serve on my thesis committee. I would also like to thank the faculty and staff of the Department of Mathematics for everything during my study at Purdue.

I have been fortunate to spend more than three years working with Prof. Vinayak Rao, and would like to thank him for his help and patience. Prof. Rao is an inspiration, both in how much he knows and how much he still wants to learn. I am always inspired by him via his thorough understanding on statistics and machine learning. I benefited immensely from his ideas, his weekly group discussions and his feedback, and from the general confidence that I was in very good hands.

I would like to give special thanks to my roommate, semi-academic brother, semi-academic advisor and friend, Prof. Tao Luo. When I was a second-year graduate student, he changed my perspective on mathematics and trained me to have a solid foundation in mathematics. I have been receiving numerous advice from him hereafter.

Finally and most importantly, I would like to thank my beloved mommy Limin Zhang, and my father Jun Li for their endless love and support. This family means everything to me, and I dedicate this thesis to them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Deep learning has become an important toolkit for data science and artificial intelligence. In contrast to its practical success across a wide range of fields, theoretical understanding of the principles behind the success of deep learning has been an issue of controversy. Optimization, as an important component of theoretical machine learning, has attracted much attention. The optimization problems induced from deep learning is often non-convex and non-smooth, which is challenging to locate the global optima. However, in practice, global convergence of first-order methods like gradient descent can be guaranteed for deep neural networks. In particular, gradient descent yields zero training loss in polynomial time for deep neural networks despite its non-convex nature. Besides that, another mysterious phenomenon is the compelling performance of Deep Residual Network (ResNet). Not only does training ResNet require weaker conditions [1], the employment of residual connections by ResNet even enables first-order methods to train the neural networks with an order of magnitude more layers [2]. Advantages arising from the usage of residual connections remain to be discovered.

In this thesis, we demystify these two phenomena accordingly. Firstly, we contribute to further understanding of gradient descent. The core of our analysis is the neural tangent hierarchy (NTH) [3] that captures the gradient descent dynamics of deep neural networks. A recent work [4] introduced the Neural Tangent Kernel (NTK) and proved that the limiting NTK describes the asymptotic behavior of neural networks trained by gradient descent in the infinite width limit. The NTH outperforms the NTK in two ways: (i) It can directly study the time variation of NTK for neural networks. (ii) It improves the result to non-asymptotic settings. Moreover, by applying NTH to ResNet with smooth and Lipschitz activation function, we reduce the requirement on the layer width m with respect to the number of training samples n from quartic to cubic, obtaining a state-of-the-art result. Secondly, we extend our scope of analysis to structural properties of deep neural networks. By making fair and consistent comparisons between fully-connected network and ResNet, we suggest strongly that the particular skip-connection architecture possessed by ResNet is the main reason for its triumph over fully-connected network.

1. INTRODUCTION

1.1 Introduction

Deep neural networks have achieved transcendent performance in a wide range of tasks such as speech recognition, computer vision, and natural language processing. There are various methods to train neural networks, such as first-order gradient based methods, which have been proven to achieve satisfactory results [5]. Experiments in [6] established that, even though with a random labeling of the training images, if one trains the state-of-the-art convolutional network for image classification using stochastic gradient descent, the network is still able to fit them well. There are numerous works trying to demystify such phenomenon theoretically. Du et al. [7] proved that gradient descent can obtain zero training loss for two-layer networks, and Zou et al. [8] analyzed the convergence of stochastic gradient descent on networks assembled with Rectified Linear Unit (ReLU) activation function. All these neural networks are heavily overparameterized: the number of learnable parameters is much larger than the number of the training samples. It is widely accepted by the machine learning community that overparameterization enables the neural network to fit all training data, and it brings no harm to the power of its generalization, i.e., the ability to predict well on unseen data [9]. In particular, the deep neural networks that evaluated positions and selected moves for the well-known program AlphaGo are highly overparameterized [10], [11].

Another advance is the outstanding performance of Deep Residual Network (ResNet) proposed by He et al. [2]. ResNet is arguably one of the most groundbreaking works in deep learning, in that it can train up to hundreds or even thousands of layers and still achieves compelling performance. Recent works have shown that ResNet can utilize the features in transfer learning with better efficiency, and its residual link structure enables faster convergence of the training loss [12], [13]. Theoretically, Hardt and Ma [14] proved that for any residual linear networks with arbitrary depth, there are no spurious local optima. Du et al. [1] showed that in the scope of the convergence of gradient descent via overparameterization for different networks, training ResNet requires weaker conditions compared with fully-connected networks. In this thesis, we make fair comparisons between ResNet and fully-connected networks by enforcing exactly same assumptions on input samples and acti-

vation functions, such settings would thus potentially enable us to explore more structural benefits of deep neural networks.

1.2 Fully-connected Neural Networks

We present two types of neural network architectures respectively in Section 1.2 and Section 1.3, one is the fully-connected network, the other is ResNet. A fully-connected neural network consists of a series of fully connected layers that connect every neuron in one layer to every neuron in the other layer. Pictorially, a fully connected layer is represented in Figure 1.1.

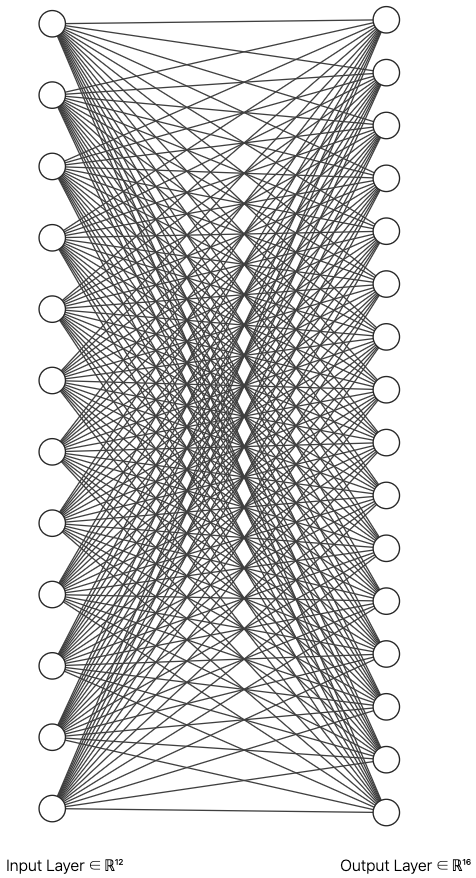

Figure 1.1. A fully connected layer in a deep network

According to the classical universal approximation theorem [15], a two-layer neural network with sigmoid-like activation functions is sufficient to represent any continuous function

on the unit cube. However, its single hidden layer might be massive and the network is prone to overfitting the data. Overfitting is a serious problem in machine learning, it happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data [16]. Therefore, the machine learning community has a common trend that network architecture needs to go deeper. Naturally, given fully connected layers, it is directly possible to form a network by stacking more of them, as depicted in Figure 1.2.

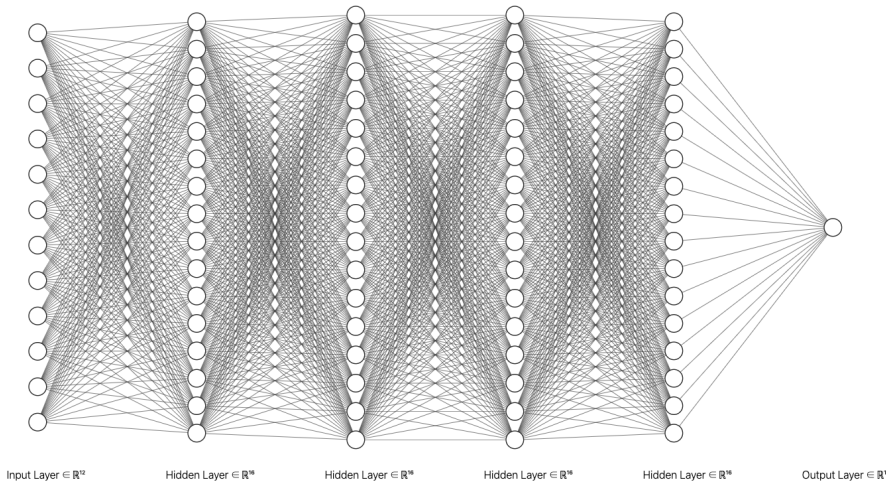

Figure 1.2. Multilayer deep fully-connected network

Let $\boldsymbol{x} \in \mathbb{R}^d$ be an input sample, then the network has d input nodes. Moreover, we have a series of weight matrices $\left\{\boldsymbol{W}^{[l]}\right\}_{l=1}^{L}$. Note that $\boldsymbol{W}^{[1]} \in \mathbb{R}^{m\times d}$ is the first weight matrix, and $\boldsymbol{W}^{[l]} \in \mathbb{R}^{m\times m}$ is the weight at the l-th layer, for $2 \le l \le L$. Let $\boldsymbol{x}^{[l]}$ be the output of layer l, with $\boldsymbol{x}^{[0]} = \boldsymbol{x}$. We consider the fully-connected network given below:

$$\begin{aligned} \boldsymbol{x}^{[l]} &= \frac{1}{\sqrt{m}}\sigma(\boldsymbol{W}^{[l]}\boldsymbol{x}^{[l-1]}), \\ f_{\mathrm{nn}}(\boldsymbol{x},\boldsymbol{\theta}) &= \boldsymbol{a}^{\intercal}\boldsymbol{x}^{[L]}, \end{aligned} \tag{1.1}$$

where $\sigma(\cdot)$ is applied coordinate-wisely to its input, and $f_{\mathrm{nn}}(\boldsymbol{x},\boldsymbol{\theta})$ is the output function. Specifically, for the case in Figure 1.2, the parameters are set to be $d = 12, m = 16, L = 4$.

The scaling factor $1/\sqrt{m}$ is key to obtaining a consistent asymptotic behavior of neural networks as the width m of the hidden layers grow to infinity [4]. In the infinite-width limit,

the output function at initialization converges to a Gaussian distribution, and it follows a linear differential equation during training. There are, of course, other scaling factors that are employed extensively. For instance, the mean-field scaling $1/m$. For two-layer networks, a line of papers [17]–[19] used optimal transport theory equipped with the mean-field scaling to establish that for infinitely wide neural networks, the empirical distribution of the neural network parameters can be described as a Wasserstein gradient flow. However, their results are limited to two-layer networks and may require an exponential amount of overparametrization. The current situation of neural network study is similar to an early era of statistical mechanics, when we observe different states of a matter at several discrete conditions without the guidance of a unified phase diagram. Therefore, inspired by that, a recently published work [20] presented a systematic and comprehensive analysis in drawing the first phase diagram for two-layer neural networks at the infinite-width limit, in pursuit of a complete characterization of its dynamical regimes and their dependence on different scaling factors.

1.3 ResNet

As mentioned both in Section 1.2 and [21], deeper and deeper network architectures are being developed nowadays. There is even a mathematical proof in [22] that reveals the utility of having deeper networks than that of wider networks. However, increasing network depth also introduce the issues of vanishing gradients [23] and degradation [2]. As the gradient is computed out by backward propagation, repeated multiplication with small weights renders it ineffectively small. Vanishing gradients, however, has been largely addressed by some normalizing tricks [24], [25].

When deeper networks are able to start converging, a degradation problem has been exposed: the training accuracy gets saturated and then degrades rapidly [26]. This is counter-intuitive in that by providing our model with more parameters, it shall be able to fit the training data at least as good as its predecessor. Surprisingly, such degradation is not due to overfitting. The degradation of training performances reveals that not all networks are similarly easy to optimize by brutal force. Moreover, the problem suggests that it might be

hard for the solvers to learn identity mappings with multiple nonlinear layers. Ultimately, this conjecture motivates the setup for learning small residuals and directly adding them to the input. The core idea of ResNet [2] is the employment of residual learning building block,

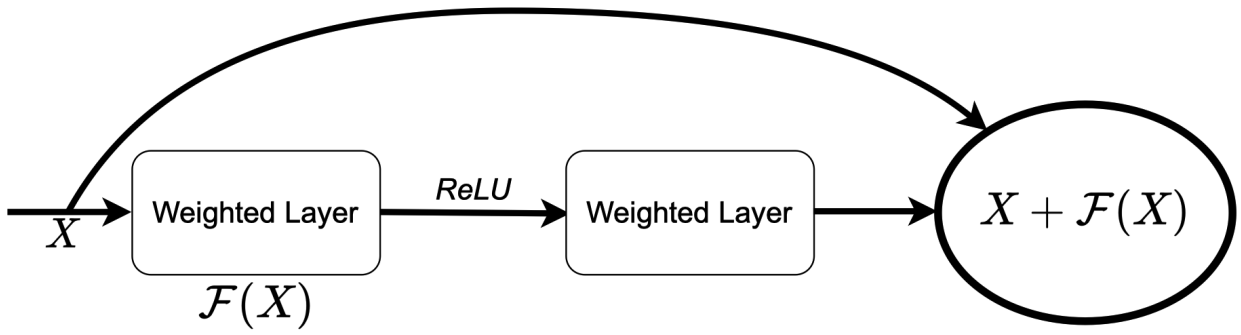

Figure 1.3. Residual learning: a building block

as shown in Figure 1.3.

We use the same notations as the fully-connected neural networks to rigorously define our version of ResNet:

$$\begin{aligned} \boldsymbol{x}^{[1]} &= \sqrt{\frac{c_\sigma}{m}}\sigma(\boldsymbol{W}^{[1]}\boldsymbol{x}), \\ \boldsymbol{x}^{[l]} &= \boldsymbol{x}^{[l-1]} + \frac{c_{\text{res}}}{L\sqrt{m}}\sigma(\boldsymbol{W}^{[l]}\boldsymbol{x}^{[l-1]}), \quad \text{for } 2 \le l \le L, \\ f_{\text{res}}(\boldsymbol{x}, \boldsymbol{\theta}) &= \boldsymbol{a}^\intercal \boldsymbol{x}^{[L]}, \end{aligned} \tag{1.2}$$

where the constant $c_\sigma = \left(\mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[\sigma(z)^2\right]\right)^{-1}$ serves as a normalizing factor, $\mathcal{N}(0,1)$ is the standard Gaussian distribution. c_{res} is a small constant satisfying $0 < c_{\text{res}} < 1$, and $f_{\text{res}}(\boldsymbol{x}, \boldsymbol{\theta})$ is the output function. Note that here we use a $\frac{c_{\text{res}}}{L}$ factor combined with the $1/\sqrt{m}$ scaling, the $\frac{c_{\text{res}}}{L}$ factor guarantees that the width per layer m does not blow up exponentially with respect to depth L, intuitively shown in Equation (3.70). Although Equation (1.2) differs by the standard ResNet architecture in [2], it will not be hard to generalize the analysis to architectures with skip-connections at every two or more layers. We also believe that the phase diagram in [20] can be drawn out similarly for ResNet.

1.4 Main Results

Owing to the non-convex nature of optimization neural networks, it is challenging to locate the global optima. A popular way to analyze such problems is to identify the geometric properties of each critical point. Some recent works have shown that for the set of functions satisfying:

- all local minima are global;
- every saddle point possesses a negative curvature (i.e. it is non-degenerate),

then gradient descent can find a global optima [27]–[29]. The objective functions of some shallow networks are in such set [14], [30]. However, even for a three-layer linear network, there exists degenerate saddle points without negative curvature [31]. So it is doubtful that global convergence of gradient descent can be ensured for deeper neural networks.

Alternatively, we directly study the dynamics of the gradient descent for specific neural network architectures. This is another approach widely taken to obtain general convergence results. Recently, it has been shown that if the network is overparameterized, gradient descent is able to find a global optima for two-layer networks [7], deep linear networks [14], [32], [33] and ResNet [34]. Jacot et al. [4] established that in the infinite width limit, the full batch gradient descent corresponds to a specific kernel regression predictor. Consequently, in the regime of infinite width, the convergence of gradient descent for neural network can be characterized by a fixed kernel [35]. This is the cornerstone upon which rests the outstanding performance of overparameterization. Inspired by the existence of such kernel, extraordinary efforts have been trying to improve it to the non-asymptotic setting, where only finite width is required.

In the regime of finite width, many works have suggested that the network can reduce training loss at exponential rate using gradient descent [8], [32]. Our thesis also belongs to this category. In this thesis, we contribute to further understanding of the gradient descent dynamics for training fully-connected networks and ResNet models. We use the same ResNet structure as in [1]. Details of the network structure are provided in Section 1.3. More importantly, we assume that the n data points are not parallel with each other. Such

assumption holds in general for standard dataset, and we focus on the empirical risk minimization problem given by the quadratic loss. We show that if $m = \Omega\left(n^3 L^2\right)$, then the empirical risk $R_S(t)$ under gradient descent decays exponentially. More precisely,

$$R_S(t) \leq R_S(0) \exp\left(-\frac{\lambda t}{n}\right),$$

where λ is the least eigenvalue of $\boldsymbol{K}^{[L+1]}$, definition of which can be found in Definition 2.4.5 Equation (2.19). It is worth noticing that:

(1). Given identical ResNet architectures, for the convergence of randomly initialized gradient descent, our results improve upon [1] (Theorem 2.5.2) in the required number of width per layer from $m = \Omega(n^4 L^2)$ to $m = \Omega(n^3 L^2)$ (Theorem 3.2.3).

(2). For fully-connected network, the required amount of overparametrization in [3] is $m = \Omega\left(n^3 2^{\mathcal{O}(L)}\right)$ (Theorem 2.6.1). We are able to reproduce the result of Du et al. [1], showing that the exponential dependence of m on the number of layers L can be eliminated for ResNet, i.e., $m = \Omega(n^3 L^2)$ (Theorem 3.2.3).

1.5 Assumptions and Notations

We introduce some assumptions and notations that will be used throughout the thesis. We set $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n\}$ for the set of input samples, and we assume that:

1. all samples are of uniform length, i.e., for any $\alpha = 1, 2, \cdots, n$, $\|\boldsymbol{x}_\alpha\|_{L_2} = 1$;
2. all samples are non-parallel with each other, i.e., $\boldsymbol{x}_{\alpha_1} \nparallel \boldsymbol{x}_{\alpha_2}$, for any $\alpha_1 \neq \alpha_2$.

We use $\sigma(\cdot)$ to denote the activation function, and we assume that:

1. $\sigma(\cdot)$ is non-polynomial, 1-Lipschitz and (infinitely) smooth;
2. its derivative of any order is also 1-Lipschitz;
3. function value at 0 satisfy $|\sigma(0)| \leq 1$.

These assumptions hold for many activation functions, including the soft-plus and sigmoid activation.

Moreover, we set n for the number of input samples, m for the width of the neural network, and L for the number of hidden layers. We denote vector L^2 norm as $\|\cdot\|_2$, vector

or function L_∞ norm as $\|\cdot\|_\infty$, matrix spectral (operator) norm as $\|\cdot\|_{2\to 2}$, matrix Frobenius norm as $\|\cdot\|_{\mathrm{F}}$, matrix infinity norm as $\|\cdot\|_{\infty\to\infty}$, and a special matrix norm, matrix 2 to infinity norm as $\|\cdot\|_{2\to\infty}$. We set a special vector $(1,1,1,\ldots,1)^\intercal \in \mathbb{R}^m$ by $\mathbf{1} := (1,1,1,\ldots,1)^\intercal$. We use $\boldsymbol{I}_m$ to signify the identity matrix in $\mathbb{R}^{m\times m}$. For a semi-positive-definite matrix $\boldsymbol{A}$, we denote its smallest eigenvalue by $\lambda_{\min}(\boldsymbol{A})$. We use $\mathcal{O}(\cdot)$ and $\Omega(\cdot)$ for the standard Big-O and Big-Omega notations. Finally, we use $\langle\cdot,\cdot\rangle$ to denote the inner product between two vectors or matrices and $\mathcal{N}(0,1)$ for the standard Gaussian distribution. These general notations are summarized in Table 1.1.

Table 1.1. Notation Table

Symbol	Representations of the symbol
$\mathcal{X}$	The set of input samples
$\sigma(\cdot)$	Nonpolynomial, 1-Lipschitz smooth function
n	Number of input samples
m	Width of a neural network
L	Number of hidden layers
$\|\cdot\|_2$	L^2 (Euclidian) norm of a vector
$\|\cdot\|_{2\to 2}$	Operator norm of a matrix
$\|\cdot\|_{\mathrm{F}}$	Frobenius norm of a matrix
$\|\cdot\|_\infty$	Infinity norm of a matrix
$\mathbf{1}$	The vector $(1,1,1,\ldots,1)^\intercal$ in $\mathbb{R}^m$
$\boldsymbol{I}_m$	The identity matrix in $\mathbb{R}^{m\times m}$
$\lambda_{\min}(\boldsymbol{A})$	Least eigenvalue of a matrix
$\mathcal{N}(0,1)$	Standard Gaussian distribution
$\langle\cdot,\cdot\rangle$	Inner product
$\mathcal{O}(\cdot)$	Big-O notation
$\Omega(\cdot)$	Big-Omega notation

Next, since we are going to perform massive computations, some useful notations shall also be introduced but not listed out in Table 1.1. We denote $\sigma\left(\boldsymbol{W}^{[l]}\boldsymbol{x}_\alpha^{[l-1]}\right)$ as $\sigma_{[l]}(\boldsymbol{x}_\alpha)$, and the diagonal matrix generated by the r-th derivatives of $\sigma_{[l]}(\boldsymbol{x}_\alpha)$, i.e., $\mathrm{diag}\left(\sigma^{(r)}(\boldsymbol{W}^{[l]}\boldsymbol{x}_\alpha^{[l-1]})\right)$ by $\boldsymbol{\sigma}_{[l]}^{(r)}(\boldsymbol{x}_\alpha)$, where $r \geq 1$. We also write the output function $f(\boldsymbol{x}_\alpha, \boldsymbol{\theta}_t)$ as $f_\alpha(t)$. Moreover, we define a series of special matrices,

$$\boldsymbol{E}_{t,\alpha}^{[l]} := \left(\boldsymbol{I}_m + \frac{c_{\mathrm{res}}}{L}\boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\alpha)\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right), \quad 2 \leq l \leq L. \tag{1.3}$$

The above matrices are termed *skip-connection matrices*. Given $\left\{\boldsymbol{E}_{t,\alpha}^{[l]}\right\}_{l=2}^{L}$, we let $\boldsymbol{E}_{t,\alpha}^{[l:L]}$ be the direct parameterization of the end-to-end mapping realized by the group of skip-connection matrices, i.e., $\boldsymbol{E}_{t,\alpha}^{[l:L]} := \boldsymbol{E}_{t,\alpha}^{[L]}\boldsymbol{E}_{t,\alpha}^{[L-1]}\cdots\boldsymbol{E}_{t,\alpha}^{[l]}$, where we set $\boldsymbol{E}_{t,\alpha}^{[\mathrm{i:j}]} := \boldsymbol{I}_m,\ \mathrm{i} > \mathrm{j}$ for the purpose of completeness.

Finally, we introduce a notion of high probability events that has been commonly used, for instance, see [3, Section 1.3]. We say that an event holds with high probability if the probability of the event is at least $1 - \exp\left(-m^{\varepsilon}\right)$ for some constant $\varepsilon > 0$. Since for a deep neural network in practice, we always have $m \lesssim \mathrm{poly}(n)$ and $n \lesssim \mathrm{poly}(m)$, then the intersection of a collection of many high probability events still has the same property as long as the number of events is at most polynomial in m and n.

2. GRADIENT DESCENT FINDS GLOBAL MINIMA

2.1 Introduction

In this chapter, we begin with an introduction to empirical risk minimization problems. We focus particularly on the empirical risk minimization problem with a quadratic loss:

$$\min_{\boldsymbol{\theta}} R_S(\boldsymbol{\theta}) = \frac{1}{2n} \sum_{\alpha=1}^{n} \|f(\boldsymbol{x}_\alpha, \boldsymbol{\theta}) - y_\alpha\|_2^2 . \tag{2.1}$$

In the above equation, $\{\boldsymbol{x}_\alpha\}_{\alpha=1}^n$ are the training inputs, $\{y_\alpha\}_{\alpha=1}^n$ are the labels. $f(\boldsymbol{x}_\alpha, \boldsymbol{\theta})$ is the prediction function, which in our case is a neural network, and $\boldsymbol{\theta}$ are the parameters to be optimized.

2.2 Gradient Descent

To learn the deep neural network, we introduce the randomly initialized gradient descent algorithm to find the global minimizer of the empirical loss Equation (2.1). The core of the algorithm consists of two steps. The first step incorporates a random initialization of parameters. As the vector containing all parameters is denoted by

$$\boldsymbol{\theta} = \left(\mathrm{vec}\left(\boldsymbol{W}^{[L]}\right), \mathrm{vec}\left(\boldsymbol{W}^{[L-1]}\right), \ldots, \mathrm{vec}\left(\boldsymbol{W}^{[1]}\right), \boldsymbol{a}\right),$$

where vec is the standard vectorization operation, we initialize the parameters following the adopted Xavier initialization scheme [25], i.e.,

$$W_{\mathrm{i,j}}^{[l]} \sim \mathcal{N}(0,1), \; a_k \sim \mathcal{N}(0,1), \qquad 1 \le l \le L, \quad 1 \le \mathrm{i}, \mathrm{j}, k \le m. \tag{2.2}$$

The second step is to train all layers of the neural network with continuous time gradient descent (gradient flow): for $l = 1, 2, \cdots, L$, and time $t \ge 0$,

$$\begin{aligned} \partial_t \boldsymbol{W}_t^{[l]} &= -\partial_{\boldsymbol{W}^{[l]}} R_S(\boldsymbol{\theta}_t), \\ \partial_t \boldsymbol{a}_t &= -\partial_{\boldsymbol{a}} R_S(\boldsymbol{\theta}_t). \end{aligned} \tag{2.3}$$

A discrete version of gradient descent can be found in other literature [7], where the parameters are updated via

$$\begin{aligned} \boldsymbol{W}^{[l]}(k) &= \boldsymbol{W}^{[l]}(k-1) - \eta \, \frac{\partial R_S(\boldsymbol{\theta}(k-1))}{\partial \boldsymbol{W}^{[l]}(k-1)}, \\ \boldsymbol{a}(k) &= \boldsymbol{a}(k-1) - \eta \, \frac{\partial R_S(\boldsymbol{\theta}(k-1))}{\partial \boldsymbol{a}(k-1)}, \end{aligned} \tag{2.4}$$

with k being the index of the step to be taken, and $\eta > 0$ being the step size.

We shall write out the dynamics (2.3) respectively for fully-connected network and ResNet. Recall that the output function $f(\boldsymbol{x}_\beta, \boldsymbol{\theta}_t)$ is denoted by $f_\beta(t)$, then for fully-connected network, dynamics (2.3) reads

$$\begin{aligned} \partial_t \boldsymbol{a}_t &= -\frac{1}{n} \sum_{\beta=1}^{n} \boldsymbol{x}_\beta^{[L]} (f_\beta(t) - y_\beta), \\ \partial_t \boldsymbol{W}_t^{[l]} &= -\frac{1}{n} \sum_{\beta=1}^{n} \left(\boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\beta) \frac{\left(\boldsymbol{W}_t^{[l+1]}\right)^\intercal}{\sqrt{m}} \cdots \boldsymbol{\sigma}_{[L]}^{(1)}(\boldsymbol{x}_\beta) \frac{\boldsymbol{a}_t}{\sqrt{m}} \right) \otimes (\boldsymbol{x}_\beta^{[l-1]})^\intercal (f_\beta(t) - y_\beta), \end{aligned} \tag{2.5}$$

and for ResNet, dynamics (2.3) reads

$$\begin{aligned} \partial_t \boldsymbol{a}_t &= -\frac{1}{n} \sum_{\beta=1}^{n} \boldsymbol{x}_\beta^{[L]} (f_\beta(t) - y_\beta), \\ \partial_t \boldsymbol{W}_t^{[L]} &= -\frac{1}{n} \sum_{\beta=1}^{n} \frac{c_{\text{res}}}{L\sqrt{m}} \operatorname{diag}\left(\boldsymbol{\sigma}_{[L]}^{(1)}(\boldsymbol{x}_\beta) \boldsymbol{a}_t \right) \mathbf{1} \otimes (\boldsymbol{x}_\beta^{[L-1]})^\intercal (f_\beta(t) - y_\beta), \\ &\text{for } l = 2, 3, \cdots, L-1, \\ \partial_t \boldsymbol{W}_t^{[l]} &= -\frac{1}{n} \sum_{\beta=1}^{n} \frac{c_{\text{res}}}{L\sqrt{m}} \operatorname{diag}\left(\boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\beta) \left(\boldsymbol{E}_{t,\beta}^{[(l+1):L]} \right)^\intercal \boldsymbol{a}_t \right) \mathbf{1} \otimes (\boldsymbol{x}_\beta^{[l-1]})^\intercal (f_\beta(t) - y_\beta), \\ \partial_t \boldsymbol{W}_t^{[1]} &= -\frac{1}{n} \sum_{\beta=1}^{n} \sqrt{\frac{c_\sigma}{m}} \operatorname{diag}\left(\boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\beta) \left(\boldsymbol{E}_{t,\beta}^{[2:L]} \right)^\intercal \boldsymbol{a}_t \right) \mathbf{1} \otimes (\boldsymbol{x}_\beta)^\intercal (f_\beta(t) - y_\beta). \end{aligned} \tag{2.6}$$

A recent line of work tries to understand the optimization process of training deep neural networks from the perspective of over-parameterization and random weight initialization. It has been observed that over-parameterization and proper random initialization can help the optimization in training neural networks, and various theoretical results have been established [32], [36]. Our results mainly build on two ideas from previous works on gradient

descent. First, we use the observation by Du et al. [1] that the required width m relies heavily on the structure of networks. Second, we applied the neural tangent hierarchy (NTH) to ResNet, a framework initially proposed by Huang and Yau [3], to directly study the change of its neural tangent kernel (NTK) [4].

With these in mind, we start with a review of NTK in Section 2.3. In Section 2.4 we introduce the concept of Gram matrices. In Section 2.5 we give out the outline of analysis and main results obtained by Du et al. [1] without proof, and we treat the counterparts in Huang and Yau [3] similarly in Section 2.6.

2.3 NTK

A flurry of recent papers in theoretical deep learning endeavor to tackle the common theme of analyzing neural networks in the infinite-width limit. At first glance, this limit may seem impractical and even pointless to study. But for mathematicians, there is a tradition of deriving insights into questions by studying them in the infinite limit, which usually tends to be easier in theory. As it turns out, neural networks in this regime simplify to linear models with a regression kernel called the NTK [4].

We shall refer the readers to the connection between infinitely wide neural networks and kernel methods [37]. Specifically, for any parametrized function $f(\boldsymbol{x}, \boldsymbol{\theta}_t)$ equipped with two inputs $\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta$ ($\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta$ could be identitcal), the corresponding kernel is

$$\mathcal{G}_{\boldsymbol{\theta}_t}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta) = \langle \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}_\alpha, \boldsymbol{\theta}_t), \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}_\beta, \boldsymbol{\theta}_t) \rangle \,. \tag{2.7}$$

The key difference between the kernel above and the one in [37] is that our kernel is defined through the inner product between the gradients of the function with respect to its parameters, while the counterpart in [37] comprises the product of the output function. Emergence of the gradient arises from the usage of gradient descent.

In the situations where $f(\boldsymbol{x}, \boldsymbol{\theta}_t)$ is the output of a fully-connected network or ResNet introduced in Chapter 1, it consists of a series of kernels $\left\{\mathcal{G}_t^{[l]}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta)\right\}_{l=1}^{L+1}$, i.e.,

$$\begin{aligned}
\mathcal{G}_{\boldsymbol{\theta}_t}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta) &= \langle \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}_\alpha, \boldsymbol{\theta}_t), \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}_\beta, \boldsymbol{\theta}_t) \rangle \\
&= \langle \partial_{\boldsymbol{a}} f(\boldsymbol{x}_\alpha, \boldsymbol{\theta}_t), \partial_{\boldsymbol{a}} f(\boldsymbol{x}_\beta, \boldsymbol{\theta}_t) \rangle + \sum_{l=1}^{L} \langle \partial_{\boldsymbol{W}^{[l]}} f(\boldsymbol{x}_\alpha, \boldsymbol{\theta}_t), \partial_{\boldsymbol{W}^{[l]}} f(\boldsymbol{x}_\beta, \boldsymbol{\theta}_t) \rangle \\
&:= \mathcal{G}_t^{[L+1]}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta) + \sum_{l=1}^{L} \mathcal{G}_t^{[l]}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta) \\
&= \sum_{l=1}^{L+1} \mathcal{G}_t^{[l]}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta).
\end{aligned} \tag{2.8}$$

For fully-connected network, $\mathcal{G}_t^{[l]}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta)$ individually reads

$$\begin{aligned}
\mathcal{G}_t^{[L+1]}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta) &= \left\langle \boldsymbol{x}_\alpha^{[L]}, \boldsymbol{x}_\beta^{[L]} \right\rangle, \\
&\quad \text{for } 1 \le l \le L \\
\mathcal{G}_t^{[l]}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta) &= \left\langle \frac{1}{\sqrt{m}} \boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\alpha) \frac{\left(\boldsymbol{W}_t^{[l+1]}\right)^\intercal}{\sqrt{m}} \cdots \boldsymbol{\sigma}_{[L]}^{(1)}(\boldsymbol{x}_\alpha) \boldsymbol{a}_t, \right. \\
&\qquad \left. \frac{1}{\sqrt{m}} \boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\beta) \frac{\left(\boldsymbol{W}_t^{[l+1]}\right)^\intercal}{\sqrt{m}} \cdots \boldsymbol{\sigma}_{[L]}^{(1)}(\boldsymbol{x}_\beta) \boldsymbol{a}_t \right\rangle \left\langle \boldsymbol{x}_\alpha^{[l-1]}, \boldsymbol{x}_\beta^{[l-1]} \right\rangle,
\end{aligned} \tag{2.9}$$

while for ResNet

$$\begin{aligned}
\mathcal{G}_t^{[L+1]}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta) &= \left\langle \boldsymbol{x}_\alpha^{[L]}, \boldsymbol{x}_\beta^{[L]} \right\rangle, \\
&\quad \text{for } 2 \le l \le L \\
\mathcal{G}_t^{[l]}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta) &= \left\langle \frac{c_{\text{res}}}{L\sqrt{m}} \boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\alpha) \left(\boldsymbol{E}_{t,\alpha}^{[(l+1):L]}\right)^\intercal \boldsymbol{a}_t, \right. \\
&\qquad \left. \frac{c_{\text{res}}}{L\sqrt{m}} \boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\beta) \left(\boldsymbol{E}_{t,\beta}^{[(l+1):L]}\right)^\intercal \boldsymbol{a}_t \right\rangle \left\langle \boldsymbol{x}_\alpha^{[l-1]}, \boldsymbol{x}_\beta^{[l-1]} \right\rangle, \\
\mathcal{G}_t^{[1]}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta) &= \left\langle \sqrt{\frac{c_\sigma}{m}} \boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\alpha) \left(\boldsymbol{E}_{t,\alpha}^{[2:L]}\right)^\intercal \boldsymbol{a}_t, \right. \\
&\qquad \left. \sqrt{\frac{c_\sigma}{m}} \boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\beta) \left(\boldsymbol{E}_{t,\beta}^{[2:L]}\right)^\intercal \boldsymbol{a}_t \right\rangle \left\langle \boldsymbol{x}_\alpha, \boldsymbol{x}_\beta \right\rangle
\end{aligned} \tag{2.10}$$

Up to this point, we have only given out the exact computations of the regression kernel, where the property of infinite width have not been used. In the large width limit, it turns

out that the time-varying kernel $\mathcal{G}_{\theta_t}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta)$ is close to a deterministic kernel $\mathcal{K}_\infty(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta)$, the limiting NTK. This property is proved in two steps. Firstly, at initial stage ($t = 0$), with appropriate scaling factor $1/\sqrt{m}$ for the parameters $\boldsymbol{\theta}_0$, there exists an infinite width limit ($m \to \infty$) of $\mathcal{G}_{\theta_0}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta)$, denoted by $\mathcal{K}_\infty(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta)$. Secondly, the kernel $\mathcal{G}_{\theta_t}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta)$ itself barely changes during the entire training process, i.e. $\mathcal{G}_{\theta_t}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta) \approx \mathcal{G}_{\theta_0}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta)$. Hence, as $m \to \infty$, $\mathcal{G}_{\theta_t}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta) \approx \mathcal{K}_\infty(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta)$ for all $t > 0$.

The results above allow us to capture the behavior of networks trained by gradient descent. In large width limit, a single output function $f(\boldsymbol{x}, \boldsymbol{\theta}_t)$ evolves as a linear differential equation

$$\partial_t \left(f(\boldsymbol{x}, \boldsymbol{\theta}_t) - y\right) = -\frac{1}{n} \sum_{\beta=1}^{n} \mathcal{K}_\infty(\boldsymbol{x}, \boldsymbol{x}_\beta)(f(\boldsymbol{x}_\beta, \boldsymbol{\theta}_t) - y_\beta), \tag{2.11}$$

where $\mathcal{K}_\infty(\cdot)$ only depends on the training inputs. More importantly, it is independent of the neural network parameters. As a direct consequence of Equation (2.11), we have

$$\begin{aligned} \partial_t \sum_{\alpha=1}^{n} \|f(\boldsymbol{x}_\alpha, \boldsymbol{\theta}_t) - y_\alpha\|_2^2 &= -\frac{2}{n} \sum_{\alpha,\beta=1}^{n} \mathcal{K}_\infty(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta)(f(\boldsymbol{x}_\alpha, \boldsymbol{\theta}_t) - y_\alpha)(f(\boldsymbol{x}_\beta, \boldsymbol{\theta}_t) - y_\beta) \\ &\leq -\frac{2}{n} \lambda_{\min}\left([\mathcal{K}_\infty(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta)]_{1\leq\alpha,\beta\leq n}\right) \sum_{\gamma=1}^{n} \|f(\boldsymbol{x}_\gamma, \boldsymbol{\theta}_t) - y_\gamma\|_2^2 . \end{aligned} \tag{2.12}$$

From Equation (2.12), we observe that the empirical loss (2.1) converges at a linear rate determined by the least eigenvalue of matrix $[\mathcal{K}_\infty(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta)]_{1\leq\alpha,\beta\leq n}$. To sum up, in the regime of infinite width, the empirical loss converges exponentially to zero regardless of the fully-connected or ResNet structure, and its convergence rate relies heavily on the limiting NTK.

2.4 Gram Matrices

In linear algebra, the Gram matrix $\boldsymbol{K}$ of a set of real vectors $\{\boldsymbol{k}_1, \boldsymbol{k}_2, \cdots, \boldsymbol{k}_n\}$ in an inner product space is the Hermitian matrix of inner products, whose entries are given by $\boldsymbol{K}_{\mathrm{ij}} := \langle \boldsymbol{k}_\mathrm{i}, \boldsymbol{k}_\mathrm{j} \rangle$. Consequently, if we denote the column matrix of $\{\boldsymbol{k}_1, \boldsymbol{k}_2, \cdots, \boldsymbol{k}_n\}$ by $\boldsymbol{D}$, then $\boldsymbol{K} = \boldsymbol{D}^\intercal \boldsymbol{D}$.

Jacot et al. [4] pointed out that convergence of the empirical loss (2.1) is related to the positive-definiteness of the limiting NTK. For any parametrized function $f(\boldsymbol{x}, \boldsymbol{\theta}_t)$, its limiting NTK reads

$$\mathcal{K}_\infty(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta) = \mathbb{E}_{\boldsymbol{\theta}_0 \sim \mathcal{W}} \left\langle \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}_\alpha, \boldsymbol{\theta}_0), \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}_\beta, \boldsymbol{\theta}_0) \right\rangle,$$

where $\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta$ are two inputs, and $\mathcal{W}$ is the initial distribution over $\boldsymbol{\theta}$.

In our cases, where $f(\boldsymbol{x}, \boldsymbol{\theta}_t)$ is the output of a fully-connected network or ResNet, and $\boldsymbol{\theta} = \left(\mathrm{vec}\left(\boldsymbol{W}^{[L]}\right), \mathrm{vec}\left(\boldsymbol{W}^{[L-1]}\right), \ldots, \mathrm{vec}\left(\boldsymbol{W}^{[1]}\right), \boldsymbol{a}\right)$, its limiting NTK is the sum of a series of kernels $\left\{\mathcal{K}^{[l]}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta)\right\}_{l=1}^{L+1}$, i.e.,

$$\mathcal{K}_\infty(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta) = \sum_{l=1}^{L+1} \mathcal{K}^{[l]}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta),$$

where

$$\begin{aligned}
\mathcal{K}^{[L+1]}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta) &= \mathbb{E}_{\boldsymbol{\theta}_0 \sim \mathcal{W}} \left\langle \partial_{\boldsymbol{a}} f(\boldsymbol{x}_\alpha, \boldsymbol{\theta}_0), \partial_{\boldsymbol{a}} f(\boldsymbol{x}_\beta, \boldsymbol{\theta}_0) \right\rangle, \\
\mathcal{K}^{[l]}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta) &= \mathbb{E}_{\boldsymbol{\theta}_0 \sim \mathcal{W}} \left\langle \partial_{\boldsymbol{W}^{[l]}} f(\boldsymbol{x}_\alpha, \boldsymbol{\theta}_0), \partial_{\boldsymbol{W}^{[l]}} f(\boldsymbol{x}_\beta, \boldsymbol{\theta}_0) \right\rangle, \qquad 1 \le l \le L.
\end{aligned}$$

We remark that at $t = 0$, respectively for all l, the limiting values of $\mathcal{G}_0^{[l]}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta)$, whose definition can be found in Equation (2.9) and Equation (2.10), equal to $\mathcal{K}^{[l]}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta)$. This is a crucial finding in [38, Corollary 2.4.]. As is shown in Equation (2.12), dynamics of the empirical loss is governed by spectral property of matrix $[\mathcal{K}_\infty(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta)]_{1 \le \alpha, \beta \le n}$. With some abuse of notations, we name matrix $\boldsymbol{K}_\infty := [\mathcal{K}_\infty(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta)]_{1 \le \alpha, \beta \le n}$ by *Gram matrix*. Also with additional abuse of terminologies, for $1 \le l \le L+1$, the matrices $\boldsymbol{K}^{[l]} := \left[\mathcal{K}^{[l]}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta)\right]_{1 \le \alpha, \beta \le n}$ are entitled *Gram matrices*, whose definitions can be traced back to [1, Definition 5.1, Definition 6.1].

As stated in [38, Corollary 2.4], to assure the convergence of NTK, $\boldsymbol{\theta}$ is required to be suitably randomized. As mentioned in Section 2.2, we initialize $\boldsymbol{\theta}$ using the adopted Xavier initialization scheme (2.2). Furthermore, it shall be noted that derivation of the Gram matrices depends on the series of matrices $\left\{\widetilde{\boldsymbol{K}}^{[l]}\right\}_{l=1}^{L}$, $\left\{\widetilde{\boldsymbol{A}}^{[l]}\right\}_{l=1}^{L+1}$ and vectors $\left\{\widetilde{\boldsymbol{b}}^{[l]}\right\}_{l=1}^{L}$ to be defined immediately in the next two sections.

2.4.1 Gram Matrices for Fully-connected Networks

In this section, as a warm up, we give out definitions of Gram matrices generated by fully-connected networks. We are able to write out the explicit formulas for $\boldsymbol{K}^{[L+1]}$ and $\boldsymbol{K}^{[L]}$. However, it is not the case for $\left\{\boldsymbol{K}^{[l]}\right\}_{l=1}^{L-1}$. A slightly different approach shall be taken to write out the expressions for them.

Definition 2.4.1. *Given input samples $\mathcal{X}$ and activation function $\sigma(\cdot)$, the matrices $\left\{\widetilde{\boldsymbol{K}}^{[l]}\right\}_{l=1}^{L}$, $\left\{\widetilde{\boldsymbol{A}}^{[l]}\right\}_{l=1}^{L+1}$ are defined recursively: for $1 \leq \mathrm{i}, \mathrm{j} \leq n, 1 \leq l \leq L$*

$$\begin{aligned}
\widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[0]} &= \langle \boldsymbol{x}_{\mathrm{i}}, \boldsymbol{x}_{\mathrm{j}} \rangle, \\
\widetilde{\boldsymbol{A}}_{\mathrm{ij}}^{[l]} &= \begin{pmatrix} \widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l-1]} & \widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[l-1]} \\ \widetilde{\boldsymbol{K}}_{\mathrm{ji}}^{[l-1]} & \widetilde{\boldsymbol{K}}_{\mathrm{jj}}^{[l-1]} \end{pmatrix}, \\
\widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[l]} &= \mathbb{E}_{(u,v)^{\intercal} \sim \mathcal{N}\left(\boldsymbol{0}, \widetilde{\boldsymbol{A}}_{\mathrm{ij}}^{[l]}\right)} \sigma(u)\sigma(v), \\
\widetilde{\boldsymbol{A}}_{\mathrm{ij}}^{[L+1]} &= \begin{pmatrix} \widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[L]} & \widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[L]} \\ \widetilde{\boldsymbol{K}}_{\mathrm{ji}}^{[L]} & \widetilde{\boldsymbol{K}}_{\mathrm{jj}}^{[L]} \end{pmatrix}.
\end{aligned} \tag{2.13}$$

Definition 2.4.2 (Gram Matrices $\boldsymbol{K}^{[L+1]}$ and $\boldsymbol{K}^{[L]}$ for Fully-connected Networks)**.** *Given $\left\{\widetilde{\boldsymbol{K}}^{[l]}\right\}_{l=1}^{L}$, $\left\{\widetilde{\boldsymbol{A}}^{[l]}\right\}_{l=1}^{L+1}$, Gram matrices $\boldsymbol{K}^{[L+1]}, \boldsymbol{K}^{[L]} \in \mathbb{R}^{n \times n}$ are defined as follows: for $1 \leq \mathrm{i}, \mathrm{j} \leq n$*

$$\boldsymbol{K}_{\mathrm{ij}}^{[L+1]} = \mathbb{E}_{(u,v)^{\intercal} \sim \mathcal{N}\left(\boldsymbol{0}, \widetilde{\boldsymbol{A}}_{\mathrm{ij}}^{[L+1]}\right)} \sigma(u)\sigma(v), \tag{2.14}$$

$$\boldsymbol{K}_{\mathrm{ij}}^{[L]} = \widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[L-1]}\, \mathbb{E}_{(u,v)^{\intercal} \sim \mathcal{N}\left(\boldsymbol{0}, \widetilde{\boldsymbol{A}}_{\mathrm{ij}}^{[L]}\right)} \left[\sigma^{(1)}(u)\sigma^{(1)}(v)\right]. \tag{2.15}$$

Definition 2.4.3. *Given $\left\{\widetilde{\boldsymbol{K}}^{[l]}\right\}_{l=1}^{L}$, $\left\{\widetilde{\boldsymbol{A}}^{[l]}\right\}_{l=1}^{L+1}$, Gram matrices $\boldsymbol{K}^{[l]} \in \mathbb{R}^{n \times n}$ are defined as follows: for $1 \leq \mathrm{i}, \mathrm{j} \leq n, 1 \leq l \leq L-1$*

$$\begin{aligned}
\boldsymbol{K}_{\mathrm{ij}}^{[l]} = \lim_{m \to \infty} \frac{1}{m} \Big\langle & \boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_{\mathrm{i}}) \frac{\left(\boldsymbol{W}_0^{[l+1]}\right)^{\intercal}}{\sqrt{m}} \cdots \boldsymbol{\sigma}_{[L]}^{(1)}(\boldsymbol{x}_{\mathrm{i}}) \boldsymbol{a}_0, \\
& \boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_{\mathrm{j}}) \frac{\left(\boldsymbol{W}_0^{[l+1]}\right)^{\intercal}}{\sqrt{m}} \cdots \boldsymbol{\sigma}_{[L]}^{(1)}(\boldsymbol{x}_{\mathrm{j}}) \boldsymbol{a}_0 \Big\rangle \left\langle \boldsymbol{x}_{\mathrm{i}}^{[l-1]}, \boldsymbol{x}_{\mathrm{j}}^{[l-1]} \right\rangle.
\end{aligned} \tag{2.16}$$

2.4.2 Gram Matrices for ResNet

We give out the derivation of Gram matrices generated by ResNet with slight abuse of notations. Similarly, only $\boldsymbol{K}^{[L+1]}$ and $\boldsymbol{K}^{[L]}$ can be written into closed forms.

Definition 2.4.4. *Given input samples $\mathcal{X}$ and activation function $\sigma(\cdot)$, the matrices $\left\{\widetilde{\boldsymbol{K}}^{[l]}\right\}_{l=1}^{L}$, $\left\{\widetilde{\boldsymbol{A}}^{[l]}\right\}_{l=1}^{L+1}$ and vectors $\left\{\widetilde{\boldsymbol{b}}^{[l]}\right\}_{l=1}^{L}$ are defined recursively: for $1 \leq \mathrm{i}, \mathrm{j} \leq n, 2 \leq l \leq L$*

$$
\begin{aligned}
\widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[0]} &= \langle \boldsymbol{x}_{\mathrm{i}}, \boldsymbol{x}_{\mathrm{j}} \rangle, \\
\widetilde{\boldsymbol{A}}_{\mathrm{ij}}^{[1]} &= \begin{pmatrix} \widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[0]} & \widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[0]} \\ \widetilde{\boldsymbol{K}}_{\mathrm{ji}}^{[0]} & \widetilde{\boldsymbol{K}}_{\mathrm{jj}}^{[0]} \end{pmatrix}, \\
\widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[1]} &= \mathbb{E}_{(u,v)^{\intercal} \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{A}_{\mathrm{ij}}^{[1]}\right)} c_{\sigma} \sigma(u) \sigma(v), \\
\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[1]} &= \sqrt{c_{\sigma}}\, \mathbb{E}_{u \sim \mathcal{N}\left(0, \widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[0]}\right)} \left[\sigma(u)\right], \\
\widetilde{\boldsymbol{A}}_{\mathrm{ij}}^{[l]} &= \begin{pmatrix} \widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l-1]} & \widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[l-1]} \\ \widetilde{\boldsymbol{K}}_{\mathrm{ji}}^{[l-1]} & \widetilde{\boldsymbol{K}}_{\mathrm{jj}}^{[l-1]} \end{pmatrix}, \\
\widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[l]} &= \widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[l-1]} + \mathbb{E}_{(u,v)^{\intercal} \sim \mathcal{N}\left(\boldsymbol{0}, \widetilde{\boldsymbol{A}}_{\mathrm{ij}}^{[l]}\right)} \left[\frac{c_{\mathrm{res}} \widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l-1]} \sigma(v)}{L} + \frac{c_{\mathrm{res}} \widetilde{\boldsymbol{b}}_{\mathrm{j}}^{[l-1]} \sigma(u)}{L} + \frac{c_{\mathrm{res}}^2 \sigma(u)\sigma(v)}{L^2}\right], \\
\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l]} &= \widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l-1]} + \frac{c_{\mathrm{res}}}{L}\, \mathbb{E}_{u \sim \mathcal{N}\left(0, \widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l-1]}\right)} \left[\sigma(u)\right], \\
\widetilde{\boldsymbol{A}}_{\mathrm{ij}}^{[L+1]} &= \begin{pmatrix} \widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[L]} & \widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[L]} \\ \widetilde{\boldsymbol{K}}_{\mathrm{ji}}^{[L]} & \widetilde{\boldsymbol{K}}_{\mathrm{jj}}^{[L]} \end{pmatrix}.
\end{aligned}
\tag{2.17}
$$

Definition 2.4.5 (Gram Matrices $\boldsymbol{K}^{[L+1]}$ and $\boldsymbol{K}^{[L]}$ for ResNet)**.** *Given $\left\{\widetilde{\boldsymbol{K}}^{[l]}\right\}_{l=1}^{L}$, $\left\{\widetilde{\boldsymbol{A}}^{[l]}\right\}_{l=1}^{L+1}$, $\left\{\widetilde{\boldsymbol{b}}^{[l]}\right\}_{l=1}^{L}$, Gram matrices $\boldsymbol{K}^{[L+1]}, \boldsymbol{K}^{[L]} \in \mathbb{R}^{n \times n}$ are defined as follows: for $1 \leq \mathrm{i}, \mathrm{j} \leq n$*

$$
\boldsymbol{K}_{\mathrm{ij}}^{[L+1]} = \widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[L]} + \mathbb{E}_{(u,v)^{\intercal} \sim \mathcal{N}\left(\boldsymbol{0}, \widetilde{\boldsymbol{A}}_{\mathrm{ij}}^{[L+1]}\right)} \left[\frac{c_{\mathrm{res}} \widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[L]} \sigma(v)}{L} + \frac{c_{\mathrm{res}} \widetilde{\boldsymbol{b}}_{\mathrm{j}}^{[L]} \sigma(u)}{L} + \frac{c_{\mathrm{res}}^2 \sigma(u)\sigma(v)}{L^2}\right], \tag{2.18}
$$

$$
\boldsymbol{K}_{\mathrm{ij}}^{[L]} = \frac{c_{\mathrm{res}}^2}{L^2} \widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[L-1]}\, \mathbb{E}_{(u,v)^{\intercal} \sim \mathcal{N}\left(\boldsymbol{0}, \widetilde{\boldsymbol{A}}_{\mathrm{ij}}^{[L]}\right)} \left[\sigma^{(1)}(u) \sigma^{(1)}(v)\right]. \tag{2.19}
$$

Definition 2.4.6. *Given* $\left\{\widetilde{\boldsymbol{K}}^{[l]}\right\}_{l=1}^{L}$, $\left\{\widetilde{\boldsymbol{A}}^{[l]}\right\}_{l=1}^{L+1}$, $\left\{\widetilde{\boldsymbol{b}}^{[l]}\right\}_{l=1}^{L}$, *Gram matrices* $\boldsymbol{K}^{[l]} \in \mathbb{R}^{n\times n}$ *are defined as follows, for* $1 \leq \mathrm{i}, \mathrm{j} \leq n, 2 \leq l \leq L-1$,

$$\boldsymbol{K}_{\mathrm{ij}}^{[l]} = \frac{c_{\mathrm{res}}^2}{L^2}\widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[l-1]} \lim_{m\to\infty} \frac{1}{m} \left\langle \boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_{\mathrm{i}}) \left(\boldsymbol{E}_{0,\mathrm{i}}^{[(l+1):L]}\right)^{\intercal} \boldsymbol{a}_0, \boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_{\mathrm{j}}) \left(\boldsymbol{E}_{0,\mathrm{j}}^{[(l+1):L]}\right)^{\intercal} \boldsymbol{a}_0 \right\rangle, \tag{2.20}$$

and for $1 \leq \mathrm{i}, \mathrm{j} \leq n, l = 1$,

$$\boldsymbol{K}_{\mathrm{ij}}^{[1]} = c_\sigma \widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[0]} \lim_{m\to\infty} \frac{1}{m} \left\langle \boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_{\mathrm{i}}) \left(\boldsymbol{E}_{0,\mathrm{i}}^{[2:L]}\right)^{\intercal} \boldsymbol{a}_0, \boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_{\mathrm{j}}) \left(\boldsymbol{E}_{0,\mathrm{j}}^{[2:L]}\right)^{\intercal} \boldsymbol{a}_0 \right\rangle. \tag{2.21}$$

We remark that the $\frac{1}{m}$ scalings in (2.16), (2.20) and (2.21) originate from the inner product between the gradients, i.e., $\frac{1}{m} = \frac{1}{\sqrt{m}} \times \frac{1}{\sqrt{m}}$. Thanks to the Strong Law of Large Numbers, the above limits (2.16), (2.20) and (2.21) exist [38, Corollary 2.4]. As we send $m \to \infty$, Gram matrices $\left\{\boldsymbol{K}^{[l]}\right\}_{l=1}^{L-1}$ only depend on the input samples and the activation patterns. Hence we conclude that all Gram matrices $\left\{\boldsymbol{K}^{[l]}\right\}_{l=1}^{L+1}$ only depend on the input samples and independent of $\boldsymbol{\theta}$.

2.5 Main Results of Du et al.

Jacot et al. [4, Proposition 2] proved the positive-definiteness of the limiting NTK when the data is supported on the sphere and the non-linearity is non-polynomial. Du et al. [1, Proposition F.1, Proposition F.2] extended their results and showed that as long as the input training data is not degenerate and supported on the unit sphere, $\lambda_{\min}\left(\boldsymbol{K}^{[L]}\right)$ is strictly positive. Since gram matrices of all orders are positive semi-definite, $\lambda_{\min}\left(\boldsymbol{K}^{[L]}\right)$ is an explicit lower bound of the least eigenvalue of the limiting NTK matrix $\boldsymbol{K}_\infty$, i.e., $\lambda_{\min}\left(\boldsymbol{K}_\infty\right) \geq \lambda_{\min}\left(\boldsymbol{K}^{[L]}\right)$. Moreover, using the contribution of all the gram matrices to the minimum eigenvalue can potentially improve the convergence rate, as is shown in [3, Corollary 2.5].

The high-level analysis framework of Du et al. consists of mainly two components. At first stage, they showed that at $t = 0$, $\left[\mathcal{G}_0^{[L]}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta)\right]_{1\leq\alpha,\beta\leq n}$ is close to $\boldsymbol{K}^{[L]}$ via repeated application of concentration inequality. Instead of sending the width m of every layer to ∞, as is the setting for Jacot et al., one only needs it to be greater than a finite threshold of

order $\Omega\left(n^2\right)$ in order for $\lambda_{\min}\left(\left[\mathcal{G}_0^{[L]}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta)\right]_{1\leq\alpha,\beta\leq n}\right)$ to maintain a lower bound with high probability. We observe that $\boldsymbol{K}^{[L]}$ is recursively defined, and so is $\left[\mathcal{G}_0^{[L]}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta)\right]_{1\leq\alpha,\beta\leq n}$. Due to the randomness inherited from the initialization scheme and the introduction of finite threshold of m, inevitably we have some perturbation in the first layer, and how this perturbation propagates to the L-th layer shall be analyzed carefully. Du et al. derived a general formulation that allows readers to analyze the initialization behavior for the fully-connected network, ResNet, and even convolutional ResNet in a unified way.

One important finding in the perturbation analysis is that ResNet architecture makes the propagation more stable. For fully-connected network, such perturbation propagates to the L-th layer exponentially, hence forcing the threshold of m to maintain exponential dependency on the depth L. Heuristically speaking, let $\mathcal{E}_1$ be the perturbation in the first layer, then $\mathcal{E}_L$, perturbation in the L-th layer, admits the form

$$\mathcal{E}_L \leq 2^{\mathcal{O}(L)}\mathcal{E}_1. \tag{2.22}$$

However, for ResNet, thanks to the skip connection structure, its counterpart reads

$$\mathcal{E}_L \leq \left(1+\mathcal{O}\left(\frac{1}{L}\right)\right)^L \mathcal{E}_1. \tag{2.23}$$

Therefore, the issue of exponential explosion can be avoided, and the above analysis sheds light on the benefit of using ResNet architecture for training.

At the second stage, for $t>0$, the averaged Frobenius norm $\frac{1}{\sqrt{m}}\left\|\boldsymbol{W}_t^{[l]}-\boldsymbol{W}_0^{[l]}\right\|_{\mathrm{F}}$ is used to control the absolute change of eigenvalues, so that the lower bound of the eigenvalue of matrix $\left[\mathcal{G}_t^{[L]}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta)\right]_{1\leq\alpha,\beta\leq n}$ can be guaranteed during the whole training process. In purpose of bounding the averaged Frobenius norm, another threshold of order $\Omega\left(n^4\right)$ is required for width m. Such analysis, whose high-level intuition is similar to (3.70), once again sheds light on the benefit of using ResNet architecture for training.

We may proceed to state the main theorems of Du et al. [1, Theorem 5.1, Theorem 6.1].

Theorem 2.5.1 (Convergence of Gradient Descent for Fully-connected Networks)**.** *Given $\mathcal{X}$, $\sigma(\cdot)$ and $\boldsymbol{K}^{[L]}$ defined in Equation (2.15), with high probability w.r.t random initialization, for width*

$$m = \Omega\left(2^{\mathcal{O}(L)} \max\left\{\frac{n^4}{\lambda_{\min}^4\left(\boldsymbol{K}^{[L]}\right)}, n, \frac{n^2 \log\left(Ln\right)}{\lambda_{\min}^2\left(\boldsymbol{K}^{[L]}\right)}\right\}\right), \tag{2.24}$$

and if we set the step size

$$\eta = \mathcal{O}\left(\frac{\lambda_{\min}\left(\boldsymbol{K}^{[L]}\right)}{n^2 2^{\mathcal{O}(L)}}\right),$$

then for $k = 0, 1, 2, \cdots$, the loss at each iteration satisfies

$$R_S(\boldsymbol{\theta}(k)) \leq \left(1 - \eta\frac{\lambda_{\min}\left(\boldsymbol{K}^{[L]}\right)}{2}\right)^k R_S(\boldsymbol{\theta}(0)).$$

Main assumption of Equation (2.24) is that a large enough width for each layer is required. We notice that the requirement of m has three terms. The first term is used to show the Gram matrix remain stable during training. The second term is used to guarantee the output in each layer is approximately normalized at initial phase. The third term is used to show the perturbation of Gram matrix at initial phase is small. However, its dependency on depth L is exponential. Such exponentiality comes from the instability of the fully-connected architecture. In the next theorem, equipped with ResNet architecture, dependency on L can be reduced from 2^L to poly(L).

Theorem 2.5.2 (Convergence of Gradient Descent for ResNet)**.** *Given $\mathcal{X}$, $\sigma(\cdot)$ and $\boldsymbol{K}^{[L]}$ defined in Equation (2.19), with high probability w.r.t random initialization, for width*

$$m = \Omega\left(\max\left\{\frac{n^4}{\lambda_{\min}^4\left(\boldsymbol{K}^{[L]}\right)}, \frac{n^2}{\lambda_{\min}^2\left(\boldsymbol{K}^{[L]}\right)L^2}, n, \frac{n^2 \log\left(Ln\right)}{\lambda_{\min}^2\left(\boldsymbol{K}^{[L]}\right)}\right\}\right), \tag{2.25}$$

and if we set the step size

$$\eta = \mathcal{O}\left(\frac{\lambda_{\min}\left(\boldsymbol{K}^{[L]}\right)L^2}{n^2}\right),$$

then for $k = 0, 1, 2, \cdots$, the loss at each iteration satisfies

$$R_S(\boldsymbol{\theta}(k)) \leq \left(1 - \eta \frac{\lambda_{\min}\left(\boldsymbol{K}^{[L]}\right)}{2}\right)^k R_S(\boldsymbol{\theta}(0)).$$

In contrast to Theorem 2.5.1, the required width m is fully polynomial in n and L. The main reason why the problematic exponential explosion can be circumvented is that the skip connection blocks enable the network structure to be more stable at both the initialization and the training phase. The requirement on m has four terms. The first two terms are used to show the stability of Gram matrix during training. The third term is used to assure that the output in each layer is approximately normalized at the initialization phase. The fourth term is used to bound the size of the perturbation of the Gram matrix at initial stage.

2.6 Main Results of Huang and Yau

Huang and Yau [3] proposed a framework in which an infinite hierarchy of ordinary differential equations, the neural tangent hierarchy (NTH), is derived.

Theorem 2.6.1 (NTH for Fully-connected Networks)**.** *Given $\mathcal{X}$ and $\sigma(\cdot)$, with high probability w.r.t random initialization, there exists an infinite family of operators $\mathcal{G}_t^{(r)} : \mathcal{X}^r \to \mathbb{R}$, where $r \geq 2$, that describes the continuous time gradient descent*:

$$\partial_t(f_\alpha(t) - y_\alpha) = -\frac{1}{n}\sum_{\beta=1}^{n} \mathcal{G}_t^{(2)}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta)(f_\beta(t) - y_\beta), \tag{2.26}$$

and for any $r \geq 2$,

$$\partial_t \mathcal{G}_t^{(r)}(\boldsymbol{x}_{\alpha_1}, \boldsymbol{x}_{\alpha_2}, \cdots, \boldsymbol{x}_{\alpha_r}) = -\frac{1}{n}\sum_{\beta=1}^{n} \mathcal{G}_t^{(r+1)}(\boldsymbol{x}_{\alpha_1}, \boldsymbol{x}_{\alpha_2}, \cdots, \boldsymbol{x}_{\alpha_r}, \boldsymbol{x}_\beta)(f_\beta(t) - y_\beta). \tag{2.27}$$

There also exists a deterministic family of operators (independent of m):

$$\mathfrak{G}^{(r)} : \mathcal{X}^r \to \mathbb{R}, \qquad 2 \leq r \leq p + 1,$$

where $\mathfrak{G}^{(r)} \equiv 0$ *if* r *is odd, and some constants* $C, C^* > 0$*, such that with high probability w.r.t random initialization,*

$$\left\| \left(\mathcal{G}_0^{(r)} - \frac{\mathfrak{G}^{(r)}}{m^{r/2-1}} \right) (\cdot) \right\|_\infty \lesssim \frac{(\ln m)^C}{m^{(r-1)/2}}, \tag{2.28}$$

and for time $0 \le t \le m^{\frac{p}{2(p+1)}} / (\ln m)^{C^*}$*,*

$$\left\| \mathcal{G}_t^{(r)} (\cdot) \right\|_\infty \lesssim \frac{(\ln m)^C}{m^{r/2-1}}. \tag{2.29}$$

For fully-connected network defined in Equation (1.1), with width $m = \Omega(2^{\mathcal{O}(L)} n^3)$, gradient descent converges at a linear rate.

Theorem 2.6.2. *Given* $\mathcal{X}$ *and* $\sigma(\cdot)$*, we further assume that there exists* $\lambda > 0$ *(might depend on* m*)*

$$\lambda_{\min} \left(\left[\mathcal{G}_0^{(2)}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta) \right]_{1 \le \alpha, \beta \le n} \right) \ge \lambda, \tag{2.30}$$

and the width m *of the neural network satisfies*

$$m \ge C^* \left(\frac{n}{\lambda} \right)^3 (\ln m)^C \ln \left(\frac{n}{\varepsilon} \right)^2, \tag{2.31}$$

for some constants $C, C^* > 0$*. Then with high probability w.r.t. random initialization, the training error decays exponentially,*

$$\sum_{\alpha=1}^n \| f_\alpha(t) - y_\alpha \|_2^2 \le \exp\left(-\frac{\lambda t}{n} \right) \sum_{\alpha=1}^n \| f_\alpha(0) - y_\alpha \|_2^2, \tag{2.32}$$

which reaches the training accuracy ε *with time complexity*

$$T = \mathcal{O}\left(\frac{n}{\lambda} \ln \left(\frac{1}{\varepsilon} \right) \right). \tag{2.33}$$

We remark that the operator $\mathcal{G}_t^{(2)}(\cdot)$ by definition is the same as the NTK $\mathcal{G}_{\boldsymbol{\theta}_t}(\cdot)$ in Equation (2.7). Also in Theorem 2.6.2, a further assumption on the least eigenvalue of the NTK $\mathcal{G}_0^{(2)}(\cdot)$ has been imposed directly, see Equation (2.30). However, as is shown in The-

orem 2.5.1, there shall be some additional requirements on width m in order for the NTK at initial phase to be positive-definite. We complement this technical issue in Chapter B, showing that the least eigenvalue of the NTK $\mathcal{G}_0^{(2)}(\cdot)$ can be ensured for a finite threshold of the width m.

3. RESNET USING NTH

3.1 Introduction

In this chapter, we study the ResNet model (1.2) with finite width using NTH. We rigorously prove that as long as no two training inputs from $\mathcal{X}$ are parallel and the width m is large enough, gradient descent achieves zero training loss at a linear convergence rate, i.e., it finds a solution $\boldsymbol{\theta}(t)$ with $R_S(\boldsymbol{\theta}(t)) \leq \varepsilon$ in time $t = \mathcal{O}\left(\ln\left(\frac{1}{\varepsilon}\right)\right)$. Thus, our results give a quantitative convergence rate involving the desired accuracy.

Our exposition mainly follows the settings of [3]. However, different from [3] in analyzing the fully-connected network, we focus on the investigation of ResNet. We exploit further benefits of using ResNet architecture for training and advantages of choosing NTH over kernel regression.

3.2 Main Results

Theorem 3.2.1 (NTH for ResNet)**.** *Given $\mathcal{X}$ and $\sigma(\cdot)$, there exists an infinite family of operators $\mathcal{G}_t^{(r)} : \mathcal{X}^r \to \mathbb{R}$, $r \geq 2$ that describes the continuous time gradient descent*

$$\partial_t(f_\alpha(t) - y_\alpha) = -\frac{1}{n}\sum_{\beta=1}^{n} \mathcal{G}_t^{(2)}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta)(f_\beta(t) - y_\beta), \tag{3.1}$$

and for any $r \geq 2$,

$$\partial_t \mathcal{G}_t^{(r)}(\boldsymbol{x}_{\alpha_1}, \boldsymbol{x}_{\alpha_2}, \cdots, \boldsymbol{x}_{\alpha_r}) = -\frac{1}{n}\sum_{\beta=1}^{n} \mathcal{G}_t^{(r+1)}(\boldsymbol{x}_{\alpha_1}, \boldsymbol{x}_{\alpha_2}, \cdots, \boldsymbol{x}_{\alpha_r}, \boldsymbol{x}_\beta)(f_\beta(t) - y_\beta), \tag{3.2}$$

then with high probability w.r.t random initialization, there exist some constants $C, C^* > 0$, such that for $r \geq 2$, time $0 \leq t \leq \sqrt{m}/(\ln m)^{C^*}$,

$$\begin{aligned} \left\|\mathcal{G}_t^{(2)}(\cdot)\right\|_\infty &\lesssim 1, \\ \left\|\mathcal{G}_t^{(r)}(\cdot)\right\|_\infty &\lesssim \frac{(\ln m)^C}{m^{r/2-1}}. \end{aligned} \tag{3.3}$$

Following Theorem 3.2.1, we shall remark that:

- The operator $\mathcal{G}_t^{(2)}(\cdot)$ is the same as the NTK $\mathcal{G}_{\boldsymbol{\theta}_t}(\cdot)$ derived in Equation (2.10);
- Constant C depends linearly on r;
- Pre-factors in Inequality (3.3) explode exponentially fast in r.

Even though the pre-factors explode exponentially, this does not exert influence on the convergence of gradient descent. Since the landscape of empirical loss $R_S(\boldsymbol{\theta}_t)$ is mainly affected by lower order kernels. As is shown in the proof of Theorem 3.2.2, we only need to analyze kernels up to order $r = 4$.

It has been proved in Theorem 3.2.1 and other literatures [7], [35], [38] that the change of NTK during the gradient descent dynamics for Deep Neural Network is bounded by $\mathcal{O}\left(\frac{1}{\sqrt{m}}\right)$. However, it was observed by Lee et al. [39] that time variation of the NTK is closer to $\mathcal{O}\left(\frac{1}{m}\right)$, indicating that there exists a performance gap between the kernel regression using the limiting NTK and neural networks. Such an observation has been confirmed by Huang and Yau in [3, Corollary 2.4], and we present a different approach to obtain similar results.

Recall that the NTK $\mathcal{G}_t^{(2)}(\cdot)$ consists of a series of kernels $\left\{\mathcal{G}_t^{[l]}(\cdot)\right\}_{l=1}^{L+1}$,

$$\mathcal{G}_t^{(2)}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta) = \sum_{l=1}^{L+1} \mathcal{G}_t^{[l]}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta),$$

and for ResNet

$$\begin{aligned}
\mathcal{G}_t^{[L+1]}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta) &= \left\langle \boldsymbol{x}_\alpha^{[L]}, \boldsymbol{x}_\beta^{[L]} \right\rangle, \\
&\quad \text{for } 2 \leq l \leq L \\
\mathcal{G}_t^{[l]}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta) &= \Bigg\langle \frac{c_{\text{res}}}{L\sqrt{m}} \boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\alpha) \left(\boldsymbol{E}_{t,\alpha}^{[(l+1):L]}\right)^\intercal \boldsymbol{a}_t, \\
&\qquad \frac{c_{\text{res}}}{L\sqrt{m}} \boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\beta) \left(\boldsymbol{E}_{t,\beta}^{[(l+1):L]}\right)^\intercal \boldsymbol{a}_t \Bigg\rangle \left\langle \boldsymbol{x}_\alpha^{[l-1]}, \boldsymbol{x}_\beta^{[l-1]} \right\rangle, \\
\mathcal{G}_t^{[1]}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta) &= \Bigg\langle \sqrt{\frac{c_\sigma}{m}} \boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\alpha) \left(\boldsymbol{E}_{t,\alpha}^{[2:L]}\right)^\intercal \boldsymbol{a}_t, \\
&\qquad \sqrt{\frac{c_\sigma}{m}} \boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\beta) \left(\boldsymbol{E}_{t,\beta}^{[2:L]}\right)^\intercal \boldsymbol{a}_t \Bigg\rangle \left\langle \boldsymbol{x}_\alpha, \boldsymbol{x}_\beta \right\rangle.
\end{aligned} \tag{3.4}$$

Theorem 3.2.2. *Given $\mathcal{X}$ and $\sigma(\cdot)$, with high probability w.r.t random initialization, there exist some constants $C, C^* > 0$, such that for time $0 \leq t \leq \sqrt{m}/(\ln m)^{C^*}$,*

$$\left\|\partial_t \mathcal{G}_t^{[L+1]}(\cdot)\right\|_\infty \lesssim \frac{(1+t)(\ln m)^C}{m}, \tag{3.5}$$

where the constant C is independent of the depth L. Moreover, the pre-factor in the inequality (3.5) is at most of order $\mathcal{O}(L^2)$.

As a direct consequence of Theorem 3.2.2, for ResNet defined in Equation (1.2), with width $m = \Omega(n^3 L^2)$, gradient descent converges at a linear rate. Precise statements are given out below.

Theorem 3.2.3. *Given $\mathcal{X}$, $\sigma(\cdot)$ and $\boldsymbol{K}^{[L+1]}$ defined in Equation (2.18), set $\lambda_0 > 0$ yielding $\lambda_0 \leq \lambda_{\min}\left(\boldsymbol{K}^{[L+1]}\right)$, there exists a small constant $\gamma_1 > 0$, such that with high probability w.r.t random initialization, for $m = \Omega\left(\left(\frac{n}{\lambda_0}\right)^{2+\gamma_1}\right)$,*

$$\lambda_{\min}\left(\left[\mathcal{G}_0^{(2)}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta)\right]_{1\leq\alpha,\beta\leq n}\right) \geq \frac{3}{4}\lambda_0. \tag{3.6}$$

Furthermore, there exists a small constant $\gamma_2 > 0$, such that for $m = \Omega\left(\left(\frac{n}{\lambda_0}\right)^{3+\gamma_2} L^2 \ln\left(\frac{1}{\varepsilon}\right)^2\right)$, the quadratic training loss $R_S(\boldsymbol{\theta}_t)$ decays exponentially

$$R_S(\boldsymbol{\theta}_t) \leq R_S(\boldsymbol{\theta}_0)\exp\left(-\frac{\lambda_0 t}{n}\right), \tag{3.7}$$

where $\varepsilon > 0$ is the desired accuracy of $R_S(\boldsymbol{\theta}_t)$.

First of all, we note that positive-definiteness of $\boldsymbol{K}^{[L+1]}$ is guaranteed from results in Chapter A. For convenience, we summarize Theorem 3.2.3 as follows. If width m satisfies that $m = \max\{\Omega(n^2), \Omega(n^3 L^2)\}$, then the continuous time gradient descent converges exponentially, and it reaches training accuracy ε with time complexity $T = \mathcal{O}\left(n \ln\left(\frac{1}{\varepsilon}\right)\right)$.

Before we end this section, we present a fair comparison of our results with others. First of all, Du et al. [1, Theorem 6.1] require $m = \Omega\left(\frac{n^4}{\lambda_{\min}\left(\boldsymbol{K}^{[L]}\right)^4 L^6}\right)$. Since there is a scaling factor $\frac{1}{L^2}$ in $\lambda_{\min}\left(\boldsymbol{K}^{[L]}\right)$, this leads to $m = \Omega(n^4 L^2)$. Their iteration complexity for discrete

time gradient descent converges with $T = \Omega\left(n^2 L^2 \ln\left(\frac{1}{\varepsilon}\right)\right)$. Our Theorem 3.2.3 improves their result in two ways:

(i) The quartic dependence on n is reduced directly to cubic dependence;

(ii) Faster convergence of the training process of gradient descent.

Secondly, our work serves as an extension of the results established in [3] from fully-connected network to ResNet. We show that for ResNet, in one hand, it is possible for us to study directly the time variation of NTK using NTH. In the other hand, compared with fully-connected network, ResNet is more stable in many aspects. Our Theorem 3.2.3 improves the results in [3] in three ways: (i) With ResNet architecture, the dependency of the amount of over-parameterization on the depth L can be reduced from $2^{\mathcal{O}(L)}$ to L^2;

(ii) While the time interval for the result in [3] takes the form $0 \leq t \leq m^{\frac{p}{2(p+1)}}/(\ln m)^{C^*}$ for some $p \geq 2$, we extend the interval to $0 \leq t \leq \sqrt{m}/(\ln m)^{C^*}$. Additionally, we are able to show even further that the results hold true for $t = \infty$;

(iii) In the proof of Corollary 2.5. of [3], further assumptions on least eigenvalue of the NTK at initial stage were imposed directly. We have rigorously shown in Chapter A and Chapter B that least eigenvalue of the NTK $\mathcal{G}_0^{(2)}(\cdot)$ stays strictly positive as long as the width m satisfies $m = \Omega(n^2)$. Moreover, for fully-connected networks, Huang and Yau asserted that adding up the whole $L+1$ kernels would give rise to the convergence rate of $R_S(\boldsymbol{\theta})$, for the belief that the sum of the least eigenvalues of all the kernels $\mathcal{G}_t^{[l]}$ is much larger than the counterpart of a single kernel, i.e.,

$$\begin{aligned}\lambda_{\min}\left(\sum_{l^*=1}^{L+1}\left[\mathcal{G}_t^{[l^*]}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta)\right]_{1\leq\alpha,\beta\leq n}\right) &= \lambda_{\min}\left(\left[\mathcal{G}_t^{(2)}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta)\right]_{1\leq\alpha,\beta\leq n}\right) \\ &\gg \lambda_{\min}\left(\left[\mathcal{G}_t^{[l]}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta)\right]_{1\leq\alpha,\beta\leq n}\right).\end{aligned}$$

However, for ResNet, even if we assume straightforward that all kernels $\mathcal{G}_t^{[l]}$ are positive definite, adding them up will not give substantial increase to the least eigenvalue. On account of the fact that there exists a scaling factor $\frac{1}{L^2}$ for kernels $\left\{\mathcal{G}_t^{[l]}\right\}_{l=2}^{L}$, heuristically, the gap of the least eigenvalues between $\mathcal{G}_t^{(2)}(\cdot)$ and $\mathcal{G}_t^{[L+1]}(\cdot) + \mathcal{G}_t^{[1]}(\cdot)$ is at most of order $\mathcal{O}\left(\frac{L-1}{L^2}\right) = \mathcal{O}\left(\frac{1}{L}\right)$. Hence for ResNet, even if it goes really deep, the least eigenvalue of its

NTK still 'concentrates' on the kernels $\mathcal{G}_t^{[L+1]}(\cdot)$ and $\mathcal{G}_t^{[1]}(\cdot)$. Thanks to that observation, we only need to bring $\mathcal{G}_t^{[L+1]}(\cdot)$ to the spotlight. Analysis of $\mathcal{G}_t^{[1]}(\cdot)$ is omitted because it is not needed for our proof.

3.3 Key Technique Number One: Kernel Structure

The core idea of this technique is to simply take derivatives.

3.3.1 Replacement Rules

We revisit the NTK

$$\mathcal{G}_t^{(2)}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta) = \sum_{l=1}^{L+1} \mathcal{G}_t^{[l]}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta),$$

where $\mathcal{G}_t^{(2)}(\cdot)$ is the sum of $L+1$ terms, with each term being the inner product of vectors containing components $\boldsymbol{a}_t$, $\boldsymbol{x}_\alpha^{[l]}$, $\boldsymbol{E}_{t,\alpha}^{[l]}$, and $\boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\alpha)$, as is shown in Equation (3.4). Following Equation (2.6), we are able to write down the gradient dynamics of $\boldsymbol{a}_t$, $\boldsymbol{x}_\alpha^{[l]}$, $\boldsymbol{E}_{t,\alpha}^{[l]}$, and $\boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\alpha)$. As is shown in Equation (2.6), whenever we take derivatives over a specific term, its expression heuristically reads

$$\partial_t \left(\textbf{Anti-Derivative}\right) = -\frac{1}{n}\sum_{\beta=1}^{n} \left(\textbf{Derivative}\right)\left(f_\beta(t) - y_\beta\right), \tag{3.8}$$

where 'Anti-Derivative' refers to the term we take derivative over. For instance, while the dynamics of $\boldsymbol{a}_t$ is written into

$$\partial_t \boldsymbol{a}_t = -\frac{1}{n}\sum_{\beta=1}^{n} \frac{1}{\sqrt{m}} \sqrt{m}\boldsymbol{x}_\beta^{[L]}(f_\beta(t) - y_\beta), \tag{3.9}$$

we refer to $\boldsymbol{a}_t$ as 'Anti-Derivative', and $\frac{1}{\sqrt{m}}\sqrt{m}\boldsymbol{x}_\beta^{[L]}$ as its 'Derivative'. For simplicity, we symbolize the dynamics (3.9) as $\boldsymbol{a}_t \to \frac{1}{\sqrt{m}}\sqrt{m}\boldsymbol{x}_\beta^{[L]}$. Similarly, for the dynamics of $\boldsymbol{x}_\alpha^{[l]}$,

$$
\begin{aligned}
&\sqrt{m}\boldsymbol{x}_\alpha^{[1]} \to \frac{c_\sigma}{\sqrt{m}}\mathrm{diag}\left(\boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\alpha)\boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[2:L]}\right)^\intercal \boldsymbol{a}_t\right)\mathbf{1}\left\langle \boldsymbol{x}_\alpha, \boldsymbol{x}_\beta\right\rangle,\\
&\text{for } 2 \le l \le L,\\
&\sqrt{m}\boldsymbol{x}_\alpha^{[l]} \to \frac{c_\sigma}{\sqrt{m}}\mathrm{diag}\left(\boldsymbol{E}_{t,\alpha}^{[2:l]}\boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\alpha)\boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[2:L]}\right)^\intercal \boldsymbol{a}_t\right)\mathbf{1}\left\langle \boldsymbol{x}_\alpha, \boldsymbol{x}_\beta\right\rangle\\
&\qquad + \sum_{k=2}^{l}\frac{c_{\mathrm{res}}^2}{L^2\sqrt{m}}\mathrm{diag}\Bigg(\boldsymbol{E}_{t,\alpha}^{[(k+1):l]}\\
&\qquad \boldsymbol{\sigma}_{[k]}^{(1)}(\boldsymbol{x}_\alpha)\boldsymbol{\sigma}_{[k]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[(k+1):L]}\right)^\intercal \boldsymbol{a}_t\Bigg)\mathbf{1}\left\langle \boldsymbol{x}_\alpha^{[k-1]}, \boldsymbol{x}_\beta^{[k-1]}\right\rangle,
\end{aligned}
\tag{3.10}
$$

of $\boldsymbol{W}_t^{[l]}$,

$$
\begin{aligned}
&\boldsymbol{W}_t^{[L]} \to \frac{c_{\mathrm{res}}}{L\sqrt{m}}\mathrm{diag}\left(\boldsymbol{\sigma}_{[L]}^{(1)}(\boldsymbol{x}_\beta)\boldsymbol{a}_t\right)\mathbf{1}\otimes(\boldsymbol{x}_\beta^{[L-1]})^\intercal,\\
&\text{for } 2 \le l \le L-1,\\
&\boldsymbol{W}_t^{[l]} \to \frac{c_{\mathrm{res}}}{L\sqrt{m}}\mathrm{diag}\left(\boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[(l+1):L]}\right)^\intercal \boldsymbol{a}_t\right)\mathbf{1}\otimes(\boldsymbol{x}_\beta^{[l-1]})^\intercal,\\
&\boldsymbol{W}_t^{[1]} \to \sqrt{\frac{c_\sigma}{m}}\mathrm{diag}\left(\boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[2:L]}\right)^\intercal \boldsymbol{a}_t\right)\mathbf{1}\otimes(\boldsymbol{x}_\beta)^\intercal,
\end{aligned}
\tag{3.11}
$$

of $\boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\alpha)$, where $r \geq 1$,

$$
\begin{aligned}
\boldsymbol{\sigma}_{[1]}^{(r)}(\boldsymbol{x}_\alpha) &\to \sqrt{\frac{c_\sigma}{m}} \boldsymbol{\sigma}_{[1]}^{(r+1)}(\boldsymbol{x}_\alpha) \mathrm{diag}\left(\boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[2:L]}\right)^\intercal \boldsymbol{a}_t\right) \langle \boldsymbol{x}_\alpha, \boldsymbol{x}_\beta \rangle, \\
\boldsymbol{\sigma}_{[2]}^{(r)}(\boldsymbol{x}_\alpha) &\to \frac{c_{\mathrm{res}}}{L\sqrt{m}} \boldsymbol{\sigma}_{[2]}^{(r+1)}(\boldsymbol{x}_\alpha) \mathrm{diag}\left(\boldsymbol{\sigma}_{[2]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[3:L]}\right)^\intercal \boldsymbol{a}_t\right) \left\langle \boldsymbol{x}_\alpha^{[1]}, \boldsymbol{x}_\beta^{[1]} \right\rangle \\
&\quad + \frac{c_\sigma}{\sqrt{m}} \boldsymbol{\sigma}_{[2]}^{(r+1)}(\boldsymbol{x}_\alpha) \mathrm{diag}\left(\frac{\boldsymbol{W}_t^{[2]}}{\sqrt{m}} \boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\alpha) \boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[2:L]}\right)^\intercal \boldsymbol{a}_t\right) \langle \boldsymbol{x}_\alpha, \boldsymbol{x}_\beta \rangle,
\end{aligned}
$$

for $2 \leq l \leq L-1$,

$$
\begin{aligned}
\boldsymbol{\sigma}_{[l+1]}^{(r)}(\boldsymbol{x}_\alpha) &\to \frac{c_{\mathrm{res}}}{L\sqrt{m}} \boldsymbol{\sigma}_{[l+1]}^{(r+1)}(\boldsymbol{x}_\alpha) \mathrm{diag}\left(\boldsymbol{\sigma}_{[l+1]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[(l+2):L]}\right)^\intercal \boldsymbol{a}_t\right) \left\langle \boldsymbol{x}_\alpha^{[l]}, \boldsymbol{x}_\beta^{[l]} \right\rangle \\
&\quad + \sum_{k=2}^{l} \frac{c_{\mathrm{res}}^2}{L^2\sqrt{m}} \boldsymbol{\sigma}_{[l+1]}^{(r+1)}(\boldsymbol{x}_\alpha) \mathrm{diag}\Bigg(\frac{\boldsymbol{W}_t^{[l+1]}}{\sqrt{m}} \boldsymbol{E}_{t,\alpha}^{[(k+1):l]} \\
&\quad \boldsymbol{\sigma}_{[k]}^{(1)}(\boldsymbol{x}_\alpha) \boldsymbol{\sigma}_{[k]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[(k+1):L]}\right)^\intercal \boldsymbol{a}_t\Bigg) \left\langle \boldsymbol{x}_\alpha^{[k-1]}, \boldsymbol{x}_\beta^{[k-1]} \right\rangle \\
&\quad + \frac{c_\sigma}{\sqrt{m}} \boldsymbol{\sigma}_{[l+1]}^{(r+1)}(\boldsymbol{x}_\alpha) \mathrm{diag}\Bigg(\frac{\boldsymbol{W}_t^{[l+1]}}{\sqrt{m}} \boldsymbol{E}_{t,\alpha}^{[2:l]} \\
&\quad \boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\alpha) \boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[2:L]}\right)^\intercal \boldsymbol{a}_t\Bigg) \langle \boldsymbol{x}_\alpha, \boldsymbol{x}_\beta \rangle,
\end{aligned}
\tag{3.12}
$$

and finally of $\boldsymbol{E}_{t,\alpha}^{[l]}$

$$
\begin{aligned}
\boldsymbol{E}_{t,\alpha}^{[2]} \to & \frac{c_{\text{res}}^2}{L^2\sqrt{m}}\text{diag}\left(\boldsymbol{\sigma}_{[2]}^{(1)}(\boldsymbol{x}_\alpha)\boldsymbol{\sigma}_{[2]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[3:L]}\right)^\intercal \boldsymbol{a}_t\right)\mathbf{1}\otimes(\frac{\sqrt{m}\boldsymbol{x}_\beta^{[1]}}{m})^\intercal \\
& + \frac{c_{\text{res}}}{L\sqrt{m}}\boldsymbol{\sigma}_{[2]}^{(2)}(\boldsymbol{x}_\alpha)\text{diag}\left(\boldsymbol{\sigma}_{[2]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[3:L]}\right)^\intercal \boldsymbol{a}_t\right)\frac{c_{\text{res}}}{L}\frac{\boldsymbol{W}_t^{[2]}}{\sqrt{m}}\left\langle \boldsymbol{x}_\alpha^{[1]}, \boldsymbol{x}_\beta^{[1]}\right\rangle \\
& + \frac{c_\sigma}{\sqrt{m}}\boldsymbol{\sigma}_{[2]}^{(2)}(\boldsymbol{x}_\alpha)\text{diag}\Bigg(\frac{\boldsymbol{W}_t^{[2]}}{\sqrt{m}} \\
& \boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\alpha)\boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[2:L]}\right)^\intercal \boldsymbol{a}_t\Bigg)\frac{c_{\text{res}}}{L}\frac{\boldsymbol{W}_t^{[2]}}{\sqrt{m}}\left\langle \boldsymbol{x}_\alpha, \boldsymbol{x}_\beta\right\rangle,
\end{aligned}
$$

for $2 \le l \le L-1$,

$$
\begin{aligned}
\boldsymbol{E}_{t,\alpha}^{[l+1]} \to & \frac{c_{\text{res}}^2}{L^2\sqrt{m}}\text{diag}\left(\boldsymbol{\sigma}_{[l+1]}^{(1)}(\boldsymbol{x}_\alpha)\boldsymbol{\sigma}_{[l+1]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[(l+2):L]}\right)^\intercal \boldsymbol{a}_t\right)\mathbf{1}\otimes(\frac{\sqrt{m}\boldsymbol{x}_\beta^{[l]}}{m})^\intercal \\
& + \frac{c_{\text{res}}^2}{L^2\sqrt{m}}\boldsymbol{\sigma}_{[l+1]}^{(2)}(\boldsymbol{x}_\alpha)\text{diag}\left(\boldsymbol{\sigma}_{[l+1]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[(l+2):L]}\right)^\intercal \boldsymbol{a}_t\right)\frac{\boldsymbol{W}_t^{[l+1]}}{\sqrt{m}}\left\langle \boldsymbol{x}_\alpha^{[l]}, \boldsymbol{x}_\beta^{[l]}\right\rangle \\
& + \sum_{k=2}^{l}\frac{c_{\text{res}}^3}{L^3\sqrt{m}}\boldsymbol{\sigma}_{[l+1]}^{(2)}(\boldsymbol{x}_\alpha)\text{diag}\Bigg(\frac{\boldsymbol{W}_t^{[l+1]}}{\sqrt{m}}\boldsymbol{E}_{t,\alpha}^{[(k+1):l]} \\
& \boldsymbol{\sigma}_{[k]}^{(1)}(\boldsymbol{x}_\alpha)\boldsymbol{\sigma}_{[k]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[(k+1):L]}\right)^\intercal \boldsymbol{a}_t\Bigg)\frac{\boldsymbol{W}_t^{[l+1]}}{\sqrt{m}}\left\langle \boldsymbol{x}_\alpha^{[k-1]}, \boldsymbol{x}_\beta^{[k-1]}\right\rangle \\
& + \frac{c_\sigma}{\sqrt{m}}\frac{c_{\text{res}}}{L}\boldsymbol{\sigma}_{[l+1]}^{(2)}(\boldsymbol{x}_\alpha)\text{diag}\Bigg(\frac{\boldsymbol{W}_t^{[l+1]}}{\sqrt{m}}\boldsymbol{E}_{t,\alpha}^{[2:l]} \\
& \boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\alpha)\boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[2:L]}\right)^\intercal \boldsymbol{a}_t\Bigg)\frac{\boldsymbol{W}_t^{[l+1]}}{\sqrt{m}}\left\langle \boldsymbol{x}_\alpha, \boldsymbol{x}_\beta\right\rangle.
\end{aligned}
\tag{3.13}
$$

Altogether Equation (3.10), Equation (3.11), Equation (3.12) and Equation (3.13) are termed the *Replacement Rules*. We instantly obtain the derivative for $f_\alpha(t)$, from Equation (1.2), we

have that $f_\alpha(t) = f_{\text{res}}(\boldsymbol{x}_\alpha, \boldsymbol{\theta}_t) = \boldsymbol{a}_t^\intercal \boldsymbol{x}_\alpha^{[L]} = \left\langle \boldsymbol{a}_t, \boldsymbol{x}_\alpha^{[L]} \right\rangle$, then by applying one of the replacement rules (Equation (3.10)) on $f_\alpha(t)$, we have

$$
\begin{aligned}
\left\langle \boldsymbol{a}_t, \boldsymbol{x}_\alpha^{[L]} \right\rangle \to & \left\langle \boldsymbol{x}_\beta^{[L]}, \boldsymbol{x}_\alpha^{[L]} \right\rangle + \left\langle \boldsymbol{a}_t, \frac{c_\sigma}{m} \operatorname{diag}\left(\boldsymbol{E}_{t,\alpha}^{[2:l]} \boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\alpha) \boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\beta) \left(\boldsymbol{E}_{t,\beta}^{[2:L]} \right)^\intercal \boldsymbol{a}_t \right) \mathbf{1} \left\langle \boldsymbol{x}_\alpha, \boldsymbol{x}_\beta \right\rangle \right\rangle \\
& + \sum_{k=2}^{L} \left\langle \boldsymbol{a}_t, \frac{c_{\text{res}}^2}{L^2 m} \operatorname{diag}\left(\boldsymbol{E}_{t,\alpha}^{[(k+1):l]} \boldsymbol{\sigma}_{[k]}^{(1)}(\boldsymbol{x}_\alpha) \boldsymbol{\sigma}_{[k]}^{(1)}(\boldsymbol{x}_\beta) \left(\boldsymbol{E}_{t,\beta}^{[(k+1):L]} \right)^\intercal \boldsymbol{a}_t \right) \mathbf{1} \left\langle \boldsymbol{x}_\alpha^{[k-1]}, \boldsymbol{x}_\beta^{[k-1]} \right\rangle \right\rangle \\
= & \left\langle \boldsymbol{x}_\alpha^{[L]}, \boldsymbol{x}_\beta^{[L]} \right\rangle + \left\langle \boldsymbol{a}_t, \frac{c_\sigma}{m} \boldsymbol{E}_{t,\alpha}^{[2:l]} \boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\alpha) \boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\beta) \left(\boldsymbol{E}_{t,\beta}^{[2:L]} \right)^\intercal \boldsymbol{a}_t \right\rangle \left\langle \boldsymbol{x}_\alpha, \boldsymbol{x}_\beta \right\rangle \\
& + \sum_{k=2}^{L} \left\langle \boldsymbol{a}_t, \frac{c_{\text{res}}^2}{L^2 m} \boldsymbol{E}_{t,\alpha}^{[(k+1):l]} \boldsymbol{\sigma}_{[k]}^{(1)}(\boldsymbol{x}_\alpha) \boldsymbol{\sigma}_{[k]}^{(1)}(\boldsymbol{x}_\beta) \left(\boldsymbol{E}_{t,\beta}^{[(k+1):L]} \right)^\intercal \boldsymbol{a}_t \right\rangle \left\langle \boldsymbol{x}_\alpha^{[k-1]}, \boldsymbol{x}_\beta^{[k-1]} \right\rangle \\
= & \left\langle \boldsymbol{x}_\alpha^{[L]}, \boldsymbol{x}_\beta^{[L]} \right\rangle + \frac{1}{m} \left\langle \sqrt{c_\sigma} \boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\alpha) \left(\boldsymbol{E}_{t,\alpha}^{[2:l]} \right)^\intercal \boldsymbol{a}_t, \sqrt{c_\sigma} \boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\beta) \left(\boldsymbol{E}_{t,\beta}^{[2:L]} \right)^\intercal \boldsymbol{a}_t \right\rangle \left\langle \boldsymbol{x}_\alpha, \boldsymbol{x}_\beta \right\rangle \\
& + \sum_{k=2}^{L} \frac{1}{m} \left\langle \frac{c_{\text{res}}}{L} \boldsymbol{\sigma}_{[k]}^{(1)}(\boldsymbol{x}_\alpha) \left(\boldsymbol{E}_{t,\alpha}^{[(k+1):l]} \right)^\intercal \boldsymbol{a}_t, \frac{c_{\text{res}}}{L} \boldsymbol{\sigma}_{[k]}^{(1)}(\boldsymbol{x}_\beta) \left(\boldsymbol{E}_{t,\beta}^{[(k+1):L]} \right)^\intercal \boldsymbol{a}_t \right\rangle \left\langle \boldsymbol{x}_\alpha^{[k-1]}, \boldsymbol{x}_\beta^{[k-1]} \right\rangle,
\end{aligned}
$$

and we notice that sum of these terms altogether reads $\mathcal{G}_t^{(2)}(\boldsymbol{x}_{\alpha_1}, \boldsymbol{x}_{\alpha_2})$, hence we obtain that

$$
\partial_t \left(f_\alpha(t) - y_\alpha \right) = \partial_t f_\alpha(t) = -\frac{1}{n} \sum_{\beta=1}^{n} \mathcal{G}_t^{(2)}(\boldsymbol{x}_{\alpha_1}, \boldsymbol{x}_{\alpha_2}) (f_\beta(t) - y_\beta).
$$

Applying the replacement rules once again, the derivative for NTK $\mathcal{G}_t^{(2)}(\cdot)$ is obtained in the following form

$$
\partial_t \mathcal{G}_t^{(2)}(\boldsymbol{x}_{\alpha_1}, \boldsymbol{x}_{\alpha_2}) = -\frac{1}{n} \sum_{\beta=1}^{n} \mathcal{G}_t^{(3)}(\boldsymbol{x}_{\alpha_1}, \boldsymbol{x}_{\alpha_2}, \boldsymbol{x}_\beta) (f_\beta(t) - y_\beta),
$$

with each term in $\mathcal{G}_t^{(3)}(\boldsymbol{x}_{\alpha_1}, \boldsymbol{x}_{\alpha_2}, \boldsymbol{x}_\beta)$ consisting of the summation of all the terms generated from $\mathcal{G}_t^{(2)}(\boldsymbol{x}_{\alpha_1}, \boldsymbol{x}_{\alpha_2})$ by performing the replacement procedure. In order to illustrate the idea, besides the above computational example where $f_\alpha(t)$ is the main object, another example is given out in the proof of Theorem 3.2.2 in Section 3.6.

By same reasoning, we could obtain higher order kernels inductively by performing all the possible replacements. For kernel $\mathcal{G}_t^{(r)}(\boldsymbol{x}_{\alpha_1}, \boldsymbol{x}_{\alpha_2}, \ldots, \boldsymbol{x}_{\alpha_r})$, where $r \geq 2$, the following Ordinary Differential Equation gives $\mathcal{G}_t^{(r+1)}(\boldsymbol{x}_{\alpha_1}, \ldots, \boldsymbol{x}_{\alpha_r}; \boldsymbol{x}_\beta)$:

$$\partial_t \mathcal{G}_t^{(r)}(\boldsymbol{x}_{\alpha_1}, \boldsymbol{x}_{\alpha_2} \ldots, \boldsymbol{x}_{\alpha_r}) = -\frac{1}{n} \sum_{\beta=1}^{n} \mathcal{G}_t^{(r+1)}(\boldsymbol{x}_{\alpha_1}, \boldsymbol{x}_{\alpha_2} \ldots, \boldsymbol{x}_{\alpha_r}, \boldsymbol{x}_\beta)(f_\beta(t) - y_\beta),$$

and it finishes the proof for Equation (3.1) and Equation (3.2) in Theorem 3.2.1.

Furthermore, in order to describe the members appearing in $\mathcal{G}_t^{(p)}(\boldsymbol{x}_{\alpha_1}, \boldsymbol{x}_{\alpha_2} \ldots, \boldsymbol{x}_{\alpha_p})$, $p \geq 3$ more systematically, some notations shall be introduced.

3.3.2 Hierarchical Sets of Kernel Expressions

Firstly, we denote $\mathbb{A}_0$ as the first set of expressions, which corresponds to the terms in $\mathcal{G}_t^{(2)}(\boldsymbol{x}_{\alpha_1}, \boldsymbol{x}_{\alpha_2})$. We define $\mathbb{A}_0$ as:

$$\mathbb{A}_0 \triangleq \{\mathrm{e}_s \mathrm{e}_{s-1} \ldots \mathrm{e}_1 \mathrm{e}_0 : 0 \leq s \leq 4L\}, \tag{3.14}$$

where $\mathrm{e_j}$ is chosen following

$$\begin{aligned} &\mathrm{e}_0 \in \left\{\boldsymbol{a}_t, \{\sqrt{m}\boldsymbol{x}_\beta^{[0]}, \sqrt{m}\boldsymbol{x}_\beta^{[1]}, \sqrt{m}\boldsymbol{x}_\beta^{[2]}, \ldots, \sqrt{m}\boldsymbol{x}_\beta^{[L]}\}_{1\leq\beta\leq n}\right\}, \\ &\text{for } 1 \leq \mathrm{j} \leq s, \\ &\mathrm{e_j} \in \left\{\left\{\boldsymbol{E}_{t,\beta}^{[2]}, \left(\boldsymbol{E}_{t,\beta}^{[2]}\right)^\intercal, \ldots, \boldsymbol{E}_{t,\beta}^{[L]}, \left(\boldsymbol{E}_{t,\beta}^{[L]}\right)^\intercal\right\}_{1\leq\beta\leq n}, \left\{\boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\beta), \ldots, \boldsymbol{\sigma}_{[L]}^{(1)}(\boldsymbol{x}_\beta)\right\}_{1\leq\beta\leq n}\right\}. \end{aligned} \tag{3.15}$$

From Equation (3.4), each term in $\mathcal{G}_t^{(2)}(\boldsymbol{x}_{\alpha_1}, \boldsymbol{x}_{\alpha_2})$ reads

$$\frac{\langle \boldsymbol{u}_1(t), \boldsymbol{u}_2(t)\rangle}{m} \quad \text{or} \quad \frac{\langle \boldsymbol{u}_1(t), \boldsymbol{u}_2(t)\rangle}{m} \frac{\langle \boldsymbol{u}_3(t), \boldsymbol{u}_4(t)\rangle}{m},$$

with vectors $\boldsymbol{u}_1(t), \boldsymbol{u}_2(t), \boldsymbol{u}_3(t), \boldsymbol{u}_4(t) \in \mathbb{A}_0$.

We remark that compared with Huang and Yau [3], their e_j $(1 \le \mathrm{j} \le s)$ is chosen differently. The counterpart in [3] is selected from the set

$$\left\{ \left\{ \frac{\boldsymbol{W}_t^{[2]}}{\sqrt{m}}, \left(\frac{\boldsymbol{W}_t^{[2]}}{\sqrt{m}}\right)^\intercal, \ldots, \frac{\boldsymbol{W}_t^{[L]}}{\sqrt{m}}, \left(\frac{\boldsymbol{W}_t^{[L]}}{\sqrt{m}}\right)^\intercal \right\}_{1\le\beta\le n}, \left\{ \boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\beta), \ldots, \boldsymbol{\sigma}_{[L]}^{(1)}(\boldsymbol{x}_\beta) \right\}_{1\le\beta\le n} \right\}.$$

Such change arises from the difference in network structures, and it will be shown further that skip-connection matrices $\boldsymbol{E}_{t,\beta}^{[l]}$ possess more stability than $\boldsymbol{W}_t^{[l]}/\sqrt{m}$.

Secondly, we shall investigate the set of expressions in higher orders. Given constructions of $\mathbb{A}_0, \mathbb{A}_1, \ldots, \mathbb{A}_r$, we denote $\mathbb{A}_{r+1}$ as the set of expressions in the following form:

$$\mathbb{A}_{r+1} \triangleq \{\mathrm{e}_s \mathrm{e}_{s-1} \ldots \mathrm{e}_1 \mathrm{e}_0 : 0 \le s \le 4L\}, \tag{3.16}$$

where e_0 is chosen from

$$\mathrm{e}_0 \in \left\{ \boldsymbol{a}_t, \mathbf{1}, \{\sqrt{m}\boldsymbol{x}_\beta^{[0]}, \sqrt{m}\boldsymbol{x}_\beta^{[1]}, \sqrt{m}\boldsymbol{x}_\beta^{[2]}, \ldots, \sqrt{m}\boldsymbol{x}_\beta^{[L]}\}_{1\le\beta\le n} \right\}, \tag{3.17}$$

while for $1 \le \mathrm{j} \le s$, each e_j is chosen from one of the three following sets

$$\begin{aligned}
&\left\{ \left\{ \boldsymbol{E}_{t,\beta}^{[2]}, \left(\boldsymbol{E}_{t,\beta}^{[2]}\right)^\intercal, \ldots, \boldsymbol{E}_{t,\beta}^{[L]}, \left(\boldsymbol{E}_{t,\beta}^{[L]}\right)^\intercal \right\}_{1\le\beta\le n}, \left\{ \boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\beta), \ldots, \boldsymbol{\sigma}_{[L]}^{(1)}(\boldsymbol{x}_\beta) \right\}_{1\le\beta\le n} \right\}, \\
&\left\{ \mathrm{diag}(\boldsymbol{g}),\ \boldsymbol{g} \in \mathbb{A}_0 \cup \mathbb{A}_1 \cup \cdots \cup \mathbb{A}_r \right\}, \\
&\text{with } 2 \le l \le L,\ 1 \le \beta \le n, 1 \le u \le r, \\
&\left\{ \boldsymbol{\sigma}_{[l]}^{(u+1)}(\boldsymbol{x}_\beta)\ \mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_1} \boldsymbol{g}_1 \right) \ldots \mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_u} \boldsymbol{g}_u \right) \left(\frac{c_{\mathrm{res}}}{L}\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_{u+1}}, \right. \\
&\left. \left(\frac{c_{\mathrm{res}}}{L}\frac{\left(\boldsymbol{W}_t^{[l]}\right)^\intercal}{\sqrt{m}}\right)^{Q_{u+1}} \boldsymbol{\sigma}_{[l]}^{(u+1)}(\boldsymbol{x}_\beta)\ \mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_1} \boldsymbol{g}_1 \right) \ldots \mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_u} \boldsymbol{g}_u \right) : \right. \\
&\left. \boldsymbol{g}_1, \boldsymbol{g}_2 \ldots \boldsymbol{g}_u \in \mathbb{A}_0 \cup \mathbb{A}_1 \cup \cdots \cup \mathbb{A}_r,\ \ Q_1, Q_2 \ldots Q_{u+1} \in \{0, 1\} \right\}.
\end{aligned} \tag{3.18}$$

We use Proposition 3.3.1 to shed light on the structures of the elements in $\mathbb{A}_r$, and consequently on the structures of each term in kernel $\mathcal{G}_t^{(r)}(\boldsymbol{x}_{\alpha_1}, \boldsymbol{x}_{\alpha_2}, \ldots, \boldsymbol{x}_{\alpha_r})$.

Proposition 3.3.1. *For any vector $\boldsymbol{u}(t) \in \mathbb{A}_r$, the new vector obtained from $\boldsymbol{u}(t)$ by applying the replacement rules is sum of the terms with the following forms*

$$\begin{aligned}
&(a). \frac{C}{\sqrt{m}}\boldsymbol{u}^*(t) : \boldsymbol{u}^*(t) \in \mathbb{A}_r,\\
&(b). \frac{C}{\sqrt{m}}\boldsymbol{u}^*(t)\frac{\langle \boldsymbol{p}, \boldsymbol{q}\rangle}{m} : \boldsymbol{u}^*(t) \in \mathbb{A}_{r+1}, \quad \boldsymbol{p}, \boldsymbol{q} \in \mathbb{A}_0,\\
&(c). \frac{C}{\sqrt{m}}\boldsymbol{u}^*(t)\frac{\langle \boldsymbol{p}, \boldsymbol{q}\rangle}{m} : \boldsymbol{u}^*(t) \in \mathbb{A}_{r-s+1}, \quad \boldsymbol{p} \in \mathbb{A}_s, \boldsymbol{q} \in \mathbb{A}_0, \text{ for some } \ s \geq 1,\\
&(d). \frac{C}{\sqrt{m}}\boldsymbol{u}^*(t)\frac{\langle \boldsymbol{p}, \boldsymbol{q}\rangle}{m} : \boldsymbol{u}^*(t) \in \mathbb{A}_s, \quad \boldsymbol{p} \in \mathbb{A}_{r-s+1}, \boldsymbol{q} \in \mathbb{A}_0, \text{ for some } s \geq 1.
\end{aligned} \tag{3.19}$$

Proof. We remark that the constant C in Equation (3.19) may change from term to term.
(i). Firstly, since $\boldsymbol{a}_t$ appears only at the position e_0, if $\boldsymbol{u}(t) \in \mathbb{A}_r$, from replacement rules

$$\boldsymbol{u}(t) = \mathrm{e}_s\mathrm{e}_{s-1}\ldots\mathrm{e}_1\boldsymbol{a}_t \to \widetilde{\boldsymbol{u}}(t) = \frac{1}{\sqrt{m}}\mathrm{e}_s\mathrm{e}_{s-1}\ldots\mathrm{e}_1\sqrt{m}\boldsymbol{x}_\beta^{[L]} = \frac{1}{\sqrt{m}}\boldsymbol{u}^*(t),$$

then $\boldsymbol{u}^*(t) = \mathrm{e}_s\mathrm{e}_{s-1}\ldots\mathrm{e}_1\sqrt{m}\boldsymbol{x}_\beta^{[L]} \in \mathbb{A}_r$.
(ii). Similarly, $\sqrt{m}\boldsymbol{x}_\alpha^{[l]}$ also appears only at e_0, then if $\boldsymbol{u}(t) \in \mathbb{A}_r$, from replacement rules

$$\begin{aligned}
\boldsymbol{u}(t) &= \mathrm{e}_s\mathrm{e}_{s-1}\ldots\mathrm{e}_1\ \sqrt{m}\boldsymbol{x}_\alpha^{[l]} \to \widetilde{\boldsymbol{u}}(t),\\
\widetilde{\boldsymbol{u}}(t) &= \sum_k \frac{C}{\sqrt{m}}\mathrm{e}_s\mathrm{e}_{s-1}\ldots\mathrm{e}_1\ \mathrm{diag}(\boldsymbol{f}_k)\ \mathbf{1}\ \frac{\left\langle \sqrt{m}\boldsymbol{x}_\alpha^{[k]}, \sqrt{m}\boldsymbol{x}_\beta^{[k]}\right\rangle}{m}\\
&= \sum_k \frac{C}{\sqrt{m}}\boldsymbol{u}_k^*(t)\frac{\left\langle \sqrt{m}\boldsymbol{x}_\alpha^{[k]}, \sqrt{m}\boldsymbol{x}_\beta^{[k]}\right\rangle}{m},
\end{aligned}$$

with $\boldsymbol{f}_k \in \mathbb{A}_0$, then $\boldsymbol{u}_k^*(t) \in \mathbb{A}_{r+1}$.

(iii). Moreover, for $\boldsymbol{\sigma}_{[l]}^{(u)}(\boldsymbol{x}_\alpha)$, which only appears at the starting or middle position. For $u=1$, $\boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\alpha)$ has no diag operations accompanied with it, and any vector $\boldsymbol{u}(t)\in\mathbb{A}_r$ could contain $\boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\alpha)$ for $r\geq 0$. From replacement rules

$$
\begin{aligned}
\boldsymbol{u}(t) &= \mathrm{e}_s\ldots\mathrm{e}_{\mathrm{j}+1}\boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\alpha)\mathrm{e}_{\mathrm{j}-1}\ldots\mathrm{e}_0\to\widetilde{\boldsymbol{u}}(t),\\
\widetilde{\boldsymbol{u}}(t) &= \frac{C}{\sqrt{m}}\mathrm{e}_s\ldots\mathrm{e}_{\mathrm{j}+1}\boldsymbol{\sigma}_{[l]}^{(2)}(\boldsymbol{x}_\alpha)\mathrm{diag}\left(\boldsymbol{f}_1\right)\mathrm{e}_{\mathrm{j}-1}\ldots\mathrm{e}_0\frac{\langle\boldsymbol{p}_1,\boldsymbol{q}_1\rangle}{m}\\
&\quad+\sum_k\frac{C}{\sqrt{m}}\mathrm{e}_s\ldots\mathrm{e}_{\mathrm{j}+1}\boldsymbol{\sigma}_{[l]}^{(2)}(\boldsymbol{x}_\alpha)\mathrm{diag}\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\boldsymbol{f}_k\right)\mathrm{e}_{\mathrm{j}-1}\ldots\mathrm{e}_0\frac{\langle\boldsymbol{p}_k,\boldsymbol{q}_k\rangle}{m}\\
&= \sum_l\frac{C}{\sqrt{m}}\boldsymbol{u}_l^*(t)\frac{\langle\boldsymbol{p}_l,\boldsymbol{q}_l\rangle}{m},
\end{aligned}
$$

with $\boldsymbol{f}_k\in\mathbb{A}_0$, then $\boldsymbol{u}_l^*(t)\in\mathbb{A}_{r+1}$, and $\boldsymbol{p}_l,\boldsymbol{q}_l\in\mathbb{A}_0$.

For $u\neq 1$, $\boldsymbol{\sigma}_{[l]}^{(u)}(\boldsymbol{x}_\alpha)$ has at most $u-1$ diag operations behind it, and only for $\boldsymbol{u}(t)\in\mathbb{A}_r$ with $r\geq u-1$ could it contain $\boldsymbol{\sigma}_{[l]}^{(u)}(\boldsymbol{x}_\alpha)$,

$$
\begin{aligned}
&\boldsymbol{u}(t) = \mathrm{e}_s\ldots\mathrm{e}_{\mathrm{j}+1}\mathrm{e}_{\mathrm{j}}\mathrm{e}_{\mathrm{j}-1}\ldots\mathrm{e}_0\to\widetilde{\boldsymbol{u}}(t),\\
&\exists\ \mathrm{e}_{\mathrm{j}} = \boldsymbol{\sigma}_{[l]}^{(u)}(\boldsymbol{x}_\alpha)\mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_1}\boldsymbol{g}_1\right)\ldots\mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_{u-1}}\boldsymbol{g}_{u-1}\right)\left(\frac{c_{\mathrm{res}}}{L}\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_u},\\
&\text{or } \mathrm{e}_{\mathrm{j}} = \left(\frac{c_{\mathrm{res}}}{L}\frac{\left(\boldsymbol{W}_t^{[l]}\right)^{\mathsf{T}}}{\sqrt{m}}\right)^{Q_u}\boldsymbol{\sigma}_{[l]}^{(u)}(\boldsymbol{x}_\alpha)\mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_1}\boldsymbol{g}_1\right)\ldots\mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_{u-1}}\boldsymbol{g}_{u-1}\right),
\end{aligned}
$$

after applying replacement rules on $\mathrm{e_j} \to \mathrm{e_j^*}$,

$$\begin{aligned}
\mathrm{e_j^*} &= \frac{C}{\sqrt{m}}\boldsymbol{\sigma}_{[l]}^{(u+1)}(\boldsymbol{x}_\alpha)\ldots\left(\frac{c_{\mathrm{res}}}{L}\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_u}\frac{\langle \boldsymbol{p}_1,\boldsymbol{q}_1\rangle}{m}\\
&+\sum_k \frac{C}{\sqrt{m}}\boldsymbol{\sigma}_{[l]}^{(u+1)}(\boldsymbol{x}_\alpha)\mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_0}\boldsymbol{f}_k\right)\ldots\left(\frac{c_{\mathrm{res}}}{L}\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_u}\frac{\langle \boldsymbol{p}_k,\boldsymbol{q}_k\rangle}{m},\\
\text{or } \mathrm{e_j^*} &= \frac{C}{\sqrt{m}}\left(\frac{c_{\mathrm{res}}}{L}\frac{\left(\boldsymbol{W}_t^{[l]}\right)^\intercal}{\sqrt{m}}\right)^{Q_u}\boldsymbol{\sigma}_{[l]}^{(u+1)}(\boldsymbol{x}_\alpha)\mathrm{diag}\left(\boldsymbol{f}_1\right)\ldots\frac{\langle \boldsymbol{p}_1,\boldsymbol{q}_1\rangle}{m}\\
&+\sum_k\frac{C}{\sqrt{m}}\left(\frac{c_{\mathrm{res}}}{L}\frac{\left(\boldsymbol{W}_t^{[l]}\right)^\intercal}{\sqrt{m}}\right)^{Q_u}\boldsymbol{\sigma}_{[l]}^{(u+1)}(\boldsymbol{x}_\alpha)\mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_0}\boldsymbol{f}_k\right)\ldots\frac{\langle \boldsymbol{p}_k,\boldsymbol{q}_k\rangle}{m},
\end{aligned}$$

hence

$$\widetilde{\boldsymbol{u}}(t) = \sum_l \frac{C}{\sqrt{m}}\boldsymbol{u}_l^*(t)\frac{\langle \boldsymbol{p}_l,\boldsymbol{q}_l\rangle}{m},$$

with $\boldsymbol{f}_k \in \mathbb{A}_0$, then $\boldsymbol{u}_l^*(t) \in \mathbb{A}_{r+1}$, and $\boldsymbol{p}_l, \boldsymbol{q}_l \in \mathbb{A}_0$.

(iv). Since $\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}$ only appears at the starting or the middle position, if $\boldsymbol{u}(t) \in \mathbb{A}_r$, from replacement rules

$$\begin{aligned}
\boldsymbol{u}(t) &= \mathrm{e}_s\mathrm{e}_{s-1}\ldots\mathrm{e}_{\mathrm{j}+1}\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\mathrm{e}_{\mathrm{j}-1}\ldots\mathrm{e}_1\mathrm{e}_0 \to \widetilde{\boldsymbol{u}}(t),\\
\widetilde{\boldsymbol{u}}(t) &= \frac{C}{m}\mathrm{e}_s\mathrm{e}_{s-1}\ldots\mathrm{e}_{\mathrm{j}+1}\mathrm{diag}(\boldsymbol{g})\ \mathbf{1}\otimes(\boldsymbol{x}_\beta^{[l-1]})^\intercal\mathrm{e}_{\mathrm{j}-1}\ldots\mathrm{e}_1\mathrm{e}_0\\
&= \frac{C}{\sqrt{m}}\mathrm{e}_s\mathrm{e}_{s-1}\ldots\mathrm{e}_{\mathrm{j}+1}\mathrm{diag}(\boldsymbol{g})\ \mathbf{1}\ \frac{\left\langle \mathrm{e}_{\mathrm{j}-1}\ldots\mathrm{e}_1\mathrm{e}_0, \sqrt{m}\boldsymbol{x}_\beta^{[l-1]}\right\rangle}{m}\\
&= \frac{C}{\sqrt{m}}\boldsymbol{u}^*(t)\frac{\langle \boldsymbol{p},\boldsymbol{q}\rangle}{m},
\end{aligned}$$

with $\boldsymbol{u}^*(t) \in \mathbb{A}_{r-s+1}$, and $\boldsymbol{p} \in \mathbb{A}_s,\ \boldsymbol{q} \in \mathbb{A}_0$, for some $s \geq 1$.

Similarly for $\frac{(\boldsymbol{W}_t^{[l]})^\intercal}{\sqrt{m}}$,

$$\begin{aligned}
\boldsymbol{u}(t) &= \mathrm{e}_s \mathrm{e}_{s-1} \dots \mathrm{e}_{\mathrm{j}+1} \frac{(\boldsymbol{W}_t^{[l]})^\intercal}{\sqrt{m}} \mathrm{e}_{\mathrm{j}-1} \dots \mathrm{e}_1 \mathrm{e}_0 \to \widetilde{\boldsymbol{u}}(t) \\
\widetilde{\boldsymbol{u}}(t) &= \frac{C}{m} \mathrm{e}_s \mathrm{e}_{s-1} \dots \mathrm{e}_{\mathrm{j}+1} \boldsymbol{x}_\beta^{[l-1]} \otimes \mathbf{1}^\intercal \operatorname{diag}(\boldsymbol{g}) \ \mathrm{e}_{\mathrm{j}-1} \dots \mathrm{e}_1 \mathrm{e}_0 \\
&= \frac{C}{\sqrt{m}} \mathrm{e}_s \mathrm{e}_{s-1} \dots \mathrm{e}_{\mathrm{j}+1} \sqrt{m} \boldsymbol{x}_\beta^{[l-1]} \ \frac{\langle \operatorname{diag}(\boldsymbol{g}) \ \mathrm{e}_{\mathrm{j}-1} \dots \mathrm{e}_1 \mathrm{e}_0, \mathbf{1} \rangle}{m} \\
&= \frac{C}{\sqrt{m}} \boldsymbol{u}^*(t) \frac{\langle \boldsymbol{p}, \boldsymbol{q} \rangle}{m},
\end{aligned}$$

with $\boldsymbol{u}^*(t) \in \mathbb{A}_{r-s}$, and $\boldsymbol{p} \in \mathbb{A}_{r-s+1}$, $\boldsymbol{q} \in \mathbb{A}_0$, for some $s \geq 1$.
(v). Finally for $\boldsymbol{E}_{t,\alpha}^{[l]}$, whose scenario is the situations combined with $\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}$ and $\boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\alpha)$, we shall skip the analysis and conclude the proof. □

From the discussion above, we apply Proposition 3.3.1 sequentially to $\mathcal{G}_t^{(2)}(\boldsymbol{x}_{\alpha_1}, \boldsymbol{x}_{\alpha_2})$ for $(r-1)$ many times to obtain kernel $\mathcal{G}_t^{(r)}(\boldsymbol{x}_{\alpha_1}, \boldsymbol{x}_{\alpha_2}, \dots, \boldsymbol{x}_{\alpha_r})$, whose individual components takes the form

$$\frac{1}{m^{r/2-1}} \prod_{\mathrm{j}=1}^{s} \frac{\langle \boldsymbol{u}_{2\mathrm{j}-1}(t), \boldsymbol{u}_{2\mathrm{j}}(t) \rangle}{m}, \qquad 1 \leq s \leq r, \qquad \boldsymbol{u}_{\mathrm{i}}(t) \in \mathbb{A}_0 \cup \mathbb{A}_1 \cup \dots \cup \mathbb{A}_{r-2}. \tag{3.20}$$

We remark that Equation (3.20) brings partial proof to Equation (3.3).

3.4 Key Technique Number Two: Apriori Estimates

Huang and Yau obtained Equation (3.20) in [3, Equation (3.8)], and they use the tensor program [38] to estimate the initial value of the kernel $\mathcal{G}_0^{(r)}(\cdot)$. They showed that for each vector $\boldsymbol{u}_{\mathrm{j}}(t)$ at $t = 0$, it is a linear combination of projections of independent Gaussian vectors. Hence, if we consider such quantity

$$\eta(t) = \left\{ \|\boldsymbol{u}(t)\|_\infty : \boldsymbol{u}(t) \in \mathbb{A}_0 \cup \mathbb{A}_1 \cdots \cup \mathbb{A}_r \right\}, \tag{3.21}$$

then with high probability,

$$\eta(0) \lesssim (\ln m)^C \tag{3.22}$$

holds, since any $\boldsymbol{u}(0)$ is a linear combination of projections of independent Gaussian vectors.

For $t > 0$, Huang and Yau derived a self-consistent Ordinary Differential Inequality for $\eta(t)$,

$$\partial_t^{(p+1)}\eta(t) \lesssim \frac{\eta(t)^{2p}}{m^{p/2}}, \tag{3.23}$$

$$\eta(0) \lesssim (\ln m)^C, \tag{3.24}$$

for some $p \geq 2$. Then for time $0 \leq t \leq m^{\frac{p}{2p+1}}/(\ln m)^{C^*}$,

$$\eta(t) \lesssim (\ln m)^C.$$

Our approach is different from theirs, instead of using tensor programs, we use a special matrix norm, the 2 to infinity matrix norm, to show Equation (3.22). Then we derive a Gronwall-type inequality for $\eta(t)$,

$$\eta(t) \lesssim (\ln m)^C + \frac{1}{\sqrt{m}} \int_0^t \eta(s)\, \mathrm{d}s,$$

Hence it follows that for time $0 \leq t \leq \sqrt{m}/\left(\ln m\right)^{C^*}$,

$$\eta(t) \lesssim (\ln m)^C$$

holds, which brings the complete proof to Equation (3.3).

3.4.1 Apriori L^2 Bounds for Expressions in $\mathbb{A}_0$

We begin our proof with several lemmas.

Lemma 3.4.1. *Given $\mathcal{X}$ and $\sigma(\cdot)$, for $t \geq 0$,*

$$R_S(\boldsymbol{\theta}_t) \leq R_S(\boldsymbol{\theta}_0) \sim \mathcal{O}(1). \tag{3.25}$$

Proof. We get inequality (3.25) by non-negative definiteness of kernel $\mathcal{G}_t^{(2)}(\cdot)$. From Equation (3.1), we obtain that

$$\partial_t \sum_{\alpha=1}^{n} \|f_\alpha(t) - y_\alpha\|_2^2 = -\frac{2}{n} \sum_{\alpha,\beta=1}^{n} \mathcal{G}_t^{(2)}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta)(f_\alpha(t) - y_\alpha)(f_\beta(t) - y_\beta) \leq 0, \tag{3.26}$$

hence

$$R_S(\boldsymbol{\theta}_t) \leq R_S(\boldsymbol{\theta}_0),$$

which finishes the proof of the lemma. □

Our next lemma is mainly on the spectral property of random matrices, which was given out in [1, Lemma G.2.], also consequence of results in [40].

Lemma 3.4.2. *Given $\boldsymbol{W} \in \mathbb{R}^{m\times m}$, with entry $W_{\mathrm{i,j}} \sim \mathcal{N}(0,1)$, then with probability at least $1 - \exp\left(-\frac{(c_{w,0}-2)^2 m}{2}\right)$, the following holds*

$$\|\boldsymbol{W}\|_{2\to 2} \leq c_{w,0}^{'} \sqrt{m}, \tag{3.27}$$

where the constant $c_{w,0}^{'} > 2$.

Our next lemma is on the tail bound of the chi-square distribution, whose proof can be found in [41].

Lemma 3.4.3. *If $Z \sim \chi^2(m)$, then we have a tail bound*

$$\mathbb{P}\left(Z \geq m + 2\sqrt{mx} + 2x\right) \leq \mathrm{e}^{-x}. \tag{3.28}$$

Our next proposition is similar to Proposition B.1. in [3].

Proposition 3.4.1. *Given $\mathcal{X}$ and $\sigma(\cdot)$, define $\xi(t)$*

$$\begin{aligned}\xi(t) = \sup_{0\leq t^* \leq t} \max \Bigg\{ 1, \frac{1}{\sqrt{m}} \Bigg\{ & \left\|\boldsymbol{W}_{t^*}^{[2]}\right\|_{2\to 2}, \left\|\left(\boldsymbol{W}_{t^*}^{[2]}\right)^{\mathsf{T}}\right\|_{2\to 2}, \dots \\ & \dots, \left\|\boldsymbol{W}_{t^*}^{[L]}\right\|_{2\to 2}, \left\|\left(\boldsymbol{W}_{t^*}^{[L]}\right)^{\mathsf{T}}\right\|_{2\to 2}, \|\boldsymbol{a}_{t^*}\|_2 \Bigg\} \Bigg\},\end{aligned} \tag{3.29}$$

then with high probability w.r.t the random initialization, for $t \lesssim \sqrt{m}$

$$\xi(t) \leq c_{w,t}, \tag{3.30}$$

where the constant $c_{w,t} > 2$ *is independent of the depth* L*. Moreover for* $t \lesssim \sqrt{m}$*,* $c_{w,t}$ *has a uniform upper bound in* t*, i.e.,*

$$c_{w,t} \leq \bar{c}, \tag{3.31}$$

where $\bar{c}$ *is independent of the depth* L *and time* t*.*

Proof. (i). Firstly, set $Z = \|\boldsymbol{a}\|_2^2$ in Lemma 3.4.3. Then if we write $2tm = m + (2t-1)m$, with $x = \frac{mt}{10}$,

$$\mathbb{P}\left(\|\boldsymbol{a}_0\|_2^2 \geq m + 2m\left(\sqrt{t/10} + t/10\right)\right) \leq \exp(-tm/10)),$$

and for $t \geq 1$, we have $2t - 1 \geq 2\left(\sqrt{t/10} + t/10\right)$. Thus, if we choose t properly, we see that such event

$$\frac{1}{\sqrt{m}} \|\boldsymbol{a}_0\|_2 \leq c_{w,0}$$

holds with high probability, where $c_{w,0}$ is the constant in Lemma 3.4.2. Hence, combined with Lemma 3.4.2 for $t = 0$, $\xi(0) \leq \max\left\{1, c'_{w,0}\right\} = c_{w,0}$.

(ii). Secondly, we derive the upper bound for $\partial_t \xi(t)$. In order for that, the L^2 bound on each output layer shall be estimated. For $l = 1$,

$$\begin{aligned} \left\|\boldsymbol{x}^{[1]}\right\|_2 &= \sqrt{\frac{c_\sigma}{m}} \left\|\sigma(\boldsymbol{W}_t^{[1]}\boldsymbol{x})\right\|_2 \leq \sqrt{c_\sigma}\left(|\sigma(0)| + \frac{1}{\sqrt{m}}\left\|\boldsymbol{W}_t^{[1]}\boldsymbol{x}\right\|_2\right) \\ &\leq \sqrt{c_\sigma}\left(1 + \xi(t)\|\boldsymbol{x}\|_2\right) \leq C\xi(t), \end{aligned} \tag{3.32}$$

and for $2 \leq l \leq L$,

$$\begin{aligned} \left\|\boldsymbol{x}^{[l]}\right\|_2 &\leq \left\|\boldsymbol{x}^{[l-1]}\right\|_2 + \frac{c_{\text{res}}}{L\sqrt{m}}\left\|\sigma\left(\boldsymbol{W}_t^{[l]}\boldsymbol{x}^{[l-1]}\right)\right\|_2 \\ &\leq \left\|\boldsymbol{x}^{[l-1]}\right\|_2 + \frac{c_{\text{res}}}{L}\left(|\sigma(0)| + \xi(t)\left\|\boldsymbol{x}^{[l-1]}\right\|_2\right) \\ &\leq \left\|\boldsymbol{x}^{[l-1]}\right\|_2 + \frac{c_{\text{res}}}{L}\left(1 + \xi(t)\left\|\boldsymbol{x}^{[l-1]}\right\|_2\right) \\ &\leq \left(1 + \frac{2c_{\text{res}}}{L}\xi(t)\right)\left\|\boldsymbol{x}^{[l-1]}\right\|_2. \end{aligned} \tag{3.33}$$

Hence an inductive relation on the 2-norm of $\boldsymbol{x}^{[l]}$ can be obtained

$$\left\|\boldsymbol{x}^{[l]}\right\|_2 \leq C\left(1+\frac{2c_{\text{res}}}{L}\xi(t)\right)^{l-1}\xi(t). \tag{3.34}$$

Based on Equation (2.6), combined with Lemma 3.4.1

$$\begin{aligned}
\partial_t\left\|\boldsymbol{W}_t^{[l]}\right\|_{2\to 2} &\leq \frac{1}{n}\sum_{\beta=1}^{n}\frac{C}{\sqrt{m}}\left\|\boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[(l+1):L]}\right)^{\intercal}\boldsymbol{a}_t\right\|_2\left\|\boldsymbol{x}_\beta^{[l-1]}\right\|_2|f_\beta(t)-y_\beta| \\
&\leq \frac{1}{n}\sum_{\beta=1}^{n}C\left(1+\frac{c_{\text{res}}}{L}\xi(t)\right)^{L-l}\xi(t)\left(1+\frac{2c_{\text{res}}}{L}\xi(t)\right)^{l-1}\xi(t)\,|f_\beta(t)-y_\beta| \\
&\leq C\left(1+\frac{2c_{\text{res}}}{L}\xi(t)\right)^{L-1}\xi(t)^2\sqrt{\frac{1}{n}\sum_{\beta=1}^{n}\|f_\beta(t)-y_\beta\|_2^2} \\
&\leq C\left(1+\frac{2c_{\text{res}}}{L}\xi(t)\right)^{L-1}\xi(t)^2\sqrt{R_S\left(\boldsymbol{\theta}_0\right)} \\
&\leq C\left(1+\frac{2c_{\text{res}}}{L}\xi(t)\right)^{L-1}\xi(t)^2 \leq C\exp\left(2c_{\text{res}}\xi(t)\right)\ \xi(t)^2, \\
\partial_t\left\|\boldsymbol{a}_t\right\|_2 &\leq \frac{1}{n}\sum_{\beta=1}^{n}\left\|\boldsymbol{x}_\beta^{[L]}\right\|_2|f_\beta(t)-y_\beta| \\
&\leq C\left(1+\frac{2c_{\text{res}}}{L}\xi(t)\right)^{L-1}\xi(t)\sqrt{\frac{1}{n}\sum_{\beta=1}^{n}\|f_\beta(t)-y_\beta\|_2^2} \\
&\leq C\left(1+\frac{2c_{\text{res}}}{L}\xi(t)\right)^{L-1}\xi(t)\sqrt{R_S\left(\boldsymbol{\theta}_0\right)} \\
&\leq C\left(1+\frac{2c_{\text{res}}}{L}\xi(t)\right)^{L-1}\xi(t) \leq C\exp(2c_{\text{res}}\xi(t))\ \xi(t).
\end{aligned} \tag{3.35}$$

Consequently, we have

$$\sqrt{m}\ \partial_t\xi(t) \leq C\exp(2c_{\text{res}}\xi(t))\xi^2(t),$$

then an integration inequality can be obtained,

$$\int_{\xi(0)}^{\xi(t)}\frac{\mathrm{d}u}{\exp(2c_{\text{res}}u)u^2} \leq \frac{Ct}{\sqrt{m}}. \tag{3.36}$$

Hence the integration term on the LHS of Equation (3.36) is

$$\begin{aligned}\int_{\xi(0)}^{\xi(t)} \frac{\mathrm{d}u}{\exp(2c_{\mathrm{res}}u)u^2} &\geq \frac{1}{\exp(2c_{\mathrm{res}}\xi(t))}\int_{\xi(0)}^{\xi(t)} \frac{\mathrm{d}u}{u^2} \\ &= \frac{1}{\exp((2c_{\mathrm{res}}\xi(t))}\left(\frac{1}{\xi(0)} - \frac{1}{\xi(t)}\right) \\ &\geq \frac{1}{\exp(2c_{\mathrm{res}}\xi(t))}\left(\frac{1}{c_{w,0}} - \frac{1}{\xi(t)}\right).\end{aligned}$$

We shall notice that, for the single variable function $f(z)$

$$f(z) = \frac{1}{\exp(2c_{\mathrm{res}}z)}(\frac{1}{c_{w,0}} - \frac{1}{z}),$$

the maximum of $f(z)$ can be achieved at point

$$z_0 = \frac{c_{w,0} + \sqrt{c_{w,0}^2 + 2c_{w,0}/c_{\mathrm{res}}}}{2},$$

and $f(z)$ is monotone increasing in the interval $[c_{w,0}, z_0]$. Thus, if we choose time t properly, say $t \leq c\sqrt{m}$, c small enough, the following holds

$$\xi(t) \leq \frac{c_{w,0} + \sqrt{c_{w,0}^2 + 2c_{w,0}/c_{\mathrm{res}}}}{2}.$$

In other words, if $t \leq c\sqrt{m}$ for some small enough $c > 0$,

$$\xi(t) \leq c_{w,t} \leq \frac{c_{w,0} + \sqrt{c_{w,0}^2 + 2c_{w,0}/c_{\mathrm{res}}}}{2},$$

where the last quantity is independent of depth L and time t, and we denote this by

$$\bar{c} = \frac{c_{w,0} + \sqrt{c_{w,0}^2 + 2c_{w,0}/c_{\mathrm{res}}}}{2}, \tag{3.37}$$

which finishes the proof of Proposition 3.4.1. □

We state the inductive relation (3.34) as a proposition.

Proposition 3.4.2. *Given $\mathcal{X}$ and $\sigma(\cdot)$, then with high probability w.r.t the random initialization, for each l, given time $t \lesssim \sqrt{m}$,*

$$\left\|\boldsymbol{x}^{[l]}\right\|_2 \leq C, \tag{3.38}$$

where $C > 0$ is a constant independent of the depth L.

We remark that the constant C in Proposition 3.4.2 only depends on c_{res} and c_σ. However, for fully-connected networks, Equation (3.38) in Proposition 3.4.2 become

$$\left\|\boldsymbol{x}^{[l]}\right\|_2 \leq C\ 2^l. \tag{3.39}$$

Hence the Euclidean norm of each output layer increases exponentially layer by layer for fully-connected networks, revealing that ResNet possesses more stability.

Finally, we make Apriori estimates on the Euclidean norm of arbitrary vector $\boldsymbol{u}(t) \in \mathbb{A}_0$.

Proposition 3.4.3. *Given $\mathcal{X}$ and $\sigma(\cdot)$, with high probability w.r.t the random initialization, uniformly for any vector $\boldsymbol{u}(t) \in \mathbb{A}_0$, given time $t \lesssim \sqrt{m}$,*

$$\|\boldsymbol{u}(t)\|_2 \leq c\sqrt{m}, \tag{3.40}$$

where $c > 0$ is a constant independent of the depth L and time t.

Proof. We shall start our analysis on the whole expressions in set $\mathbb{A}_0$. For any vector $\boldsymbol{u}(t) \in \mathbb{A}_0$, it can be written into

$$\boldsymbol{u}(t) = \mathrm{e}_s \mathrm{e}_{s-1} \ldots \mathrm{e}_1 \mathrm{e}_0, \qquad 0 \leq s \leq 4L.$$

We start with the estimate on e_0, since e_0 is chosen from

$$\mathrm{e}_0 \in \left\{\boldsymbol{a}_t, \{\sqrt{m}\boldsymbol{x}_\beta^{[0]}, \sqrt{m}\boldsymbol{x}_\beta^{[1]}, \sqrt{m}\boldsymbol{x}_\beta^{[2]}, \ldots, \sqrt{m}\boldsymbol{x}_\beta^{[L]}\}_{1\leq\beta\leq n}\right\}.$$

- (a). If $\mathrm{e}_0 = \boldsymbol{a}_t$, then by Lemma 3.4.3, for $t \lesssim \sqrt{m}$,

$$\|\boldsymbol{a}_t\|_2 \leq c_{w,t}\sqrt{m} \leq c\sqrt{m}.$$

- (b). If $\mathrm{e}_0 = \sqrt{m}\boldsymbol{x}_\beta^{[l]}$, then based on Proposition 3.4.2, for $t \lesssim \sqrt{m}$,

$$\left\|\sqrt{m}\boldsymbol{x}_\beta^{[l]}\right\|_2 = \sqrt{m}\left\|\boldsymbol{x}_\beta^{[l]}\right\|_2 \leq c\sqrt{m}.$$

Now we proceed to $\mathrm{e_j}$, $\mathrm{j} \geq 1$.

- (i). If $\mathrm{e_j} = \boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\beta)$, then we have

$$\begin{aligned}\|\boldsymbol{u}(t)\|_2 &= \|\mathrm{e}_s\mathrm{e}_{s-1}\ldots\mathrm{e}_1\mathrm{e}_0\|_2 \\ &= \|\mathrm{e}_s\|_{2\to2}\|\mathrm{e}_{s-1}\|_{2\to2}\ldots\|\mathrm{e}_1\|_{2\to2}\|\mathrm{e}_0\|_2.\end{aligned}$$

  Since $\left\|\boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\beta)\right\|_{2\to2} \leq 1$, thus for all $\mathrm{j} \geq 1$ with $\mathrm{e_j} = \boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\beta)$

$$\|\boldsymbol{u}(t)\|_2 \leq c\sqrt{m}.$$

- (ii). If $\mathrm{e_j} = \boldsymbol{E}_{t,\beta}^{[l]}$ or $\mathrm{e_j} = \left(\boldsymbol{E}_{t,\beta}^{[l]}\right)^\intercal$, then based on Proposition 3.4.1

$$\|\boldsymbol{u}(t)\|_2 = \|\mathrm{e}_s\|_{2\to2}\|\mathrm{e}_{s-1}\|_{2\to2}\ldots\|\mathrm{e}_1\|_{2\to2}\|\mathrm{e}_0\|_2.$$

  Since

$$\left\|\boldsymbol{E}_{t,\beta}^{[l]}\right\|_{2\to2} = \left\|\left(\boldsymbol{E}_{t,\beta}^{[l]}\right)^\intercal\right\|_{2\to2} \leq \left(1+\frac{c_{\mathrm{res}}}{L}\xi(t)\right) \leq \left(1+\frac{c_{\mathrm{res}}c_{w,t}}{L}\right),$$

  thus for all $\mathrm{j} \geq 1$ with $\mathrm{e_j} = \boldsymbol{E}_{t,\beta}^{[l]}$ or $\mathrm{e_j} = \left(\boldsymbol{E}_{t,\beta}^{[l]}\right)^\intercal$,

$$\|\boldsymbol{u}(t)\|_2 \leq \left(1+\frac{c_{\mathrm{res}}c_{w,t}}{L}\right)^s\|\mathrm{e}_0\|_2, \tag{3.41}$$

then by taking supreme on $0 \le s \le 4L$, we have

$$\begin{aligned}\|\boldsymbol{u}(t)\|_2 &\le \left(1 + \frac{c_{\text{res}} c_{w,t}}{L}\right)^{4L} \|\mathrm{e}_0\|_2 \\ &\le c \exp(4 c_{\text{res}} c_{w,t}) \sqrt{m} \le c\sqrt{m}.\end{aligned}$$

Combining these two observations, we finish the proof. □

To sum up, we define

$$\xi_{\infty,0}(t) = \sup_{0 \le t^* \le t} \left\{ \|\boldsymbol{u}(t^*)\|_2 : \boldsymbol{u}(t^*) \in \mathbb{A}_0 \right\}. \tag{3.42}$$

Then directly from Proposition 3.4.3, for time $0 \le t \le \sqrt{m}/(\ln m)^{C^*}$,

$$\xi_{\infty,0}(t) \le c\sqrt{m}. \tag{3.43}$$

3.4.2 Apriori L^∞ Bounds for Expressions in $\mathbb{A}_0$

In this part, we shall make estimate on such quantity

$$\eta_{\infty,0}(t) = \sup_{0 \le t^* \le t} \left\{ \|\boldsymbol{u}(t^*)\|_\infty : \boldsymbol{u}(t^*) \in \mathbb{A}_0 \right\}. \tag{3.44}$$

We begin with a lemma on the $\|\cdot\|_\infty$ norm of a standard Gaussian vector.

Lemma 3.4.4. *For any i.i.d. standard normal distribution $X_1, X_2, \ldots, X_m$, it holds with high probability that L^∞-norm of the Gaussian vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_m)^\intercal$ is upper bounded by*

$$\|\boldsymbol{X}\|_\infty \le (\ln m)^C,$$

for some large constant $C > 0$.

Proof. For any $X_{\mathrm{i}} \sim \mathcal{N}(0,1)$, for some $\varepsilon, \lambda > 0$,

$$\begin{aligned}\mathbb{P}\left(X_{\mathrm{i}} \geq \varepsilon\right) &= \mathbb{P}\left(\exp\left(\lambda X_{\mathrm{i}}\right) \geq \exp\left(\lambda\varepsilon\right)\right) \\ &\leq \frac{\mathbb{E}\left(\lambda X_{\mathrm{i}}\right)}{\exp\left(\lambda\varepsilon\right)} = \frac{\exp\left(\frac{1}{2}\lambda^2\right)}{\exp\left(\lambda\varepsilon\right)} = \exp\left(\frac{1}{2}\lambda^2 - \lambda\varepsilon\right).\end{aligned}$$

We shall optimize over λ,

$$\mathbb{P}\left(X_{\mathrm{i}} \geq \varepsilon\right) \leq \min_{\lambda>0} \exp\left(\frac{1}{2}\lambda^2 - \lambda\varepsilon\right) = \exp\left(-\frac{\varepsilon^2}{2}\right),$$

from symmetry of Gaussian variables

$$\mathbb{P}\left(\left|X_{\mathrm{i}}\right| \geq \varepsilon\right) \leq 2\exp\left(-\frac{\varepsilon^2}{2}\right),$$

hence by taking over m unions

$$\mathbb{P}\left(\left\|\boldsymbol{X}\right\|_{\infty} \geq \varepsilon\right) \leq 2m\exp\left(-\frac{\varepsilon^2}{2}\right).$$

Set $\varepsilon = \left(\ln m\right)^C$, we have

$$\mathbb{P}\left(\left\|\boldsymbol{X}\right\|_{\infty} \leq \left(\ln m\right)^C\right) \geq 1 - 2m\exp\left(-\frac{\left(\ln m\right)^{2C}}{2}\right).$$

We remark that when $C > 0$ is large enough, $\left(\ln m\right)^C \approx m^{\varepsilon}$ for some small $\varepsilon > 0$. □

Our next lemma concerns the matrix two to infinity norm.

Lemma 3.4.5. *Given $\boldsymbol{W} \in \mathbb{R}^{m\times m}$ with entry $W_{\mathrm{i,j}} \sim \mathcal{N}(0,1)$, then with high probability,*

$$\left\|\boldsymbol{W}\right\|_{2\to\infty} = \sup_{\left\|\boldsymbol{x}\right\|_2=1} \left\|\boldsymbol{W}\boldsymbol{x}\right\|_{\infty} \leq \left(\ln m\right)^C, \tag{3.45}$$

for some large constant $C > 0$.

Proof. Note that $\boldsymbol{W}\boldsymbol{x}$ shares the same distribution as the Gaussian vector $\boldsymbol{X}$ in Lemma 3.4.4, i.e. $\boldsymbol{W}\boldsymbol{x} \sim \boldsymbol{X}$. Then apply Lemma 3.4.4 directly, we finish the proof. □

To evaluate $\eta_{\infty,0}(t)$, another lemma shall be stated.

Lemma 3.4.6. *Given $\mathcal{X}$ and $\sigma(\cdot)$, for vectors in $\mathbb{A}_0$, define*

$$\eta_{q,0}(t) := \sup_{0\le t^*\le t} \left\{ \|\boldsymbol{u}_q(t^*)\|_\infty : \boldsymbol{u}_q(t^*) = \mathrm{e}_q \mathrm{e}_{q-1} \dots \mathrm{e}_1 \mathrm{e}_0,\ \boldsymbol{u}_q(t^*) \in \mathbb{A}_0 \right\}. \tag{3.46}$$

Moreover, define

$$\omega(t) := \sup_{0\le t^*\le t} \max \left\{ \left\| \boldsymbol{W}_{t^*}^{[2]} \right\|_{2\to\infty}, \left\| \left(\boldsymbol{W}_{t^*}^{[2]} \right)^\intercal \right\|_{2\to\infty}, \dots, \left\| \boldsymbol{W}_{t^*}^{[L]} \right\|_{2\to\infty}, \left\| \left(\boldsymbol{W}_{t^*}^{[L]} \right)^\intercal \right\|_{2\to\infty} \right\},$$

then with high probability w.r.t the random initialization, for $t \lesssim \sqrt{m}$,

$$\eta_{q,0}(t) \le \eta_{0,0}(t) + c\ \omega(t) \left(1 + \frac{c_{\mathrm{res}} c_{w,t}}{L} \right)^q, \tag{3.47}$$

where constants $c_{w,t} > 2, c > 0$ are independent of the depth L.

Proof. For any vector $\boldsymbol{u}_q(t) \in \mathbb{A}_0$ of length $q,\ 0 \le q \le 4L$, it can be written into

$$\boldsymbol{u}_q(t) = \mathrm{e}_q \mathrm{e}_{q-1} \dots \mathrm{e}_1 \mathrm{e}_0.$$

We shall prove Equation (3.47) by performing induction on q. Firstly, for $q = 0$, Equation (3.47) is trivial. For $q \ge 1$, we shall investigate on the terms $\mathrm{e_j}$ in the expression $\boldsymbol{u}_q(t)$.

- (i). If $\mathrm{e_j} = \boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\beta)$, then

$$\begin{aligned} \|\boldsymbol{u}_q(t)\|_\infty &= \|\mathrm{e}_q \mathrm{e}_{q-1} \dots \mathrm{e}_1 \mathrm{e}_0\|_\infty \\ &= \|\mathrm{e}_q\|_{\infty\to\infty} \|\mathrm{e}_{q-1}\|_{\infty\to\infty} \dots \|\mathrm{e}_1\|_{\infty\to\infty} \|\mathrm{e}_0\|_\infty, \end{aligned}$$

since $\left\| \boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\beta) \right\|_{\infty\to\infty} \le 1$, we have

$$\|\boldsymbol{u}_q(t)\|_\infty \le \|\mathrm{e}_0\|_\infty,$$

then

$$\eta_{q,0}(t) \leq \eta_{0,0}(t).$$

- (ii). If $\mathrm{e_j} = \boldsymbol{E}_{t,\beta}^{[l]}$ or $\mathrm{e_j} = \left(\boldsymbol{E}_{t,\beta}^{[l]}\right)^{\mathsf{T}}$, $\|\mathrm{e_j}\|_{\infty\to\infty} \geq 1$, so we need to tackle it differently.

$$\begin{aligned}
\|\boldsymbol{u}_q(t)\|_\infty &= \left\|\boldsymbol{E}_{t,\beta}^{[l]}\boldsymbol{u}_{q-1}(t)\right\|_\infty \\
&= \left\|\boldsymbol{u}_{q-1}(t) + \frac{c_{\mathrm{res}}}{L}\boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\beta)\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\boldsymbol{u}_{q-1}(t)\right\|_\infty \\
&\leq \|\boldsymbol{u}_{q-1}(t)\|_\infty + \frac{c_{\mathrm{res}}}{L}\left\|\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right\|_{2\to\infty}\|\boldsymbol{u}_{q-1}(t)\|_2\,, \\
\text{or } \|\boldsymbol{u}_q(t)\|_\infty &= \left\|\left(\boldsymbol{E}_{t,\beta}^{[l]}\right)^{\mathsf{T}}\boldsymbol{u}_{q-1}(t)\right\|_\infty \\
&= \left\|\boldsymbol{u}_{q-1}(t) + \frac{c_{\mathrm{res}}}{L}\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{\mathsf{T}}\boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\beta)\boldsymbol{u}_{q-1}(t)\right\|_\infty \\
&\leq \|\boldsymbol{u}_{q-1}(t)\|_\infty + \frac{c_{\mathrm{res}}}{L}\left\|\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{\mathsf{T}}\right\|_{2\to\infty}\|\boldsymbol{u}_{q-1}(t)\|_2\,,
\end{aligned}$$

recall definition of $\omega(t)$, we have

$$\|\boldsymbol{u}_q(t)\|_\infty \leq \|\boldsymbol{u}_{q-1}(t)\|_\infty + \frac{c_{\mathrm{res}}}{L\sqrt{m}}\omega(t)\,\|\boldsymbol{u}_{q-1}(t)\|_2\,.$$

Based on Proposition 3.4.1

$$\|\boldsymbol{u}_{q-1}(t)\|_2 \leq c\left(1 + \frac{c_{\mathrm{res}}c_{w,t}}{L}\right)^{q-1}\sqrt{m},$$

then

$$\begin{aligned}
\|\boldsymbol{u}_q(t)\|_\infty &\leq \|\boldsymbol{u}_{q-1}(t)\|_\infty + \frac{c_{\mathrm{res}}}{L\sqrt{m}}\omega(t)\,\|\boldsymbol{u}_{q-1}(t)\|_2 \\
&\leq \|\boldsymbol{u}_{q-1}(t)\|_\infty + \frac{c\;c_{\mathrm{res}}}{L}\omega(t)\left(1 + \frac{c_{\mathrm{res}}c_{w,t}}{L}\right)^{q-1},
\end{aligned}$$

inductively

$$\begin{aligned}\|\boldsymbol{u}_q(t)\|_\infty &\leq \|\boldsymbol{u}_0(t)\|_\infty + \frac{c}{c_{w,t}}\omega(t)\left(1+\frac{c_{\text{res}}c_{w,t}}{L}\right)^q \\ &\leq \|\boldsymbol{u}_0(t)\|_\infty + c\ \omega(t)\left(1+\frac{c_{\text{res}}c_{w,t}}{L}\right)^q,\end{aligned}$$

where properties of geometric sums are used. By taking supreme, we have

$$\eta_{q,0}(t) \leq \eta_{0,0}(t) + c\ \omega(t)\left(1+\frac{c_{\text{res}}c_{w,t}}{L}\right)^q.$$

□

These lemmas above enables us to state a proposition on the quantity $\eta_{\infty,0}(0)$.

Proposition 3.4.4. *Given $\mathcal{X}$ and $\sigma(\cdot)$, then with high probability w.r.t the random initialization,*

$$\eta_{\infty,0}(0) \leq c(\ln m)^C, \tag{3.48}$$

where $c, C > 0$ are constants independent of the depth L.

Proof. As always, any vector $\boldsymbol{u}(t) \in \mathbb{A}_0$ can be written into

$$\boldsymbol{u}(t) = \mathrm{e}_s\mathrm{e}_{s-1}\ldots\mathrm{e}_1\mathrm{e}_0, 0 \leq s \leq 4L.$$

We start with the estimate on $\eta_{0,0}(0)$, since e_0 is chosen from

$$\mathrm{e}_0 \in \left\{\boldsymbol{a}_t, \{\sqrt{m}\boldsymbol{x}_\beta^{[0]}, \sqrt{m}\boldsymbol{x}_\beta^{[1]}, \sqrt{m}\boldsymbol{x}_\beta^{[2]}, \ldots, \sqrt{m}\boldsymbol{x}_\beta^{[L]}\}_{1\leq\beta\leq n}\right\}.$$

- (a). If $\mathrm{e}_0 = \boldsymbol{a}_t$, then by Lemma 3.4.4,

$$\|\boldsymbol{a}_0\|_\infty \leq (\ln m)^C.$$

- (b). If $\mathrm{e}_0 = \sqrt{m}\boldsymbol{x}_\beta^{[l]}$, for $l = 1$,

$$\begin{aligned}\left\|\sqrt{m}\boldsymbol{x}_\beta^{[1]}\right\|_\infty &= \sqrt{c_\sigma}\left\|\sigma\left(\boldsymbol{W}_0^{[1]}\boldsymbol{x}_\beta\right)\right\|_\infty \\ &\le \sqrt{c_\sigma}\left(|\sigma(0)| + \left\|\boldsymbol{W}_0^{[1]}\boldsymbol{x}_\beta\right\|_\infty\right) \\ &\le \sqrt{c_\sigma}\left(1 + \left\|\boldsymbol{W}_0^{[1]}\right\|_{2\to\infty}\|\boldsymbol{x}_\beta\|_2\right) \\ &\le \sqrt{c_\sigma}\left(1 + (\ln m)^C\right) \le c(\ln m)^C.\end{aligned}$$

  Moreover, for $l \ge 1$, from Proposition 3.4.2,

$$\begin{aligned}\left\|\sqrt{m}\boldsymbol{x}_\beta^{[l]}\right\|_\infty &\le \left\|\sqrt{m}\boldsymbol{x}_\beta^{[l-1]}\right\|_\infty + \frac{c_{\mathrm{res}}}{L}\left\|\sigma\left(\boldsymbol{W}_0^{[l]}\boldsymbol{x}_\beta^{[l-1]}\right)\right\|_\infty \\ &\le \left\|\sqrt{m}\boldsymbol{x}_\beta^{[l-1]}\right\|_\infty + \frac{c_{\mathrm{res}}}{L}\left(1 + \left\|\boldsymbol{W}_0^{[l]}\right\|_{2\to\infty}\left\|\boldsymbol{x}_\beta^{[l-1]}\right\|_2\right) \\ &\le \left\|\sqrt{m}\boldsymbol{x}_\beta^{[l-1]}\right\|_\infty + \frac{c_{\mathrm{res}}}{L}\left(1 + C(\ln m)^C\right) \\ &\le \left\|\sqrt{m}\boldsymbol{x}_\beta^{[l-1]}\right\|_\infty + \frac{c}{L}(\ln m)^C,\end{aligned}$$

  inductively,

$$\left\|\sqrt{m}\boldsymbol{x}_\beta^{[l]}\right\|_\infty \le c\left(1 + \frac{l}{L}\right)(\ln m)^C \le c(\ln m)^C, \tag{3.49}$$

  where c is independent of the depth L.

Consequently,

$$\eta_{0,0}(0) \le c(\ln m)^C. \tag{3.50}$$

Directly from Lemma 3.4.6

$$\begin{aligned}\eta_{q,0}(0) &\le \eta_{0,0}(0) + c(\ln m)^C\left(1 + \frac{c_{\mathrm{res}}c_{w,0}}{L}\right)^q \\ &\le c(\ln m)^C + c(\ln m)^C\exp(4c_{\mathrm{res}}c_{w,0}) \le c(\ln m)^C,\end{aligned}$$

by taking supreme on $0 \le q \le 4L$, we finish our proof.

□

Our next proposition is on the estimate of $\eta_{\infty,0}(t)$ for time $0 \le t \le \sqrt{m}/(\ln m)^{C^*}$, it is one of the most important propositions in this thesis.

Proposition 3.4.5. *Given $\mathcal{X}$ and $\sigma(\cdot)$, then with high probability w.r.t the random initialization, for time $0 \le t \le \sqrt{m}/(\ln m)^{C^*}$,*

$$\eta_{\infty,0}(t) \le c(\ln m)^C, \tag{3.51}$$

where $c, C, C^* > 0$ are constants independent of the depth L.

Proof. We start with the estimate on $\eta_{0,0}(t)$, since e_0 is chosen from

$$\mathrm{e}_0 \in \left\{ \boldsymbol{a}_t, \{\sqrt{m}\boldsymbol{x}_\beta^{[0]}, \sqrt{m}\boldsymbol{x}_\beta^{[1]}, \sqrt{m}\boldsymbol{x}_\beta^{[2]}, \ldots, \sqrt{m}\boldsymbol{x}_\beta^{[L]}\}_{1\le\beta\le n} \right\}.$$

We observe that from the replacement rules given in Section 3.3.1

$$\begin{aligned}
\boldsymbol{a}_t &\to \frac{1}{\sqrt{m}}\sqrt{m}\boldsymbol{x}_\beta^{[L]}, \\
\sqrt{m}\boldsymbol{x}_\alpha^{[1]} &\to \frac{c_\sigma}{\sqrt{m}}\mathrm{diag}\left(\boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\alpha)\boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[2:L]}\right)^\mathsf{T}\boldsymbol{a}_t\right)\mathbf{1}\left\langle \boldsymbol{x}_\alpha, \boldsymbol{x}_\beta\right\rangle, \\
&\text{for } 2 \le l \le L, \\
\sqrt{m}\boldsymbol{x}_\alpha^{[l]} &\to \frac{c_\sigma}{\sqrt{m}}\mathrm{diag}\left(\boldsymbol{E}_{t,\alpha}^{[2:l]}\boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\alpha)\boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[2:L]}\right)^\mathsf{T}\boldsymbol{a}_t\right)\mathbf{1}\left\langle \boldsymbol{x}_\alpha, \boldsymbol{x}_\beta\right\rangle \\
&\quad + \sum_{k=2}^{l}\frac{c_{\mathrm{res}}^2}{L^2\sqrt{m}}\mathrm{diag}\Bigg(\boldsymbol{E}_{t,\alpha}^{[(k+1):l]} \\
&\quad \boldsymbol{\sigma}_{[k]}^{(1)}(\boldsymbol{x}_\alpha)\boldsymbol{\sigma}_{[k]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[(k+1):L]}\right)^\mathsf{T}\boldsymbol{a}_t\Bigg)\mathbf{1}\left\langle \boldsymbol{x}_\alpha^{[k-1]}, \boldsymbol{x}_\beta^{[k-1]}\right\rangle,
\end{aligned}$$

then for $0 \le t \le \sqrt{m}/(\ln m)^{C^*}$, from Proposition 3.4.2

$$\begin{aligned}
\partial_t \left\|\boldsymbol{a}_t\right\|_\infty &\le \frac{C}{\sqrt{m}}\left\|\sqrt{m}\boldsymbol{x}_\beta^{[L]}\right\|_\infty, \\
\partial_t \left\|\sqrt{m}\boldsymbol{x}_\alpha^{[l]}\right\|_\infty &\le \sum_{k=1}^{l}\frac{C}{\sqrt{m}}\left\|\boldsymbol{E}_{t,\alpha}^{[(k+1):l]}\boldsymbol{\sigma}_{[k]}^{(1)}(\boldsymbol{x}_\alpha)\boldsymbol{\sigma}_{[k]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[(k+1):L]}\right)^\mathsf{T}\boldsymbol{a}_t\right\|_\infty,
\end{aligned}$$

by taking supreme on time $0 \le t \le \sqrt{m}/(\ln m)^{C^*}$,

$$\eta_{0,0}(t) \le c(\ln m)^C + \frac{C}{\sqrt{m}} \int_0^t \eta_{\infty,0}(s)\,\mathrm{d}s. \tag{3.52}$$

Secondly, for the estimate of $\omega(t)$, from the replacement rules,

$$\begin{aligned}
\boldsymbol{W}_t^{[l]} &\to \frac{c_{\mathrm{res}}}{L\sqrt{m}} \boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\beta) \left(\boldsymbol{E}_{t,\beta}^{[(l+1):L]}\right)^\intercal \boldsymbol{a}_t \otimes (\boldsymbol{x}_\beta^{[l-1]})^\intercal, \\
\left(\boldsymbol{W}_t^{[l]}\right)^\intercal &\to \frac{c_{\mathrm{res}}}{L\sqrt{m}} \boldsymbol{x}_\beta^{[l-1]} \otimes \left(\boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\beta) \left(\boldsymbol{E}_{t,\beta}^{[(l+1):L]}\right)^\intercal \boldsymbol{a}_t\right)^\intercal,
\end{aligned}$$

then from Proposition 3.4.3,

$$\begin{aligned}
\partial_t \left\|\boldsymbol{W}_t^{[l]}\right\|_{2\to\infty} &\le \frac{C}{\sqrt{m}} \left\|\boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\beta) \left(\boldsymbol{E}_{t,\beta}^{[(l+1):L]}\right)^\intercal \boldsymbol{a}_t\right\|_\infty, \\
\partial_t \left\|\left(\boldsymbol{W}_t^{[l]}\right)^\intercal\right\|_{2\to\infty} &\le \frac{C}{\sqrt{m}} \left\|\sqrt{m}\boldsymbol{x}_\beta^{[l-1]}\right\|_\infty,
\end{aligned}$$

hence by taking supreme on time $0 \le t \le \sqrt{m}/(\ln m)^{C^*}$,

$$\omega(t) \le (\ln m)^C + \frac{C}{\sqrt{m}} \int_0^t \eta_{\infty,0}(s)\,\mathrm{d}s. \tag{3.53}$$

Directly from Lemma 3.4.6

$$\begin{aligned}
\eta_{q,0}(t) &\le \eta_{0,0}(t) + c\,\omega(t) \left(1 + \frac{c_{\mathrm{res}} c_{w,t}}{L}\right)^q \\
&\le \left(c(\ln m)^C + \frac{C}{\sqrt{m}} \int_0^t \eta_{\infty,0}(s)\,\mathrm{d}s\right) \left(1 + \left(1 + \frac{c_{\mathrm{res}} c_{w,t}}{L}\right)^q\right).
\end{aligned}$$

Finally, by taking supreme on $0 \le q \le 4L$,

$$\eta_{\infty,0}(t) \le c(\ln m)^C + \frac{C}{\sqrt{m}} \int_0^t \eta_{\infty,0}(s)\,\mathrm{d}s. \tag{3.54}$$

We notice that Equation (3.54) gives us a Gronwall-type inequality, hence

$$\eta_{\infty,0}(t) \le c(\ln m)^C \exp\left(\frac{Ct}{\sqrt{m}}\right). \tag{3.55}$$

To sum up, for $t \leq \sqrt{m}/(\ln m)^{C^*}$,

$$\eta_{\infty,0}(t) \leq c(\ln m)^C. \tag{3.56}$$

□

3.4.3 Apriori L^2 and L^∞ Bounds for Expressions in $\mathbb{A}_r$, $r \geq 1$

In this part, we shall make estimates on the norms $\|\cdot\|_\infty$ and $\|\cdot\|_2$ of vectors belonging to higher order sets $\mathbb{A}_r$, $r \geq 1$. Naturally, several quantities for vectors in $\mathbb{A}_r$ with length q shall be defined

$$\xi_{q,r}(t) := \sup_{0\leq t^*\leq t} \left\{ \|\boldsymbol{u}_q(t^*)\|_2 : \boldsymbol{u}_q(t^*) = \mathrm{e}_q\mathrm{e}_{q-1}\ldots\mathrm{e}_1\mathrm{e}_0,\ \boldsymbol{u}_q(t^*) \in \mathbb{A}_r \right\}, \tag{3.57}$$

Specifically, from Proposition 3.4.1 and Proposition 3.4.3,

$$\xi_{q,0}(t) \leq c\left(1 + \frac{c_{\mathrm{res}}c_{w,t}}{L}\right)^q \sqrt{m}. \tag{3.58}$$

Moreover, we define

$$\xi_{\infty,r}(t) = \sup_{0\leq q\leq 4L} \{\xi_{q,r}(t)\}, \tag{3.59}$$

and we remark that $\xi_{\infty,0}(t)$ is consistent with $\xi_{\infty,r}(t)$ for $r = 0$.

Similarly we define

$$\eta_{q,r}(t) := \sup_{0\leq t^*\leq t} \left\{ \|\boldsymbol{u}_q(t^*)\|_\infty : \boldsymbol{u}_q(t^*) = \mathrm{e}_q\mathrm{e}_{q-1}\ldots\mathrm{e}_1\mathrm{e}_0,\ \boldsymbol{u}_q(t^*) \in \mathbb{A}_r \right\}, \tag{3.60}$$

and

$$\eta_{\infty,r}(t) = \sup_{0\leq q\leq 4L} \{\eta_{q,r}(t)\}. \tag{3.61}$$

Once again, for any vector $\boldsymbol{u}(t) \in \mathbb{A}_r$, it can be written into

$$\boldsymbol{u}(t) = \mathrm{e}_s\mathrm{e}_{s-1}\ldots\mathrm{e}_1\mathrm{e}_0, \qquad 0 \leq s \leq 4L,$$

Since e_0 is chosen from

$$\mathrm{e}_0 \in \left\{\boldsymbol{a}_t, \mathbf{1}, \{\sqrt{m}\boldsymbol{x}_\beta^{[0]}, \sqrt{m}\boldsymbol{x}_\beta^{[1]}, \sqrt{m}\boldsymbol{x}_\beta^{[2]}, \ldots, \sqrt{m}\boldsymbol{x}_\beta^{[L]}\}_{1\le\beta\le n}\right\}.$$

Since e_0 is chosen from

$$\mathrm{e}_0 \in \left\{\boldsymbol{a}_t, \mathbf{1}, \{\sqrt{m}\boldsymbol{x}_\beta^{[0]}, \sqrt{m}\boldsymbol{x}_\beta^{[1]}, \sqrt{m}\boldsymbol{x}_\beta^{[2]}, \ldots, \sqrt{m}\boldsymbol{x}_\beta^{[L]}\}_{1\le\beta\le n}\right\}.$$

and $\|\mathbf{1}\|_\infty = 1, \|\mathbf{1}\|_2 = \sqrt{m}$, then from Proposition 3.4.1 and Proposition 3.4.5, for time $0 \le t \le \sqrt{m}/(\ln m)^{C^*}$,

$$\xi_{0,r}(t) \le c\sqrt{m}, \quad \eta_{0,r}(t) \le c(\ln m)^C.$$

Next we proceed to $\mathrm{e_j}$, $\mathrm{j} \ge 1$. For each $\mathrm{e_j}$,

- (i) $\mathrm{e_j} = \boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\beta)$, $\mathrm{e_j} = \boldsymbol{E}_{t,\beta}^{[l]}$ or $\mathrm{e_j} = \left(\boldsymbol{E}_{t,\beta}^{[l]}\right)^\intercal$, $\quad 2 \le l \le L$.
- (ii) $\mathrm{e_j} = \mathrm{diag}(\boldsymbol{g}), \boldsymbol{g} \in \mathbb{A}_0 \cup \mathbb{A}_1 \cup \cdots \cup \mathbb{A}_{r-1}$.
- (iii)

$$\mathrm{e_j} = \boldsymbol{\sigma}_{[l]}^{(u+1)}(\boldsymbol{x}_\beta)\mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_1}\boldsymbol{g}_1\right)\ldots\mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_u}\boldsymbol{g}_u\right)\left(\frac{c_{\mathrm{res}}}{L}\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_{u+1}},$$

or

$$\mathrm{e_j} = \left(\frac{c_{\mathrm{res}}}{L}\frac{\left(\boldsymbol{W}_t^{[l]}\right)^\intercal}{\sqrt{m}}\right)^{Q_{u+1}}\boldsymbol{\sigma}_{[l]}^{(u+1)}(\boldsymbol{x}_\beta)\mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_1}\boldsymbol{g}_1\right)\ldots\mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_u}\boldsymbol{g}_u\right).$$

We observe that the total number of diag operations in $\boldsymbol{u}(t) \in \mathbb{A}_r$ is at most r, and that is how a vector belonging to different hierarchical sets is characterized. Especially if $\exists$ $\mathrm{e_j}$ belonging to case (iii), there are two scenarios:

- $Q_{u+1} = 0$, then $\mathrm{e_j}$ is just multiplication of several diagonal matrices, a special situation for case (ii).

- $Q_{u+1} = 1$, since diagonal matrices commute, $\mathrm{e_j}$ reads

$$\mathrm{e_j} = \mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_1}\boldsymbol{g}_1\right)\ldots\mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_u}\boldsymbol{g}_u\right)\boldsymbol{\sigma}_{[l]}^{(u+1)}(\boldsymbol{x}_\beta)\frac{c_{\mathrm{res}}}{L}\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}},$$

or

$$\mathrm{e_j} = \left(\frac{c_{\mathrm{res}}}{L}\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{\intercal}\boldsymbol{\sigma}_{[l]}^{(u+1)}(\boldsymbol{x}_\beta)\mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_1}\boldsymbol{g}_1\right)\ldots\mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_u}\boldsymbol{g}_u\right),$$

we shall take advantage of the special structure of $\mathrm{e_j}$. Define a new type of skip-connection matrix

$$\widetilde{\boldsymbol{E}}_{t,\beta}^{[l,r]} := \left(\boldsymbol{I}_m + \frac{c_{\mathrm{res}}}{L}\boldsymbol{\sigma}_{[l]}^{(r)}(\boldsymbol{x}_\beta)\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right) \qquad r \geq 2. \tag{3.62}$$

Then $\mathrm{e_j}$ can be written into

$$\begin{aligned}
\mathrm{e_j} &= \mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_1}\boldsymbol{g}_1\right)\ldots\mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_u}\boldsymbol{g}_u\right)\boldsymbol{\sigma}_{[l]}^{(u+1)}(\boldsymbol{x}_\beta)\frac{c_{\mathrm{res}}}{L}\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}} \\
&= \mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_1}\boldsymbol{g}_1\right)\ldots\mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_u}\boldsymbol{g}_u\right)\widetilde{\boldsymbol{E}}_{t,\beta}^{[l,u+1]} \\
&\quad - \mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_1}\boldsymbol{g}_1\right)\ldots\mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_u}\boldsymbol{g}_u\right),
\end{aligned}$$

or

$$\begin{aligned}
\mathrm{e_j} &= \left(\frac{c_{\mathrm{res}}}{L}\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{\intercal}\boldsymbol{\sigma}_{[l]}^{(u+1)}(\boldsymbol{x}_\beta)\mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_1}\boldsymbol{g}_1\right)\ldots\mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_u}\boldsymbol{g}_u\right) \\
&= \left(\widetilde{\boldsymbol{E}}_{t,\beta}^{[l,u+1]}\right)^{\intercal}\mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_1}\boldsymbol{g}_1\right)\ldots\mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_u}\boldsymbol{g}_u\right) \\
&\quad - \mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_1}\boldsymbol{g}_1\right)\ldots\mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_u}\boldsymbol{g}_u\right).
\end{aligned}$$

To illustrate such relation, if some vector $\bar{\boldsymbol{u}}(t)$ contains e_{j} belonging to case (iii), it can be decomposed into

$$\begin{aligned}\bar{\boldsymbol{u}}(t) &= \mathrm{e}_s\mathrm{e}_{s-1}\cdots\mathrm{e}_{\mathrm{j}+1}\mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_1}\boldsymbol{g}_1\right)\ldots\mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_u}\boldsymbol{g}_u\right)\boldsymbol{\sigma}_{[l]}^{(u+1)}(\boldsymbol{x}_\beta)\frac{c_{\mathrm{res}}}{L}\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\mathrm{e}_{\mathrm{j}-1}\cdots\mathrm{e}_0\\ &= \mathrm{e}_s\mathrm{e}_{s-1}\cdots\mathrm{e}_{\mathrm{j}+1}\mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_1}\boldsymbol{g}_1\right)\ldots\mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_u}\boldsymbol{g}_u\right)\widetilde{\boldsymbol{E}}_{t,\beta}^{[l,u+1]}\mathrm{e}_{\mathrm{j}-1}\cdots\mathrm{e}_0\\ &\quad-\mathrm{e}_s\mathrm{e}_{s-1}\cdots\mathrm{e}_{\mathrm{j}+1}\mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_1}\boldsymbol{g}_1\right)\ldots\mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_u}\boldsymbol{g}_u\right)\mathrm{e}_{\mathrm{j}-1}\cdots\mathrm{e}_0.\end{aligned}$$

From the analysis above, we are able to characterize an element in set $\mathbb{A}_r$. If $\boldsymbol{u}(t)\in\mathbb{A}_r$, then there exists $\mathrm{e}_{\mathrm{j}_1},\mathrm{e}_{\mathrm{j}_2},\cdots,\mathrm{e}_{\mathrm{j}_k}$, such that

$$\begin{aligned}\mathrm{e}_{\mathrm{j}_1} &= \mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l_1]}}{\sqrt{m}}\right)^{Q_1}\boldsymbol{g}_1\right),\ \boldsymbol{g}_1\in\mathbb{A}_{r_1-1},\\ \mathrm{e}_{\mathrm{j}_2} &= \mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l_2]}}{\sqrt{m}}\right)^{Q_2}\boldsymbol{g}_2\right),\ \boldsymbol{g}_2\in\mathbb{A}_{r_2-1},\\ &\vdots\\ \mathrm{e}_{\mathrm{j}_k} &= \mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l_k]}}{\sqrt{m}}\right)^{Q_k}\boldsymbol{g}_k\right),\ \boldsymbol{g}_k\in\mathbb{A}_{r_k-1},\end{aligned}$$

with

$$r_1+r_2+\cdots+r_k=r,\qquad r_1,r_2,\cdots,r_k\in\mathbb{N}^+, \tag{3.63}$$

then Equation (3.63) serves as the counting of the number of diag operations contained in $\boldsymbol{u}(t)$, while for other e_{j} ($\mathrm{j}\notin\{\mathrm{j}_1,\mathrm{j}_2,\cdots,\mathrm{j}_k,0\}$), chosen from

$$\left\{\boldsymbol{E}_{t,\beta}^{[l]},\left(\boldsymbol{E}_{t,\beta}^{[l]}\right)^{\intercal}:2\le l\le L\right\}_{1\le\beta\le n}, \tag{3.64}$$

$$\left\{\boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\beta):1\le l\le L\right\}_{1\le\beta\le n}, \tag{3.65}$$

$$\left\{\widetilde{\boldsymbol{E}}_{t,\beta}^{[l,p]},\left(\widetilde{\boldsymbol{E}}_{t,\beta}^{[l,p]}\right)^{\intercal}:2\le l\le L,\ p\ge 2\right\}_{1\le\beta\le n}, \tag{3.66}$$

note that the elements in set (3.64) and set (3.66) share the same matrix properties, thanks to the assumptions concerning the activation function.

Hence, in order to make estimates on $\xi_{q,r}(t)$ and $\eta_{q,r}(t)$, we shall perform induction on the number of diag operations, which is shown in the proof of Proposition 3.4.6. We remark that Proposition 3.4.6 is also one of the most important proposition in this thesis.

Proposition 3.4.6. *Given $\mathcal{X}$ and $\sigma(\cdot)$, then with high probability w.r.t the random initialization, for some finite $r \geq 1$ and time $0 \leq t \leq \sqrt{m}/(\ln m)^{C^*}$,*

$$\begin{aligned} \xi_{\infty,r}(t) &\leq c(\ln m)^C \sqrt{m}, \\ \eta_{\infty,r}(t) &\leq c(\ln m)^C, \end{aligned} \tag{3.67}$$

where $c, C, C^* > 0$ are constants independent of the depth L.

Proof. We recall the definition of $\omega(t)$, $\eta_{\infty,0}(t)$ and $\xi_{\infty,0}(t)$, then with high probability, for time $0 \leq t \leq \sqrt{m}/(\ln m)^{C^*}$,

$$\begin{aligned} \omega(t) &\leq c(\ln m)^C, \\ \eta_{\infty,0}(t) &\leq c(\ln m)^C, \\ \xi_{\infty,0}(t) &\leq c\sqrt{m}. \end{aligned}$$

(i). Let us start out induction with $r = 1$. For $\boldsymbol{u}(t) \in \mathbb{A}_1$, since there exists only one solution to Equation (3.63), then there is one and it is the only index i, such that $\mathrm{e_i} = \mathrm{diag}\,(\boldsymbol{g})$, or $\mathrm{e_i} = \mathrm{diag}\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\boldsymbol{g}\right)$, with $\boldsymbol{g} \in \mathbb{A}_0$. Consequently,

$$\begin{aligned} \xi_{\mathrm{i},1}(t) &\leq \sup_{\boldsymbol{g}\in\mathbb{A}_0} \|\mathrm{diag}\,(\boldsymbol{g})\|_{2\to 2}\, \xi_{\mathrm{i}-1,0}(t) \\ &\leq \sup_{\boldsymbol{g}\in\mathbb{A}_0} \|\boldsymbol{g}\|_\infty\, \xi_{\mathrm{i}-1,0}(t) \leq \eta_{\infty,0}(t)\xi_{\mathrm{i}-1,0}(t) \leq c(\ln m)^C \xi_{\mathrm{i}-1,0}(t), \\ \text{or}\quad \xi_{\mathrm{i},1}(t) &\leq \sup_{\boldsymbol{g}\in\mathbb{A}_0} \left\|\mathrm{diag}\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\boldsymbol{g}\right)\right\|_{2\to 2} \xi_{\mathrm{i}-1,0}(t) \\ &\leq \sup_{\boldsymbol{g}\in\mathbb{A}_0} \left\|\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\boldsymbol{g}\right\|_\infty \xi_{\mathrm{i}-1,0}(t) \leq \frac{\omega(t)}{\sqrt{m}}\xi_{\infty,0}(t)\xi_{\mathrm{i}-1,0}(t) \leq c(\ln m)^C \xi_{\mathrm{i}-1,0}(t), \end{aligned}$$

and for $q > \mathrm{i}$,

$$\xi_{q,1}(t) \leq \left(1 + \frac{c_{\mathrm{res}} c_{w,t}}{L}\right) \xi_{q-1,1}(t),$$

then inductively,

$$\xi_{q,1}(t) \leq \left(1 + \frac{c_{\mathrm{res}} c_{w,t}}{L}\right)^{q-\mathrm{i}} \xi_{\mathrm{i},1}(t).$$

By taking supreme on q and i,

$$\xi_{\infty,1}(t) \leq \exp(4c_{\mathrm{res}} c_{w,t}) c(\ln m)^C \xi_{\infty,0}(t) \leq c(\ln m)^C \sqrt{m}. \tag{3.68}$$

(ii). For $\eta_{\mathrm{i},1}(t)$, we have

$$\begin{aligned}
\eta_{\mathrm{i},1}(t) &\leq \sup_{\boldsymbol{g} \in \mathbb{A}_0} \|\mathrm{diag}\,(\boldsymbol{g})\|_{\infty\to\infty}\, \eta_{\mathrm{i}-1,0}(t) \\
&\leq \sup_{\boldsymbol{g} \in \mathbb{A}_0} \|\boldsymbol{g}\|_\infty\, \eta_{\mathrm{i}-1,0}(t) \leq \eta_{\infty,0}(t)\eta_{\mathrm{i}-1,0}(t) \leq c(\ln m)^C \eta_{\mathrm{i}-1,0}(t), \\
\text{or}\quad \eta_{\mathrm{i},1}(t) &\leq \sup_{\boldsymbol{g} \in \mathbb{A}_0} \left\| \mathrm{diag}\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}} \boldsymbol{g}\right) \right\|_{\infty\to\infty} \eta_{\mathrm{i}-1,0}(t) \\
&\leq \sup_{\boldsymbol{g} \in \mathbb{A}_0} \left\| \frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}} \boldsymbol{g} \right\|_\infty \eta_{\mathrm{i}-1,0}(t) \leq \frac{\omega(t)}{\sqrt{m}} \xi_{\infty,0}(t) \eta_{\mathrm{i}-1,0}(t) \leq c(\ln m)^C \eta_{\mathrm{i}-1,0}(t),
\end{aligned}$$

and for $q > \mathrm{i}$, inductively

$$\begin{aligned}
\eta_{q,1}(t) &\leq \eta_{q-1,1}(t) + \frac{c_{\mathrm{res}}}{L\sqrt{m}} \omega(t) \xi_{q-1,1}(t) \\
&\leq \eta_{\mathrm{i},1}(t) + \frac{c_{\mathrm{res}}}{L\sqrt{m}} \omega(t) \xi_{q-1,1}(t) + \frac{c_{\mathrm{res}}}{L\sqrt{m}} \omega(t) \xi_{q-2,1}(t) + \cdots + \frac{c_{\mathrm{res}}}{L\sqrt{m}} \omega(t) \xi_{\mathrm{i},1}(t),
\end{aligned}$$

by taking supreme on q and i, combined with Equation (3.68),

$$\begin{aligned}
\eta_{\infty,1}(t) &\leq c(\ln m)^C \eta_{\infty,0}(t) + \frac{4c_{\mathrm{res}}}{\sqrt{m}} \omega(t) \xi_{\infty,1}(t) \\
&\leq c(\ln m)^C + c(\ln m)^C \leq c(\ln m)^C.
\end{aligned}$$

(iii). In the following we assume that Equation (3.67) holds for $1, 2, \cdots, r-1$ and we will prove it for r.

If $\boldsymbol{u}(t) \in \mathbb{A}_r$, there exists $\mathrm{e}_{\mathrm{j}_1}, \mathrm{e}_{\mathrm{j}_2}, \cdots, \mathrm{e}_{\mathrm{j}_k}$, such that

$$\begin{aligned}
\mathrm{e}_{\mathrm{j}_1} &= \operatorname{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l_1]}}{\sqrt{m}}\right)^{Q_1} \boldsymbol{g}_1\right), \ \boldsymbol{g}_1 \in \mathbb{A}_{r_1-1}, \\
\mathrm{e}_{\mathrm{j}_2} &= \operatorname{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l_2]}}{\sqrt{m}}\right)^{Q_2} \boldsymbol{g}_2\right), \ \boldsymbol{g}_2 \in \mathbb{A}_{r_2-1}, \\
&\vdots \\
\mathrm{e}_{\mathrm{j}_k} &= \operatorname{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l_k]}}{\sqrt{m}}\right)^{Q_k} \boldsymbol{g}_k\right), \ \boldsymbol{g}_k \in \mathbb{A}_{r_k-1},
\end{aligned}$$

with

$$r_1 + r_2 + \cdots + r_k = r, \quad r_1, r_2, \cdots, r_k \in \mathbb{N}^+.$$

Let i be the largest index among $\mathrm{j}_1, \mathrm{j}_2, \cdots, \mathrm{j}_k$, i.e.

$$\mathrm{i} = \max\{\mathrm{j}_1, \mathrm{j}_2, \cdots, \mathrm{j}_k\},$$

and without loss of generality, let $\mathrm{i} = \mathrm{j}_1$, we have $\mathrm{e}_{\mathrm{i}} = \operatorname{diag}(\boldsymbol{g}_1)$, or $\mathrm{e}_{\mathrm{i}} = \operatorname{diag}\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\boldsymbol{g}_1\right)$ with $\boldsymbol{g}_1 \in \mathbb{A}_{r_1-1}$, then

$$\begin{aligned}
\xi_{\mathrm{i},r}(t) &\le \sup_{\boldsymbol{g}\in\mathbb{A}_{r_1-1}} \|\operatorname{diag}(\boldsymbol{g})\|_{2\to 2}\, \xi_{\mathrm{i}-1,r-r_1}(t) \\
&\le \sup_{\boldsymbol{g}\in\mathbb{A}_{r_1-1}} \|\boldsymbol{g}\|_\infty\, \xi_{\mathrm{i}-1,r-r_1}(t) \le \eta_{\infty,r_1-1}(t)\xi_{\mathrm{i}-1,r-r_1}(t) \le c(\ln m)^C \xi_{\mathrm{i}-1,r-r_1}(t), \\
\text{or} \quad \xi_{\mathrm{i},r}(t) &\le \sup_{\boldsymbol{g}\in\mathbb{A}_{r_1-1}} \left\|\operatorname{diag}\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\boldsymbol{g}\right)\right\|_{2\to 2} \xi_{\mathrm{i}-1,r-r_1}(t) \le \sup_{\boldsymbol{g}\in\mathbb{A}_{r_1-1}} \left\|\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\boldsymbol{g}\right\|_\infty \xi_{\mathrm{i}-1,r-r_1}(t) \\
&\le \frac{\omega(t)}{\sqrt{m}}\xi_{\infty,r_1-1}(t)\xi_{\mathrm{i}-1,r-r_1}(t) \le c(\ln m)^C \xi_{\mathrm{i}-1,r-r_1}(t),
\end{aligned}$$

inductively

$$\xi_{q,r}(t) \le \left(1 + \frac{c_{\text{res}} c_{w,t}}{L}\right)^{q-\mathrm{i}} \xi_{\mathrm{i},r-r_1}(t),$$

by taking supreme on q and i,

$$\xi_{\infty,r}(t) \le \exp(4c_{\text{res}} c_{w,t}) c(\ln m)^C \xi_{\infty,r-r_1}(t) \le c(\ln m)^C \sqrt{m}. \tag{3.69}$$

(iv). For $\eta_{\mathrm{i},r}(t)$,

$$\begin{aligned}
\eta_{\mathrm{i},r}(t) &\le \sup_{\boldsymbol{g}\in\mathbb{A}_{r_1-1}} \left\|\mathrm{diag}\,(\boldsymbol{g})\right\|_{\infty\to\infty} \eta_{\mathrm{i}-1,r-r_1}(t) \\
&\le \sup_{\boldsymbol{g}\in\mathbb{A}_{r_1-1}} \|\boldsymbol{g}\|_\infty \eta_{\mathrm{i}-1,r-r_1}(t) \le \eta_{\infty,r_1-1}(t)\eta_{\mathrm{i}-1,r-r_1}(t) \le c(\ln m)^C \eta_{\mathrm{i}-1,r-r_1}(t), \\
\text{or}\quad \eta_{\mathrm{i},r}(t) &\le \sup_{\boldsymbol{g}\in\mathbb{A}_{r_1-1}} \left\|\mathrm{diag}\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\boldsymbol{g}\right)\right\|_{\infty\to\infty} \eta_{\mathrm{i}-1,r-r_1}(t) \le \sup_{\boldsymbol{g}\in\mathbb{A}_{r_1-1}} \left\|\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\boldsymbol{g}\right\|_\infty \eta_{\mathrm{i}-1,r-r_1}(t) \\
&\le \frac{\omega(t)}{\sqrt{m}} \xi_{\infty,r_1-1}(t)\eta_{\mathrm{i}-1,r-r_1}(t) \le c(\ln m)^C \eta_{\mathrm{i}-1,r-r_1}(t),
\end{aligned}$$

and for $q > \mathrm{i}$,

$$\begin{aligned}
\eta_{q,r}(t) &\le \eta_{q-1,r}(t) + \frac{c_{\text{res}}}{L\sqrt{m}}\omega(t)\xi_{q-1,r}(t) \\
&\le \eta_{\mathrm{i},r}(t) + \frac{c_{\text{res}}}{L\sqrt{m}}\omega(t)\xi_{q-1,r}(t) + \frac{c_{\text{res}}}{L\sqrt{m}}\omega(t)\xi_{q-2,r}(t) + \cdots + \frac{c_{\text{res}}}{L\sqrt{m}}\omega(t)\xi_{\mathrm{i},r}(t),
\end{aligned}$$

by taking supreme on q and i,

$$\begin{aligned}
\eta_{\infty,r}(t) &\le c(\ln m)^C \eta_{\infty,r-r_1}(t) + \frac{4c_{\text{res}}}{\sqrt{m}}\omega(t)\xi_{\infty,r}(t) \\
&\le c(\ln m)^C + c(\ln m)^C \le c(\ln m)^C.
\end{aligned}$$

□

We observe from the above proof that for different r, the constant c grows exponentially in r, while the growth rate of C is linear.

3.5 Proof of Theorem 3.2.1

Since for each term in kernel $\mathcal{G}_t^{(r)}(\boldsymbol{x}_{\alpha_1}, \boldsymbol{x}_{\alpha_2}, \ldots, \boldsymbol{x}_{\alpha_r})$, $r \geq 2$, it takes the form

$$\frac{1}{m^{r/2-1}} \prod_{\mathrm{j}=1}^{s} \frac{\langle \boldsymbol{u}_{2\mathrm{j}-1}(t), \boldsymbol{u}_{2\mathrm{j}}(t) \rangle}{m}, \quad 1 \leq s \leq r, \qquad \boldsymbol{u}_{\mathrm{i}}(t) \in \mathbb{A}_0 \cup \mathbb{A}_1 \cup \cdots \cup \mathbb{A}_{r-2}, \quad 1 \leq \mathrm{i} \leq s.$$

Firstly, from Equation (3.43), for time $0 \leq t \leq \sqrt{m}/(\ln m)^{C^*}$,

$$\left\| \mathcal{G}_t^{(2)}(\cdot) \right\|_\infty \lesssim \left(\frac{\xi_{\infty,0}(t)^2}{m} \right)^2 \lesssim 1,$$

and for $r \geq 3$, from Proposition 3.4.6, for time $0 \leq t \leq \sqrt{m}/(\ln m)^{C^*}$,

$$\left\| \mathcal{G}_t^{(r)}(\cdot) \right\|_\infty \lesssim \frac{1}{m^{r/2-1}} \left(\frac{\xi_{\infty,r-2}(t)^2}{m} \right)^s \lesssim \frac{1}{m^{r/2-1}} \left(\frac{\left(c(\ln m)^C \sqrt{m} \right)^2}{m} \right)^r \lesssim \frac{(\ln m)^{2rC}}{m^{r/2-1}}.$$

3.6 Proof of Theorem 3.2.2

Since there exists a $\frac{1}{L^2}$ scaling in some kernels, we use $C(r, L)$ to denote the 'effective terms' in each kernel. We need to investigate the order of $C(r, L)$ for $r = 2, 3, 4$ respectively.

Firstly, $\mathcal{G}_t^{[L+1]}(\boldsymbol{x}_{\alpha_1}, \boldsymbol{x}_{\alpha_2}) = \left\langle \boldsymbol{x}_{\alpha_1}^{[L]}, \boldsymbol{x}_{\alpha_2}^{[L]} \right\rangle$, then $C(2, L) = \mathcal{O}(1)$, since there is only one term.

Secondly, from the replacement rule, all possible terms generated from $\mathcal{G}_t^{[L+1]}(\cdot)$ are

$$\begin{aligned}
&\mathcal{G}_t^{[L+1]}(\boldsymbol{x}_{\alpha_1}, \boldsymbol{x}_{\alpha_2}) = \left\langle \boldsymbol{x}_{\alpha_1}^{[L]}, \boldsymbol{x}_{\alpha_2}^{[L]} \right\rangle \to \mathcal{G}_t^{[L+1],(1)}(\boldsymbol{x}_{\alpha_1}, \boldsymbol{x}_{\alpha_2}, \boldsymbol{x}_\beta) \\
&\mathcal{G}_t^{[L+1],(1)}(\boldsymbol{x}_{\alpha_1}, \boldsymbol{x}_{\alpha_2}, \boldsymbol{x}_\beta) = \frac{c_\sigma}{m} \underbrace{\left\langle \operatorname{diag}\left(\boldsymbol{E}_{t,\alpha_1}^{[2:L]} \boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_{\alpha_1}) \boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\beta) \left(\boldsymbol{E}_{t,\beta}^{[2:L]} \right)^\intercal \boldsymbol{a}_t \right) \mathbf{1}, \boldsymbol{x}_{\alpha_2}^{[L]} \right\rangle \langle \boldsymbol{x}_{\alpha_1}, \boldsymbol{x}_\beta \rangle}_{\mathrm{I}} \\
&+ \sum_{k=2}^{L} \frac{c_{\mathrm{res}}^2}{L^2 m} \underbrace{\left\langle \operatorname{diag}\left(\boldsymbol{E}_{t,\alpha_1}^{[(k+1):L]} \boldsymbol{\sigma}_{[k]}^{(1)}(\boldsymbol{x}_{\alpha_1}) \boldsymbol{\sigma}_{[k]}^{(1)}(\boldsymbol{x}_\beta) \left(\boldsymbol{E}_{t,\beta}^{[(k+1):L]} \right)^\intercal \boldsymbol{a}_t \right) \mathbf{1}, \boldsymbol{x}_{\alpha_2}^{[L]} \right\rangle \left\langle \boldsymbol{x}_{\alpha_1}^{[k-1]}, \boldsymbol{x}_\beta^{[k-1]} \right\rangle}_{\mathrm{II}} \\
&+ \frac{c_\sigma}{m} \left\langle \operatorname{diag}\left(\boldsymbol{E}_{t,\alpha_2}^{[2:L]} \boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_{\alpha_2}) \boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\beta) \left(\boldsymbol{E}_{t,\beta}^{[2:L]} \right)^\intercal \boldsymbol{a}_t \right) \mathbf{1}, \boldsymbol{x}_{\alpha_1}^{[L]} \right\rangle \langle \boldsymbol{x}_{\alpha_2}, \boldsymbol{x}_\beta \rangle \\
&+ \sum_{k=2}^{L} \frac{c_{\mathrm{res}}^2}{L^2 m} \left\langle \operatorname{diag}\left(\boldsymbol{E}_{t,\alpha_2}^{[(k+1):L]} \boldsymbol{\sigma}_{[k]}^{(1)}(\boldsymbol{x}_{\alpha_2}) \boldsymbol{\sigma}_{[k]}^{(1)}(\boldsymbol{x}_\beta) \left(\boldsymbol{E}_{t,\beta}^{[(k+1):L]} \right)^\intercal \boldsymbol{a}_t \right) \mathbf{1}, \boldsymbol{x}_{\alpha_1}^{[L]} \right\rangle \left\langle \boldsymbol{x}_{\alpha_2}^{[k-1]}, \boldsymbol{x}_\beta^{[k-1]} \right\rangle.
\end{aligned}$$

Thanks to the $\frac{1}{L^2}$ scaling, we obtain that

$$C(3,L) = \mathcal{O}\left(2\left(1+\frac{L-1}{L^2}\right)\right) = \mathcal{O}\left(1+\frac{1}{L}\right).$$

In order to analyze the dynamics for $\mathcal{G}_t^{[L+1],(1)}(\cdot)$, we need the information of its derivative. Hence, we apply replacement rules once again to $\mathcal{G}_t^{[L+1],(1)}(\cdot)$ to obtain $\mathcal{G}_t^{[L+1],(2)}(\cdot)$, i.e.,

$$\mathcal{G}_t^{[L+1],(1)}(\boldsymbol{x}_{\alpha_1},\boldsymbol{x}_{\alpha_2},\boldsymbol{x}_{\alpha_3}) \to \mathcal{G}_t^{[L+1],(2)}(\boldsymbol{x}_{\alpha_1},\boldsymbol{x}_{\alpha_2},\boldsymbol{x}_{\alpha_3},\boldsymbol{x}_{\beta}).$$

Finally for $\mathcal{G}_t^{[L+1],(2)}(\cdot)$, by symmetry, only terms I and II need to be analyzed.

- There are at most $(2L+2)$ symbols in term I to be replaced, and each replacement operation will bring about up to $(L+1)$ many terms.
- For term II, for each summand, there are also at most $(2L+2)$ symbols to be replaced. Since there are $L-1$ summands in II, and each replacement will bring about up to $(L+1)$ many terms, then

$$C(4,L) = \mathcal{O}\left(2\left((2L+2)(L+1)+\frac{1}{L^2}(L-1)(2L+2)(L+1)\right)\right) = \mathcal{O}\left(L^2\right). \quad (3.70)$$

Next, we turn to the proof of Equation (3.5). From Proposition 3.4.6, for time $0 \le t \le \sqrt{m}/(\ln m)^{C^*}$,

$$\begin{aligned}
\left\|\mathcal{G}_t^{[L+1],(2)}(\cdot)\right\|_\infty &\le \frac{C(4,L)}{m}\left(\frac{\xi_{\infty,2}(t)^2}{m}\right)^4 \\
&\le \frac{C(4,L)}{m}\left(\frac{\left(c(\ln m)^C\sqrt{m}\right)^2}{m}\right)^4 \\
&\le C(4,L)\frac{(\ln m)^C}{m},
\end{aligned}$$

then

$$\begin{aligned}\left|\partial_t \mathcal{G}_t^{[L+1],(1)}\left(\boldsymbol{x}_{\alpha_1}, \boldsymbol{x}_{\alpha_2}, \boldsymbol{x}_{\alpha_3}\right)\right| &\leq \sup_{1\leq\beta\leq n}\left|\mathcal{G}_t^{[L+1],(2)}\left(\boldsymbol{x}_{\alpha_1}, \boldsymbol{x}_{\alpha_2}, \boldsymbol{x}_{\alpha_3}, \boldsymbol{x}_{\beta}\right)\right| \sqrt{\frac{\sum_{\beta=1}^{n}\left|f_\beta(t)-y_\beta\right|^2}{n}} \\ &\leq \left\|\mathcal{G}_t^{[L+1],(2)}(\cdot)\right\|_\infty \sqrt{R_S(\boldsymbol{\theta}_0)} \\ &\leq C(4,L)\frac{(\ln m)^C}{m},\end{aligned}$$

moreover, for $1 \leq \alpha_1, \alpha_2, \alpha_3 \leq n$, and time $0 \leq t \leq \sqrt{m}/(\ln m)^{C^*}$,

$$\begin{aligned}\left|\mathcal{G}_t^{[L+1],(1)}\left(\boldsymbol{x}_{\alpha_1}, \boldsymbol{x}_{\alpha_2}, \boldsymbol{x}_{\alpha_3}\right)\right| &\leq \left|\mathcal{G}_0^{[L+1],(1)}\left(\boldsymbol{x}_{\alpha_1}, \boldsymbol{x}_{\alpha_2}, \boldsymbol{x}_{\alpha_3}\right)\right| + t\sup_{0\leq s\leq t}\left|\partial_s \mathcal{G}_s^{[L+1],(1)}\left(\boldsymbol{x}_{\alpha_1}, \boldsymbol{x}_{\alpha_2}, \boldsymbol{x}_{\alpha_3}\right)\right| \\ &\leq \left\|\mathcal{G}_0^{[L+1],(1)}\left(\cdot\right)\right\|_\infty + t\ C(4,L)\frac{(\ln m)^C}{m}.\end{aligned}$$

Finally, estimates shall be made on $\left\|\mathcal{G}_0^{[L+1],(1)}\left(\cdot\right)\right\|_\infty$. We rewrite $\mathcal{G}_t^{[L+1],(1)}\left(\cdot\right)$ into

$$\begin{aligned}\mathcal{G}_t^{[L+1],(1)}(\boldsymbol{x}_{\alpha_1}, \boldsymbol{x}_{\alpha_2}, \boldsymbol{x}_{\beta}) &= \frac{c_\sigma}{m}\left\langle \boldsymbol{E}_{t,\alpha_1}^{[2:L]}\boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_{\alpha_1})\boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_{\beta})\left(\boldsymbol{E}_{t,\beta}^{[2:L]}\right)^\intercal \boldsymbol{a}_t, \boldsymbol{x}_{\alpha_2}^{[L]}\right\rangle\left\langle \boldsymbol{x}_{\alpha_1}, \boldsymbol{x}_{\beta}\right\rangle \\ &+ \sum_{k=2}^{L}\frac{c_{\text{res}}^2}{L^2 m}\left\langle \boldsymbol{E}_{t,\alpha_1}^{[(k+1):L]}\boldsymbol{\sigma}_{[k]}^{(1)}(\boldsymbol{x}_{\alpha_1})\boldsymbol{\sigma}_{[k]}^{(1)}(\boldsymbol{x}_{\beta})\left(\boldsymbol{E}_{t,\beta}^{[(k+1):L]}\right)^\intercal \boldsymbol{a}_t, \boldsymbol{x}_{\alpha_2}^{[L]}\right\rangle\left\langle \boldsymbol{x}_{\alpha_1}^{[k-1]}, \boldsymbol{x}_{\beta}^{[k-1]}\right\rangle \\ &+ \frac{c_\sigma}{m}\left\langle \boldsymbol{E}_{t,\alpha_2}^{[2:L]}\boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_{\alpha_2})\boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_{\beta})\left(\boldsymbol{E}_{t,\beta}^{[2:L]}\right)^\intercal \boldsymbol{a}_t, \boldsymbol{x}_{\alpha_1}^{[L]}\right\rangle\left\langle \boldsymbol{x}_{\alpha_2}, \boldsymbol{x}_{\beta}\right\rangle \\ &+ \sum_{k=2}^{L}\frac{c_{\text{res}}^2}{L^2 m}\left\langle \boldsymbol{E}_{t,\alpha_2}^{[(k+1):L]}\boldsymbol{\sigma}_{[k]}^{(1)}(\boldsymbol{x}_{\alpha_2})\boldsymbol{\sigma}_{[k]}^{(1)}(\boldsymbol{x}_{\beta})\left(\boldsymbol{E}_{t,\beta}^{[(k+1):L]}\right)^\intercal \boldsymbol{a}_t, \boldsymbol{x}_{\alpha_1}^{[L]}\right\rangle\left\langle \boldsymbol{x}_{\alpha_2}^{[k-1]}, \boldsymbol{x}_{\beta}^{[k-1]}\right\rangle.\end{aligned}$$

We observe that each term in $\mathcal{G}_t^{[L+1],(1)}\left(\cdot\right)$ is of the form

$$\frac{c}{m}\left\langle \boldsymbol{B}\boldsymbol{a}_t, \boldsymbol{x}_{\alpha_1}^{[L]}\right\rangle\left\langle \boldsymbol{x}_{\alpha_2}^{[l]}, \boldsymbol{x}_{\beta}^{[l]}\right\rangle, \tag{3.71}$$

where $\boldsymbol{B}$ is some matrix that varies from term to term, c is a constant that also changes term by term. Since

$$\left\langle \boldsymbol{a}_t, \boldsymbol{B}^\intercal \boldsymbol{x}_{\alpha_1}^{[L]}\right\rangle\left\langle \boldsymbol{x}_{\alpha_2}^{[l]}, \boldsymbol{x}_{\beta}^{[l]}\right\rangle = \left\langle \boldsymbol{B}\boldsymbol{a}_t, \boldsymbol{x}_{\alpha_1}^{[L]}\right\rangle\left\langle \boldsymbol{x}_{\alpha_2}^{[l]}, \boldsymbol{x}_{\beta}^{[l]}\right\rangle \tag{3.72}$$

holds, then at $t = 0$, we shall focus on the term

$$\left\langle \boldsymbol{a}_0, \boldsymbol{B}^\intercal \boldsymbol{x}_{\alpha_1}^{[L]} \right\rangle \left\langle \boldsymbol{x}_{\alpha_2}^{[l]}, \boldsymbol{x}_{\beta}^{[l]} \right\rangle. \tag{3.73}$$

Note that $\boldsymbol{a}_0$ is a standard Gaussian vector, from Proposition 3.4.1 and Proposition 3.4.2, with high probability w.r.t random initialization, there exists a uniform constant $C > 0$, such that

$$\|\boldsymbol{B}^\intercal\|_{2\to 2}, \quad \left\|\boldsymbol{x}_{\alpha_1}^{[L]}\right\|_2, \quad \left\|\boldsymbol{x}_{\alpha_2}^{[l]}\right\|_2, \quad \left\|\boldsymbol{x}_{\beta}^{[l]}\right\|_2 \le C,$$

after taking conditional expectation,

$$\left\langle \boldsymbol{a}_0, \boldsymbol{B}^\intercal \boldsymbol{x}_{\alpha_1}^{[L]} \right\rangle \left\langle \boldsymbol{x}_{\alpha_2}^{[l]}, \boldsymbol{x}_{\beta}^{[l]} \right\rangle \sim \mathcal{N}\left(0, \left(\left\langle \boldsymbol{x}_{\alpha_2}^{[l]}, \boldsymbol{x}_{\beta}^{[l]} \right\rangle\right)^2 \left\|\boldsymbol{B}^\intercal \boldsymbol{x}_{\alpha_1}^{[L]}\right\|_2^2\right), \tag{3.74}$$

apply Lemma 3.4.4 directly, with high probability

$$\frac{c}{m} \left\langle \boldsymbol{a}_0, \boldsymbol{B}^\intercal \boldsymbol{x}_{\alpha_1}^{[L]} \right\rangle \left\langle \boldsymbol{x}_{\alpha_2}^{[l]}, \boldsymbol{x}_{\beta}^{[l]} \right\rangle \le c \frac{(\ln m)^C}{m}. \tag{3.75}$$

Consequently,

$$\left\|\mathcal{G}_0^{[L+1],(1)}(\cdot)\right\|_\infty \le C(3, L) \frac{(\ln m)^C}{m}, \tag{3.76}$$

then for $1 \le \alpha_1, \alpha_2, \alpha_3 \le n$, and time $0 \le t \le \sqrt{m}/(\ln m)^{C^*}$,

$$\begin{aligned} \left|\mathcal{G}_t^{[L+1],(1)}(\boldsymbol{x}_{\alpha_1}, \boldsymbol{x}_{\alpha_2}, \boldsymbol{x}_{\alpha_3})\right| &\le \left\|\mathcal{G}_0^{[L+1],(1)}(\cdot)\right\|_\infty + tC(4, L) \frac{(\ln m)^C}{m} \\ &\le C(3, L) \frac{(\ln m)^C}{m} + tC(4, L) \frac{(\ln m)^C}{m}. \end{aligned}$$

Set $\boldsymbol{x}_\beta = \boldsymbol{x}_{\alpha_3}$,

$$\begin{aligned} \left|\partial_t \mathcal{G}_t^{[L+1]}(\boldsymbol{x}_{\alpha_1}, \boldsymbol{x}_{\alpha_2})\right| &\le \left(C(3, L) \frac{(\ln m)^C}{m} + tC(4, L) \frac{(\ln m)^C}{m}\right) \sqrt{\frac{\sum_{\beta=1}^n |f_\beta(t) - y_\beta|^2}{n}} \\ &\le (C(3, L) + tC(4, L)) \frac{(\ln m)^C}{m}, \end{aligned} \tag{3.77}$$

which finishes the proof of Theorem 3.2.2.

3.7 Proof of Theorem 3.2.3

Firstly, based on Proposition B.3.2, for $m = \Omega\left(\left(\frac{n}{\lambda_0}\right)^{2+\varepsilon}\right)$, then with high probability w.r.t random initialization,

$$\lambda_{\min}\left[\mathcal{G}_0^{(2)}\left(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta\right)\right]_{1\leq\alpha,\beta\leq n} \geq \lambda_{\min}\left(\boldsymbol{G}^{[L+1]}(0)\right) \geq \frac{3\lambda_0}{4},$$

set $\gamma_1 = \varepsilon$, we finish the first part of the proof.

We move on to the change of the least eigenvalue of the NTK. Recall Equation (3.77) in the proof of Theorem 3.2.2 (Section 3.6), for time $0 \leq t \leq \sqrt{m}/(\ln m)^{C^*}$,

$$\left|\partial_t \mathcal{G}_t^{[L+1]}\left(\boldsymbol{x}_{\alpha_1}, \boldsymbol{x}_{\alpha_2}\right)\right| \leq \left(C(3,L) + tC(4,L)\right)\frac{(\ln m)^C}{m},$$

consequently,

$$\left|\mathcal{G}_t^{[L+1]}\left(\boldsymbol{x}_{\alpha_1}, \boldsymbol{x}_{\alpha_2}\right) - \mathcal{G}_0^{[L+1]}\left(\boldsymbol{x}_{\alpha_1}, \boldsymbol{x}_{\alpha_2}\right)\right| \leq t\left(C(3,L) + tC(4,L)\right)\frac{(\ln m)^C}{m}.$$

The inequality above can be used to derive an upper bound of the change of the least eigenvalue of the $\mathcal{G}_t^{[L+1]}(\cdot)$.

$$\begin{aligned}\left\|\left(\mathcal{G}_t^{[L+1]} - \mathcal{G}_0^{[L+1]}\right)(\cdot)\right\|_{2\to 2} \leq \left\|\left(\mathcal{G}_t^{[L+1]} - \mathcal{G}_0^{[L+1]}\right)(\cdot)\right\|_{\mathrm{F}} &\leq n\left\|\left(\mathcal{G}_t^{[L+1]} - \mathcal{G}_0^{[L+1]}\right)(\cdot)\right\|_{\infty} \\ &\leq nt\left(C(3,L) + tC(4,L)\right)\frac{(\ln m)^C}{m},\end{aligned}$$

set t^* satisfying

$$nt^*\left(C(3,L) + t^*C(4,L)\right)\frac{(\ln m)^C}{m} = \frac{\lambda_0}{4},$$

rewrite the equation above,

$$C(4,L)(t^*)^2 + C(3,L)t^* = \frac{\lambda_0 m}{4(\ln m)^C n}, \tag{3.78}$$

solving (3.78), we obtain that

$$t^* = \frac{-C(3,L) + \sqrt{\left(C(3,L)\right)^2 + C(4,L)\lambda_0 \frac{m}{(\ln m)^C n}}}{2C(4,L)}, \tag{3.79}$$

since we are in the regime of over-parametrization, for m large enough,

$$t^* \geq \frac{1}{4}\sqrt{\frac{\lambda_0 m}{C(4,L)(\ln m)^C n}}. \tag{3.80}$$

Moreover

$$\begin{aligned}
\lambda_{\min}\left(\left[\mathcal{G}_t^{(2)}\left(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta\right)\right]_{1\leq\alpha,\beta\leq n}\right) &\geq \lambda_{\min}\left(\left[\mathcal{G}_t^{[L+1]}\left(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta\right)\right]_{1\leq\alpha,\beta\leq n}\right) \\
&\geq \lambda_{\min}\left(\left[\mathcal{G}_0^{[L+1]}\left(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta\right)\right]_{1\leq\alpha,\beta\leq n}\right) - \left\|\left(\mathcal{G}_t^{[L+1]} - \mathcal{G}_0^{[L+1]}\right)(\cdot)\right\|_{2\to 2},
\end{aligned}$$

set $\bar{t} := \inf\left\{t : \lambda_{\min}\left[\mathcal{G}_t^{(2)}\left(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta\right)\right]_{1\leq\alpha,\beta\leq n} \geq \lambda_0/2\right\}$, then

$$t^* \leq \bar{t}, \tag{3.81}$$

for any $0 \leq t \leq \bar{t}$,

$$\begin{aligned}
\partial_t \sum_{\alpha=1}^n \|f_\alpha(t) - y_\alpha\|_2^2 &= -\frac{2}{n}\sum_{\alpha,\beta=1}^n \mathcal{G}_t^{(2)}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta)(f_\alpha(t) - y_\alpha)(f_\beta(t) - y_\beta) \\
&\leq -\frac{\lambda_0}{n}\sum_{\gamma=1}^n \|f_\gamma(t) - y_\gamma\|_2^2,
\end{aligned}$$

then

$$R_S(\boldsymbol{\theta}_t) \leq \exp\left(-\frac{\lambda_0 t}{n}\right) R_S(\boldsymbol{\theta}_0). \tag{3.82}$$

Set $R_S(\boldsymbol{\theta}_t) = \varepsilon$, it takes time

$$t \leq \frac{n}{\lambda_0} \ln(\frac{C^*}{\varepsilon})$$

for $R_S(\boldsymbol{\theta}_t)$ to reach accuracy ε, hence if

$$t \leq \frac{n}{\lambda} \ln\left(\frac{C^*}{\varepsilon}\right) \leq t^* \leq \bar{t}, \tag{3.83}$$

then width m is required to yield the lower bound for t^* derived in Equation (3.80),

$$\frac{n}{\lambda_0} \ln\left(\frac{C^*}{\varepsilon}\right) \leq \frac{1}{4}\sqrt{\frac{\lambda_0 m}{C(4,L)(\ln m)^C n}}. \tag{3.84}$$

then for

$$m \geq C(4,L)\left(\frac{n}{\lambda_0}\right)^3 (\ln m)^C \ln\left(\frac{C^*}{\varepsilon}\right)^2,$$

since $C(4,L) = \mathcal{O}\left(L^2\right)$, we conclude that the required width m should be

$$m = \Omega\left(\left(\frac{n}{\lambda_0}\right)^3 L^2 (\ln m)^C \ln\left(\frac{C^*}{\varepsilon}\right)^2\right), \tag{3.85}$$

where ε is the desired training accuracy.

4. SUMMARY AND FUTURE WORK

In this thesis, we described a framework for analyzing the training behavior of ResNet in continuous time gradient descent dynamics. At a high level, our approach is illustrated using Figure 4.1, which builds upon sharp analysis of the least eigenvalue of randomly initialized Gram matrix, and upon uniform estimates for kernels of all orders in the NTH. Finally, we would like to re-emphasize the significance of the $\frac{c_{\text{res}}}{L}$ scaling placed in ResNet, which has been proven to be successful in the stabilization of random propagation (Proposition B.3.2, Equation (B.36)), in the uniform estimate of $\xi(t)$ (Proposition 3.4.1), $\xi_{\infty,0}(t)$ (Proposition 3.4.3, Equation (3.43)), and $\eta_{\infty,r}(t)$ (Proposition 3.4.6) with $r \geq 1$, and most importantly, in lowering the magnitude of $C(4, L)$ (Equation (3.70)).

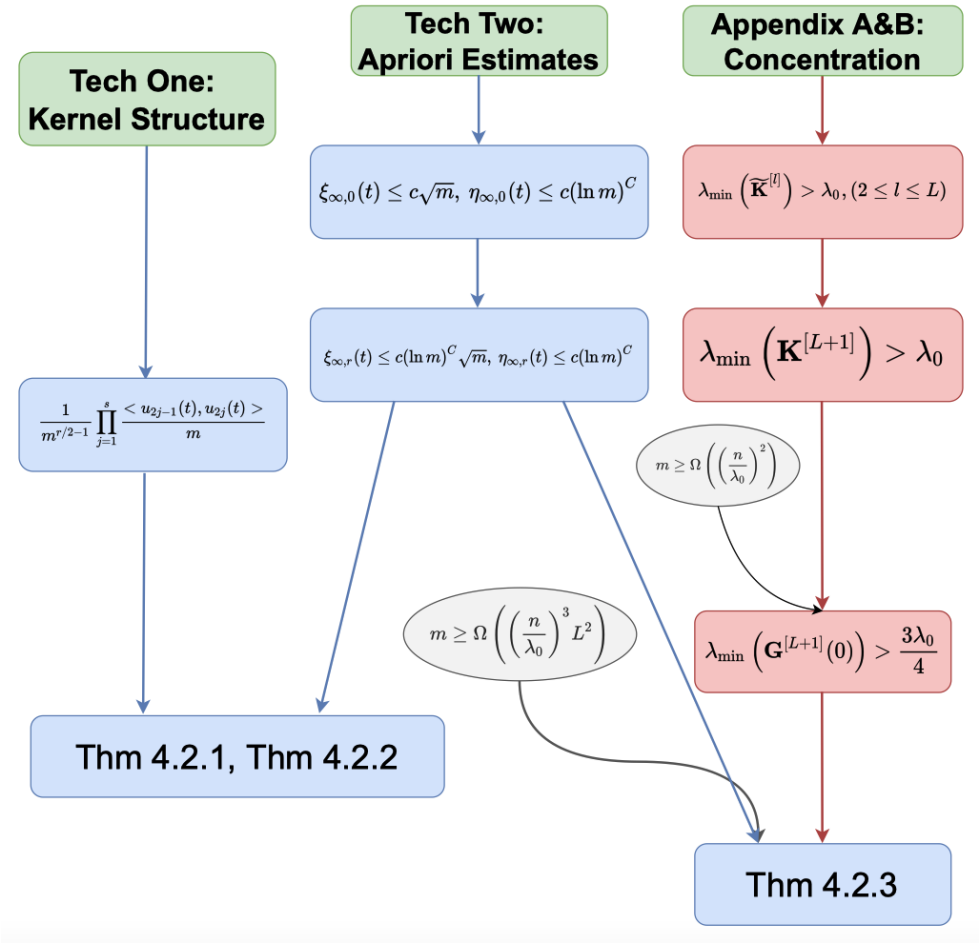

Figure 4.1. Diagram for the Proof of Main Theorems

Some future directions are listed out for our research:

- The NTH is an infinite sequence of relationship. Huang and Yau showed that under certain conditions, the NTH can be truncated and its truncated version is still able to approximate the original dynamics up to any precision. We have faith in that for ResNet, such technical conditions can be loosened.

- In Theorem 3.2.3, the dependence of m on the depth L is quadratic, we believe that the dependence can be reduced even further. We conjecture that m depends linearly in L.

- In this thesis, we focus on gradient descent, and we believe that it can be extended to stochastic gradient descent, while maintaining the linear convergence rate.

- The test loss has not been addressed in our work. To further investigate the generalization power of ResNet, some Apriori estimates for its generalization error would be useful [42].

REFERENCES

[1] S. Du, J. Lee, H. Li, L. Wang, and X. Zhai, "Gradient descent finds global minima of deep neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, K. Chaudhuri and R. Salakhutdinov, Eds., ser. Proceedings of Machine Learning Research, vol. 97, PMLR, Sep. 2019, pp. 1675–1685. [Online]. Available: http://proceedings.mlr.press/v97/du19c.html.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[3] J. Huang and H.-T. Yau, "Dynamics of deep neural networks and neural tangent hierarchy," in *Proceedings of the 37th International Conference on Machine Learning*, H. D. III and A. Singh, Eds., ser. Proceedings of Machine Learning Research, vol. 119, PMLR, 13–18 Jul 2020, pp. 4542–4551. [Online]. Available: http://proceedings.mlr.press/v119/huang20l.html.

[4] A. Jacot, F. Gabriel, and C. Hongler, "Neural tangent kernel: Convergence and generalization in neural networks," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31, Curran Associates, Inc., 2018, pp. 8571–8580. [Online]. Available: https://proceedings.neurips.cc/paper/2018/file/5a4be1fa34e62bb8a6ec6b91d2462f5a-Paper.pdf.

[5] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT press, 2016.

[6] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, OpenReview.net, 2017. [Online]. Available: https://openreview.net/forum?id=Sy8gdB9xx.

[7] S. S. Du, X. Zhai, B. Poczos, and A. Singh, "Gradient descent provably optimizes over-parameterized neural networks," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=S1eK3i09YQ.

[8] D. Zou, Y. Cao, D. Zhou, and Q. Gu, "Gradient descent optimizes over-parameterized deep relu networks," *Machine Learning*, vol. 109, no. 3, pp. 467–492, 2020.

[9] S. Arora, R. Ge, B. Neyshabur, and Y. Zhang, "Stronger generalization bounds for deep nets via a compression approach," in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy and A. Krause, Eds., ser. Proceedings of Machine Learning Research, vol. 80, Stockholmsmässan, Stockholm Sweden: PMLR, Oct. 2018, pp. 254–263. [Online]. Available: http://proceedings.mlr.press/v80/arora18b.html.

[10] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, *et al.*, "Mastering the Game of Go with Deep Neural Networks and Tree Search," *Nature*, vol. 529, no. 7587, p. 484, 2016.

[11] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, *et al.*, "Mastering the Game of Go without Human Knowledge," *Nature*, vol. 550, no. 7676, pp. 354–359, 2017.

[12] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *Proceedings of the British Machine Vision Conference (BMVC)*, E. R. H. Richard C. Wilson and W. A. P. Smith, Eds., BMVA Press, Sep. 2016, pp. 87.1–87.12, ISBN: 1-901725-59-6. DOI: 10.5244/C.30.87. [Online]. Available: https://dx.doi.org/10.5244/C.30.87.

[13] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, ser. AAAI'17, San Francisco, California, USA: AAAI Press, 2017, pp. 4278–4284.

[14] M. Hardt and T. Ma, "Identity matters in deep learning," *ICLR*, 2017. [Online]. Available: https://arxiv.org/abs/1611.04231.

[15] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of control, signals and systems*, vol. 2, no. 4, pp. 303–314, 1989.

[16] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014. [Online]. Available: http://jmlr.org/papers/v15/srivastava14a.html.

[17] S. Mei, A. Montanari, and P.-M. Nguyen, "A mean field view of the landscape of two-layer neural networks," *Proceedings of the National Academy of Sciences*, vol. 115, no. 33, E7665–E7671, 2018, ISSN: 0027-8424. DOI: 10.1073/pnas.1806579115. eprint: https://www.pnas.org/content/115/33/E7665.full.pdf. [Online]. Available: https://www.pnas.org/content/115/33/E7665.

[18] G. Rotskoff and E. Vanden-Eijnden, "Parameters as interacting particles: Long time convergence and asymptotic error scaling of neural networks," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31, Curran Associates, Inc., 2018, pp. 7146–7155. [Online]. Available: https://proceedings.neurips.cc/paper/2018/file/196f5641aa9dc87067da4ff90fd81e7b-Paper.pdf.

[19] L. Chizat and F. Bach, "On the global convergence of gradient descent for over-parameterized models using optimal transport," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31, Curran Associates, Inc., 2018, pp. 3036–3046. [Online]. Available: https://arxiv.org/pdf/1805.09545.pdf.

[20] T. Luo, Z.-Q. J. Xu, Z. Ma, and Y. Zhang, "Phase diagram for two-layer relu neural networks at infinite-width limit," *CoRR*, vol. abs/2007.07497, 2020. arXiv: 2007.07497. [Online]. Available: https://arxiv.org/abs/2007.07497.

[21] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, "Deep networks with stochastic depth," in *European conference on computer vision*, Springer, 2016, pp. 646–661.

[22] R. Eldan and O. Shamir, "The power of depth for feedforward neural networks," in *29th Annual Conference on Learning Theory*, V. Feldman, A. Rakhlin, and O. Shamir, Eds., ser. Proceedings of Machine Learning Research, vol. 49, Columbia University, New York, New York, USA: PMLR, 23–26 Jun 2016, pp. 907–940. [Online]. Available: http://proceedings.mlr.press/v49/eldan16.html.

[23] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994. DOI: 10.1109/72.279181.

[24] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning*, F. Bach and D. Blei, Eds., ser. Proceedings of Machine Learning Research, vol. 37, Lille, France: PMLR, Jul. 2015, pp. 448–456. [Online]. Available: http://proceedings.mlr.press/v37/ioffe15.html.

[25] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, Y. W. Teh and M. Titterington, Eds., ser. Proceedings of Machine Learning Research, vol. 9, Chia Laguna Resort, Sardinia, Italy: JMLR Workshop and Conference Proceedings, 13–15 May 2010, pp. 249–256. [Online]. Available: http://proceedings.mlr.press/v9/glorot10a.html.

[26] K. He and J. Sun, "Convolutional neural networks at constrained time cost," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 5353–5360. DOI: 10.1109/CVPR.2015.7299173.

[27] S. S. Du, C. Jin, J. D. Lee, M. I. Jordan, A. Singh, and B. Poczos, "Gradient descent can take exponential time to escape saddle points," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30, Curran Associates, Inc., 2017, pp. 1067–1077. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/f79921bbae40a577928b76d2fc3edc2a-Paper.pdf.

[28] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, "How to escape saddle points efficiently," in *Proceedings of the 34th International Conference on Machine Learning*, D. Precup and Y. W. Teh, Eds., ser. Proceedings of Machine Learning Research, vol. 70, International Convention Centre, Sydney, Australia: PMLR, Jun. 2017, pp. 1724–1732. [Online]. Available: http://proceedings.mlr.press/v70/jin17a.html.

[29] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping from saddle points — online stochastic gradient for tensor decomposition," in *Proceedings of The 28th Conference on Learning Theory*, P. Grünwald, E. Hazan, and S. Kale, Eds., ser. Proceedings of Machine Learning Research, vol. 40, Paris, France: PMLR, Mar. 2015, pp. 797–842. [Online]. Available: http://proceedings.mlr.press/v40/Ge15.html.

[30] Q. Nguyen and M. Hein, "The loss surface of deep and wide neural networks," in *Proceedings of the 34th International Conference on Machine Learning*, D. Precup and Y. W. Teh, Eds., ser. Proceedings of Machine Learning Research, vol. 70, International Convention Centre, Sydney, Australia: PMLR, Jun. 2017, pp. 2603–2612. [Online]. Available: http://proceedings.mlr.press/v70/nguyen17a.html.

[31] K. Kawaguchi, "Deep learning without poor local minima," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29, Curran Associates, Inc., 2016, pp. 586–594. [Online]. Available: https://proceedings.neurips.cc/paper/2016/file/f2fc990265c712c49d51a18a32b39f0c-Paper.pdf.

[32] S. Arora, N. Cohen, N. Golowich, and W. Hu, "A convergence analysis of gradient descent for deep linear neural networks," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=SkMQg3C5K7.

[33] P. Bartlett, D. Helmbold, and P. Long, "Gradient descent with identity initialization efficiently learns positive definite linear transformations by deep residual networks," in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy and A. Krause, Eds., ser. Proceedings of Machine Learning Research, vol. 80, Stockholmsmässan, Stockholm Sweden: PMLR, Oct. 2018, pp. 521–530. [Online]. Available: http://proceedings.mlr.press/v80/bartlett18a.html.

[34] Z. Allen-Zhu, Y. Li, and Z. Song, "A convergence theory for deep learning via over-parameterization," in *Proceedings of the 36th International Conference on Machine Learning*, K. Chaudhuri and R. Salakhutdinov, Eds., ser. Proceedings of Machine Learning Research, vol. 97, PMLR, Sep. 2019, pp. 242–252. [Online]. Available: http://proceedings.mlr.press/v97/allen-zhu19a.html.

[35] S. Arora, S. S. Du, W. Hu, Z. Li, R. R. Salakhutdinov, and R. Wang, "On exact computation with an infinitely wide neural net," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, Curran Associates, Inc., 2019, pp. 8141–8150. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/dbc4d84bfcfe2284ba11beffb853a8c4-Paper.pdf.

[36] Y. Li and Y. Liang, "Learning overparameterized neural networks via stochastic gradient descent on structured data," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31, Curran Associates, Inc., 2018, pp. 8157–8166. [Online]. Available: https://proceedings.neurips.cc/paper/2018/file/54fe976ba170c19ebae453679b362263-Paper.pdf.

[37] R. M. Neal, *Bayesian learning for neural networks*. Springer Science & Business Media, 2012, vol. 118.

[38] G. Yang, "Scaling Limits of Wide Neural Networks with Weight Sharing: Gaussian Process Behavior, Gradient Independence, and Neural Tangent Kernel Derivation," *arXiv preprint arXiv:1902.04760*, 2019.

[39] J. Lee, L. Xiao, S. Schoenholz, Y. Bahri, R. Novak, J. Sohl-Dickstein, and J. Pennington, "Wide neural networks of any depth evolve as linear models under gradient descent," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/0d1a9651497a38d8b1c3871c84528bd4-Paper.pdf.

[40] R. Vershynin, "Introduction to the Non-asymptotic Analysis of Random Matrices," *arXiv preprint arXiv:1011.3027*, 2010.

[41] B. Laurent and P. Massart, "Adaptive Estimation of a Quadratic Functional by Model Selection," *Annals of Statistics*, pp. 1302–1338, 2000.

[42] C. Ma, Q. Wang, and E. Weinan, "A Priori Estimates of the Population Risk for Residual Networks," *arXiv preprint arXiv:1903.02154*, 2019.

[43] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

A. LEAST EIGENVALUE OF GRAM MATRICES

A.1 Introduction

We shall recall the series of matrices $\left\{\widetilde{\boldsymbol{K}}^{[l]}\right\}_{l=1}^{L}$, $\left\{\widetilde{\boldsymbol{A}}^{[l]}\right\}_{l=1}^{L+1}$ and vectors $\left\{\widetilde{\boldsymbol{b}}^{[l]}\right\}_{l=1}^{L}$ given in Definition 2.4.4:

$$
\begin{aligned}
\widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[0]} &= \langle \boldsymbol{x}_{\mathrm{i}}, \boldsymbol{x}_{\mathrm{j}} \rangle , \\
\widetilde{\boldsymbol{A}}_{\mathrm{ij}}^{[1]} &= \begin{pmatrix} \widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[0]} & \widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[0]} \\ \widetilde{\boldsymbol{K}}_{\mathrm{ji}}^{[0]} & \widetilde{\boldsymbol{K}}_{\mathrm{jj}}^{[0]} \end{pmatrix} , \\
\widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[1]} &= \mathbb{E}_{(u,v)^{\intercal} \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{A}_{\mathrm{ij}}^{[1]}\right)} c_{\sigma} \sigma(u) \sigma(v), \\
\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[1]} &= \sqrt{c_{\sigma}}\, \mathbb{E}_{u \sim \mathcal{N}\left(0, \widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[0]}\right)} \left[\sigma(u)\right] , \\
\widetilde{\boldsymbol{A}}_{\mathrm{ij}}^{[l]} &= \begin{pmatrix} \widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l-1]} & \widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[l-1]} \\ \widetilde{\boldsymbol{K}}_{\mathrm{ji}}^{[l-1]} & \widetilde{\boldsymbol{K}}_{\mathrm{jj}}^{[l-1]} \end{pmatrix} , \\
\widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[l]} &= \widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[l-1]} + \mathbb{E}_{(u,v)^{\intercal} \sim \mathcal{N}\left(\boldsymbol{0}, \widetilde{\boldsymbol{A}}_{\mathrm{ij}}^{[l]}\right)} \left[\frac{c_{\mathrm{res}} \widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l-1]} \sigma(v)}{L} + \frac{c_{\mathrm{res}} \widetilde{\boldsymbol{b}}_{\mathrm{j}}^{[l-1]} \sigma(u)}{L} + \frac{c_{\mathrm{res}}^2 \sigma(u) \sigma(v)}{L^2} \right] , \\
\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l]} &= \widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l-1]} + \frac{c_{\mathrm{res}}}{L}\, \mathbb{E}_{u \sim \mathcal{N}\left(0, \widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l-1]}\right)} \left[\sigma(u)\right] , \\
\widetilde{\boldsymbol{A}}_{\mathrm{ij}}^{[L+1]} &= \begin{pmatrix} \widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[L]} & \widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[L]} \\ \widetilde{\boldsymbol{K}}_{\mathrm{ji}}^{[L]} & \widetilde{\boldsymbol{K}}_{\mathrm{jj}}^{[L]} \end{pmatrix} .
\end{aligned}
$$

Given expressions above, we shall recall definitions of $\boldsymbol{K}^{[L+1]}$ and $\boldsymbol{K}^{[L]}$ (Definition 2.4.5):

$$
\begin{aligned}
\boldsymbol{K}_{\mathrm{ij}}^{[L+1]} &= \widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[L]} + \mathbb{E}_{(u,v)^{\intercal} \sim \mathcal{N}\left(\boldsymbol{0}, \widetilde{\boldsymbol{A}}_{\mathrm{ij}}^{[L+1]}\right)} \left[\frac{c_{\mathrm{res}} \widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[L]} \sigma(v)}{L} + \frac{c_{\mathrm{res}} \widetilde{\boldsymbol{b}}_{\mathrm{j}}^{[L]} \sigma(u)}{L} + \frac{c_{\mathrm{res}}^2 \sigma(u) \sigma(v)}{L^2} \right] , \\
\boldsymbol{K}_{\mathrm{ij}}^{[L]} &= \frac{c_{\mathrm{res}}^2}{L^2} \widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[L-1]}\, \mathbb{E}_{(u,v)^{\intercal} \sim \mathcal{N}\left(\boldsymbol{0}, \widetilde{\boldsymbol{A}}_{\mathrm{ij}}^{[L]}\right)} \left[\sigma^{(1)}(u) \sigma^{(1)}(v)\right] .
\end{aligned}
$$

We shall state two lemmas concerning full rankness of $\boldsymbol{K}^{[L+1]}$ and $\boldsymbol{K}^{[L]}$, which have been stated as [1, Lemma F.1, Lemma F.2].

Lemma A.1.1 (Lemma F.1 in [1]). *Assume $\sigma(\cdot)$ is analytic and not a polynomial function. Consider input data set $\mathcal{U} = \{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_n\}$ of n non-parallel points, i.e., $\boldsymbol{u}_{\mathrm{j}} \notin \operatorname{span}(\boldsymbol{u}_k)$ for any $\mathrm{j} \neq k$. Define*

$$\boldsymbol{G}(\mathcal{U})_{\mathrm{ij}} := \mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}\left[\sigma(\boldsymbol{w}^{\intercal}\boldsymbol{u}_{\mathrm{i}})\sigma(\boldsymbol{w}^{\intercal}\boldsymbol{u}_{\mathrm{j}})\right], \tag{A.1}$$

then $\lambda_{\min}(\boldsymbol{G}(\mathcal{U})) > 0$.

Lemma A.1.2 (Lemma F.2 in [1]). *Assume $\sigma(\cdot)$ is analytic and not a polynomial function. Consider input data set $\mathcal{U} = \{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_n\}$ of n non-parallel points, i.e., $\boldsymbol{u}_{\mathrm{j}} \notin \operatorname{span}(\boldsymbol{u}_k)$ for any $\mathrm{j} \neq k$. Define*

$$\boldsymbol{G}(\mathcal{U})_{\mathrm{ij}} := \mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}\left[\sigma^{(1)}(\boldsymbol{w}^{\intercal}\boldsymbol{u}_{\mathrm{i}})\sigma^{(1)}(\boldsymbol{w}^{\intercal}\boldsymbol{u}_{\mathrm{j}})\,(\boldsymbol{u}_{\mathrm{i}}^{\intercal}\boldsymbol{u}_{\mathrm{j}})\right], \tag{A.2}$$

then $\lambda_{\min}(\boldsymbol{G}(\mathcal{U})) > 0$.

Given Lemma A.1.1 and Lemma A.1.2, we may proceed to quantify the least eigenvalues of $\boldsymbol{K}^{[L+1]}$ and $\boldsymbol{K}^{[L]}$.

A.2 Full Rankness for $(L+1)$-th Gram matrix

Lemma A.2.1. *Given $\mathcal{X}$, $\sigma(\cdot)$, $\left\{\widetilde{\boldsymbol{K}}^{[l]}\right\}_{l=1}^{L}$, and $\left\{\widetilde{\boldsymbol{b}}^{[l]}\right\}_{l=1}^{L}$, then for each l, each diagonal entry of $\widetilde{\boldsymbol{K}}^{[l]}$ is the same with each other, and each element of the vector $\widetilde{\boldsymbol{b}}^{[l]}$ is also the same, i.e.,*

$$\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l_1]} = \widetilde{\boldsymbol{K}}_{\mathrm{jj}}^{[l_1]},\ \widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l_2]} = \widetilde{\boldsymbol{b}}_{\mathrm{j}}^{[l_2]}, \qquad \text{for } \mathrm{i} \neq \mathrm{j}.$$

Moreover

$$\left(1 - \frac{l}{L}\frac{c}{\sqrt{c_\sigma}}\right)^2 \leq \widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]} \leq \left(1 + \frac{l}{L}\frac{c}{\sqrt{c_\sigma}}\right)^2, \tag{A.3}$$

and

$$\left(\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l]}\right)^2 < \widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]}, \tag{A.4}$$

where $c > 0$ and depends solely on c_{res} and activation function $\sigma(\cdot)$.

Proof. We shall prove it by induction on l. Firstly, we notice that $\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[0]} = \widetilde{\boldsymbol{K}}_{\mathrm{jj}}^{[0]}$ for any $\mathrm{i} \neq \mathrm{j}$, which is obvious since $\|\boldsymbol{x}_{\mathrm{i}}\|_2 = 1$, then $\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[0]} = \widetilde{\boldsymbol{K}}_{\mathrm{jj}}^{[0]} = 1$.

Next we need to show that (A.3), (A.4) hold true for $l = 1$. Based on definition, we recall that $c_\sigma = \left(\mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[\sigma(x)^2\right]\right)^{-1}$,

$$\begin{aligned}
\boldsymbol{K}_{\mathrm{ii}}^{[1]} &= c_\sigma \, \mathbb{E}_{u\sim\mathcal{N}\left(0,\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[0]}\right)}\left(\sigma(u)^2\right) = c_\sigma \, \mathbb{E}_{u\sim\mathcal{N}(0,1)}\left(\sigma(u)^2\right) = 1, \\
\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[1]} &= \sqrt{c_\sigma}\, \mathbb{E}_{u\sim\mathcal{N}\left(0,\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[0]}\right)}\left[\sigma(u)\right] = \sqrt{c_\sigma}\, \mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma(u)\right],
\end{aligned}$$

then

$$\left(\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[1]}\right)^2 = c_\sigma \left(\mathbb{E}_{u\sim\mathcal{N}\left(0,\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[0]}\right)}\left[\sigma(u)\right]\right)^2 < 1, \tag{A.5}$$

Equation (A.5) holds because

$$\left(\mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[\sigma(x)\right]\right)^2 < \mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[\sigma(x)^2\right],$$

since the quantity is independent of choice of the index i, then $\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[1]} = \widetilde{\boldsymbol{K}}_{\mathrm{jj}}^{[1]}$, $\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[1]} = \widetilde{\boldsymbol{b}}_{\mathrm{j}}^{[1]}$, for any $\mathrm{i} \neq \mathrm{j}$.

Now we assume that (A.3), (A.4) hold for $1, 2, \cdots, l-1$, and we want to show that it is also the case for l. First of all, since

$$\begin{aligned}
\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]} &= \widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l-1]} + \mathbb{E}_{u\sim\mathcal{N}\left(0,\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l-1]}\right)}\left[\frac{c_{\mathrm{res}}\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l-1]}\sigma(u)}{L} + \frac{c_{\mathrm{res}}\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l-1]}\sigma(u)}{L} + \frac{c_{\mathrm{res}}^2\sigma(u)\sigma(u)}{L^2}\right], \\
\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l]} &= \widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l-1]} + \frac{c_{\mathrm{res}}}{L}\, \mathbb{E}_{u\sim\mathcal{N}\left(0,\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l-1]}\right)}\left[\sigma(u)\right],
\end{aligned}$$

then such quantities $\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]}$, $\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l]}$ are independent of choice of the index i.

Secondly, concerning $\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l]}$, we assume Equation (A.4) holds for $l = 1, 2, \cdots, l-1$, then

$$\begin{aligned}
\left(\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l]}\right)^2 &= \left(\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l-1]}\right)^2 + 2\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l-1]} \frac{c_{\mathrm{res}}}{L} \mathbb{E}_{u\sim\mathcal{N}\left(0,\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l-1]}\right)}\left[\sigma(u)\right] + \left(\frac{c_{\mathrm{res}}}{L} \mathbb{E}_{u\sim\mathcal{N}\left(0,\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l-1]}\right)}\left[\sigma(u)\right]\right)^2 \\
&< \left(\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l-1]}\right)^2 + 2\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l-1]} \frac{c_{\mathrm{res}}}{L} \mathbb{E}_{u\sim\mathcal{N}\left(0,\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l-1]}\right)}\left[\sigma(u)\right] + \frac{c_{\mathrm{res}}^2}{L^2} \mathbb{E}_{u\sim\mathcal{N}\left(0,\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l-1]}\right)}\left[\sigma(u)^2\right] \\
&< \widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l-1]} + 2\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l-1]} \frac{c_{\mathrm{res}}}{L} \mathbb{E}_{u\sim\mathcal{N}\left(0,\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l-1]}\right)}\left[\sigma(u)\right] + \frac{c_{\mathrm{res}}^2}{L^2} \mathbb{E}_{u\sim\mathcal{N}\left(0,\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l-1]}\right)}\left[\sigma(u)^2\right] = \widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]}.
\end{aligned}$$

Finally, for $\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]}$, we note that the following inequality holds

$$\begin{aligned}
&\left(\sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l-1]}} - \frac{c_{\mathrm{res}}}{L}\sqrt{\mathbb{E}_{u\sim\mathcal{N}\left(0,\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l-1]}\right)}\left[\sigma(u)^2\right]}\right)^2 \\
&\leq \widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]} \leq \left(\sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l-1]}} + \frac{c_{\mathrm{res}}}{L}\sqrt{\mathbb{E}_{u\sim\mathcal{N}\left(0,\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l-1]}\right)}\left[\sigma(u)^2\right]}\right)^2.
\end{aligned}$$

Since $\sigma(\cdot)$ is 1-Lipschitz, then for any $1/2 \leq \alpha \leq 2$, we have

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

then

$$\mathbb{E}_{X\sim\mathcal{N}(0,1)}\left[\sigma(\alpha X)^2\right] \leq \frac{1}{c_\sigma} + \frac{C}{c_\sigma}|\alpha - 1|.$$

Based on our induction hypothesis,

$$1 - \frac{l-1}{L}\frac{c}{\sqrt{c_\sigma}} \leq \sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l-1]}} \leq 1 + \frac{l-1}{L}\frac{c}{\sqrt{c_\sigma}},$$

set $\alpha = \sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l-1]}}$, we obtain

$$\mathbb{E}_{X\sim\mathcal{N}\left(0,\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l-1]}\right)}\left[\sigma(X)^2\right] \leq \frac{1}{c_\sigma} + \frac{C}{c_\sigma}\frac{l-1}{L}\frac{c}{\sqrt{c_\sigma}},$$

we shall choose c wisely, for instance, set c as

$$c = \frac{Cc_{\mathrm{res}}^2}{2\sqrt{c_\sigma}} + \sqrt{\frac{C^2c_{\mathrm{res}}^4}{4c_\sigma} + c_{\mathrm{res}}^2},$$

by our choice of c, combined with Equation (A.6), we have

$$\left(\sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l-1]}} - \frac{1}{L}\frac{c}{\sqrt{c_\sigma}}\right)^2 \leq \widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]} \leq \left(\sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l-1]}} + \frac{1}{L}\frac{c}{\sqrt{c_\sigma}}\right)^2,$$

then

$$\left(1 - \frac{l}{L}\frac{c}{\sqrt{c_\sigma}}\right)^2 \leq \widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]} \leq \left(1 + \frac{l}{L}\frac{c}{\sqrt{c_\sigma}}\right)^2,$$

which finishes our proof. □

Our next lemma is crucial in that it revels a covariance type structure for the series of matrices $\left\{\widetilde{\boldsymbol{K}}^{[l]}\right\}_{l=1}^{L}$. A standard notation related to matrices shall be introduced. We denote that $\boldsymbol{A} \succeq \boldsymbol{B}$ if and only if $\boldsymbol{A} - \boldsymbol{B}$ is a semi-positive definite matrix, and $\boldsymbol{A} \succ \boldsymbol{B}$ if and only if $\boldsymbol{A} - \boldsymbol{B}$ is strictly positive definite.

Proposition A.2.1. *Given $\mathcal{X}$, $\sigma(\cdot)$, $\left\{\widetilde{\boldsymbol{K}}^{[l]}\right\}_{l=1}^{L}$, and $\left\{\widetilde{\boldsymbol{b}}^{[l]}\right\}_{l=1}^{L}$, then for each l,*

$$\widetilde{\boldsymbol{K}}^{[l]} - \widetilde{\boldsymbol{b}}^{[l]} \otimes \left(\widetilde{\boldsymbol{b}}^{[l]}\right)^{\intercal} \succeq \widetilde{\boldsymbol{K}}^{[l-1]} - \widetilde{\boldsymbol{b}}^{[l-1]} \otimes \left(\widetilde{\boldsymbol{b}}^{[l-1]}\right)^{\intercal}. \tag{A.6}$$

Set $\lambda_0 > 0$ as

$$\lambda_{\min}\left(\widetilde{\boldsymbol{K}}^{[1]} - \widetilde{\boldsymbol{b}}^{[1]} \otimes \left(\widetilde{\boldsymbol{b}}^{[1]}\right)^{\intercal}\right) = \lambda_0,$$

then for all $2 \leq l \leq L$,

$$\lambda_{\min}\left(\widetilde{\boldsymbol{K}}^{[l]}\right) \geq \lambda_0. \tag{A.7}$$

Proof. Since for $1 \le \mathrm{i}, \mathrm{j} \le n$, and $1 \le l \le L$,

$$
\begin{aligned}
&\widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[l]} - \widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l]} \widetilde{\boldsymbol{b}}_{\mathrm{j}}^{[l]} \\
&= \widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[l-1]} + \mathbb{E}_{(u,v)^\intercal \sim \mathcal{N}\left(0, \widetilde{\boldsymbol{A}}_{\mathrm{ij}}^{[l]}\right)} \left[\frac{c_{\mathrm{res}} \widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l-1]} \sigma(v)}{L} + \frac{c_{\mathrm{res}} \widetilde{\boldsymbol{b}}_{\mathrm{j}}^{[l-1]} \sigma(u)}{L} + \frac{c_{\mathrm{res}}^2 \sigma(u)\sigma(v)}{L^2} \right] \\
&\quad - \left(\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l-1]} + \frac{c_{\mathrm{res}}}{L} \mathbb{E}_{u \sim \mathcal{N}\left(0, \widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l-1]}\right)} [\sigma(u)] \right) \left(\widetilde{\boldsymbol{b}}_{\mathrm{j}}^{[l-1]} + \frac{c_{\mathrm{res}}}{L} \mathbb{E}_{v \sim \mathcal{N}\left(0, \widetilde{\boldsymbol{K}}_{\mathrm{jj}}^{[l-1]}\right)} [\sigma(v)] \right) \\
&= \widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[l-1]} - \widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l-1]} \widetilde{\boldsymbol{b}}_{\mathrm{j}}^{[l-1]} + \mathbb{E}_{(u,v)^\intercal \sim \mathcal{N}\left(0, \widetilde{\boldsymbol{A}}_{\mathrm{ij}}^{[l]}\right)} \left[\frac{c_{\mathrm{res}}^2 \sigma(u)\sigma(v)}{L^2} \right] \\
&\quad - \frac{c_{\mathrm{res}}}{L} \mathbb{E}_{u \sim \mathcal{N}\left(0, \widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l-1]}\right)} [\sigma(u)] \frac{c_{\mathrm{res}}}{L} \mathbb{E}_{v \sim \mathcal{N}\left(0, \widetilde{\boldsymbol{K}}_{\mathrm{jj}}^{[l-1]}\right)} [\sigma(v)] \\
&= \widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[l-1]} - \widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l-1]} \widetilde{\boldsymbol{b}}_{\mathrm{j}}^{[l-1]} + \frac{c_{\mathrm{res}}^2}{L^2} \operatorname{Cov}_{(u,v)^\intercal \sim \mathcal{N}\left(0, \widetilde{\boldsymbol{A}}_{\mathrm{ij}}^{[l]}\right)} [\sigma(u)\sigma(v)] .
\end{aligned}
$$

We define another series of covariance matrices $\left\{ \boldsymbol{P}^{[s]} : 1 \le s \le L \right\}$,

$$
\boldsymbol{P}_{\mathrm{ij}}^{[s]} := \frac{c_{\mathrm{res}}^2}{L^2} \operatorname{Cov}_{(u,v)^\intercal \sim \mathcal{N}\left(0, \widetilde{\boldsymbol{A}}_{\mathrm{ij}}^{[s+1]}\right)} [\sigma(u)\sigma(v)] , \qquad 1 \le s \le L.
$$

We notice that $\boldsymbol{P}^{[s]}$ are covariance matrices, hence naturally $\boldsymbol{P}^{[s]} \succeq 0$. Inductively, we have for all l,

$$
\begin{aligned}
\widetilde{\boldsymbol{K}}^{[l]} &\succeq \widetilde{\boldsymbol{K}}^{[l]} - \widetilde{\boldsymbol{b}}^{[l]} \otimes \left(\widetilde{\boldsymbol{b}}^{[l]} \right)^\intercal \\
&= \widetilde{\boldsymbol{K}}^{[l-1]} - \widetilde{\boldsymbol{b}}^{[l-1]} \otimes \left(\widetilde{\boldsymbol{b}}^{[l-1]} \right)^\intercal + \boldsymbol{P}^{[l-1]} \\
&\succeq \widetilde{\boldsymbol{K}}^{[l-1]} - \widetilde{\boldsymbol{b}}^{[l-1]} \otimes \left(\widetilde{\boldsymbol{b}}^{[l-1]} \right)^\intercal \\
&= \widetilde{\boldsymbol{K}}^{[l-2]} - \widetilde{\boldsymbol{b}}^{[l-2]} \otimes \left(\widetilde{\boldsymbol{b}}^{[l-2]} \right)^\intercal + \boldsymbol{P}^{[l-2]} \\
&\quad \vdots \\
&\succeq \widetilde{\boldsymbol{K}}^{[1]} - \widetilde{\boldsymbol{b}}^{[1]} \otimes \left(\widetilde{\boldsymbol{b}}^{[1]} \right)^\intercal ,
\end{aligned}
\tag{A.8}
$$

the last line brings us to notice of the entry of $\widetilde{\boldsymbol{K}}^{[1]} - \widetilde{\boldsymbol{b}}^{[1]} \otimes \left(\widetilde{\boldsymbol{b}}^{[1]} \right)^\intercal$, which reads respectively

$$
\left[\widetilde{\boldsymbol{K}}^{[1]} - \widetilde{\boldsymbol{b}}^{[1]} \otimes \left(\widetilde{\boldsymbol{b}}^{[1]} \right)^\intercal \right]_{\mathrm{ij}} = c_\sigma \operatorname{Cov}_{(u,v)^\intercal \sim \mathcal{N}\left(0, \widetilde{\boldsymbol{A}}_{\mathrm{ij}}^{[1]}\right)} [\sigma(u)\sigma(v)] .
$$

We notice that $\left[\widetilde{\boldsymbol{K}}^{[1]} - \widetilde{\boldsymbol{b}}^{[1]} \otimes \left(\widetilde{\boldsymbol{b}}^{[1]}\right)^{\intercal}\right]$ is a covariance matrix, hence naturally

$$\left[\widetilde{\boldsymbol{K}}^{[1]} - \widetilde{\boldsymbol{b}}^{[1]} \otimes \left(\widetilde{\boldsymbol{b}}^{[1]}\right)^{\intercal}\right] \succeq 0,$$

and $\left[\widetilde{\boldsymbol{K}}^{[1]} - \widetilde{\boldsymbol{b}}^{[1]} \otimes \left(\widetilde{\boldsymbol{b}}^{[1]}\right)^{\intercal}\right] \succ 0$ except when one of the samples is an exact linear combination of the others. From Lemma A.1.1, we see that $\widetilde{\boldsymbol{K}}^{[1]} \succ 0$, consequently

$$\lambda_{\min}\left(\widetilde{\boldsymbol{K}}^{[1]} - \widetilde{\boldsymbol{b}}^{[1]} \otimes \left(\widetilde{\boldsymbol{b}}^{[1]}\right)^{\intercal}\right) = \lambda_0 > 0,$$

then $\lambda_0 > 0$, and it depends solely on the input samples and activation function. □

Proposition A.2.2. *Given* $\mathcal{X}$, $\sigma(\cdot)$, $\left\{\widetilde{\boldsymbol{K}}^{[l]}\right\}_{l=1}^{L}$, *and* $\left\{\widetilde{\boldsymbol{b}}^{[l]}\right\}_{l=1}^{L}$, *if we set* $\lambda_0 > 0$ *as*

$$\lambda_{\min}\left(\widetilde{\boldsymbol{K}}^{[1]} - \widetilde{\boldsymbol{b}}^{[1]} \otimes \left(\widetilde{\boldsymbol{b}}^{[1]}\right)^{\intercal}\right) = \lambda_0,$$

then

$$\lambda_{\min}\left(\boldsymbol{K}^{[L+1]}\right) \geq \lambda_0. \tag{A.9}$$

Proof. Proof of Proposition A.2.2 is quite similar to the proof of Proposition A.2.1. We recall that

$$\boldsymbol{K}_{\mathrm{ij}}^{[L+1]} = \widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[L]} + \mathbb{E}_{(u,v)^{\intercal}\sim\mathcal{N}\left(\boldsymbol{0},\widetilde{\boldsymbol{A}}_{\mathrm{ij}}^{[L+1]}\right)}\left[\frac{c_{\mathrm{res}}\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[L]}\sigma(v)}{L} + \frac{c_{\mathrm{res}}\widetilde{\boldsymbol{b}}_{\mathrm{j}}^{[L]}\sigma(u)}{L} + \frac{c_{\mathrm{res}}^2\sigma(u)\sigma(v)}{L^2}\right],$$

and we define further that

$$\boldsymbol{b}_{\mathrm{i}}^{[L+1]} := \widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[L]} + \frac{c_{\mathrm{res}}}{L}\,\mathbb{E}_{u\sim\mathcal{N}(0,\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[L]})}\left[\sigma(u)\right],$$

then

$$\boldsymbol{K}_{\mathrm{ij}}^{[L+1]} - \boldsymbol{b}_{\mathrm{i}}^{[L+1]}\boldsymbol{b}_{\mathrm{j}}^{[L+1]} = \widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[L]} - \widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[L]}\widetilde{\boldsymbol{b}}_{\mathrm{j}}^{[L]} + \frac{c_{\mathrm{res}}^2}{L^2}\,\mathrm{Cov}_{(u,v)^{\intercal}\sim\sim\mathcal{N}\left(\boldsymbol{0},\widetilde{\boldsymbol{A}}_{\mathrm{ij}}^{[L+1]}\right)}\left[\sigma(u)\sigma(v)\right],$$

hence by same reasoning in Equation (A.8),

$$\begin{aligned}\boldsymbol{K}^{[L+1]} &\succeq \boldsymbol{K}^{[L+1]} - \boldsymbol{b}^{[L+1]} \otimes \left(\boldsymbol{b}^{[L+1]}\right)^{\intercal} \\ &\succeq \widetilde{\boldsymbol{K}}^{[L]} - \widetilde{\boldsymbol{b}}^{[L]} \otimes \left(\widetilde{\boldsymbol{b}}^{[L]}\right)^{\intercal} \\ &\succeq \widetilde{\boldsymbol{K}}^{[1]} - \widetilde{\boldsymbol{b}}^{[1]} \otimes \left(\widetilde{\boldsymbol{b}}^{[1]}\right)^{\intercal},\end{aligned}$$

apply Proposition A.2.1 directly, we are able to finish the proof. □

From Proposition A.2.2, we observe that $\lambda_{\min}\left(\boldsymbol{K}^{[L+1]}\right) \sim \Omega(1)$.

A.3 Full Rankness for the L-th Gram matrix

Our next proposition concerns the least eigenvalue of $\boldsymbol{K}^{[L]}$, which has been stated as Proposition F.2 in Du et al. [1].

Proposition A.3.1. *Given $\mathcal{X}$, $\sigma(\cdot)$, $\left\{\widetilde{\boldsymbol{K}}^{[l]}\right\}_{l=1}^{L}$, and $\left\{\widetilde{\boldsymbol{b}}^{[l]}\right\}_{l=1}^{L}$, then*

$$\lambda_{\min}\left(\boldsymbol{K}^{[L]}\right) \geq \frac{c_{\text{res}}^2}{L^2}\kappa, \tag{A.10}$$

where κ is independent of depth L.

Proof. Based on Lemma A.2.1, there exits a constant $c > 0$, such that

$$1/c \leq \widetilde{\boldsymbol{K}}_{\text{ii}}^{[L]} \leq c.$$

We define function $\boldsymbol{G} : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$:

$$\boldsymbol{G}(\boldsymbol{U})_{\text{ij}} := \boldsymbol{U}_{\text{ij}}\, \mathbb{E}_{(u,v)^{\intercal}\sim\mathcal{N}\left(\boldsymbol{0}, \begin{pmatrix}\boldsymbol{U}_{\text{ii}} & \boldsymbol{U}_{\text{ij}} \\ \boldsymbol{U}_{\text{ji}} & \boldsymbol{U}_{\text{jj}}\end{pmatrix}\right)} \sigma^{(1)}(u)\sigma^{(1)}(v),$$

from which a scalar function $g(\lambda)$ could be defined as

$$g(\lambda) := \min_{\boldsymbol{U}:\boldsymbol{U}\succ 0, 1/c\leq \boldsymbol{U}_{\text{ii}}\leq c, \lambda_{\min}(\boldsymbol{U})\geq\lambda} \lambda_{\min}\left(\boldsymbol{G}(\boldsymbol{U})\right),$$

and Lemma A.1.2 guarantees that

$$g(\lambda_0) > 0.$$

From Proposition A.2.1, we observe that

$$\lambda_{\min}\left(\widetilde{\boldsymbol{K}}^{[L-1]}\right) \geq \lambda_0,$$

hence

$$\lambda_{\min}\left(\boldsymbol{K}^{[L]}\right) \geq \frac{c_{\text{res}}^2}{L^2} g(\lambda_0). \tag{A.11}$$

Set $\kappa = g(\lambda_0) > 0$, since κ is independent of depth L, we finish our proof. □

From Proposition A.3.1, we remark that $\lambda_{\min}\left(\boldsymbol{K}^{[L]}\right) \sim \Omega(\frac{1}{L^2})$.

B. RANDOM INITIALIZATION OF NTK

In this chapter, we are going to show that

$$\lambda_{\min}\left(\left[\mathcal{G}_0^{[L+1]}\left(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta\right)\right]_{1\leq\alpha,\beta\leq n}\right) \geq \frac{3\lambda_0}{4}, \tag{B.1}$$

holds with high probability, where $\lambda_0 > 0$ is set to be the least eigenvalue of matrix $\widetilde{\boldsymbol{K}}^{[1]} - \widetilde{\boldsymbol{b}}^{[1]} \otimes \left(\widetilde{\boldsymbol{b}}^{[1]}\right)^\intercal$, i.e.,

$$\lambda_0 = \lambda_{\min}\left(\widetilde{\boldsymbol{K}}^{[1]} - \widetilde{\boldsymbol{b}}^{[1]} \otimes \left(\widetilde{\boldsymbol{b}}^{[1]}\right)^\intercal\right).$$

B.1 Several Lemmas on Gaussian Concentration and Other Aspects

We investigate the concentration properties of Lipschitz functions of Gaussian variables. Let us say that a function $f(\cdot) : \mathbb{R}^p \to \mathbb{R}$ is C_L-Lipschitz function with respect to the Euclidean norm $\|\cdot\|_2$, if for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$,

$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq C_L \|\boldsymbol{x} - \boldsymbol{y}\|_2.$$

Our next lemma reveals that any Lipschitz function of Gaussian variables is itself a sub-Gaussian variable.

Lemma B.1.1 (Gaussian Concentration Inequality)**.** *Let $\boldsymbol{X} = (X_1, \cdots X_p)^\intercal \in \mathbb{R}^p$, whose components $X_1, X_2, \cdots, X_p$ are i.i.d. Gaussian variables drawn from $\mathcal{N}(0, \sigma^2)$, and $f(\cdot) : \mathbb{R}^p \to \mathbb{R}$ is a C_L-Lipschitz function with respect to the Euclidean norm $\|\cdot\|_2$, then for all $t \geq 0$:*

$$\mathbb{P}\left(|f(\boldsymbol{X}) - \mathbb{E}\, f(\boldsymbol{X})| \geq t\right) \leq 2\exp(-\frac{t^2}{2{C_L}^2\sigma^2}). \tag{B.2}$$

We refer to [43, Theorem 5.6] for the proof of Lemma B.1.1, and remarkably, this is a dimension free inequality.

Next we shall state two lemmas, which have been stated as [1, Lemma G.3, Lemma G.4]

Lemma B.1.2 (Lemma G.3 in [1])**.** *If $\sigma(\cdot)$ is C_L-Lipschitz, then for $a, b > 0$, satisfying*

$$1/c \leq \min(a, b), \max(a, b) \leq c,$$

for some $c > 0$, we have

$$\left|\mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[\sigma(az)\right] - \mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[\sigma(bz)\right]\right| \leq C\left|a-b\right|, \tag{B.3}$$

where $C > 0$ only depends on c and Lipschitz constant C_L.

Lemma B.1.3 (Lemma G.4 in [1])**.** *If $\sigma(\cdot)$ is C_L-Lipschitz, define $F(\boldsymbol{K}) : \mathbb{R}^{2\times 2} \to \mathbb{R}$*

$$F(\boldsymbol{K}) = \mathbb{E}_{(u,v)^\intercal\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{K})}\left[\sigma(u)\sigma(v)\right],$$

then for any two matrices $\boldsymbol{A}, \boldsymbol{B}$, with

$$\boldsymbol{A} = \begin{pmatrix} a_1^2 & \rho_1 a_1 b_1 \\ \rho_1 a_1 b_1 & b_1^2 \end{pmatrix},$$
$$\boldsymbol{B} = \begin{pmatrix} a_2^2 & \rho_2 a_2 b_2 \\ \rho_2 a_2 b_2 & b_2^2 \end{pmatrix},$$

whose entries satisfy

$$1/c \leq \min(a_1, b_1), \min(a_2, b_2), \max(a_1, b_1), \max(a_2, b_2) \leq c,$$
$$-1 \leq \rho_1, \rho_2 \leq 1$$

for some $c > 0$. Then, we have

$$\left|F(\boldsymbol{A}) - F(\boldsymbol{B})\right| \leq C\left\|\boldsymbol{A}-\boldsymbol{B}\right\|_{\mathrm{F}} \leq 2C\left\|\boldsymbol{A}-\boldsymbol{B}\right\|_{\infty},$$

where $C > 0$ only depends on c and Lipschitz constant C_L.

B.2 Analysis of Random Propagation

We recall readers once again the series of matrices $\left\{\widetilde{\boldsymbol{K}}^{[l]}\right\}_{l=1}^{L}$, $\left\{\widetilde{\boldsymbol{A}}^{[l]}\right\}_{l=1}^{L+1}$ and vectors $\left\{\widetilde{\boldsymbol{b}}^{[l]}\right\}_{l=1}^{L}$ given in Definition 2.4.4:

$$
\begin{aligned}
\widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[0]} &= \langle \boldsymbol{x}_{\mathrm{i}}, \boldsymbol{x}_{\mathrm{j}} \rangle \,, \\
\widetilde{\boldsymbol{A}}_{\mathrm{ij}}^{[1]} &= \begin{pmatrix} \widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[0]} & \widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[0]} \\ \widetilde{\boldsymbol{K}}_{\mathrm{ji}}^{[0]} & \widetilde{\boldsymbol{K}}_{\mathrm{jj}}^{[0]} \end{pmatrix}, \\
\widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[1]} &= \mathbb{E}_{(u,v)^\intercal \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{A}_{\mathrm{ij}}^{[1]}\right)} c_\sigma \sigma(u)\sigma(v), \\
\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[1]} &= \sqrt{c_\sigma}\, \mathbb{E}_{u \sim \mathcal{N}\left(0, \widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[0]}\right)} \left[\sigma(u)\right], \\
\widetilde{\boldsymbol{A}}_{\mathrm{ij}}^{[l]} &= \begin{pmatrix} \widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l-1]} & \widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[l-1]} \\ \widetilde{\boldsymbol{K}}_{\mathrm{ji}}^{[l-1]} & \widetilde{\boldsymbol{K}}_{\mathrm{jj}}^{[l-1]} \end{pmatrix}, \\
\widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[l]} &= \widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[l-1]} + \mathbb{E}_{(u,v)^\intercal \sim \mathcal{N}\left(\boldsymbol{0}, \widetilde{\boldsymbol{A}}_{\mathrm{ij}}^{[l]}\right)} \left[\frac{c_{\mathrm{res}} \widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l-1]} \sigma(v)}{L} + \frac{c_{\mathrm{res}} \widetilde{\boldsymbol{b}}_{\mathrm{j}}^{[l-1]} \sigma(u)}{L} + \frac{c_{\mathrm{res}}^2 \sigma(u)\sigma(v)}{L^2} \right], \\
\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l]} &= \widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l-1]} + \frac{c_{\mathrm{res}}}{L}\, \mathbb{E}_{u \sim \mathcal{N}\left(0, \widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l-1]}\right)} \left[\sigma(u)\right], \\
\widetilde{\boldsymbol{A}}_{\mathrm{ij}}^{[L+1]} &= \begin{pmatrix} \widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[L]} & \widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[L]} \\ \widetilde{\boldsymbol{K}}_{\mathrm{ji}}^{[L]} & \widetilde{\boldsymbol{K}}_{\mathrm{jj}}^{[L]} \end{pmatrix}.
\end{aligned}
$$

We shall begin with a proposition on the tail probabilities concerning the outputs of each layer at initial state, denoted by $\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)$ for all $1 \leq \mathrm{i} \leq n$ and $1 \leq l \leq L$.

Proposition B.2.1. *Given $\mathcal{X}$, $\sigma(\cdot)$, $\left\{\widetilde{\boldsymbol{K}}^{[l]}\right\}_{l=1}^{L}$, and $\left\{\widetilde{\boldsymbol{b}}^{[l]}\right\}_{l=1}^{L}$, then for all $t > 0$, $1 \leq \mathrm{i} \leq n$, $1 \leq l \leq L$:*

$$
\mathbb{P}\left(\left| \left\| \boldsymbol{x}_{\mathrm{i}}^{[l]}(0) \right\|_2 - \sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]}} \right| \geq t \right) \leq \exp\left(-cmt^2\right), \tag{B.4}
$$

$$
\mathbb{P}\left(\left| \left\langle \frac{\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)}{\sqrt{m}}, \boldsymbol{1} \right\rangle - \widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l]} \right| \geq t \right) \leq \exp\left(-cmt^2\right), \tag{B.5}
$$

where $c > 0$ is independent of depth L.

Proof. (i). For $l = 1$, we have

$$\left\|\boldsymbol{x}_{\mathrm{i}}^{[1]}(0)\right\|_2^2 = \frac{c_\sigma}{m}\sum_{\mathrm{j}=1}^{m}\sigma\left((\boldsymbol{W}_0^{[1]}\boldsymbol{x}_{\mathrm{i}})_{\mathrm{j}}\right)^2,$$

with

$$\mathbb{E}\left[\left\|\boldsymbol{x}_{\mathrm{i}}^{[1]}(0)\right\|_2^2\right] = c_\sigma\mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[\sigma(x)^2\right] = \widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[1]} = 1.$$

Since for all i, $\|\boldsymbol{x}_{\mathrm{i}}\|_2 = 1$, then for each $\mathrm{j} = 1, 2, \cdots, m$, $\left(\boldsymbol{W}_0^{[1]}\boldsymbol{x}_{\mathrm{i}}\right)_{\mathrm{j}}$ is a standard Gaussian, i.e., $\left(\boldsymbol{W}_0^{[1]}\boldsymbol{x}_{\mathrm{i}}\right)_{\mathrm{j}} \sim \mathcal{N}(0,1)$. Moreover, $\sigma(\cdot)$ is 1-Lipschitz, then $\sigma\left((\boldsymbol{W}_0^{[1]}\boldsymbol{x}_{\mathrm{i}})_{\mathrm{j}}\right)$ is sub-Gaussian, hence $\left\{\sigma\left((\boldsymbol{W}_0^{[1]}\boldsymbol{x}_{\mathrm{i}})\right)_{\mathrm{j}}^2\right\}_{\mathrm{j}=1}^{m}$ is a collection of i.i.d. sub-exponential variables. Consequently, for all $t > 0$,

$$\mathbb{P}\left(\left|\left\|\boldsymbol{x}_{\mathrm{i}}^{[1]}(0)\right\|_2^2 - \widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[1]}\right| \geq t\right) \leq \exp\left(-cmt^2\right),$$

hence,

$$\mathbb{P}\left(\left|\left\|\boldsymbol{x}_{\mathrm{i}}^{[1]}(0)\right\|_2^2 - \widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[1]}\right| \geq 2t\right) \leq \mathbb{P}\left(\left|\left\|\boldsymbol{x}_{\mathrm{i}}^{[1]}(0)\right\|_2 - \sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[1]}}\right| \geq t\right) \leq \exp\left(-cmt^2\right).$$

We have shown already for $l = 1$, Equation (B.4) holds. We need to show further that Equation (B.5) holds.

We shall note that

$$\left\langle \frac{\boldsymbol{x}_{\mathrm{i}}^{[1]}(0)}{\sqrt{m}}, \mathbf{1}\right\rangle = \frac{\sqrt{c_\sigma}}{m}\sum_{\mathrm{j}=1}^{m}\sigma\left((\boldsymbol{W}_0^{[1]}\boldsymbol{x}_{\mathrm{i}})_{\mathrm{j}}\right),$$

and

$$\mathbb{E}\left[\left\langle \frac{\boldsymbol{x}_{\mathrm{i}}^{[1]}(0)}{\sqrt{m}}, \mathbf{1}\right\rangle\right] = \widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[1]}.$$

Since $\boldsymbol{x}_{\mathrm{i}}^{[1]}(0)$ can be written into the form

$$\boldsymbol{x}_{\mathrm{i}}^{[1]}(0) = \sqrt{\frac{c_\sigma}{m}}\sigma\left(\boldsymbol{X}\right),$$

where $\boldsymbol{X}$ is a standard normal Gaussian vector.

We shall focus on the inner product function $g^{[1]}(\cdot) : \mathbb{R}^m \to \mathbb{R}$,

$$g^{[1]}(\boldsymbol{X}) = \frac{\sqrt{c_\sigma}}{m} \langle \sigma(\boldsymbol{X}), \mathbf{1} \rangle .$$

We show that $g^{[1]}(\cdot)$ is also Lipschitz with respect to the Euclidean norm $\|\cdot\|_2$:

For any $\boldsymbol{X}_1, \boldsymbol{X}_2 \in \mathbb{R}^m$:

$$\begin{aligned} \left| g^{[1]}(\boldsymbol{X}_1) - g^{[1]}(\boldsymbol{X}_2) \right| &\leq \left| \frac{\sqrt{c_\sigma}}{m} \langle \sigma(\boldsymbol{X}_1), \mathbf{1} \rangle - \frac{\sqrt{c_\sigma}}{m} \langle \sigma(\boldsymbol{X}_2), \mathbf{1} \rangle \right| \\ &\leq \frac{\sqrt{c_\sigma}}{m} |\langle \boldsymbol{X}_1 - \boldsymbol{X}_2, \mathbf{1} \rangle| \leq \sqrt{\frac{c_\sigma}{m}} \|\boldsymbol{X}_1 - \boldsymbol{X}_2\|_2 . \end{aligned}$$

Hence $g^{[1]}(\cdot)$ is $\frac{C}{\sqrt{m}}$-Lipschitz. Apply Lemma B.1.1 directly, we have

$$\mathbb{P}\left(\left| g^{[1]}(\boldsymbol{X}) - \mathbb{E}\, g^{[1]}(\boldsymbol{X}) \right| \geq t \right) \leq \exp(-cmt^2),$$

then

$$\mathbb{P}\left(\left| \left\langle \frac{\boldsymbol{x}_{\mathrm{i}}^{[1]}(0)}{\sqrt{m}}, \mathbf{1} \right\rangle - \widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[1]} \right| \geq t \right) \leq \exp(-cmt^2).$$

(ii). Our next step is to prove that Equation (B.4) and Equation (B.5) hold for $l \geq 2$, and we will prove it by induction. Assume that Equation (B.4) and Equation (B.5) hold true for $1, 2, 3, \cdots, l$, and we show further that it is also the case for $l+1$, i.e.,

$$\mathbb{P}\left(\left| \left\| \boldsymbol{x}_{\mathrm{i}}^{[l+1]}(0) \right\|_2 - \sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l+1]}} \right| \geq t \right) \leq \exp\left(-cmt^2\right), \tag{B.6}$$

$$\mathbb{P}\left(\left| \left\langle \frac{\boldsymbol{x}_{\mathrm{i}}^{[l+1]}(0)}{\sqrt{m}}, \mathbf{1} \right\rangle - \widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l+1]} \right| \geq t \right) \leq \exp\left(-cmt^2\right). \tag{B.7}$$

The structure of outputs of each layer at initial state shall be recalled as follows

$$\boldsymbol{x}_{\mathrm{i}}^{[l+1]}(0) = \boldsymbol{x}_{\mathrm{i}}^{[l]}(0) + \frac{c_{\mathrm{res}}}{L\sqrt{m}} \sigma\left(\boldsymbol{W}_0^{[l+1]} \boldsymbol{x}_{\mathrm{i}}^{[l]}(0) \right),$$

and the counterpart of $\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]}$ and $\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l]}$

$$\begin{aligned}
\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l+1]} &= \widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]} + \mathbb{E}_{u\sim\mathcal{N}(0,\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]})}\left[\frac{c_{\mathrm{res}}\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l]}\sigma(u)}{L} + \frac{c_{\mathrm{res}}\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l]}\sigma(u)}{L} + \frac{c_{\mathrm{res}}^2\sigma(u)\sigma(u)}{L^2}\right], \\
\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l+1]} &= \widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l]} + \frac{c_{\mathrm{res}}}{L}\,\mathbb{E}_{u\sim\mathcal{N}(0,\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]})}\left[\sigma(u)\right],
\end{aligned}$$

then

$$\begin{aligned}
\left\|\boldsymbol{x}_{\mathrm{i}}^{[l+1]}(0)\right\|_2^2 = \left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2^2 + \frac{2c_{\mathrm{res}}}{L}\underbrace{\left\langle\frac{\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)}{\sqrt{m}}, \sigma\left(\boldsymbol{W}_0^{[l+1]}\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right)\right\rangle}_{\mathrm{I}} \\
+ \frac{c_{\mathrm{res}}^2}{L^2}\underbrace{\frac{1}{m}\left\langle\sigma\left(\boldsymbol{W}_0^{[l+1]}\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right), \sigma\left(\boldsymbol{W}_0^{[l+1]}\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right)\right\rangle}_{\mathrm{II}}.
\end{aligned}$$

Based on our induction hypothesis, there exits estimates on $\left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2^2$. We need to focus on terms I and II. Firstly for term I, there is a $\frac{1}{\sqrt{m}}$ scaling factor contained in $\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)$, and $\sigma\left(\boldsymbol{W}_0^{[l+1]}\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right)$ has distribution

$$\sigma\left(\boldsymbol{W}_0^{[l+1]}\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right) \sim \sigma\left(\left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2 \boldsymbol{Y}\right),$$

with $\boldsymbol{Y}$ being a standard normal Gaussian vector, then

$$\mathbb{E}\left[\frac{1}{\sqrt{m}}\left\langle\boldsymbol{x}_{\mathrm{i}}^{[l]}(0), \sigma\left(\boldsymbol{W}_0^{[l+1]}\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right)\right\rangle\right] = \frac{1}{\sqrt{m}}\left\langle\boldsymbol{x}_{\mathrm{i}}^{[l]}(0), \mathbb{E}\left[\sigma\left(\left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2 \boldsymbol{Y}\right)\right]\right\rangle.$$

We shall focus on the inner product function $g^{[l]}(\cdot) : \mathbb{R}^m \to \mathbb{R}$,

$$g^{[l]}(\boldsymbol{Y}) = \frac{1}{\sqrt{m}}\left\langle\boldsymbol{x}_{\mathrm{i}}^{[l]}(0), \sigma\left(\left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2 \boldsymbol{Y}\right)\right\rangle,$$

then for any $\boldsymbol{Y}_1, \boldsymbol{Y}_2 \in \mathbb{R}^m$:

$$\begin{aligned}
\left|g^{[l]}(\boldsymbol{Y}_1) - g^{[l]}(\boldsymbol{Y}_2)\right| &\le \frac{1}{\sqrt{m}}\left|\left\langle\boldsymbol{x}_{\mathrm{i}}^{[l]}(0), \sigma\left(\left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2 \boldsymbol{Y}_1\right)\right\rangle - \left\langle\boldsymbol{x}_{\mathrm{i}}^{[l]}(0), \sigma\left(\left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2 \boldsymbol{Y}_2\right)\right\rangle\right| \\
&\le \frac{1}{\sqrt{m}}\left|\left\langle\boldsymbol{x}_{\mathrm{i}}^{[l]}(0), \left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2 |\boldsymbol{Y}_1 - \boldsymbol{Y}_2|\right\rangle\right| \le \frac{1}{\sqrt{m}}\left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2^2 \|\boldsymbol{Y}_1 - \boldsymbol{Y}_2\|_2.
\end{aligned}$$

Based on our induction hypothesis, with high probability, there exists a uniform constant $C > 0$, such that for $1, 2, \cdots, l$, $\left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2 \leq C$. Hence $g^{[l]}(\cdot)$ is $\frac{C}{\sqrt{m}}$-Lipschitz, apply Lemma B.1.1 once again, we have

$$\begin{aligned}&\mathbb{P}\left(\left|\frac{1}{\sqrt{m}}\left\langle \boldsymbol{x}_{\mathrm{i}}^{[l]}(0), \sigma\left(\boldsymbol{W}_0^{[l+1]}\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right)\right\rangle - \frac{1}{\sqrt{m}}\left\langle \boldsymbol{x}_{\mathrm{i}}^{[l]}(0), \mathbb{E}\left[\sigma\left(\left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2 \boldsymbol{Y}\right)\right]\right\rangle\right| \geq t\right) \\ &\leq \exp(-cmt^2).\end{aligned} \tag{B.8}$$

From our induction hypothesis

$$\mathbb{P}\left(\left|\left\langle \frac{\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)}{\sqrt{m}}, \mathbf{1}\right\rangle - \widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l]}\right| \geq t\right) \leq \exp\left(-cmt^2\right),$$

then differ by a multiplication of constant, the inequality

$$\begin{aligned}&\mathbb{P}\left(\left|\frac{1}{\sqrt{m}}\left\langle \boldsymbol{x}_{\mathrm{i}}^{[l]}(0), \mathbb{E}\left[\sigma\left(\left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2 \boldsymbol{Y}\right)\right]\right\rangle - \widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l]}\, \mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2 u\right)\right]\right| \geq t\right) \\ &\leq \exp(-cmt^2)\end{aligned} \tag{B.9}$$

holds. From Lemma B.1.2,

$$\left|\mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2 u\right)\right] - \mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]}}u\right)\right]\right| \leq C\left|\left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2 - \sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]}}\right|,$$

naturally, we have

$$\begin{aligned}&\left|\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l]}\, \mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2 u\right)\right] - \widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l]}\, \mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]}}u\right)\right]\right| \\ &\leq C\left|\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l]}\right|\left|\left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2 - \sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]}}\right|.\end{aligned} \tag{B.10}$$

Once again from our induction hypothesis,

$$\mathbb{P}\left(\left|\left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2 - \sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]}}\right| \geq t\right) \leq \exp\left(-cmt^2\right)$$

holds, then from Equation (B.10), we obtain that

$$\begin{aligned}&\mathbb{P}\left(\left|\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l]}\,\mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2 u\right)\right]-\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l]}\,\mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]}}u\right)\right]\right|\geq t\right)\\&\leq\mathbb{P}\left(\left|C\left|\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l]}\right|\left|\left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2-\sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]}}\right|\right|\geq t\right)\leq\exp(-cmt^2).\end{aligned}\tag{B.11}$$

Combine altogether Equation (B.8), Equation (B.9) and Equation (B.11),

$$\begin{aligned}&\mathbb{P}\left(\left|\frac{1}{\sqrt{m}}\left\langle\boldsymbol{x}_{\mathrm{i}}^{[l]}(0),\sigma\left(\boldsymbol{W}_0^{[l+1]}\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right)\right\rangle-\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l]}\,\mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]}}u\right)\right]\right|\geq t\right)\\&\leq\mathbb{P}\left(\left|\frac{1}{\sqrt{m}}\left\langle\boldsymbol{x}_{\mathrm{i}}^{[l]}(0),\sigma\left(\boldsymbol{W}_0^{[l+1]}\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right)\right\rangle-\frac{1}{\sqrt{m}}\left\langle\boldsymbol{x}_{\mathrm{i}}^{[l]}(0),\mathbb{E}\left[\sigma\left(\left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2\boldsymbol{Y}\right)\right]\right\rangle\right|\geq\frac{t}{3}\right)\\&+\mathbb{P}\left(\left|\frac{1}{\sqrt{m}}\left\langle\boldsymbol{x}_{\mathrm{i}}^{[l]}(0),\mathbb{E}\left[\sigma\left(\left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2\boldsymbol{Y}\right)\right]\right\rangle-\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l]}\,\mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2 u\right)\right]\right|\geq\frac{t}{3}\right)\\&+\mathbb{P}\left(\left|\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l]}\,\mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2 u\right)\right]-\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l]}\,\mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]}}u\right)\right]\right|\geq\frac{t}{3}\right)\\&\leq\exp(-cmt^2).\end{aligned}\tag{B.12}$$

Secondly for term II, notice that $\sigma\left(\boldsymbol{W}_0^{[l+1]}\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right)$ has distribution

$$\sigma\left(\boldsymbol{W}_0^{[l+1]}\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right)\sim\sigma\left(\left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2\boldsymbol{Y}\right),$$

with $\boldsymbol{Y}$ being a standard normal Gaussian vector, then

$$\mathbb{E}\left[\frac{1}{m}\left\langle\sigma\left(\boldsymbol{W}_0^{[l+1]}\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right),\sigma\left(\boldsymbol{W}_0^{[l+1]}\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right)\right\rangle\right]=\mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2 u\right)^2\right],$$

we recall once again that $\left\{\sigma\left(\left(\boldsymbol{W}_0^{[1]}\boldsymbol{x}_{\mathrm{i}}\right)_{\mathrm{j}}\right\}_{\mathrm{j}=1}^{m}$ is a collection of i.i.d. sub-Gaussian variables, $\left\{\sigma\left(\left(\boldsymbol{W}_0^{[1]}\boldsymbol{x}_{\mathrm{i}}\right)_{\mathrm{j}}\right)^2\right\}_{\mathrm{j}=1}^{m}$ sub-exponential. Consequently, for all $t>0$,

$$\begin{aligned}&\mathbb{P}\left(\left|\frac{1}{m}\left\langle\sigma\left(\boldsymbol{W}_0^{[l+1]}\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right),\sigma\left(\boldsymbol{W}_0^{[l+1]}\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right)\right\rangle-\mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2 u\right)^2\right]\right|\geq t\right)\\&\leq\exp\left(-cmt^2\right)\end{aligned}\tag{B.13}$$

holds. From Lemma B.1.3,

$$\begin{aligned}&\left|\mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2 u\right)^2\right]-\mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]}}u\right)^2\right]\right|\\&\leq C\left|\left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2-\sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]}}\right|.\end{aligned}\tag{B.14}$$

Based on our induction hypothesis,

$$\mathbb{P}\left(\left|\left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2-\sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]}}\right|\geq t\right)\leq\exp\left(-cmt^2\right)$$

holds, then from Equation (B.14), we obtain that

$$\begin{aligned}&\mathbb{P}\left(\left|\mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2 u\right)^2\right]-\mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]}}u\right)^2\right]\right|\geq t\right)\\&\leq\mathbb{P}\left(\left|C\left|\left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2-\sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]}}\right|\right|\geq t\right)\leq\exp\left(-cmt^2\right).\end{aligned}\tag{B.15}$$

Combine altogether Equation (B.13) and Equation (B.15), since

$$\begin{aligned}&\mathbb{P}\left(\left|\frac{1}{m}\left\langle\sigma\left(\boldsymbol{W}_0^{[l+1]}\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right),\sigma\left(\boldsymbol{W}_0^{[l+1]}\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right)\right\rangle-\mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]}}u\right)^2\right]\right|\geq t\right)\\&\leq\mathbb{P}\left(\left|\frac{1}{m}\left\langle\sigma\left(\boldsymbol{W}_0^{[l+1]}\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right),\sigma\left(\boldsymbol{W}_0^{[l+1]}\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right)\right\rangle-\mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2 u\right)^2\right]\right|\geq\frac{t}{2}\right)\\&+\mathbb{P}\left(\left|\mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2 u\right)^2\right]-\mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]}}u\right)^2\right]\right|\geq\frac{t}{2}\right),\end{aligned}$$

then

$$\begin{aligned}&\mathbb{P}\left(\left|\frac{1}{m}\left\langle\sigma\left(\boldsymbol{W}_0^{[l+1]}\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right),\sigma\left(\boldsymbol{W}_0^{[l+1]}\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right)\right\rangle-\mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]}}u\right)^2\right]\right|\geq t\right)\\&\leq\exp\left(-cmt^2\right).\end{aligned}\tag{B.16}$$

We recall once again that

$$\left\|\boldsymbol{x}_{\mathrm{i}}^{[l+1]}(0)\right\|_2^2=\left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2^2+\frac{2c_{\mathrm{res}}}{L}\underbrace{\left\langle\frac{\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)}{\sqrt{m}},\sigma\left(\boldsymbol{W}_0^{[l+1]}\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right)\right\rangle}_{\mathrm{I}}$$
$$+\frac{c_{\mathrm{res}}^2}{L^2}\underbrace{\frac{1}{m}\left\langle\sigma\left(\boldsymbol{W}_0^{[l+1]}\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right),\sigma\left(\boldsymbol{W}_0^{[l+1]}\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right)\right\rangle}_{\mathrm{II}},$$

where term I is close to $\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l]}\,\mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]}}u\right)\right]$, as is shown in Equation (B.12)

$$\begin{aligned}&\mathbb{P}\left(\left|\frac{1}{\sqrt{m}}\left\langle\boldsymbol{x}_{\mathrm{i}}^{[l]}(0),\sigma\left(\boldsymbol{W}_0^{[l+1]}\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right)\right\rangle-\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l]}\,\mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]}}u\right)\right]\right|\geq t\right)\\&\leq\exp(-cmt^2),\end{aligned}\tag{B.17}$$

and term II is close to $\mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]}}u\right)^2\right]$, as is shown in Equation (B.16)

$$\begin{aligned}&\mathbb{P}\left(\left|\frac{1}{m}\left\langle\sigma\left(\boldsymbol{W}_0^{[l+1]}\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right),\sigma\left(\boldsymbol{W}_0^{[l+1]}\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right)\right\rangle-\mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]}}u\right)^2\right]\right|\geq t\right)\\&\leq\exp\left(-cmt^2\right),\end{aligned}\tag{B.18}$$

then

$$\begin{aligned}&\mathbb{P}\left(\left|\left\|\boldsymbol{x}_{\mathrm{i}}^{[l+1]}(0)\right\|_2^2-\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l+1]}\right|\geq t\left(1+\frac{c_{\mathrm{res}}}{L}\right)^2\right)\\&\leq\mathbb{P}\left(\left|\left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2^2-\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]}\right|\geq t\right)\\&+\mathbb{P}\left(2\frac{c_{\mathrm{res}}}{L}\left|\left\langle\frac{\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)}{\sqrt{m}},\sigma\left(\boldsymbol{W}_0^{[l+1]}\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right)\right\rangle-\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l]}\,\mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]}}u\right)\right]\right|\geq 2\frac{c_{\mathrm{res}}}{L}t\right)\\&+\mathbb{P}\left(\frac{c_{\mathrm{res}}^2}{L^2}\left|\frac{1}{m}\left\langle\sigma\left(\boldsymbol{W}_0^{[l+1]}\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right),\sigma\left(\boldsymbol{W}_0^{[l+1]}\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right)\right\rangle-\mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]}}u\right)^2\right]\right|\geq\frac{c_{\mathrm{res}}^2}{L^2}t\right)\\&\leq\exp\left(-cmt^2\right).\end{aligned}$$

We shall remark that thanks to the $\frac{c_{\text{res}}}{L}$ structure, with high probability, the difference $\left|\left\|\boldsymbol{x}_{\text{i}}^{[l]}(0)\right\|_2^2 - \widetilde{\boldsymbol{K}}_{\text{ii}}^{[l]}\right|$ does not explode exponentially with respect to the number of layer. From inequality above, we have

$$\mathbb{P}\left(\left|\left\|\boldsymbol{x}_{\text{i}}^{[l+1]}(0)\right\|_2^2 - \widetilde{\boldsymbol{K}}_{\text{ii}}^{[l+1]}\right| \geq t\left(1+\frac{c_{\text{res}}}{L}\right)^2\right) \leq \exp\left(-cmt^2\right),$$

if we choose c_{res} and t smartly, set $t^* = t\left(1+\frac{c_{\text{res}}}{L}\right)^2$, then $t = t^*\left(1+\frac{c_{\text{res}}}{L}\right)^{-2}$, there exists a uniform constant c^*, such that

$$1/c^* \leq \left(1+\frac{c_{\text{res}}}{L}\right)^{-2L} \leq c^*,$$

then

$$\mathbb{P}\left(\left|\left\|\boldsymbol{x}_{\text{i}}^{[l+1]}(0)\right\|_2^2 - \widetilde{\boldsymbol{K}}_{\text{ii}}^{[l+1]}\right| \geq t^*\right) \leq \exp\left(-cmt^{*2}\right), \tag{B.19}$$

which finishes the proof of Equation (B.6).

Finally, we need to prove Equation (B.7) for $\widetilde{\boldsymbol{b}}_{\text{i}}^{[l+1]}$. The structure of outputs of each layer at initial state shall be recalled once again,

$$\boldsymbol{x}_{\text{i}}^{[l+1]}(0) = \boldsymbol{x}_{\text{i}}^{[l]}(0) + \frac{c_{\text{res}}}{L\sqrt{m}}\sigma\left(\boldsymbol{W}_0^{[l+1]}\boldsymbol{x}_{\text{i}}^{[l]}(0)\right),$$

and the counterpart of $\widetilde{\boldsymbol{b}}_{\text{i}}^{[l]}$

$$\widetilde{\boldsymbol{b}}_{\text{i}}^{[l+1]} = \widetilde{\boldsymbol{b}}_{\text{i}}^{[l]} + \frac{c_{\text{res}}}{L}\,\mathbb{E}_{u\sim\mathcal{N}(0,\widetilde{\boldsymbol{K}}_{\text{ii}}^{[l]})}\left[\sigma(u)\right].$$

Apply Lemma B.1.1,

$$\mathbb{P}\left(\left|\left\langle\frac{\sigma\left(\boldsymbol{W}_0^{[l+1]}\boldsymbol{x}_{\text{i}}^{[l]}(0)\right)}{m}, \mathbf{1}\right\rangle - \mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\left\|\boldsymbol{x}_{\text{i}}^{[l]}(0)\right\|_2 u\right)\right]\right| \geq t\right) \leq \exp\left(-cmt^2\right). \tag{B.20}$$

Apply Lemma B.1.3,

$$\begin{aligned}&\left|\mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2 u\right)^2\right]-\mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]}}u\right)^2\right]\right|\\&\leq C\left|\left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2-\sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]}}\right|,\end{aligned}\tag{B.21}$$

from induction hypothesis, since

$$\mathbb{P}\left(\left|\left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2-\sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]}}\right|\geq t\right)\leq\exp\left(-cmt^2\right)$$

holds, then from Equation (B.21), we obtain that

$$\begin{aligned}&\mathbb{P}\left(\left|\mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2 u\right)^2\right]-\mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]}}u\right)^2\right]\right|\geq t\right)\\&\leq\mathbb{P}\left(\left|C\left|\left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2-\sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]}}\right|\right|\geq t\right)\leq\exp\left(-cmt^2\right).\end{aligned}\tag{B.22}$$

combine Equation (B.20) and Equation (B.22),

$$\begin{aligned}&\mathbb{P}\left(\left|\left\langle\frac{\sigma\left(\boldsymbol{W}_0^{[l+1]}\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right)}{m},\mathbf{1}\right\rangle-\mathbb{E}\left[\sigma\left(\sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]}}\boldsymbol{Y}\right)\right]\right|\geq t\right)\\&\leq\mathbb{P}\left(\left|\left\langle\frac{\sigma\left(\boldsymbol{W}_0^{[l+1]}\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right)}{m},\mathbf{1}\right\rangle-\mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2 u\right)\right]\right|\geq\frac{t}{2}\right)\\&+\mathbb{P}\left(\left|\mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\left\|\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right\|_2 u\right)^2\right]-\mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]}}u\right)^2\right]\right|\geq\frac{t}{2}\right)\\&\leq\exp\left(-cmt^2\right),\end{aligned}\tag{B.23}$$

then

$$\begin{aligned}
&\mathbb{P}\left(\left|\left\langle \frac{\boldsymbol{x}_{\mathrm{i}}^{[l+1]}(0)}{\sqrt{m}}, \mathbf{1}\right\rangle - \widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l+1]}\right| \geq t\left(1+\frac{c_{\mathrm{res}}}{L}\right)\right) \\
&\leq \mathbb{P}\left(\left|\left\langle \frac{\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)}{\sqrt{m}}, \mathbf{1}\right\rangle - \widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l]}\right| \geq t\right) \\
&+ \mathbb{P}\left(\frac{c_{\mathrm{res}}}{L}\left|\left\langle \frac{\sigma\left(\boldsymbol{W}_0^{[l+1]}\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)\right)}{m}, \mathbf{1}\right\rangle - \mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[l]}}u\right)^2\right]\right| \geq \frac{c_{\mathrm{res}}}{L}t\right) \\
&\leq \exp\left(-cmt^2\right).
\end{aligned} \tag{B.24}$$

We shall see once again that thanks to the $\frac{c_{\mathrm{res}}}{L}$ structure, with high probability, the difference $\left|\left\langle \frac{\boldsymbol{x}_{\mathrm{i}}^{[l]}(0)}{\sqrt{m}}, \mathbf{1}\right\rangle - \widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l]}\right|$ only has linear increment with respect to the number of layer. If we choose c_{res} and t smartly, set $t^* = t\left(1+\frac{c_{\mathrm{res}}}{L}\right)$, then $t = t^*\left(1+\frac{c_{\mathrm{res}}}{L}\right)^{-1}$, there exists a uniform constant c^*, such that

$$1/c^* \leq \left(1+\frac{c_{\mathrm{res}}}{L}\right)^{-L} \leq c^*,$$

then

$$\mathbb{P}\left(\left|\left\langle \frac{\boldsymbol{x}_{\mathrm{i}}^{[l+1]}(0)}{\sqrt{m}}, \mathbf{1}\right\rangle - \widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[l+1]}\right| \geq t^*\right) \leq \exp\left(-cmt^{*2}\right), \tag{B.25}$$

which finishes the proof of Equation (B.7). □

B.3 Analysis on Random Initialization

Our next proposition is on the least eigenvalue of the randomly initialized Gram matrix. First, we shall denote the series of randomly initialized Gram matrices by $\left\{\boldsymbol{G}^{[l]}(0)\right\}_{l=1}^{L+1}$, whose components read:

$$\boldsymbol{G}_{\mathrm{ij}}^{[l]}(0) = \left\langle \boldsymbol{x}_{\mathrm{i}}^{[1]}(0), \boldsymbol{x}_{\mathrm{j}}^{[1]}(0)\right\rangle. \tag{B.26}$$

Proposition B.3.1. *Given $\mathcal{X}$, $\sigma(\cdot)$, $\left\{\widetilde{\boldsymbol{K}}^{[l]}\right\}_{l=1}^{L}$, and $\left\{\widetilde{\boldsymbol{b}}^{[l]}\right\}_{l=1}^{L}$, set $\lambda_0 > 0$ as*

$$\lambda_{\min}\left(\widetilde{\boldsymbol{K}}^{[1]} - \widetilde{\boldsymbol{b}}^{[1]} \otimes \left(\widetilde{\boldsymbol{b}}^{[1]}\right)^{\mathsf{T}}\right) = \lambda_0,$$

there exists a small constant $\varepsilon > 0$, such that with high probability w.r.t random initialization, for $m = \Omega\left(\left(\frac{n}{\lambda_0}\right)^{2+\varepsilon}\right)$, then

$$\lambda_{\min}\left(\boldsymbol{G}^{[1]}(0)\right) \geq \frac{3\lambda_0}{4}. \tag{B.27}$$

Proof. We have that

$$\begin{aligned}
\boldsymbol{G}_{\mathrm{ij}}^{[1]}(0) &= \left\langle \boldsymbol{x}_{\mathrm{i}}^{[1]}(0), \boldsymbol{x}_{\mathrm{j}}^{[1]}(0) \right\rangle \\
\widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[1]} &= c_\sigma \, \mathbb{E}_{(u,v)^\intercal \sim \mathcal{N}\left(\mathbf{0}, \widetilde{\boldsymbol{A}}_{\mathrm{ij}}^{[1]}\right)} \left[\sigma(u)\sigma(v)\right].
\end{aligned}$$

Now we need to apply Lemma B.1.1 again, except that we are going to apply it to the inner product function $h^{[1]}(\cdot) : \mathbb{R}^{2m} \to \mathbb{R}$,

$$h^{[1]}(\boldsymbol{Z}) = \frac{c_\sigma}{m} \left\langle \sigma(\boldsymbol{X}), \sigma(\rho \boldsymbol{X} + \sqrt{1-\rho^2}\boldsymbol{Y}) \right\rangle,$$

where $\boldsymbol{Z}^\intercal = (\boldsymbol{X}^\intercal, \boldsymbol{Y}^\intercal)$, $\boldsymbol{X}, \boldsymbol{Y}$ are standard normal Gaussian vectors, and $-1 \leq \rho \leq 1$.

For any $\boldsymbol{Z}_1, \boldsymbol{Z}_2 \in \mathbb{R}^m$,

$$\begin{aligned}
\left|h^{[1]}(\boldsymbol{Z}_1) - h^{[1]}(\boldsymbol{Z}_2)\right| \leq{}& \sqrt{\frac{c_\sigma}{m}} \left\|\sigma(\rho \boldsymbol{X}_1 + \sqrt{1-\rho^2}\boldsymbol{Y}_1)\right\|_2 \sqrt{\frac{c_\sigma}{m}} \left\|\boldsymbol{X}_1 - \boldsymbol{X}_2\right\|_2 \\
&+ \sqrt{\frac{c_\sigma}{m}} \left\|\sigma(\boldsymbol{X}_2)\right\|_2 \sqrt{\frac{c_\sigma}{m}} \left(|\rho| \left\|\boldsymbol{X}_1 - \boldsymbol{X}_2\right\|_2 + \sqrt{1-\rho^2} \left\|\boldsymbol{Y}_1 - \boldsymbol{Y}_2\right\|_2\right),
\end{aligned}$$

combined with Proposition B.2.1, with high probability, there exists a uniform constant $C > 0$, such that

$$\sqrt{\frac{c_\sigma}{m}} \left\|\sigma(\rho \boldsymbol{X}_1 + \sqrt{1-\rho^2}\boldsymbol{Y}_1)\right\|_2, \ \sqrt{\frac{c_\sigma}{m}} \left\|\sigma(\boldsymbol{X}_2)\right\|_2 \leq C.$$

So we have

$$\left|h^{[1]}(\boldsymbol{Z}_1) - h^{[1]}(\boldsymbol{Z}_2)\right| \leq 4C\sqrt{\frac{c_\sigma}{m}} \left\|\boldsymbol{Z}_1 - \boldsymbol{Z}_2\right\|_2,$$

hence $h^{[1]}(\boldsymbol{Z})$ is $4C\sqrt{\frac{c_\sigma}{m}}$-Lipschitz. Then set $\rho = \widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[0]}$, for all $1 \leq \mathrm{i}, \mathrm{j} \leq m$,

$$\mathbb{P}\left(\left|\boldsymbol{G}_{\mathrm{ij}}^{[1]}(0) - \widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[1]}\right| \geq t\right) \leq \exp(-cmt^2).$$

Note that based on Proposition A.2.1, $\lambda_{\min}\left(\widetilde{\boldsymbol{K}}^{[1]}\right) \geq \lambda_0$, also we have

$$\left\|\boldsymbol{G}^{[1]}(0) - \widetilde{\boldsymbol{K}}^{[1]}\right\|_{2\to 2} \leq \left\|\boldsymbol{G}^{[1]}(0) - \widetilde{\boldsymbol{K}}^{[1]}\right\|_{\mathrm{F}} \leq n\left\|\boldsymbol{G}^{[1]}(0) - \widetilde{\boldsymbol{K}}^{[1]}\right\|_{\infty},$$

then if we choose $t = \frac{\lambda_0}{4n}$ and with a union of m^2 such events, we have with probability $1 - m^2\exp\left(-cm\lambda_0^2/n^2\right)$,

$$\left\|\boldsymbol{G}^{[1]}(0) - \widetilde{\boldsymbol{K}}^{[1]}\right\|_{2\to 2} \leq \frac{\lambda_0}{4}$$

holds. Hence if there exists a small constant $\varepsilon > 0$, such that $m = \Omega\left(\left(\frac{n}{\lambda_0}\right)^{2+\varepsilon}\right)$, then with probability $1 - \exp(-m^{\varepsilon})$,

$$\lambda_{\min}(\boldsymbol{G}^{[1]}(0)) \geq \lambda_{\min}(\widetilde{\boldsymbol{K}}^{[1]}) - \left\|\boldsymbol{G}^{[1]}(0) - \widetilde{\boldsymbol{K}}^{[1]}\right\|_{2\to 2} \geq \frac{3\lambda_0}{4}. \tag{B.28}$$

□

Our next proposition is on the least eigenvalue of other randomly initialized Gram matrices $\boldsymbol{G}^{[l]}(0),\ l \neq 1$.

Proposition B.3.2. *Given $\mathcal{X}$, $\sigma(\cdot)$, $\left\{\widetilde{\boldsymbol{K}}^{[l]}\right\}_{l=1}^{L}$, and $\left\{\widetilde{\boldsymbol{b}}^{[l]}\right\}_{l=1}^{L}$, set $\lambda_0 > 0$ as*

$$\lambda_{\min}\left(\widetilde{\boldsymbol{K}}^{[1]} - \widetilde{\boldsymbol{b}}^{[1]} \otimes \left(\widetilde{\boldsymbol{b}}^{[1]}\right)^{\intercal}\right) = \lambda_0,$$

there exists a small constant $\varepsilon > 0$, such that with high probability w.r.t random initialization, for $m = \Omega\left(\left(\frac{n}{\lambda_0}\right)^{2+\varepsilon}\right)$, then

$$\lambda_{\min}\left(\boldsymbol{G}^{[l]}(0)\right) \geq \frac{3\lambda_0}{4}, \quad 2 \leq l \leq L. \tag{B.29}$$

Proof. (i). For $l=2$, we shall make estimate on norm $\left\|\boldsymbol{G}^{[2]}(0)-\widetilde{\boldsymbol{K}}^{[2]}\right\|_{\infty}$. Since by definition,

$$\begin{aligned}
\boldsymbol{G}_{\mathrm{ij}}^{[2]}(0) &= \left\langle \boldsymbol{x}_{\mathrm{i}}^{[2]}(0), \boldsymbol{x}_{\mathrm{j}}^{[2]}(0)\right\rangle = \boldsymbol{G}_{\mathrm{ij}}^{[1]}(0) + \frac{c_{\mathrm{res}}}{L}\underbrace{\frac{1}{\sqrt{m}}\left\langle \boldsymbol{x}_{\mathrm{i}}^{[1]}(0), \sigma\left(\boldsymbol{W}_0^{[2]}\boldsymbol{x}_{\mathrm{j}}^{[1]}(0)\right)\right\rangle}_{\mathrm{I}} \\
&\quad + \frac{c_{\mathrm{res}}}{L}\underbrace{\frac{1}{\sqrt{m}}\left\langle \boldsymbol{x}_{\mathrm{j}}^{[1]}(0), \sigma\left(\boldsymbol{W}_0^{[2]}\boldsymbol{x}_{\mathrm{i}}^{[1]}(0)\right)\right\rangle}_{\mathrm{II}} + \frac{c_{\mathrm{res}}^2}{L^2}\underbrace{\frac{1}{m}\left\langle \sigma\left(\boldsymbol{W}_0^{[2]}\boldsymbol{x}_{\mathrm{i}}^{[1]}(0)\right), \sigma\left(\boldsymbol{W}_0^{[2]}\boldsymbol{x}_{\mathrm{j}}^{[1]}(0)\right)\right\rangle}_{\mathrm{III}} \\
\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[1]} &= \sqrt{c_\sigma}\,\mathbb{E}_{u\sim\mathcal{N}(0,\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[0]})}\left[\sigma(u)\right], \\
\widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[2]} &= \widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[1]} + \mathbb{E}_{(u,v)^\intercal\sim\mathcal{N}\left(\boldsymbol{0},\widetilde{\boldsymbol{A}}_{\mathrm{ij}}^{[2]}\right)}\left[\frac{c_{\mathrm{res}}}{L}\underbrace{\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[1]}\sigma(v)}_{\mathrm{I'}} + \frac{c_{\mathrm{res}}}{L}\underbrace{\widetilde{\boldsymbol{b}}_{\mathrm{j}}^{[1]}\sigma(u)}_{\mathrm{II'}} + \frac{c_{\mathrm{res}}^2}{L^2}\underbrace{\sigma(u)\sigma(v)}_{\mathrm{III'}}\right].
\end{aligned}$$

We need to tackle the difference between I, II, III and I', II', III'. For I and I', we need to write the difference into:

$$\begin{aligned}
&\left|\frac{1}{\sqrt{m}}\left\langle \boldsymbol{x}_{\mathrm{i}}^{[1]}(0), \sigma\left(\boldsymbol{W}_0^{[2]}\boldsymbol{x}_{\mathrm{j}}^{[1]}(0)\right)\right\rangle - \widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[1]}\,\mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{jj}}^{[1]}}u\right)\right]\right| \\
&\le \left|\frac{1}{\sqrt{m}}\left\langle \boldsymbol{x}_{\mathrm{i}}^{[1]}(0), \sigma\left(\boldsymbol{W}_0^{[2]}\boldsymbol{x}_{\mathrm{j}}^{[1]}(0)\right)\right\rangle - \frac{1}{\sqrt{m}}\left\langle \boldsymbol{x}_{\mathrm{i}}^{[1]}(0), \mathbb{E}\left[\sigma\left(\left\|\boldsymbol{x}_{\mathrm{j}}^{[1]}(0)\right\|_2\boldsymbol{Y}\right)\right]\right\rangle\right| \\
&+ \left|\frac{1}{\sqrt{m}}\left\langle \boldsymbol{x}_{\mathrm{i}}^{[1]}(0), \mathbb{E}\left[\sigma\left(\left\|\boldsymbol{x}_{\mathrm{j}}^{[1]}(0)\right\|_2\boldsymbol{Y}\right)\right]\right\rangle - \widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[1]}\,\mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\left\|\boldsymbol{x}_{\mathrm{j}}^{[1]}(0)\right\|_2 u\right)\right]\right| \\
&+ \left|\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[1]}\,\mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\left\|\boldsymbol{x}_{\mathrm{j}}^{[1]}(0)\right\|_2 u\right)\right] - \widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[1]}\,\mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{jj}}^{[1]}}u\right)\right]\right|,
\end{aligned}$$

similar to the proof in Proposition B.2.1, with $\boldsymbol{Y}$ being a standard normal Gaussian vector,

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{\sqrt{m}}\left\langle \boldsymbol{x}_{\mathrm{i}}^{[1]}(0), \sigma\left(\boldsymbol{W}_0^{[2]}\boldsymbol{x}_{\mathrm{j}}^{[1]}(0)\right)\right\rangle - \widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[1]}\,\mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{jj}}^{[1]}}u\right)\right]\right| \ge t\right) \\
&\le \exp\left(-cmt^2\right).
\end{aligned} \tag{B.30}$$

By symmetry, for II and II',

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{\sqrt{m}}\left\langle \boldsymbol{x}_{\mathrm{j}}^{[1]}(0), \sigma\left(\boldsymbol{W}_0^{[2]}\boldsymbol{x}_{\mathrm{i}}^{[1]}(0)\right)\right\rangle - \widetilde{\boldsymbol{b}}_{\mathrm{j}}^{[1]}\,\mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[1]}}u\right)\right]\right| \ge t\right) \\
&\le \exp\left(-cmt^2\right).
\end{aligned} \tag{B.31}$$

For the difference between III and III', we define another inner product function $h^{[2]}(\cdot) : \mathbb{R}^{2m} \to \mathbb{R}$:

$$h^{[2]}(\boldsymbol{Z}) = \frac{1}{m} \left\langle \sigma\left(\left\|\boldsymbol{x}_{\mathrm{i}}^{[1]}(0)\right\|_2 \boldsymbol{X}\right), \sigma\left(\left(\left\|\boldsymbol{x}_{\mathrm{i}}^{[1]}(0)\right\|_2 \left(\rho \boldsymbol{X} + \sqrt{1-\rho^2}\boldsymbol{Y}\right)\right)\right)\right\rangle,$$

where $\boldsymbol{Z}^\intercal = (\boldsymbol{X}^\intercal, \boldsymbol{Y}^\intercal)$, $\boldsymbol{X}, \boldsymbol{Y}$ are standard normal Gaussian vectors, and $-1 \le \rho \le 1$.

For any $\boldsymbol{Z}_1, \boldsymbol{Z}_2 \in \mathbb{R}^m$,

$$\begin{aligned}\left|h^{[2]}(\boldsymbol{Z}_1) - h^{[2]}(\boldsymbol{Z}_2)\right| \le{}& \frac{1}{\sqrt{m}} \left\|\sigma\left(\left(\left\|\boldsymbol{x}_{\mathrm{i}}^{[1]}(0)\right\|_2 \left(\rho \boldsymbol{X}_1 + \sqrt{1-\rho^2}\boldsymbol{Y}_1\right)\right)\right)\right\|_2 \frac{1}{\sqrt{m}} \left\|\boldsymbol{X}_1 - \boldsymbol{X}_2\right\|_2 \\ &+ \frac{1}{\sqrt{m}} \left\|\sigma\left(\left\|\boldsymbol{x}_{\mathrm{i}}^{[1]}(0)\right\|_2 \boldsymbol{X}_2\right)\right\|_2 \frac{\left\|\boldsymbol{x}_{\mathrm{i}}^{[1]}(0)\right\|_2}{\sqrt{m}} \left(|\rho| \left\|\boldsymbol{X}_1 - \boldsymbol{X}_2\right\|_2 + \sqrt{1-\rho^2} \left\|\boldsymbol{Y}_1 - \boldsymbol{Y}_2\right\|_2\right),\end{aligned}$$

combined with Proposition B.2.1, with high probability, there exists a uniform constant $C > 0$, such that

$$\left\|\boldsymbol{x}_{\mathrm{i}}^{[1]}(0)\right\|_2 \le C,$$

consequently,

$$\begin{aligned}&\frac{1}{\sqrt{m}} \left\|\sigma\left(\left(\left\|\boldsymbol{x}_{\mathrm{i}}^{[1]}(0)\right\|_2 \left(\rho \boldsymbol{X}_1 + \sqrt{1-\rho^2}\boldsymbol{Y}_1\right)\right)\right)\right\|_2 \le 10C, \\ &\frac{1}{\sqrt{m}} \left\|\sigma\left(\left\|\boldsymbol{x}_{\mathrm{i}}^{[1]}(0)\right\|_2 \boldsymbol{X}_2\right)\right\|_2 \le 10C,\end{aligned}$$

with abuse of notations, $h^{[2]}(\cdot)$ is $\frac{C}{\sqrt{m}}$-Lipschitz, then

$$\mathbb{P}\left(\left|h^{[2]}(\boldsymbol{Z}) - \mathbb{E}\, h^{[2]}(\boldsymbol{Z})\right| \ge t\right) \le \exp(-cmt^2),$$

hence we have

$$\begin{aligned}&\mathbb{P}\left(\left|\frac{1}{m}\left\langle \sigma\left(\boldsymbol{W}_0^{[2]} \boldsymbol{x}_{\mathrm{i}}^{[1]}(0)\right), \sigma\left(\boldsymbol{W}_0^{[2]} \boldsymbol{x}_{\mathrm{j}}^{[1]}(0)\right)\right\rangle - \mathbb{E}_{(u,v)^\intercal \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{A}_{\mathrm{ij}}^{[2]}\right)} \left[\sigma(u)\sigma(v)\right]\right| \ge t\right) \\ &\le \exp(-cmt^2),\end{aligned} \tag{B.32}$$

with

$$\boldsymbol{A}_{\mathrm{ij}}^{[2]}=\begin{pmatrix}\left\langle \boldsymbol{x}_{\mathrm{i}}^{[1]}(0),\boldsymbol{x}_{\mathrm{i}}^{[1]}(0)\right\rangle & \left\langle \boldsymbol{x}_{\mathrm{i}}^{[1]}(0),\boldsymbol{x}_{\mathrm{j}}^{[1]}(0)\right\rangle \\ \left\langle \boldsymbol{x}_{\mathrm{j}}^{[1]}(0),\boldsymbol{x}_{\mathrm{i}}^{[1]}(0)\right\rangle & \left\langle \boldsymbol{x}_{\mathrm{j}}^{[1]}(0),\boldsymbol{x}_{\mathrm{j}}^{[1]}(0)\right\rangle \end{pmatrix}.$$

Combine Lemma B.1.3 and Proposition B.2.1

$$\begin{aligned}&\mathbb{P}\left(\left|\mathbb{E}_{(u,v)^{\intercal}\sim\mathcal{N}\left(\mathbf{0},\boldsymbol{A}_{\mathrm{ij}}^{[2]}\right)}\left[\sigma(u)\sigma(v)\right]-\mathbb{E}_{(u,v)^{\intercal}\sim\mathcal{N}\left(\mathbf{0},\widetilde{\boldsymbol{A}}_{\mathrm{ij}}^{[2]}\right)}\left[\sigma(u)\sigma(v)\right]\right|\geq t\right)\\&\leq\exp(-cmt^{2}),\end{aligned}\tag{B.33}$$

consequently, we have

$$\begin{aligned}&\mathbb{P}\left(\left|\frac{1}{m}\left\langle\sigma\left(\boldsymbol{W}_{0}^{[2]}\boldsymbol{x}_{\mathrm{i}}^{[1]}(0)\right),\sigma\left(\boldsymbol{W}_{0}^{[2]}\boldsymbol{x}_{\mathrm{j}}^{[1]}(0)\right)\right\rangle-\mathbb{E}_{(u,v)^{\intercal}\sim\mathcal{N}\left(\mathbf{0},\widetilde{\boldsymbol{A}}_{\mathrm{ij}}^{[2]}\right)}\left[\sigma(u)\sigma(v)\right]\right|\geq t\right)\\&\leq\mathbb{P}\left(\left|\frac{1}{m}\left\langle\sigma\left(\boldsymbol{W}_{0}^{[2]}\boldsymbol{x}_{\mathrm{i}}^{[1]}(0)\right),\sigma\left(\boldsymbol{W}_{0}^{[2]}\boldsymbol{x}_{\mathrm{j}}^{[1]}(0)\right)\right\rangle-\mathbb{E}_{(u,v)^{\intercal}\sim\mathcal{N}\left(\mathbf{0},\boldsymbol{A}_{\mathrm{ij}}^{[2]}\right)}\left[\sigma(u)\sigma(v)\right]\right|\geq\frac{t}{2}\right)\\&+\mathbb{P}\left(\left|\mathbb{E}_{(u,v)^{\intercal}\sim\mathcal{N}\left(\mathbf{0},\boldsymbol{A}_{\mathrm{ij}}^{[2]}\right)}\left[\sigma(u)\sigma(v)\right]-\mathbb{E}_{(u,v)^{\intercal}\sim\mathcal{N}\left(\mathbf{0},\widetilde{\boldsymbol{A}}_{\mathrm{ij}}^{[2]}\right)}\left[\sigma(u)\sigma(v)\right]\right|\geq\frac{t}{2}\right)\\&\leq\exp(-cmt^{2}).\end{aligned}\tag{B.34}$$

To sum up, we have

$$\begin{aligned}&\mathbb{P}\left(\left|\boldsymbol{G}_{\mathrm{ij}}^{[2]}(0)-\widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[2]}\right|\geq t\left(1+\frac{c_{\mathrm{res}}}{L}\right)^{2}\right)\\&\leq\mathbb{P}\left(\left|\boldsymbol{G}_{\mathrm{ij}}^{[1]}(0)-\widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[1]}\right|\geq t\right)\\&+\mathbb{P}\left(\left|\frac{1}{\sqrt{m}}\left\langle\boldsymbol{x}_{\mathrm{i}}^{[1]}(0),\sigma\left(\boldsymbol{W}_{0}^{[2]}\boldsymbol{x}_{\mathrm{j}}^{[1]}(0)\right)\right\rangle-\widetilde{\boldsymbol{b}}_{\mathrm{i}}^{[1]}\,\mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{jj}}^{[1]}}u\right)\right]\right|\geq t\right)\\&+\mathbb{P}\left(\left|\frac{1}{\sqrt{m}}\left\langle\boldsymbol{x}_{\mathrm{j}}^{[1]}(0),\sigma\left(\boldsymbol{W}_{0}^{[2]}\boldsymbol{x}_{\mathrm{i}}^{[1]}(0)\right)\right\rangle-\widetilde{\boldsymbol{b}}_{\mathrm{j}}^{[1]}\,\mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma\left(\sqrt{\widetilde{\boldsymbol{K}}_{\mathrm{ii}}^{[1]}}u\right)\right]\right|\geq t\right)\\&+\mathbb{P}\left(\left|\frac{1}{m}\left\langle\sigma\left(\boldsymbol{W}_{0}^{[2]}\boldsymbol{x}_{\mathrm{i}}^{[1]}(0)\right),\sigma\left(\boldsymbol{W}_{0}^{[2]}\boldsymbol{x}_{\mathrm{j}}^{[1]}(0)\right)\right\rangle-\mathbb{E}_{(u,v)^{\intercal}\sim\mathcal{N}\left(\mathbf{0},\widetilde{\boldsymbol{A}}_{\mathrm{ij}}^{[2]}\right)}\left[\sigma(u)\sigma(v)\right]\right|\geq t\right)\end{aligned}\tag{B.35}$$

Inductively, for $2 \leq l \leq L$,

$$\mathbb{P}\left(\left|\boldsymbol{G}_{\mathrm{ij}}^{[l]}(0) - \widetilde{\boldsymbol{K}}_{\mathrm{ij}}^{[l]}\right| \geq t\left(1 + \frac{c_{\mathrm{res}}}{L}\right)^{2l-2}\right) \leq \exp\left(-cmt^2\right), \tag{B.36}$$

specifically, for $l = L$,

$$\mathbb{P}\left(\left|\boldsymbol{G}_{\mathrm{ij}}^{[L+1]}(0) - \boldsymbol{K}_{\mathrm{ij}}^{[L+1]}\right| \geq t\left(1 + \frac{c_{\mathrm{res}}}{L}\right)^{2L}\right) \leq \exp\left(-cmt^2\right). \tag{B.37}$$

(ii). We recall the matrix inequality once again,

$$\left\|\boldsymbol{G}^{[L+1]}(0) - \boldsymbol{K}^{[L+1]}\right\|_{2\to 2} \leq \left\|\boldsymbol{G}^{[L+1]}(0) - \boldsymbol{K}^{[L+1]}\right\|_{\mathrm{F}} \leq n\left\|\boldsymbol{G}^{[L+1]}(0) - \boldsymbol{K}^{[L+1]}\right\|_{\infty},$$

based on Proposition A.2.1, $\lambda_{\min}(\boldsymbol{K}^{[L+1]}) > \lambda_0$. Then if we choose $t = \frac{\lambda_0}{4n\exp(2c_{\mathrm{res}})}$, then for all l, with probability $1 - \exp\left(-cm\lambda_0^2/n^2\right)$,

$$\left\|\boldsymbol{G}^{[l]}(0) - \widetilde{\boldsymbol{K}}^{[l]}\right\|_{2\to 2} \leq \frac{\lambda_0}{4} \tag{B.38}$$

holds. Hence if there exists a small constant $\varepsilon > 0$, such that $m = \Omega\left(\left(\frac{n}{\lambda_0}\right)^{2+\varepsilon}\right)$, then with probability $1 - \exp(-m^{\varepsilon})$,

$$\lambda_{\min}(\boldsymbol{G}^{[l]}(0)) \geq \lambda_{\min}(\widetilde{\boldsymbol{K}}^{[l]}) - \left\|\boldsymbol{G}^{[l]}(0) - \widetilde{\boldsymbol{K}}^{[l]}\right\|_{2\to 2} > \frac{3\lambda_0}{4}. \tag{B.39}$$

In particular, with probability $1 - \exp\left(-cm\lambda_0^2/n^2\right)$,

$$\left\|\boldsymbol{G}^{[L+1]}(0) - \boldsymbol{K}^{[L+1]}\right\|_{2\to 2} \leq \frac{\lambda_0}{4}, \tag{B.40}$$

hence if $m = \Omega\left(\left(\frac{n}{\lambda_0}\right)^{2+\varepsilon}\right)$, with probability $1 - \exp(-m^{\varepsilon})$

$$\lambda_{\min}(\boldsymbol{G}^{[L+1]}(0)) \geq \lambda_{\min}(\boldsymbol{K}^{[L+1]}) - \left\|\boldsymbol{G}^{[L+1]}(0) - \boldsymbol{K}^{[L+1]}\right\|_{2\to 2} > \frac{3\lambda_0}{4}. \tag{B.41}$$

□

We remark that

$$\left[\mathcal{G}_0^{[L+1]}\left(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta\right)\right]_{1\leq\alpha,\beta\leq n} = \boldsymbol{G}^{[L+1]}(0),$$

then directly from Proposition B.3.2, for $m = \Omega\left(\left(\frac{n}{\lambda_0}\right)^{2+\varepsilon}\right)$,

$$\lambda_{\min}\left(\left[\mathcal{G}_0^{[L+1]}\left(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta\right)\right]_{1\leq\alpha,\beta\leq n}\right) \geq \frac{3\lambda_0}{4} \tag{B.42}$$

holds with high probability w.r.t random initialization.