CSIAM Trans. Appl. Math.
doi: 10.4208/csiam-am.SO-2021-0053

Vol. **3**, No. 4, pp. 692-760
December 2022

Towards an Understanding of Residual Networks Using Neural Tangent Hierarchy (NTH)

Yuqing Li[1], Tao Luo[2,*] and Nung Kwan Yip[3]

[1] *School of Mathematical Sciences, CMA-Shanghai, Shanghai Jiao Tong University, Shanghai, 200240, P.R. China.*
[2] *School of Mathematical Sciences, CMA-Shanghai, Institute of Natural Sciences, MOE-LSC, and Qing Yuan Research Institute, Shanghai Jiao Tong University, Shanghai, 200240, P.R. China.*
[3] *Department of Mathematics, Purdue University, IN, 47907, USA.*

Received 7 December 2021; Accepted 11 April 2022

Abstract. Gradient descent yields zero training loss in polynomial time for deep neural networks despite non-convex nature of the objective function. The behavior of network in the infinite width limit trained by gradient descent can be described by the Neural Tangent Kernel (NTK) introduced in [25]. In this paper, we study dynamics of the NTK for finite width Deep Residual Network (ResNet) using the neural tangent hierarchy (NTH) proposed in [24]. For a 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. Our analysis suggests strongly that the particular skip-connection structure of ResNet is the main reason for its triumph over fully-connected network.

AMS subject classifications: 68U99, 90C26, 34A45

Key words: Residual networks, training process, neural tangent kernel, neural tangent hierarchy.

1 Introduction

Deep neural networks have achieved transcendent performance in a wide range of tasks such as speech recognition [9], computer vision [39], and natural language processing [8]. There are various methods to train neural networks, such as first-order gradient based methods like Gradient Descent (**GD**) and Stochastic Gradient Descent (**SGD**), which have been proven to achieve satisfactory results [20]. Experiments in [49] established

*Corresponding author. *Email addresses:* liyuqing_551@sjtu.edu.cn (Y. Li), luotao41@sjtu.edu.cn (T. Luo), yipn@purdue.edu (N. K. Yip)

 ©2022 Global-Science Press

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 SGD, the network is still able to fit them well. There are numerous works trying to demystify such phenomenon theoretically. Du et al. [12, 15] proved that GD can obtain zero training loss for deep and shallow neural networks, and Zou et al. [52] analyzed the convergence of SGD on networks assembled with Rectified Linear Unit (**ReLU**) activation function. All these results are built upon the over-parameterized regime, and it is widely accepted that over-parameterization enables the neural network to fit all training data and bring no harm to the power of its generalization [49]. In particular, the deep neural networks that evaluated positions and selected moves for the well-known program AlphaGo are highly over-parameterized [41, 42].

Another advance is the outstanding performance of Deep Residual Network (**ResNet**), initially proposed by He et al. [22]. 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 [23]. 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 [45, 48]. Theoretically, Hardt and Ma [21] proved that for any residual linear networks with arbitrary depth, there are no spurious local optima. Du et al. [12] showed that in the scope of the convergence of GD via over-parameterization for different networks, training ResNet requires weaker conditions compared with fully-connected networks. Apart from that, the advantages of using residual connections remain to be discovered.

In this paper, we contribute to the further understanding of the above two aspects and make improvements in the analysis of their performance. We use the same ResNet structure as in [12]. (Details of the network structure are provided in Section 3.2.) The ResNet has L layers with width m. We will assume that the n data points are not parallel with each other. Such an assumption holds in general for a standard dataset [15]. We focus on the empirical risk minimization problem given by the quadratic loss and the activation function is 1-Lipschitz and analytic. We show that if $m = \Omega\left(n^3L^2\right)$, then the empirical risk $R_S(\boldsymbol{\theta}_t)$ under GD decays exponentially. More precisely,

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

where λ is the least eigenvalue of $\boldsymbol{K}^{[L+1]}$, definition of which can be found in (4.2).

It is worth noticing that

- Given identical ResNet architectures, for the convergence of randomly initialized GD, our results improve upon [12] in the required number of width per layer from $m = \Omega(n^4L^2)$ to $m = \Omega(n^3L^2)$ (Corollary 4.1).
- For fully-connected network, the required amount of over-parameterization in [24] is $m = \Omega\left(n^3 2^{\mathcal{O}(L)}\right)$. We are able to reproduce the result of Du et al. [12], showing

that the exponential dependence of m on the number of layers L can be eliminated for ResNet.

Our work is mainly motivated by the framework proposed by Huang and Yau [24], in which an infinite hierarchy of ordinary differential equations, the neural tangent hierarchy (NTH) is derived. Huang and Yau applied NTH to a fully-connected feedforward network and showed that it is possible for us to directly study the change of the neural tangent kernel (NTK) [25], and NTH outperforms kernel regressions using the corresponding limiting NTK.

Different from Huang and Yau's work in analyzing the fully-connected network, ResNet is investigated in our paper. We exploit the benefits of using ResNet architecture for training and the advantage of choosing NTH over kernel regression. In Section 5, an overview of our technique is provided.

The organization of the paper is listed as follows. In Section 2, we discuss some related works. In Section 3, we give some preliminary introductions to our problem. In Section 4, we state our main results for ResNet using NTH. In Section 5, we present an outline of our approach. We give some conclusions and future direction in Section 6. All the details of the proof are deferred to the Appendix.

2 Related works

In this section, we survey some previous works related to optimization aspect of neural networks.

Due to the non-convex nature of optimizing a neural network, it is challenging to locate the global optima. A popular way to analyze such optimization problems is to identify the geometric properties of each critical point. Some recent works have shown that for the set of functions satisfying: (i) all local minima are global and (ii) every saddle point possesses a negative curvature (i.e., it is non-degenerate), then GD can find a global optima [14, 17, 26, 31]. The objective functions of some shallow networks are in such set [11,21,38,51]. The work [27] indicates that even for a three-layer linear network, there exists degenerate saddle points without negative curvature. So it is doubtful that global convergence of first-order methods can be guaranteed for deep neural networks.

Here we directly study the dynamics of the GD for a specific neural network architecture. This is another approach widely taken to obtain convergence results. Recently, it has been shown that if the network is over-parameterized, the SGD is able to find a global optima for two-layer networks [6, 13, 15, 16, 18, 33], deep linear networks [2, 5, 21] and ResNet [1, 12]. Jacot et al. [25] established that in the infinite width limit, the full batch GD corresponds to kernel regression predictor using the limiting NTK. Consequently, the convergence of GD for any 'infinite-width' neural network can be characterized by a fixed kernel [4,25]. This is the cornerstone upon which rests the compelling performance of over-parameterization. In the regime of finite width, many works have suggested that the network can reduce training loss at exponential rate using GD [2, 12, 15, 16, 24]. As

the width increases, there are going to be small changes in the parameters during the whole training process [10,52]. Such a variation of the parameters is crucial to the results we present, where the NTK of our ResNet behaves linearly in terms of its parameters throughout training (Theorem 4.2). Specifically, we use the results concerning the stability of the Gram matrices in [12] to demonstrate the benefits of choosing ResNet over fully-connected networks (Proposition C.3).

Some other works used optimal transport theory to analyze the mean field SGD dynamics of training neural networks in the large-width limit [7,37,40,43]. However, their results are limited to one hidden layer networks, and their normalization factor $1/m$ is different from our $1/\sqrt{m}$ which is commonly employed in modern networks [19,22].

3 Preliminaries

3.1 Notations

We begin this section by introducing some notations that will be used in the rest of this paper. We set n for the number of input samples and m for the width of the neural network, and a special vector $(1,1,1,\cdots,1)^\intercal \in \mathbb{R}^m$ by $\mathbf{1} := (1,1,1,\cdots,1)^\intercal$. 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 $\|\cdot\|_{2\to\infty}$, which was shown to be useful in [15]. 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. We take C and c for some universal constants, which might vary from line to line.

Next we introduce a notion of high probability events that was also used in Huang and Yau [24, Section 1.3]. We say that an event holds with high probability if the probability of the event is at least $1-\exp(-m^\varepsilon)$ for some constant $\varepsilon > 0$. Since for a deep neural network in practice, we always have $m \lesssim \mathrm{poly}(n)$ and $n \lesssim \mathrm{poly}(m)$ [1,28], 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.

3.2 Problem setup

We shall focus on the empirical risk minimization problem given by quadratic loss:

$$\min_{\boldsymbol{\theta}} R_S(\boldsymbol{\theta}) = \frac{1}{2n}\sum_{\alpha=1}^{n} \|f(\boldsymbol{x}_\alpha,\boldsymbol{\theta}) - y_\alpha\|_2^2. \tag{3.1}$$

In the above $\{\boldsymbol{x}_\alpha\}_{\alpha=1}^n$ are the training inputs, $\{y_\alpha\}_{\alpha=1}^n$ are the labels, $f(\boldsymbol{x}_\alpha,\boldsymbol{\theta})$ is the prediction function, and $\boldsymbol{\theta}$ are the parameters to be optimized, and their dependence is modeled by ResNet with L hidden layers, each of which has m neurons. Let $\boldsymbol{x} \in \mathbb{R}^d$ be an input

sample, then the network has d input nodes. Let $\boldsymbol{x}^{[l]}$ be the output of layer l with $\boldsymbol{x}^{[0]}=\boldsymbol{x}$. We consider the ResNet given below:

$$\begin{aligned}\boldsymbol{x}^{[1]}&=\sqrt{\frac{c_\sigma}{m}}\sigma(\boldsymbol{W}^{[1]}\boldsymbol{x}),\\ \boldsymbol{x}^{[l]}&=\boldsymbol{x}^{[l-1]}+\frac{c_{\rm res}}{L\sqrt{m}}\sigma(\boldsymbol{W}^{[l]}\boldsymbol{x}^{[l-1]}),\quad \text{for } 2\le l\le L,\end{aligned}\tag{3.2}$$

where $\sigma(\cdot)$ is the activation function applied coordinate-wisely to its input. We assume that $\sigma(\cdot)$ is 1-Lipschitz and smooth. The constant $c_\sigma=\left(\mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[\sigma(x)^2\right]\right)^{-1}$ is a scaling factor serving for the purpose of normalization, and $0<c_{\rm res}<1$ is a small constant. Moreover, we have a series of weight matrices $\left\{\boldsymbol{W}^{[l]}\right\}_{l=1}^{L}$. Note that $\boldsymbol{W}^{[l]}\in\mathbb{R}^{m\times d}$ for $l=1$, and $\boldsymbol{W}^{[l]}\in\mathbb{R}^{m\times m}$ for $2\le l\le L$. The output function of ResNet is

$$f_{\rm res}(\boldsymbol{x},\boldsymbol{\theta})=\boldsymbol{a}^{\intercal}\boldsymbol{x}^{[L]},\tag{3.3}$$

where $\boldsymbol{a}\in\mathbb{R}^m$ is the weight vector of the output layer. We denote the vector containing all parameters by $\boldsymbol{\theta}=\left(\mathrm{vec}(\boldsymbol{W}^{[L]}),\mathrm{vec}(\boldsymbol{W}^{[L-1]}),\cdots,\mathrm{vec}(\boldsymbol{W}^{[1]}),\boldsymbol{a}\right)$. Such a parameterization has been employed widely, see [12, 15, 30]. We shall initialize the parameter vector $\boldsymbol{\theta}_0$ following the adopted Xavier initialization scheme [19], i.e., $W_{i,j}^{[l]}\sim\mathcal{N}(0,1)$, $a_k\sim\mathcal{N}(0,1)$, where $\mathcal{N}(0,1)$ denotes the standard Gaussian distribution. Applying the continuous time GD for the loss function (3.1), we have for any time $t\ge 0$:

$$\partial_t\boldsymbol{W}_t^{[l]}=-\partial_{\boldsymbol{W}^{[l]}}R_S(\boldsymbol{\theta}_t),\quad l=1,2,\cdots,L,\tag{3.4}$$

$$\partial_t\boldsymbol{a}_t=-\partial_{\boldsymbol{a}}R_S(\boldsymbol{\theta}_t).\tag{3.5}$$

We use $\mathcal{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$ for the set of input samples, $\sigma\left(\boldsymbol{W}^{[l]}\boldsymbol{x}_\alpha^{[l-1]}\right)$ 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 $r\ge 1$. We also write the output function $f_{\rm res}(\boldsymbol{x}_\alpha,\boldsymbol{\theta}_t)$ as $f_\alpha(t)$. Moreover, we shall define a series of special matrices. Using $\boldsymbol{I}_m$ to signify the identity matrix in $\mathbb{R}^{m\times m}$, we define for $2\le l\le L$:

$$\boldsymbol{E}_{t,\alpha}^{[l]}:=\left(\boldsymbol{I}_m+\frac{c_{\rm res}}{L}\boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\alpha)\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right).\tag{3.6}$$

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 $\boldsymbol{E}_{t,\alpha}^{[i:j]}:=\boldsymbol{I}_m$, $i>j$ for completeness.

With the above notations, the continuous time GD dynamics (3.4) and (3.5) can be

written as

$$\partial_t \boldsymbol{a}_t = -\frac{1}{n}\sum_{\beta=1}^{n} \boldsymbol{x}_\beta^{[L]}(f_\beta(t)-y_\beta), \tag{3.7}$$

$$\partial_t \boldsymbol{W}_t^{[L]} = -\frac{1}{n}\sum_{\beta=1}^{n} \frac{c_{\text{res}}}{L\sqrt{m}}\text{diag}\left(\boldsymbol{\sigma}_{[L]}^{(1)}(\boldsymbol{x}_\beta)\boldsymbol{a}_t\right)\mathbf{1}\otimes(\boldsymbol{x}_\beta^{[L-1]})^\intercal(f_\beta(t)-y_\beta), \tag{3.8}$$

$$\partial_t \boldsymbol{W}_t^{[l]} = -\frac{1}{n}\sum_{\beta=1}^{n} \frac{c_{\text{res}}}{L\sqrt{m}}\text{diag}\left(\boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[(l+1):L]}\right)^\intercal \boldsymbol{a}_t\right)\mathbf{1}\otimes(\boldsymbol{x}_\beta^{[l-1]})^\intercal(f_\beta(t)-y_\beta), \tag{3.9}$$

$$\text{for } l=2,3,\cdots,L-1,$$

$$\partial_t \boldsymbol{W}_t^{[1]} = -\frac{1}{n}\sum_{\beta=1}^{n} \sqrt{\frac{c_\sigma}{m}}\text{diag}\left(\boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[2:L]}\right)^\intercal \boldsymbol{a}_t\right)\mathbf{1}\otimes(\boldsymbol{x}_\beta)^\intercal(f_\beta(t)-y_\beta). \tag{3.10}$$

3.3 Neural tangent kernel

The Neural Tangent Kernel (NTK) is introduced in Jacot et al. [25]. For any parametrized function $f(\boldsymbol{x},\boldsymbol{\theta}_t)$, it is defined as:

$$\mathcal{K}_{\boldsymbol{\theta}_t}(\boldsymbol{x}_\alpha,\boldsymbol{x}_\beta) = \left\langle \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}_\alpha,\boldsymbol{\theta}_t), \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}_\beta,\boldsymbol{\theta}_t)\right\rangle.$$

In the situations where $f(\boldsymbol{x},\boldsymbol{\theta}_t)$ is the output of a fully-connected feedforward network with appropriate scaling factor $1/\sqrt{m}$ for the parameters, there is an infinite width limit ($m\to\infty$) of $\mathcal{K}_{\boldsymbol{\theta}_t}(\boldsymbol{x}_\alpha,\boldsymbol{x}_\beta)$, denoted by $\mathcal{K}_\infty(\boldsymbol{x}_\alpha,\boldsymbol{x}_\beta)$. This result allows them to capture the behavior of fully-connected feedforward network trained by GD in the infinite width limit. More precisely, the output function $f(\boldsymbol{x},\boldsymbol{\theta}_t)$ evolves as a linear differential equation:

$$\partial_t f(\boldsymbol{x},\boldsymbol{\theta}_t) = -\frac{1}{n}\sum_{\beta=1}^{n}\mathcal{K}_\infty(\boldsymbol{x},\boldsymbol{x}_\beta)(f(\boldsymbol{x}_\beta,\boldsymbol{\theta}_t)-y_\beta). \tag{3.11}$$

Note that the training dynamic is identical to the dynamics of kernel regression under gradient flow. Also we note that $\mathcal{K}_\infty(\cdot)$ only depends on the training inputs. More importantly, $\mathcal{K}_\infty(\cdot)$ is independent of the neural network parameters $\boldsymbol{\theta}$ [4, 12, 15]. Similar result holds for our ResNet structure.

The finding above is groundbreaking in that it provides us an analytically tractable equation to predict the behavior of GD. However, the convergence $\mathcal{K}_{\boldsymbol{\theta}_t}(\boldsymbol{x}_\alpha,\boldsymbol{x}_\beta)$ to $\mathcal{K}_\infty(\boldsymbol{x}_\alpha,\boldsymbol{x}_\beta)$ is proved in the regime of infinite width. This is unrealistic in nature. Some concurrent works concerning various network structures [2,3,12,15,32,44] have extended the result in [25] to the regime of finite width. For a two-layer network with ReLU, the required width m in Song and Yang [44] is $m=\Omega(n^2\text{poly}(\log(n)))$ under some strong assumptions on the input data. For fully-connected feedforward network, Huang and Yau requires width $m=\Omega\left(n^3\log(n)2^{\mathcal{O}(L)}\right)$. Finally, for ResNet which is the main focus of

our paper, the required width for Du et al. [12] is $m=\Omega(n^4L^2)$ with iteration complexity $T=\Omega(n^2\log(\frac{1}{\varepsilon}))$. Our Corollary 4.1 only requires $m=\Omega(n^3L^2)$ and $T=\Omega\left(n\log(\frac{1}{\varepsilon})\right)$.

We now write out the NTK for ResNet:

$$\begin{aligned}\partial_t(f_\alpha(t)-y_\alpha)&=-\nabla_{\boldsymbol{\theta}}f_\alpha(t)\cdot\nabla_{\boldsymbol{\theta}}R_S(\boldsymbol{\theta}_t)\\&=-\frac{1}{n}\nabla_{\boldsymbol{\theta}}f_\alpha(t)\cdot\sum_{\beta=1}^{n}\nabla_{\boldsymbol{\theta}}f_\beta(t)(f_\beta(t)-y_\beta)\\&=-\frac{1}{n}\sum_{\beta=1}^{n}\mathcal{K}_{\boldsymbol{\theta}_t}(\boldsymbol{x}_\alpha,\boldsymbol{x}_\beta)(f_\beta(t)-y_\beta),\end{aligned}\tag{3.12}$$

using Eqs. (3.7), (3.8), (3.9) and (3.10), the NTK $\mathcal{K}_{\boldsymbol{\theta}_t}(\cdot)$ is given below

$$\mathcal{K}_{\boldsymbol{\theta}_t}(\boldsymbol{x}_\alpha,\boldsymbol{x}_\beta)=\left\langle\nabla_{\boldsymbol{\theta}}f_\alpha(t),\nabla_{\boldsymbol{\theta}}f_\beta(t)\right\rangle=\sum_{l=1}^{L+1}\mathcal{G}_t^{[l]}(\boldsymbol{x}_\alpha,\boldsymbol{x}_\beta),\tag{3.13}$$

where

$$\begin{aligned}\mathcal{G}_t^{[1]}(\boldsymbol{x}_\alpha,\boldsymbol{x}_\beta)&=\left\langle\partial_{\boldsymbol{W}^{[1]}}f_\alpha(t),\partial_{\boldsymbol{W}^{[1]}}f_\beta(t)\right\rangle\\&=\left\langle\sqrt{\frac{c_\sigma}{m}}\boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\alpha)\left(\boldsymbol{E}_{t,\alpha}^{[2:L]}\right)^{\mathsf{T}}\boldsymbol{a}_t,\sqrt{\frac{c_\sigma}{m}}\boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[2:L]}\right)^{\mathsf{T}}\boldsymbol{a}_t\right\rangle\left\langle\boldsymbol{x}_\alpha,\boldsymbol{x}_\beta\right\rangle,\end{aligned}$$

for $2\le l\le L$,

$$\begin{aligned}&\mathcal{G}_t^{[l]}(\boldsymbol{x}_\alpha,\boldsymbol{x}_\beta)\\&=\left\langle\partial_{\boldsymbol{W}^{[l]}}f_\alpha(t),\partial_{\boldsymbol{W}^{[l]}}f_\beta(t)\right\rangle\\&=\left\langle\frac{c_{\mathrm{res}}}{L\sqrt{m}}\boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\alpha)\left(\boldsymbol{E}_{t,\alpha}^{[(l+1):L]}\right)^{\mathsf{T}}\boldsymbol{a}_t,\frac{c_{\mathrm{res}}}{L\sqrt{m}}\boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[(l+1):L]}\right)^{\mathsf{T}}\boldsymbol{a}_t\right\rangle\left\langle\boldsymbol{x}_\alpha^{[l-1]},\boldsymbol{x}_\beta^{[l-1]}\right\rangle,\end{aligned}$$

and finally

$$\mathcal{G}_t^{[L+1]}(\boldsymbol{x}_\alpha,\boldsymbol{x}_\beta)=\left\langle\partial_{\boldsymbol{a}}f_\alpha(t),\partial_{\boldsymbol{a}}f_\beta(t)\right\rangle=\left\langle\boldsymbol{x}_\alpha^{[L]},\boldsymbol{x}_\beta^{[L]}\right\rangle.$$

We remark that all the $\mathcal{G}_t^{[l]}$ depends on $\boldsymbol{\theta}_t$ but for simplicity it is not explicitly written.

4 Main results

4.1 Activation function and input samples

In this paper, we will impose some following technical conditions on the activation function and input samples.

Assumption 4.1. The activation function $\sigma(\cdot)$ is smooth, and there exists a universal constant $0<C_{\text{Lip}}\leq 1$ such that for any $r\geq 1$, its r-th derivative and the function value at 0 satisfy

$$|\sigma(0)|, \left\|\sigma^{(r)}(\cdot)\right\|_{\infty}\leq C_{\text{Lip}}. \tag{4.1}$$

We use C_{Lip} to avoid confusion with the usage of constant C in the Appendix. Here the subscript 'Lip' stands for Lipschitz condition. Note that Assumption 4.1 can be satisfied by using the softplus activation:

$$\sigma(x)=\ln(1+\exp(x)).$$

Some other functions also satisfy this assumption, for instance, the sigmoid activation:

$$\sigma(x)=\frac{1}{1+\exp(-x)}.$$

Assumption 4.2. The training inputs and labels satisfy $\|\boldsymbol{x}_\alpha\|_2=1, |y_\alpha|\leq 1$, for any $\boldsymbol{x}_\alpha\in\mathcal{X}$. All training inputs are non-parallel with each other, i.e., $\boldsymbol{x}_{\alpha_1}\nparallel\boldsymbol{x}_{\alpha_2}$, for any $\alpha_1\neq\alpha_2$.

Assumption 4.2 guarantees that some of the Gram matrices defined in Section 4.2 are strictly positive definite.

4.2 Gram matrices

Recent works [15, 44, 50] have shown that the convergence of the outputs of neural networks are determined by the spectral property of Gram matrices. Here we define the key Gram matrices $\{\boldsymbol{K}^{[l]}\}_{l=1}^{L+1}$ below. We more or less follow the definition of the Gram matrices partially from [12, Definition 6.1]. Also we note that the Gram matrices depend on the series of matrices $\{\widetilde{\boldsymbol{K}}^{[l]}\}_{l=1}^{L}$, $\{\widetilde{\boldsymbol{A}}^{[l]}\}_{l=1}^{L+1}$, and the series of vectors $\{\widetilde{\boldsymbol{b}}^{[l]}\}_{l=1}^{L}$, which are listed out as follows, for $2\leq l\leq L$

$$\begin{aligned}
\widetilde{\boldsymbol{K}}_{ij}^{[0]}&=\langle \boldsymbol{x}_i,\boldsymbol{x}_j\rangle,\\
\widetilde{\boldsymbol{K}}_{ij}^{[1]}&=\mathbb{E}_{(u,v)^{\intercal}\sim\mathcal{N}\left(\boldsymbol{0},\begin{pmatrix}\widetilde{\boldsymbol{K}}_{ii}^{[0]} & \widetilde{\boldsymbol{K}}_{ij}^{[0]}\\ \widetilde{\boldsymbol{K}}_{ji}^{[0]} & \widetilde{\boldsymbol{K}}_{jj}^{[0]}\end{pmatrix}\right)} c_\sigma\sigma(u)\sigma(v),\\
\widetilde{\boldsymbol{b}}_i^{[1]}&=\sqrt{c_\sigma}\mathbb{E}_{u\sim\mathcal{N}(0,\widetilde{\boldsymbol{K}}_{ii}^{[0]})}[\sigma(u)],
\end{aligned}$$

$$\widetilde{\boldsymbol{A}}_{ij}^{[l]}=\begin{pmatrix}\widetilde{\boldsymbol{K}}_{ii}^{[l-1]} & \widetilde{\boldsymbol{K}}_{ij}^{[l-1]}\\ \widetilde{\boldsymbol{K}}_{ji}^{[l-1]} & \widetilde{\boldsymbol{K}}_{jj}^{[l-1]}\end{pmatrix},$$

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

$$\widetilde{\boldsymbol{b}}_i^{[l]}=\widetilde{\boldsymbol{b}}_i^{[l-1]}+\frac{c_{\text{res}}}{L}\mathbb{E}_{u\sim\mathcal{N}(0,\widetilde{\boldsymbol{K}}_{ii}^{[l-1]})}[\sigma(u)],$$

$$\widetilde{\boldsymbol{A}}_{ij}^{[L+1]}=\begin{pmatrix}\widetilde{\boldsymbol{K}}_{ii}^{[L]} & \widetilde{\boldsymbol{K}}_{ij}^{[L]}\\ \widetilde{\boldsymbol{K}}_{ji}^{[L]} & \widetilde{\boldsymbol{K}}_{jj}^{[L]}\end{pmatrix},$$

then we may proceed to the definitions of Gram matrices for $l=L+1$ and L.

Definition 4.1. *Given the input samples $\mathcal{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$, the Gram matrix $\boldsymbol{K}^{[L+1]}\in\mathbb{R}^{n\times n}$ is defined as follows, for $1\le i,j\le n$,*

$$\boldsymbol{K}_{ij}^{[L+1]}=\widetilde{\boldsymbol{K}}_{ij}^{[L]}+\mathbb{E}_{(u,v)^\intercal\sim\mathcal{N}\left(\boldsymbol{0},\widetilde{\boldsymbol{A}}_{ij}^{[L+1]}\right)}\left[\frac{c_{\text{res}}\widetilde{\boldsymbol{b}}_i^{[L]}\sigma(v)}{L}+\frac{c_{\text{res}}\widetilde{\boldsymbol{b}}_j^{[L]}\sigma(u)}{L}+\frac{c_{\text{res}}^2\sigma(u)\sigma(v)}{L^2}\right],\tag{4.2}$$

and Gram matrix $\boldsymbol{K}^{[L]}\in\mathbb{R}^{n\times n}$ is also defined as follows, for $1\le i,j\le n$,

$$\boldsymbol{K}_{ij}^{[L]}=\frac{c_{\text{res}}^2}{L^2}\widetilde{\boldsymbol{K}}_{ij}^{[L-1]}\mathbb{E}_{(u,v)^\intercal\sim\mathcal{N}\left(\boldsymbol{0},\widetilde{\boldsymbol{A}}_{ij}^{[L]}\right)}\left[\sigma^{(1)}(u)\sigma^{(1)}(v)\right].\tag{4.3}$$

Note that matrix $\boldsymbol{K}^{[L]}$ coincides with $\boldsymbol{K}^{[H]}$ given by [12, Definition 6.1]. Now that given the definition of $\boldsymbol{K}^{[L+1]}$ and $\boldsymbol{K}^{[L]}$, we need to move forward to the definition of other Gram matrices $\{\boldsymbol{K}^{[l]}\}_{l=1}^{L-1}$. Since it is challenging to give an explicit formula for the series of matrices $\{\boldsymbol{K}^{[l]}\}_{l=1}^{L-1}$, we shall use a slightly different approach to write out the definitions for these matrices.

Definition 4.2. *Given the input samples $\mathcal{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$, the Gram matrices $\boldsymbol{K}^{[l]}\in\mathbb{R}^{n\times n}$ are defined as follows, for $1\le i,j\le n$, $2\le l\le L-1$,*

$$\boldsymbol{K}_{ij}^{[l]}=\frac{c_{\text{res}}^2}{L^2}\widetilde{\boldsymbol{K}}_{ij}^{[l-1]}\lim_{m\to\infty}\frac{1}{m}\left\langle\boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_i)\left(\boldsymbol{E}_{0,i}^{[(l+1):L]}\right)^\intercal\boldsymbol{a}_0,\boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_j)\left(\boldsymbol{E}_{0,j}^{[(l+1):L]}\right)^\intercal\boldsymbol{a}_0\right\rangle,\tag{4.4}$$

and for $1\le i,j\le n$, $l=1$,

$$\boldsymbol{K}_{ij}^{[1]}=c_\sigma\widetilde{\boldsymbol{K}}_{ij}^{[0]}\lim_{m\to\infty}\frac{1}{m}\left\langle\boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_i)\left(\boldsymbol{E}_{0,i}^{[2:L]}\right)^\intercal\boldsymbol{a}_0,\boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_j)\left(\boldsymbol{E}_{0,j}^{[2:L]}\right)^\intercal\boldsymbol{a}_0\right\rangle.\tag{4.5}$$

Remark 4.1. Thanks to the Strong Law of Large Numbers, the above limit exists [4]. Since we send $m\to\infty$, the Gram matrices $\{\boldsymbol{K}^{[l]}\}_{l=1}^{L-1}$ only depend on the input samples and the activation patterns.

Moreover, in Section B, we show that under Assumption 4.2 and width $m\sim n^2$, $\boldsymbol{K}^{[L+1]}$ and $\boldsymbol{K}^{[L]}$ are strictly positive definite.

4.3 Convergence of gradient descent

Here we state our main theorems for the NTH of ResNet.

Theorem 4.1. *Under Assumptions* 4.1 *and* 4.2*, there exists an infinite family of operators* $\mathcal{K}_t^{(r)}: \mathcal{X}^r \to \mathbb{R}$, $r \geq 2$ *that describes the continuous time GD:*

$$\partial_t (f_\alpha(t) - y_\alpha) = -\frac{1}{n} \sum_{\beta=1}^{n} \mathcal{K}_t^{(2)}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta)(f_\beta(t) - y_\beta), \tag{4.6}$$

and for $r \geq 2$*, we have*

$$\partial_t \mathcal{K}_t^{(r)}(\boldsymbol{x}_{\alpha_1}, \boldsymbol{x}_{\alpha_2}, \cdots, \boldsymbol{x}_{\alpha_r}) = -\frac{1}{n} \sum_{\beta=1}^{n} \mathcal{K}_t^{(r+1)}(\boldsymbol{x}_{\alpha_1}, \boldsymbol{x}_{\alpha_2}, \cdots, \boldsymbol{x}_{\alpha_r}, \boldsymbol{x}_\beta)(f_\beta(t) - y_\beta). \tag{4.7}$$

Moreover, with high probability with respect to random initialization, there exist some constants $C, C' > 0$*, such that for time* $0 \leq t \leq \sqrt{m}/(\ln m)^{C'}$*, the following holds:*

$$\left\| \mathcal{K}_t^{(2)}(\cdot) \right\|_\infty \lesssim 1, \tag{4.8}$$

and for $r \geq 3$*,*

$$\left\| \mathcal{K}_t^{(r)}(\cdot) \right\|_\infty \lesssim \frac{(\ln m)^C}{m^{r/2-1}}. \tag{4.9}$$

Remark 4.2. The operator $\mathcal{K}_t^{(2)}(\cdot)$ by definition is the same as the NTK $\mathcal{K}_{\boldsymbol{\theta}_t}(\cdot)$ derived in (3.12).

We note that as r increases, the pre-factor in (4.9) explodes exponentially fast in r. However, this will not significantly affect the convergence of GD. Firstly, only some lower order kernels need to be analyzed. As is shown in the proof of Corollary 4.1, only kernels up to order $r=4$ will be used. Secondly, we shall recall the NTK $\mathcal{K}_t^{(2)}(\cdot)$ derived in (3.13) :

$$\mathcal{K}_t^{(2)}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta) = \left\langle \nabla_{\boldsymbol{\theta}} f_\alpha(t), \nabla_{\boldsymbol{\theta}} f_\beta(t) \right\rangle = \sum_{l=1}^{L+1} \mathcal{G}_t^{[l]}(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta), \tag{4.10}$$

in the case of Huang and Yau [24], for a fully-connected feedforward network, since all those kernels $\mathcal{G}_t^{[l]}$ are positive definite, then the sum of the least eigenvalue of all the kernels $\mathcal{G}_t^{[l]}$ is much larger than the counterpart of a single kernel, i.e.,

$$\lambda_{\min}\left[\mathcal{K}_t^{(2)}\left(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta\right)\right]_{1 \leq \alpha, \beta \leq n} \gg \lambda_{\min}\left[\mathcal{G}_t^{[l]}\left(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta\right)\right]_{1 \leq \alpha, \beta \leq n}.$$

However, adding up all the kernels will not give substantial increase to the least eigenvalue for the ResNet. Since there exists a scaling factor $\frac{1}{L^2}$ for the kernels $\mathcal{G}_t^{[l]}$, where

$2 \le l \le L$, then heuristically, the gap of the least eigenvalues between $\mathcal{K}_t^{(2)}(\cdot)$ 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, we shall see that even if the depth L gets larger, the least eigenvalue of the NTK is still concentrated on the kernels $\mathcal{G}_t^{[L+1]}(\cdot)$ and $\mathcal{G}_t^{[1]}(\cdot)$. Thanks to that observation, we only need to bring the kernel $\mathcal{G}_t^{[L+1]}(\cdot)$ to the spotlight. We omit the analysis of $\mathcal{G}_t^{[1]}(\cdot)$ because it is not needed in our proof.

It was proven in Theorem 4.1 and other literatures [4, 12, 47] that the change of NTK during the dynamics for Deep Neural Network is bounded by $\mathcal{O}\left(\frac{1}{\sqrt{m}}\right)$. However, it was observed by Lee et al. [30] that the 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 [24] and listed out as Corollary 2.4. in their paper. We use a different approach to obtain similar results and state them as Theorem 4.2.

Theorem 4.2. *Under Assumptions* 4.1 *and* 4.2*, with high probability with respect to random initialization, for time* $0 \le t \le \sqrt{m}/(\ln m)^{C'}$*, the following holds:*

$$\left\|\partial_t \mathcal{G}_t^{[L+1]}(\cdot)\right\|_\infty \lesssim \frac{(1+t)(\ln m)^C}{m}, \tag{4.11}$$

where the constant C is independent of the depth L. Moreover, the pre-factor in (4.11) *is at most of order* $\mathcal{O}\left(L^2\right)$.

As a direct consequence of Theorem 4.1, for the ResNet defined in (3.2), with width $m \sim n^3$, the GD converges to zero training loss at a linear rate. The precise statement is given in the following.

Corollary 4.1. *Under Assumptions* 4.1 *and* 4.2*, with* $\boldsymbol{K}^{[L+1]}$ *defined in* (4.2)*, we have that for some* $\lambda_0 > 0$, $\lambda_{\min}\left(\boldsymbol{K}^{[L+1]}\right) > \lambda_0$*. Equipped with this, we have the following two statements.*

There exists a small constant $\gamma_1 > 0$*, such that for* $m = \Omega\left(\left(\frac{n}{\lambda_0}\right)^{2+\gamma_1}\right)$*, with high probability with respect to random initialization, the following holds:*

$$\lambda_{\min}\left[\mathcal{K}_0^{(2)}\left(\boldsymbol{x}_\alpha, \boldsymbol{x}_\beta\right)\right]_{1\le\alpha,\beta\le n} \ge \frac{3}{4}\lambda_0. \tag{4.12}$$

Furthermore, there exists a small constant $\gamma_2 > 0$*, such that for*

$$m = \Omega\left(\left(\frac{n}{\lambda_0}\right)^{3+\gamma_2} L^2 \ln\left(\frac{1}{\varepsilon}\right)^2\right),$$

where $\varepsilon > 0$ *is the desired accuracy for* $R_S(\boldsymbol{\theta}_t)$*, then the training loss* $R_S(\boldsymbol{\theta}_t)$ *decays exponentially with respect to time* $t \ge 0$*,*

$$R_S(\boldsymbol{\theta}_t) \le R_S(\boldsymbol{\theta}_0)\exp\left(-\frac{\lambda t}{n}\right). \tag{4.13}$$

For convenience, we summarize the above statement in the following manner. If

$$m=\max\left\{\Omega\left(\left(\frac{n}{\lambda}\right)^{2+\gamma_1}\right),\Omega\left(\left(\frac{n}{\lambda}\right)^{3+\gamma_2}L^2\ln\left(\frac{1}{\varepsilon}\right)^2\right)\right\},\tag{4.14}$$

then the continuous GD converges exponentially and reaches the training accuracy ε with time complexity

$$T=\mathcal{O}\left(\frac{n}{\lambda}\ln\left(\frac{1}{\varepsilon}\right)\right).\tag{4.15}$$

Before we end this section, we present a fair comparison of our result with others. First of all, Du et al. [12, Theorem 6.1.] required $m=\Omega\big(\frac{n^4}{\lambda_{\min}\left(\boldsymbol{K}^{[L]}\right)^4L^6}\big)$. Since there is a scaling factor $\frac{1}{L^2}$ in $\lambda_{\min}\big(\boldsymbol{K}^{[L]}\big)$, this leads to $m=\Omega\big(n^4L^2\big)$. Then their GD converges with iteration complexity $T=\Omega\left(n^2L^2\ln\left(\frac{1}{\varepsilon}\right)\right)$. Our Corollary 4.1 improves this result in two ways: (i) The quartic dependence on n is reduced directly to cubic dependence. (ii) A faster convergence of the training process of GD.

Second, our work serves as an extension of the NTH proposed by Huang and Yau [24], which captures the GD dynamics for a fully-connected feedforward network. We show that not only it is possible to study directly the time variation of NTK for ResNet using NTH, but that ResNet possesses more stability in many aspects than fully-connected network. In particular, we improve their results in three aspects: (i) With ResNet architecture, the dependency of the amount of over-parameterization on the depth L can be reduced from their $2^{\mathcal{O}(L)}$ to L^2. (ii) While the time interval for the result in [24] takes the form $0\le t\le m^{\frac{p}{2(p+1)}}/(\ln m)^{C'}$ for some $p\ge 2$, we extend the interval to $0\le t\le\sqrt{m}/(\ln m)^{C'}$. Moreover, we are able to show even further that the results hold true for $t\to\infty$ using techniques from [36]. (iii) In the proof of Corollary 2.5. in [24], a further assumption on the least eigenvalue of the NTK $\mathcal{K}_t^{(2)}(\cdot)$ has been imposed directly, we show in Appendix C that the least eigenvalue of the NTK $\mathcal{K}_t^{(2)}(\cdot)$ can be guaranteed with high probability as long as the width m satisfies $m=\Omega(n^2)$.

5 Technique overview

In this part we first describe some technical tools and present the sketch of proofs for Theorems 4.1 and 4.2 and Corollary 4.1.

5.1 Replacement rules

We revisit the NTK (3.12) derived in Section 3.3,

$$\mathcal{K}_{\boldsymbol{\theta}_t}(\boldsymbol{x}_\alpha,\boldsymbol{x}_\beta)=\sum_{l=1}^{L+1}\mathcal{G}_t^{[l]}(\boldsymbol{x}_\alpha,\boldsymbol{x}_\beta),\tag{5.1}$$

Notice that $\mathcal{K}_{\boldsymbol{\theta}_t}(\cdot)$ coincides with $\mathcal{K}_t^{(2)}(\cdot)$ in (3.12), and $\mathcal{K}_t^{(2)}(\cdot)$ is the sum of $L+1$ terms, with each term being the inner product of vectors containing the quantities $\boldsymbol{a}_t, \boldsymbol{x}_\alpha^{[l]}, \boldsymbol{E}_{t,\alpha}^{[l]}$ and $\boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\alpha)$. We are able to write down the dynamics of $\boldsymbol{a}_t, \boldsymbol{x}_\alpha^{[l]}, \boldsymbol{E}_{t,\alpha}^{[l]}$ and $\boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\alpha)$ following GD, using Eqs. (3.7), (3.8) (3.9), (3.10) and chain rules. In order to shorten the space, we perform a similar replacement rule as in Huang and Yau [24]. For instance, the dynamics of $\boldsymbol{a}_t$ is written as

$$\partial_t \boldsymbol{a}_t = -\frac{1}{n}\sum_{\beta=1}^{n} \frac{1}{\sqrt{m}} \sqrt{m}\boldsymbol{x}_\beta^{[L]}(f_\beta(t)-y_\beta). \tag{5.2}$$

For simplicity, we symbolize the dynamics (5.2) as $\boldsymbol{a}_t \to \frac{1}{\sqrt{m}}\sqrt{m}\boldsymbol{x}_\beta^{[L]}$. Similarly, for the dynamics of $\boldsymbol{x}_\alpha^{[l]}$, $2 \le l \le L$, we have

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

and of $\boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\alpha)$, for $2 \le l \le L-1$, $r \ge 1$,

$$\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)^\mathsf{T}\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)^\mathsf{T}\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)^\mathsf{T}\boldsymbol{a}_t\right)\langle \boldsymbol{x}_\alpha, \boldsymbol{x}_\beta\rangle,\\
\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)^\mathsf{T}\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}\left(\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)^\mathsf{T}\boldsymbol{a}_t\right)\\
&\qquad\qquad \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}\left(\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)^\mathsf{T}\boldsymbol{a}_t\right)\langle \boldsymbol{x}_\alpha, \boldsymbol{x}_\beta\rangle,
\end{aligned}$$

and finally of $\boldsymbol{E}_{t,\alpha}^{[l]}, 2\le l\le L-1$,

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

We notice that the constant $\frac{c_{\rm res}}{L}$ plays an important part in our proof, so that the width per layer m does not depend exponentially in depth L.

Using the above rules, the derivative for NTK $\mathcal{K}_t^{(2)}(\cdot)$ is obtained in the following form

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

where each term in $\mathcal{K}_t^{(3)}(\boldsymbol{x}_{\alpha_1},\boldsymbol{x}_{\alpha_2},\boldsymbol{x}_\beta)$ is the summation of all the terms generated from $\mathcal{K}_t^{(2)}(\boldsymbol{x}_{\alpha_1},\boldsymbol{x}_{\alpha_2})$ by performing the replacement procedure. In order to illustrate the idea, we give out an example in the proof of Theorem 4.2 in Section 5.4.

By the same reasoning, we could obtain the higher order kernels inductively by performing all the possible replacements. For instance, for kernel $\mathcal{K}_t^{(r)}(\boldsymbol{x}_{\alpha_1},\boldsymbol{x}_{\alpha_2},\cdots,\boldsymbol{x}_{\alpha_r})$, we could obtain $\mathcal{K}_t^{(r+1)}(\boldsymbol{x}_{\alpha_1},\cdots,\boldsymbol{x}_{\alpha_r};\boldsymbol{x}_\beta)$ given by the following Ordinary Differential Equation

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

In order to describe the vectors appearing in $\mathcal{K}_t^{(r)}(\boldsymbol{x}_{\alpha_1},\boldsymbol{x}_{\alpha_2}\cdots,\boldsymbol{x}_{\alpha_r})$, we need to introduce some systematic notations.

5.2 Hierarchical sets of kernel expressions

The hierarchy of sets are proposed originally by Huang and Yau in [24]. We denote $\mathbb{A}_0$ the first set of expressions in the following form, which corresponds to the terms in $\mathcal{K}_t^{(2)}(\boldsymbol{x}_{\alpha_1},\boldsymbol{x}_{\alpha_2})$. We define $\mathbb{A}_0$ as:

$$\mathbb{A}_0 \triangleq \{\boldsymbol{e}_s\boldsymbol{e}_{s-1}\cdots\boldsymbol{e}_1\boldsymbol{e}_0 \colon 0\leq s\leq 4L\}, \tag{5.3}$$

where $\boldsymbol{e}_j$ is chosen following the rules:

$$\boldsymbol{e}_0 \in \left\{\boldsymbol{a}_t, \{\sqrt{m}\boldsymbol{x}_\beta^{[1]}, \sqrt{m}\boldsymbol{x}_\beta^{[2]}, \cdots, \sqrt{m}\boldsymbol{x}_\beta^{[L]}\}_{1\leq\beta\leq n}\right\}, \tag{5.4}$$

and for $1\leq j\leq s$,

$$\boldsymbol{e}_j \in \left\{\left\{\boldsymbol{E}_{t,\beta}^{[2]}, \left(\boldsymbol{E}_{t,\beta}^{[2]}\right)^{\mathsf{T}}, \cdots, \boldsymbol{E}_{t,\beta}^{[L]}, \left(\boldsymbol{E}_{t,\beta}^{[L]}\right)^{\mathsf{T}}\right\}_{1\leq\beta\leq n}, \left\{\boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\beta), \cdots, \boldsymbol{\sigma}_{[L]}^{(1)}(\boldsymbol{x}_\beta)\right\}_{1\leq\beta\leq n}\right\}. \tag{5.5}$$

Hence, each term in $\mathcal{K}_t^{(2)}(\boldsymbol{x}_{\alpha_1},\boldsymbol{x}_{\alpha_2})$ writes as

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

where $\boldsymbol{v}_1(t),\boldsymbol{v}_2(t),\boldsymbol{v}_3(t),\boldsymbol{v}_4(t)\in\mathbb{A}_0$. Note that $\boldsymbol{v}_i(t)$ can take the value of $\boldsymbol{v}_i(t)=\sqrt{m}\boldsymbol{x}_\alpha$, which are not contained in $\mathbb{A}_0$. However, such a singularity is not a big issue-see Appendix A.2. We remark that compared with [24], $\boldsymbol{e}_j$ is chosen in a way different from ours, the counterpart in [24] is chosen from the set

$$\left\{\left\{\frac{\boldsymbol{W}_t^{[2]}}{\sqrt{m}}, \left(\frac{\boldsymbol{W}_t^{[2]}}{\sqrt{m}}\right)^{\mathsf{T}}, \cdots, \frac{\boldsymbol{W}_t^{[L]}}{\sqrt{m}}, \left(\frac{\boldsymbol{W}_t^{[L]}}{\sqrt{m}}\right)^{\mathsf{T}}\right\}_{1\leq\beta\leq n}, \left\{\boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\beta), \cdots, \boldsymbol{\sigma}_{[L]}^{(1)}(\boldsymbol{x}_\beta)\right\}_{1\leq\beta\leq n}\right\}.$$

Such changes arise from the change of the network structure, and it has been shown in Appendix A.2 that the group of skip-connection matrices $\boldsymbol{E}_{t,\beta}^{[l]}$ possesses more stability than $\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}$.

Moreover, given the construction of $\mathbb{A}_0,\mathbb{A}_1,\cdots,\mathbb{A}_r$, we denote $\mathbb{A}_{r+1}$ to be the set of expressions in the following form:

$$\mathbb{A}_{r+1} \triangleq \{\boldsymbol{e}_s\boldsymbol{e}_{s-1}\cdots\boldsymbol{e}_1\boldsymbol{e}_0 \colon 0\leq s\leq 4L\},$$

where e_j is chosen from the following sets:

$$e_0 \in \left\{\boldsymbol{a}_t, \mathbf{1}, \{\sqrt{m}\boldsymbol{x}_\beta^{[1]}, \sqrt{m}\boldsymbol{x}_\beta^{[2]}, \cdots, \sqrt{m}\boldsymbol{x}_\beta^{[L]}\}_{1\leq\beta\leq n}\right\}, \tag{5.6}$$

and for $1\le j\le s$, we have that each $\boldsymbol{e}_j$ comes from one of the three following sets

$$\left\{\left\{\boldsymbol{E}_{t,\beta}^{[2]},\left(\boldsymbol{E}_{t,\beta}^{[2]}\right)^{\mathsf{T}},\cdots,\boldsymbol{E}_{t,\beta}^{[L]},\left(\boldsymbol{E}_{t,\beta}^{[L]}\right)^{\mathsf{T}}\right\}_{1\le\beta\le n},\left\{\boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\beta),\cdots,\boldsymbol{\sigma}_{[L]}^{(1)}(\boldsymbol{x}_\beta)\right\}_{1\le\beta\le n}\right\},$$

$$\{\mathrm{diag}(\boldsymbol{g}),\ \boldsymbol{g}\in\mathbb{A}_0\cup\mathbb{A}_1\cup\cdots\cup\mathbb{A}_r\},$$

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

$$\left.\left(\frac{c_{\mathrm{res}}}{L}\frac{\left(\boldsymbol{W}_t^{[l]}\right)^{\mathsf{T}}}{\sqrt{m}}\right)^{Q_{u+1}}\boldsymbol{\sigma}_{[l]}^{(u+1)}(\boldsymbol{x}_\beta)\mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_1}\boldsymbol{g}_1\right)\cdots\mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_u}\boldsymbol{g}_u\right):\ 2\le l\le L,\right.$$

$$\left.1\le\beta\le n,1\le u\le r,\ \boldsymbol{g}_1,\boldsymbol{g}_2\cdots\boldsymbol{g}_u\in\mathbb{A}_0\cup\mathbb{A}_1\cup\cdots\cup\mathbb{A}_r\text{ and }Q_1,Q_2\cdots Q_{u+1}\in\{0,1\}\right\}.$$

The maximum possible total number of diag operations for any element in $\mathbb{A}_r$ is r, i.e., if $\boldsymbol{v}(t)\in\mathbb{A}_r$, it contains at most r diag operations. We observe from the replacement rules that there will be a scaling of $\frac{1}{\sqrt{m}}$ whenever we take derivatives. Hence inductively, each term in kernel $\mathcal{K}_t^{(p)}(\boldsymbol{x}_{\alpha_1},\boldsymbol{x}_{\alpha_2},\cdots,\boldsymbol{x}_{\alpha_p})$ takes the form

$$\frac{1}{m^{p/2-1}}\prod_{j=1}^{s}\frac{\left\langle\boldsymbol{v}_{2j-1}(t),\boldsymbol{v}_{2j}(t)\right\rangle}{m},\quad 1\le s\le p,\quad \boldsymbol{v}_i(t)\in\mathbb{A}_0\cup\mathbb{A}_1\cup\cdots\cup\mathbb{A}_{p-2},\tag{5.7}$$

which is a direct consequence of Proposition A.1. Note that $\boldsymbol{v}_i(t)$ can still take the value of $\boldsymbol{v}_i(t)=\sqrt{m}\boldsymbol{x}_\alpha$, which are not in the set $\mathbb{A}_{r+1}$. Huang and Yau also obtained (5.7) in equation (3.8) in [24], and they use the tensor program proposed by Yang [47] to estimate the initial value of the kernel $\mathcal{K}_0^{(p)}(\boldsymbol{x}_{\alpha_1},\boldsymbol{x}_{\alpha_2},\cdots,\boldsymbol{x}_{\alpha_p})$. They showed that for each vector $\boldsymbol{v}_j(t)$ in (5.7) at $t=0$, it is a linear combination of projections of independent Gaussian vectors. Hence, if we consider such quantity

$$\eta(t)=\{\|\boldsymbol{v}(t)\|_\infty:\ \boldsymbol{v}(t)\in\mathbb{A}_0\cup\mathbb{A}_1\cdots\cup\mathbb{A}_r\},$$

at $t=0$, since $\boldsymbol{v}(0)$ is a linear combination of projections of independent Gaussian vectors, then with high probability, $\eta(0)\lesssim(\ln m)^C$. For $t>0$, Huang and Yau derived a self-consistent Ordinary Differential Inequality for $\eta(t)$ leading to the following estimates:

$$\partial_t^{(p+1)}\eta(t)\lesssim\frac{\eta(t)^{2p}}{m^{p/2}},\tag{5.8}$$

$$\eta(0)\lesssim(\ln m)^C,\tag{5.9}$$

then it holds that $\eta(t)\lesssim(\ln m)^C$ for time $0\le t\le m^{\frac{p}{2p+1}}/(\ln m)^{C'}$.

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 that $\eta(0)\lesssim(\ln m)^C$. We can then obtain a Gronwall-type inequality for $\eta(t)$:

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

for which it follows that for time $0\le t\le\sqrt{m}/(\ln m)^{C'}$, the estimate $\eta(t)\lesssim(\ln m)^C$ holds. Then (4.6) and (4.7) in Theorem 4.1 holds, and we are able to show that the kernels of higher order vary slowly, which brings us the proof of Theorem 4.2.

5.3 Least eigenvalue for randomly initialized matrix

Firstly, since $\boldsymbol{K}^{[L]}$ is a recursively defined matrix, we use the results in Du et al. [12] to show that the Gram matrix $\boldsymbol{K}^{[L]}$ is positive definite. Secondly, we need to analyze how the difference between $\boldsymbol{G}^{[1]}$ and $\boldsymbol{K}^{[1]}$, termed the perturbation by Du et al. [12], propagates from the lower layers to the L-th layer. We quantitatively characterize how large such propagation dynamics would be and rediscover that ResNet architecture serves as a stabilizer for such propagation (Proposition C.1). Our proof is slightly different from [12], where we use the concentration inequality for Lipschitz functions. We refer readers to Appendix C for details.

5.4 Sketch of proof

We use Fig. 1 to illustrate the ideas of the proofs. Due to space constraints, all the proofs of the technical Lemmas and Propositions are provided in Supplementary Material. Note that for the quantities in Fig. 1, λ_0 is the least eigenvalue of $\widetilde{\boldsymbol{K}}_{ij}^{[1]}$, $\xi_{\infty,r}(t)=\sup_{0\le t'\le t}\{\|\boldsymbol{v}(t')\|_2:\boldsymbol{v}(t')\in\mathbb{A}_r\}$, and $\eta_{\infty,r}(t)=\sup_{0\le t'\le t}\{\|\boldsymbol{v}(t')\|_\infty:\boldsymbol{v}(t')\in\mathbb{A}_r\}$, where $r\ge0$. Now we proceed to the Proof of Theorem 4.1.

Proof of Theorem 4.1. Since each term in kernel $\mathcal{K}_t^{(r)}(\boldsymbol{x}_{\alpha_1},\boldsymbol{x}_{\alpha_2},\cdots,\boldsymbol{x}_{\alpha_r})$, it takes the form

$$\frac{1}{m^{r/2-1}}\prod_{j=1}^{s}\frac{\langle\boldsymbol{v}_{2j-1}(t),\boldsymbol{v}_{2j}(t)\rangle}{m},\quad 1\le s\le r,\quad \boldsymbol{v}_i(t)\in\mathbb{A}_0\cup\mathbb{A}_1\cup\cdots\cup\mathbb{A}_{r-2},$$

then for time $0\le t\le\sqrt{m}/(\ln m)^{C'}$, $r\ge3$

$$\left\|\mathcal{K}_t^{(2)}(\cdot)\right\|_\infty\lesssim\left(\frac{\xi_{\infty,0}(t)^2}{m}\right)^2\lesssim1,$$

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

This completes the proof. □

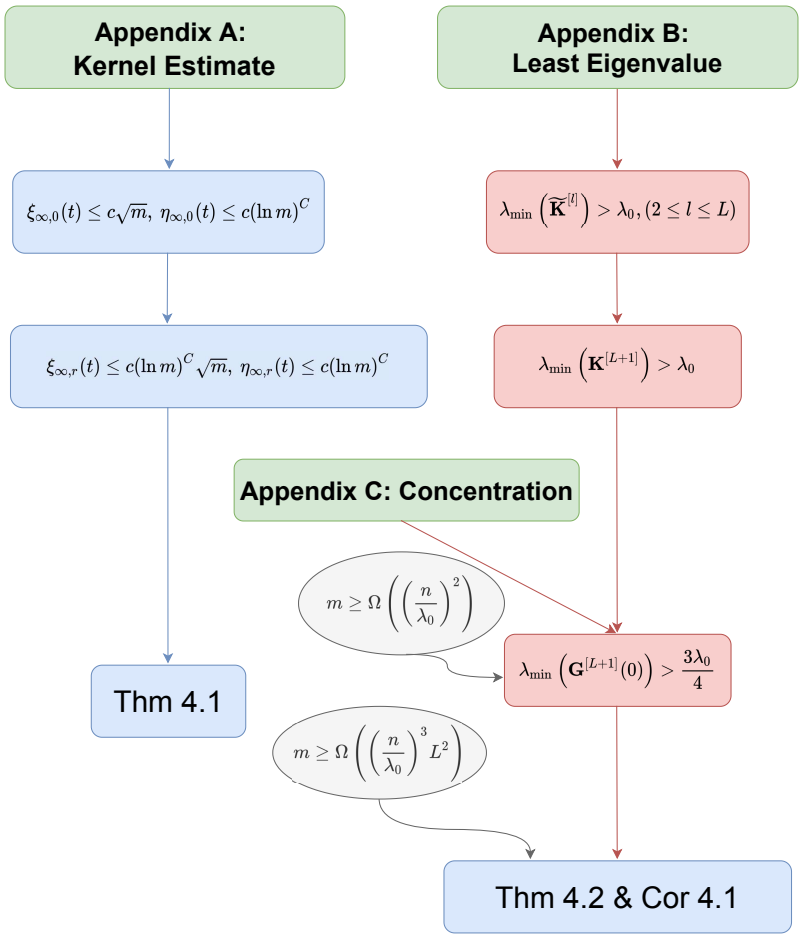

Figure 1: Flow chart for the proof of main theorems.

Now we sketch the proofs for Theorem 4.2 and Corollary 4.1. Details can be found in Appendix D.

Sketch of the proof of Theorem 4.2. Since there exists $\frac{1}{L^2}$ scaling in some kernels, we use $C(r,L)$ to denote the 'effective terms' in each kernel. We denote $\mathcal{G}_t^{[L+1]}(\cdot)$ by $\mathcal{G}_t^{[2]}(\cdot)$, i.e., $\mathcal{G}_t^{(2)}(\cdot) := \mathcal{G}_t^{[L+1]}(\cdot)$, it's natural for us to get that $C(2,L) = \mathcal{O}(1)$.

Next, we apply the replacement rule, all the possible terms generated from $\mathcal{G}_t^{(2)}(\cdot)$ are

$$\mathcal{G}_t^{(2)}(\boldsymbol{x}_{\alpha_1},\boldsymbol{x}_{\alpha_2}) = \left\langle \boldsymbol{x}_{\alpha_1}^{[L]},\boldsymbol{x}_{\alpha_2}^{[L]} \right\rangle \to \mathcal{G}_t^{(3)}(\boldsymbol{x}_{\alpha_1},\boldsymbol{x}_{\alpha_2},\boldsymbol{x}_{\beta}),$$

$$
\begin{aligned}
&\mathcal{G}_t^{(3)}(\boldsymbol{x}_{\alpha_1},\boldsymbol{x}_{\alpha_2},\boldsymbol{x}_\beta)\\
=&\frac{c_\sigma}{m}\left\langle \operatorname{diag}\left(\boldsymbol{E}_{t,\alpha_1}^{[2:L]}\boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_{\alpha_1})\boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[2:L]}\right)^{\mathsf{T}}\boldsymbol{a}_t\right)\mathbf{1},\boldsymbol{x}_{\alpha_2}^{[L]}\right\rangle\langle \boldsymbol{x}_{\alpha_1},\boldsymbol{x}_\beta\rangle \qquad \text{(I)}\\
&+\sum_{k=2}^{L}\frac{c_{\text{res}}^2}{L^2m}\left\langle \operatorname{diag}\left(\boldsymbol{E}_{t,\alpha_1}^{[(k+1):L]}\boldsymbol{\sigma}_{[k]}^{(1)}(\boldsymbol{x}_{\alpha_1})\boldsymbol{\sigma}_{[k]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[(k+1):L]}\right)^{\mathsf{T}}\boldsymbol{a}_t\right)\mathbf{1},\boldsymbol{x}_{\alpha_2}^{[L]}\right\rangle\left\langle \boldsymbol{x}_{\alpha_1}^{[k-1]},\boldsymbol{x}_\beta^{[k-1]}\right\rangle \qquad \text{(II)}\\
&+\frac{c_\sigma}{m}\left\langle \operatorname{diag}\left(\boldsymbol{E}_{t,\alpha_2}^{[2:L]}\boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_{\alpha_2})\boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[2:L]}\right)^{\mathsf{T}}\boldsymbol{a}_t\right)\mathbf{1},\boldsymbol{x}_{\alpha_1}^{[L]}\right\rangle\langle \boldsymbol{x}_{\alpha_2},\boldsymbol{x}_\beta\rangle\\
&+\sum_{k=2}^{L}\frac{c_{\text{res}}^2}{L^2m}\left\langle \operatorname{diag}\left(\boldsymbol{E}_{t,\alpha_2}^{[(k+1):L]}\boldsymbol{\sigma}_{[k]}^{(1)}(\boldsymbol{x}_{\alpha_2})\boldsymbol{\sigma}_{[k]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[(k+1):L]}\right)^{\mathsf{T}}\boldsymbol{a}_t\right)\mathbf{1},\boldsymbol{x}_{\alpha_1}^{[L]}\right\rangle\left\langle \boldsymbol{x}_{\alpha_2}^{[k-1]},\boldsymbol{x}_\beta^{[k-1]}\right\rangle,
\end{aligned}
$$

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

Finally for $\mathcal{G}_t^{(4)}(\cdot)$, by symmetry, we are only going to analyze terms I and II above. Since there are at most $(2L+2)$ symbols in term I to be replaced, each replacement 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, we have that

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

It holds that for time $0\le t\le\sqrt{m}/(\ln m)^{C'}$

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

Finally, we need to estimate $\|\mathcal{G}_0^{(3)}(\cdot)\|_\infty$. Each term in $\mathcal{G}_0^{(3)}(\cdot)$ is of the form $\frac{c}{m}\langle \boldsymbol{B}\boldsymbol{a}_0,\boldsymbol{x}_{\alpha_1}^{[L]}\rangle\langle \boldsymbol{x}_{\alpha_2}^{[l]},\boldsymbol{x}_\beta^{[l]}\rangle$ where $\boldsymbol{B}$ is some specific matrix that changes from term to term. After taking conditional expectation with respect to the random variable $\boldsymbol{a}_0$, we have with high probability that

$$
\frac{c}{m}\left\langle \boldsymbol{a}_0,\boldsymbol{B}^{\mathsf{T}}\boldsymbol{x}_{\alpha_1}^{[L]}\right\rangle\left\langle \boldsymbol{x}_{\alpha_2}^{[l]},\boldsymbol{x}_\beta^{[l]}\right\rangle\le c\frac{(\ln m)^C}{m}. \tag{5.10}
$$

Consequently, for time $0\le t\le\sqrt{m}/(\ln m)^{C'}$,

$$
\begin{aligned}
\left|\mathcal{G}_t^{(3)}(\boldsymbol{x}_{\alpha_1},\boldsymbol{x}_{\alpha_2},\boldsymbol{x}_{\alpha_3})\right|&\le C(3,L)\frac{(\ln m)^C}{m}+tC(4,L)\frac{(\ln m)^C}{m},\\
\left|\partial_t\mathcal{G}_t^{(2)}(\boldsymbol{x}_{\alpha_1},\boldsymbol{x}_{\alpha_2})\right|&\le\left\|\mathcal{G}_t^{(3)}(\cdot)\right\|_\infty\sqrt{R_S(\boldsymbol{\theta}_0)}\le(C(3,L)+tC(4,L))\frac{(\ln m)^C}{m},
\end{aligned}
$$

which finishes the proof of Theorem 4.2. □

Sketch of the proof of Corollary 4.1. If $m=\Omega\big((\frac{n}{\lambda_0})^{2+\varepsilon}\big)$, with high probability with respect to random initialization, $\lambda_{\min}\big[\mathcal{K}_0^{(2)}\left(\boldsymbol{x}_\alpha,\boldsymbol{x}_\beta\right)\big]_{1\le\alpha,\beta\le n}>\lambda_{\min}\big(\boldsymbol{G}^{[L+1]}(0)\big)>\frac{3\lambda_0}{4}$. We finish the proof of (4.12) by setting $\lambda=\frac{3\lambda_0}{4}$.

Concerning the change of the least eigenvalue of the NTK, from the Sketch Proof of Theorem 4.2, for time $0\le t\le\sqrt{m}/(\ln m)^{C'}$,

$$\begin{aligned}\left\|\left(\mathcal{G}_t^{(2)}-\mathcal{G}_0^{(2)}\right)(\cdot)\right\|_{2\to2}&\le\left\|\left(\mathcal{G}_t^{(2)}-\mathcal{G}_0^{(2)}\right)(\cdot)\right\|_{\mathrm{F}}\\&\le n\left\|\left(\mathcal{G}_t^{(2)}-\mathcal{G}_0^{(2)}\right)(\cdot)\right\|_{\infty}\le nt(C(3,L)+tC(4,L))\frac{(\ln m)^C}{m}.\end{aligned}$$

Let t^* satisfy

$$C(4,L)(t^*)^2+C(3,L)t^*=\frac{\lambda m}{2(\ln m)^C n},\tag{5.11}$$

which gives

$$t^*=\frac{-C(3,L)+\sqrt{(C(3,L))^2+2C(4,L)\frac{\lambda m}{(\ln m)^C n}}}{2C(4,L)}\ge\frac{1}{2}\sqrt{\frac{\lambda m}{C(4,L)(\ln m)^C n}}.$$

Next, let $\bar{t}:=\inf\big\{t:\lambda_{\min}\big[\mathcal{K}_t^{(2)}\left(\boldsymbol{x}_\alpha,\boldsymbol{x}_\beta\right)\big]_{1\le\alpha,\beta\le n}\ge\lambda/2\big\}$. Naturally, we have $t^*\le\bar{t}$. Using (4.6), we have that for any $0\le t\le\bar{t}$, $R_S(\boldsymbol{\theta}_t)\le\exp(-\lambda t/n)R_S(\boldsymbol{\theta}_0)$.

Set $R_S(\boldsymbol{\theta}_t)=\varepsilon$. It takes time $t\le(n/\lambda)\ln(R_S(\boldsymbol{\theta}_0)/\varepsilon)$ for the loss $R_S(\boldsymbol{\theta}_t)$ to reach accuracy ε. Hence, if $t\le(n/\lambda)\ln(R_S(\boldsymbol{\theta}_0)/\varepsilon)\le t^*\le\bar{t}$, then width m is required to be

$$\frac{n}{\lambda}\ln\left(\frac{R_S(\boldsymbol{\theta}_0)}{\varepsilon}\right)\le\frac{1}{2}\sqrt{\frac{\lambda m}{C(4,L)(\ln m)^C n}}.\tag{5.12}$$

Thus we have

$$m\ge C(4,L)\left(\frac{n}{\lambda}\right)^3(\ln m)^C\ln\left(\frac{R_S(\boldsymbol{\theta}_0)}{\varepsilon}\right)^2.$$

Since $C(4,L)=\mathcal{O}\left(L^2\right)$, we conclude the proof. □

6 Discussion

In this paper, we show that the GD on ResNet can obtain zero training loss and its training dynamic is given by an infinite hierarchy of ordinary differential equations, i.e., the NTH, which makes it possible to study the change of the NTK directly for deep neural

networks. Our proof builds on a careful analysis of the least eigenvalue of randomly initialized Gram matrix, and the uniform upper bound on kernels of higher order in the NTH.

We list some future directions for research:

- The NTH is an infinite sequence of relationship. However, Huang and Yau [24] showed that under certain conditions on the width and the data set dimension, the NTH can be truncated and the truncated version of NTH is still able to approximate the original dynamic up to any precision. We believe that for ResNet, such technical conditions can be loosened based on our result.

- In Corollary 4.1, the dependence of m on the depth L is quadratic, we believe that the dependence can be reduced even further. We conjecture that m is independent of L.

- In this paper, we focus on the GD, and we believe that it can be extended to SGD, while maintaining the linear convergence rate.

- We focus here on the training loss, but does not address the test loss. To further investigate the generalization power of ResNet, we believe that some apriori estimates for the generalization error of ResNet may be useful [34, 35].

Acknowledgments

This work is supported by National Natural Science Foundation of China Grant No. 12101401 (T.L.), Shanghai Municipal Science and Technology Key Project No. 22JC1401500 (T.L.), and Shanghai Municipal of Science and Technology Major Project No. 2021SHZDZX0102 (T.L.).

Appendices

A Estimates on the kernel

A.1 Structure on hierarchical sets of kernel expressions

Since we have mentioned the replacement rules in Section 5.1, we haven't rigorously justified it yet. Hence we use Proposition A.1 to shed light on the structures of the elements in $\mathbb{A}_r$, and consequently on the structures of each term in kernel $\mathcal{K}_t^{(r)}(\boldsymbol{x}_{\alpha_1}, \boldsymbol{x}_{\alpha_2}, \cdots, \boldsymbol{x}_{\alpha_r})$.

Proposition A.1. *For any vector $\boldsymbol{v}(t)\in\mathbb{A}_r$, the new vector obtained from $\boldsymbol{v}(t)$ by performing the replacement rules are the sum of terms of the following forms:*

$$(a).\ \frac{C}{\sqrt{m}}\boldsymbol{v}'(t):\ \boldsymbol{v}'(t)\in\mathbb{A}_r,$$
$$(b).\ \frac{C}{\sqrt{m}}\boldsymbol{v}'(t)\frac{\langle \boldsymbol{p},\boldsymbol{q}\rangle}{m}:\ \boldsymbol{v}'(t)\in\mathbb{A}_{r+1},\ \boldsymbol{p},\boldsymbol{q}\in\mathbb{A}_0,$$
$$(c).\ \frac{C}{\sqrt{m}}\boldsymbol{v}'(t)\frac{\langle \sqrt{m}\boldsymbol{x}_\alpha,\sqrt{m}\boldsymbol{x}_\beta\rangle}{m}:\ \boldsymbol{v}'(t)\in\mathbb{A}_{r+1},\ \boldsymbol{p},\boldsymbol{q}\in\mathbb{A}_0,$$
$$(d).\ \frac{C}{\sqrt{m}}\boldsymbol{v}'(t)\frac{\langle \boldsymbol{p},\boldsymbol{q}\rangle}{m}:\ \boldsymbol{v}'(t)\in\mathbb{A}_{r-s+1},\ \boldsymbol{p}\in\mathbb{A}_s,\boldsymbol{q}\in\mathbb{A}_0,\ \textit{for some}\ \ s\geq 1,$$
$$(e).\ \frac{C}{\sqrt{m}}\boldsymbol{v}'(t)\frac{\langle \boldsymbol{p},\boldsymbol{q}\rangle}{m}:\ \boldsymbol{v}'(t)\in\mathbb{A}_s,\ \boldsymbol{p}\in\mathbb{A}_{r-s+1},\boldsymbol{q}\in\mathbb{A}_0,\ \textit{for some}\ s\geq 1.$$

Proof. The proof comes as follows. Note that the constant C listed out below might keep changing from term to term.

Since $\boldsymbol{a}_t$ appears only at the position $\boldsymbol{e}_0$, if $\boldsymbol{v}(t)\in\mathbb{A}_r$, based on the replacement rule

$$\boldsymbol{v}(t)=\boldsymbol{e}_s\boldsymbol{e}_{s-1}\cdots\boldsymbol{e}_1\boldsymbol{a}_t\to\widetilde{\boldsymbol{v}}(t)=\frac{1}{\sqrt{m}}\boldsymbol{e}_s\boldsymbol{e}_{s-1}\cdots\boldsymbol{e}_1\sqrt{m}\boldsymbol{x}_\beta^{[L]}=\frac{1}{\sqrt{m}}\boldsymbol{v}'(t),$$

then $\boldsymbol{v}'(t)=\boldsymbol{e}_s\boldsymbol{e}_{s-1}\cdots\boldsymbol{e}_1\sqrt{m}\boldsymbol{x}_\beta^{[L]}\in\mathbb{A}_r$.

Similarly, $\sqrt{m}\boldsymbol{x}_\alpha^{[l]}$ also appears only at $\boldsymbol{e}_0$, then if $\boldsymbol{v}(t)\in\mathbb{A}_r$, by the replacement rule

$$\boldsymbol{v}(t)=\boldsymbol{e}_s\boldsymbol{e}_{s-1}\cdots\boldsymbol{e}_1\sqrt{m}\boldsymbol{x}_\alpha^{[l]}\to\widetilde{\boldsymbol{v}}(t),$$
$$\widetilde{\boldsymbol{v}}(t)=\sum_k\frac{C}{\sqrt{m}}\boldsymbol{e}_s\boldsymbol{e}_{s-1}\cdots\boldsymbol{e}_1\ \mathrm{diag}(\boldsymbol{f}_k)\mathbf{1}\frac{\left\langle\sqrt{m}\boldsymbol{x}_\alpha^{[k]},\sqrt{m}\boldsymbol{x}_\beta^{[k]}\right\rangle}{m}=\sum_k\frac{C}{\sqrt{m}}\boldsymbol{v}_k'(t)\frac{\left\langle\sqrt{m}\boldsymbol{x}_\alpha^{[k]},\sqrt{m}\boldsymbol{x}_\beta^{[k]}\right\rangle}{m},$$

given that $\boldsymbol{f}_k\in\mathbb{A}_0$, then $\boldsymbol{v}_k'(t)\in\mathbb{A}_{r+1}$.

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

$$\begin{aligned}
\boldsymbol{v}(t)&=\boldsymbol{e}_s\cdots\boldsymbol{e}_{j+1}\boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\alpha)\boldsymbol{e}_{j-1}\cdots\boldsymbol{e}_0\to\widetilde{\boldsymbol{v}}(t),\\
\widetilde{\boldsymbol{v}}(t)&=\frac{C}{\sqrt{m}}\boldsymbol{e}_s\cdots\boldsymbol{e}_{j+1}\boldsymbol{\sigma}_{[l]}^{(2)}(\boldsymbol{x}_\alpha)\mathrm{diag}(\boldsymbol{f}_1)\boldsymbol{e}_{j-1}\cdots\boldsymbol{e}_0\frac{\langle\boldsymbol{p}_1,\boldsymbol{q}_1\rangle}{m}\\
&\quad+\sum_k\frac{C}{\sqrt{m}}\boldsymbol{e}_s\cdots\boldsymbol{e}_{j+1}\boldsymbol{\sigma}_{[l]}^{(2)}(\boldsymbol{x}_\alpha)\mathrm{diag}\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\boldsymbol{f}_k\right)\boldsymbol{e}_{j-1}\cdots\boldsymbol{e}_0\frac{\langle\boldsymbol{p}_k,\boldsymbol{q}_k\rangle}{m}\\
&=\sum_l\frac{C}{\sqrt{m}}\boldsymbol{v}_l'(t)\frac{\langle\boldsymbol{p}_l,\boldsymbol{q}_l\rangle}{m},
\end{aligned}$$

since $f_k \in \mathbb{A}_0$, then $v_l'(t) \in \mathbb{A}_{r+1}$, and $p_l, q_l \in \mathbb{A}_0$.

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

$$v(t) = e_s \cdots e_{j+1} e_j e_{j-1} \cdots e_0 \to \widetilde{v}(t),$$

with

$$e_j = \sigma_{[l]}^{(u)}(x_\alpha)\mathrm{diag}\left(\left(\frac{W_t^{[l]}}{\sqrt{m}}\right)^{Q_1} g_1\right)\cdots\mathrm{diag}\left(\left(\frac{W_t^{[l]}}{\sqrt{m}}\right)^{Q_{u-1}} g_{u-1}\right)\left(\frac{c_{\mathrm{res}}}{L}\frac{W_t^{[l]}}{\sqrt{m}}\right)^{Q_u},$$

$$\text{or} \quad e_j = \left(\frac{c_{\mathrm{res}}}{L}\frac{\left(W_t^{[l]}\right)^{\mathsf{T}}}{\sqrt{m}}\right)^{Q_u} \sigma_{[l]}^{(u)}(x_\alpha)\mathrm{diag}\left(\left(\frac{W_t^{[l]}}{\sqrt{m}}\right)^{Q_1} g_1\right)\cdots\mathrm{diag}\left(\left(\frac{W_t^{[l]}}{\sqrt{m}}\right)^{Q_{u-1}} g_{u-1}\right),$$

and after applying replacement rules on $e_j \to e_j'$,

$$\begin{aligned} e_j' = {} & \frac{C}{\sqrt{m}}\sigma_{[l]}^{(u+1)}(x_\alpha)\mathrm{diag}(f_1)\mathrm{diag}\left(\left(\frac{W_t^{[l]}}{\sqrt{m}}\right)^{Q_1} g_1\right)\cdots\left(\frac{c_{\mathrm{res}}}{L}\frac{W_t^{[l]}}{\sqrt{m}}\right)^{Q_u}\frac{\langle p_1, q_1\rangle}{m} \\ & + \sum_k \frac{C}{\sqrt{m}}\sigma_{[l]}^{(u+1)}(x_\alpha)\mathrm{diag}\left(\left(\frac{W_t^{[l]}}{\sqrt{m}}\right)^{Q_0} f_k\right)\mathrm{diag}\left(\left(\frac{W_t^{[l]}}{\sqrt{m}}\right)^{Q_1} g_1\right) \\ & \cdots\left(\frac{c_{\mathrm{res}}}{L}\frac{W_t^{[l]}}{\sqrt{m}}\right)^{Q_u}\frac{\langle p_k, q_k\rangle}{m}, \end{aligned}$$

$$\begin{aligned} \text{or} \quad e_j' = {} & \frac{C}{\sqrt{m}}\left(\frac{c_{\mathrm{res}}}{L}\frac{\left(W_t^{[l]}\right)^{\mathsf{T}}}{\sqrt{m}}\right)^{Q_u}\sigma_{[l]}^{(u+1)}(x_\alpha)\mathrm{diag}(f_1)\cdots\mathrm{diag}\left(\left(\frac{W_t^{[l]}}{\sqrt{m}}\right)^{Q_{u-1}} g_{u-1}\right)\frac{\langle p_1, q_1\rangle}{m} \\ & + \sum_k \frac{C}{\sqrt{m}}\left(\frac{c_{\mathrm{res}}}{L}\frac{\left(W_t^{[l]}\right)^{\mathsf{T}}}{\sqrt{m}}\right)^{Q_u}\sigma_{[l]}^{(u+1)}(x_\alpha)\mathrm{diag}\left(\left(\frac{W_t^{[l]}}{\sqrt{m}}\right)^{Q_0} f_k\right) \\ & \cdots\mathrm{diag}\left(\left(\frac{W_t^{[l]}}{\sqrt{m}}\right)^{Q_{u-1}} g_{u-1}\right)\frac{\langle p_k, q_k\rangle}{m}, \end{aligned}$$

then

$$\widetilde{v}(t) = \sum_l \frac{C}{\sqrt{m}} v_l'(t)\frac{\langle p_l, q_l\rangle}{m},$$

since $f_k \in \mathbb{A}_0$, then $v_l'(t) \in \mathbb{A}_{r+1}$, and $p_l, q_l \in \mathbb{A}_0$.

Since $\frac{W_t^{[l]}}{\sqrt{m}}$ only appears at the starting or the middle position $\boldsymbol{e}_j$, we have that if $\boldsymbol{v}(t)\in\mathbb{A}_r$, then based on the replacement rules

$$\begin{aligned}\boldsymbol{v}(t)&=\boldsymbol{e}_s\boldsymbol{e}_{s-1}\cdots\boldsymbol{e}_{j+1}\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\boldsymbol{e}_{j-1}\cdots\boldsymbol{e}_1\boldsymbol{e}_0\to\widetilde{\boldsymbol{v}}(t),\\ \widetilde{\boldsymbol{v}}(t)&=\frac{C}{m}\boldsymbol{e}_s\boldsymbol{e}_{s-1}\cdots\boldsymbol{e}_{j+1}\mathrm{diag}(\boldsymbol{g})\,\mathbf{1}\otimes(\boldsymbol{x}_\beta^{[l-1]})^\intercal\boldsymbol{e}_{j-1}\cdots\boldsymbol{e}_1\boldsymbol{e}_0\\ &=\frac{C}{\sqrt{m}}\boldsymbol{e}_s\boldsymbol{e}_{s-1}\cdots\boldsymbol{e}_{j+1}\mathrm{diag}(\boldsymbol{g})\,\mathbf{1}\frac{\left\langle\boldsymbol{e}_{j-1}\cdots\boldsymbol{e}_1\boldsymbol{e}_0,\sqrt{m}\boldsymbol{x}_\beta^{[l-1]}\right\rangle}{m}\\ &=\frac{C}{\sqrt{m}}\boldsymbol{v}'(t)\frac{\langle\boldsymbol{p},\boldsymbol{q}\rangle}{m},\end{aligned}$$

with $\boldsymbol{v}'(t)\in\mathbb{A}_{r-s+1}$, and $\boldsymbol{p}\in\mathbb{A}_s$, $\boldsymbol{q}\in\mathbb{A}_0$, for some $s\geq 1$.

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

$$\begin{aligned}\boldsymbol{v}(t)&=\boldsymbol{e}_s\boldsymbol{e}_{s-1}\cdots\boldsymbol{e}_{j+1}\frac{(\boldsymbol{W}_t^{[l]})^\intercal}{\sqrt{m}}\boldsymbol{e}_{j-1}\cdots\boldsymbol{e}_1\boldsymbol{e}_0\to\widetilde{\boldsymbol{v}}(t),\\ \widetilde{\boldsymbol{v}}(t)&=\frac{C}{m}\boldsymbol{e}_s\boldsymbol{e}_{s-1}\cdots\boldsymbol{e}_{j+1}\boldsymbol{x}_\beta^{[l-1]}\otimes\mathbf{1}^\intercal\mathrm{diag}(\boldsymbol{g})\boldsymbol{e}_{j-1}\cdots\boldsymbol{e}_1\boldsymbol{e}_0\\ &=\frac{C}{\sqrt{m}}\boldsymbol{e}_s\boldsymbol{e}_{s-1}\cdots\boldsymbol{e}_{j+1}\sqrt{m}\boldsymbol{x}_\beta^{[l-1]}\frac{\left\langle\mathrm{diag}(\boldsymbol{g})\boldsymbol{e}_{j-1}\cdots\boldsymbol{e}_1\boldsymbol{e}_0,\mathbf{1}\right\rangle}{m}\\ &=\frac{C}{\sqrt{m}}\boldsymbol{v}'(t)\frac{\langle\boldsymbol{p},\boldsymbol{q}\rangle}{m},\end{aligned}$$

with $\boldsymbol{v}'(t)\in\mathbb{A}_{r-s}$, and $\boldsymbol{p}\in\mathbb{A}_{r-s+1}$, $\boldsymbol{q}\in\mathbb{A}_0$, for some $s\geq 1$. Since $\boldsymbol{E}_{t,\alpha}^{[l]}$ is situations combined with $\frac{W_t^{[l]}}{\sqrt{m}}$ and $\boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\alpha)$, so we will skip the analysis. □

From the discussion above, if we apply Proposition A.1 to $\mathcal{K}_t^{(r)}(\boldsymbol{x}_{\alpha_1},\boldsymbol{x}_{\alpha_2})$ inductively $(r-1)$ times, for each term in kernel $\mathcal{K}_t^{(r)}(\boldsymbol{x}_{\alpha_1},\boldsymbol{x}_{\alpha_2},\cdots,\boldsymbol{x}_{\alpha_r})$, it takes the form:

$$\frac{1}{m^{r/2-1}}\prod_{j=1}^{s}\frac{\left\langle\boldsymbol{v}_{2j-1}(t),\boldsymbol{v}_{2j}(t)\right\rangle}{m},\quad 1\leq s\leq r,\quad \boldsymbol{v}_i(t)\in\mathbb{A}_0\cup\mathbb{A}_1\cup\cdots\cup\mathbb{A}_{r-2}. \tag{A.1}$$

A.2 Apriori L^2 bounds for expressions in $\mathbb{A}_0$

We begin with an estimate on the empirical risk $R_S(\boldsymbol{\theta}_t)$.

Proposition A.2. *Under Assumptions* 4.1 *and* 4.2, *we have for* $t\geq 0$,

$$R_S(\boldsymbol{\theta}_t)\leq R_S(\boldsymbol{\theta}_0)\sim\mathcal{O}(1). \tag{A.2}$$

Proof. We get inequality (A.2) by non-negative definiteness of kernel $K_t^{(2)}(\cdot)$. From (4.6), we obtain that

$$\partial_t \sum_{\alpha=1}^{n} \|f_\alpha(t)-y_\alpha\|_2^2 = -\frac{2}{n}\sum_{\alpha,\beta=1}^{n} K_t^{(2)}(\boldsymbol{x}_\alpha,\boldsymbol{x}_\beta)(f_\alpha(t)-y_\alpha)(f_\beta(t)-t_\beta) \leq 0, \tag{A.3}$$

hence

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

which finish the proof of Proposition A.2. □

Our next proposition is mainly on the spectral property of the skip-connection matrices. This proposition is similar to Proposition B.1. in [24].

Proposition A.3. *Under Assumptions* 4.1 *and* 4.2*, we define* $\xi(t)$ *as follows*

$$\begin{aligned}\xi(t) = \sup_{0\leq t'\leq t} \max\Bigg\{1, \frac{1}{\sqrt{m}}\Bigg\{ &\left\|\boldsymbol{W}_{t'}^{[2]}\right\|_{2\to 2}, \left\|\left(\boldsymbol{W}_{t'}^{[2]}\right)^{\intercal}\right\|_{2\to 2}, \cdots \\ &\cdots, \left\|\boldsymbol{W}_{t'}^{[L]}\right\|_{2\to 2}, \left\|\left(\boldsymbol{W}_{t'}^{[L]}\right)^{\intercal}\right\|_{2\to 2}, \|\boldsymbol{a}_{t'}\|_2 \Bigg\}\Bigg\},\end{aligned} \tag{A.4}$$

then with high probability with respect to the random initialization, for $t \lesssim \sqrt{m}$

$$\xi(t) \leq c_{w,t}, \tag{A.5}$$

where $c_{w,t} > 2$ *is a constant independent of the depth of the network L.*

Moreover for $t \lesssim \sqrt{m}$*,* $c_{w,t}$ *has a uniform upper bound in t, i.e.,*

$$c_{w,t} \leq \bar{c}, \tag{A.6}$$

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

Proof. For the purpose of proving the proposition, we shall state two lemmas, Lemmas A.1 and A.2. Lemma A.1 is given out as Lemma G.2. in Du et al. [12], also consequence of the results in [46].

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

$$\|\boldsymbol{W}\|_{2\to 2} \leq c'_{w,0}\sqrt{m}, \tag{A.7}$$

where $c'_{w,0} > 2$ *is a constant.*

Remark A.1. This event is an event that holds with high probability.

Next concerning the term $\frac{1}{\sqrt{m}}\|\boldsymbol{a}_0\|_2$, we shall state a lemma on the tail bound of the chi-square distribution, using Lemma 1 from [29]

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

$$\mathbb{P}\left(Z\geq m+2\sqrt{mx}+2x\right)\leq e^{-x}. \tag{A.8}$$

Remark A.2. This event is also an event that holds with high probability.

Then if we write $2tm=m+(2t-1)m$, letting $x=\frac{mt}{10}$, we can obtain that

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

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. Hence, for $t=0$, $\xi(0)\leq\max\left\{1,c_{w,0}^{'},c_{w,0}^{''}\right\}$. We set $c_{w,0}$ as $c_{w,0}=\max\left\{1,c_{w,0}^{'},c_{w,0}^{''}\right\}$, then

$$\xi(0)\leq c_{w,0}. \tag{A.9}$$

In the following we are going to show the upper bound of $\partial_t\xi(t)$. In order to do that, we need to estimate L^2 bound on each output layer. For $l=1$,

$$\begin{aligned}\left\|\boldsymbol{x}^{[1]}\right\|_2&=\sqrt{\frac{c_\sigma}{m}}\left\|\sigma(\boldsymbol{W}_t^{[1]}\boldsymbol{x})\right\|_2\leq\sqrt{c_\sigma}\left(|\sigma(0)|+\frac{C_{\mathrm{Lip}}}{\sqrt{m}}\left\|\boldsymbol{W}_t^{[1]}\boldsymbol{x}\right\|_2\right)\\&\leq\sqrt{c_\sigma}C_{\mathrm{Lip}}\left(1+\xi(t)\|\boldsymbol{x}\|_2\right)\leq C\xi(t),\end{aligned} \tag{A.10}$$

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

$$\begin{aligned}\left\|\boldsymbol{x}^{[l]}\right\|_2&\leq\left\|\boldsymbol{x}^{[l-1]}\right\|_2+\frac{c_{\mathrm{res}}}{L\sqrt{m}}\left\|\sigma\left(\boldsymbol{W}_t^{[l]}\boldsymbol{x}^{[l-1]}\right)\right\|_2\\&\leq\left\|\boldsymbol{x}^{[l-1]}\right\|_2+\frac{c_{\mathrm{res}}}{L}\left(|\sigma(0)|+C_{\mathrm{Lip}}\xi(t)\left\|\boldsymbol{x}^{[l-1]}\right\|_2\right)\\&\leq\left\|\boldsymbol{x}^{[l-1]}\right\|_2+\frac{c_{\mathrm{res}}}{L}\left(C_{\mathrm{Lip}}+C_{\mathrm{Lip}}\xi(t)\left\|\boldsymbol{x}^{[l-1]}\right\|_2\right)\\&\leq\left(1+\frac{2c_{\mathrm{res}}}{L}\xi(t)\right)\left\|\boldsymbol{x}^{[l-1]}\right\|_2.\end{aligned} \tag{A.11}$$

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

$$\left\|\boldsymbol{x}^{[l]}\right\|_2\leq C\left(1+\frac{2c_{\mathrm{res}}}{L}\xi(t)\right)^{l-1}\xi(t). \tag{A.12}$$

Based on (3.7), (3.8), (3.9) and (3.10), combined with Proposition A.2

$$\begin{aligned}
\partial_t \left\| \boldsymbol{W}_t^{[l]} \right\|_{2\to 2} &\le \frac{1}{n}\sum_{\beta=1}^{n} \frac{C}{\sqrt{m}} \left\| \boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\beta) \left(\boldsymbol{E}_{t,\beta}^{[(l+1):L]} \right)^{\intercal} \boldsymbol{a}_t \right\|_2 \left\| \boldsymbol{x}_\beta^{[l-1]} \right\|_2 \left| f_\beta(t) - y_\beta \right| \\
&\le \frac{1}{n}\sum_{\beta=1}^{n} C C_{\mathrm{Lip}} \left(1 + \frac{c_{\mathrm{res}} C_{\mathrm{Lip}}}{L} \xi(t) \right)^{L-l} \xi(t) \left(1 + \frac{2c_{\mathrm{res}}}{L} \xi(t) \right)^{l-1} \xi(t) \left| f_\beta(t) - y_\beta \right| \\
&\le C \left(1 + \frac{2c_{\mathrm{res}}}{L} \xi(t) \right)^{L-1} \xi(t)^2 \sqrt{\frac{1}{n}\sum_{\beta=1}^{n} \left\| f_\beta(t) - y_\beta \right\|_2^2} \\
&\le C \left(1 + \frac{2c_{\mathrm{res}}}{L} \xi(t) \right)^{L-1} \xi(t)^2 \sqrt{R_S(\boldsymbol{\theta}_0)} \\
&\le C \left(1 + \frac{2c_{\mathrm{res}}}{L} \xi(t) \right)^{L-1} \xi(t)^2 \le C \exp(2c_{\mathrm{res}} \xi(t))\ \xi(t)^2, \qquad \text{(A.13)}
\end{aligned}$$

$$\begin{aligned}
\partial_t \left\| \boldsymbol{a}_t \right\|_2 &\le \frac{1}{n}\sum_{\beta=1}^{n} \left\| \boldsymbol{x}_\beta^{[L]} \right\|_2 \left| f_\beta(t) - y_\beta \right| \le C \left(1 + \frac{2c_{\mathrm{res}}}{L} \xi(t) \right)^{L-1} \xi(t) \sqrt{\frac{1}{n}\sum_{\beta=1}^{n} \left\| f_\beta(t) - y_\beta \right\|_2^2} \\
&\le C \left(1 + \frac{2c_{\mathrm{res}}}{L} \xi(t) \right)^{L-1} \xi(t) \sqrt{R_S(\boldsymbol{\theta}_0)} \\
&\le C \left(1 + \frac{2c_{\mathrm{res}}}{L} \xi(t) \right)^{L-1} \xi(t) \le C \exp(2c_{\mathrm{res}} \xi(t))\ \xi(t). \qquad \text{(A.14)}
\end{aligned}$$

Based on (A.13) and (A.14), we have

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

we can obtain an integration inequality,

$$\int_{\xi(0)}^{\xi(t)} \frac{\mathrm{d}u}{\exp(2c_{\mathrm{res}} u) u^2} \le \frac{Ct}{\sqrt{m}}. \qquad \text{(A.15)}$$

Hence the integration term on the LHS of (A.15) is

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

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

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

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

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

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

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

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

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

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

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

which finishes the proof of Proposition A.3. □

We state the inductive relation (A.12) as a proposition.

Proposition A.4. *Under Assumptions* 4.1 *and* 4.2*, we have with high probability with respect to the random initialization, for time* $t \lesssim \sqrt{m}$ *with* $0 \le l \le L$,

$$\left\| \boldsymbol{x}^{[l]} \right\|_2 \le C, \tag{A.16}$$

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

Remark A.3. We shall note that the constant C in Proposition A.4 only depends on $c_{\text{res}}, c_{w,0}$ and c_σ. However, for a fully-connected feedforward network, (A.16) in Proposition A.4 become

$$\left\| \boldsymbol{x}^{[l]} \right\|_2 \le C\, 2^l. \tag{A.17}$$

Note that the 2-norm for each output layer increase exponentially layer by layer for fully-connected network, showing that ResNet possesses more stability compared with fully-connected network.

Next we end this part by making an Apriori estimate on the L^2-norm for arbitrary vector $\boldsymbol{v}(t) \in \mathbb{A}_0$.

Proposition A.5. *Under Assumptions* 4.1 *and* 4.2*, with high probability with respect to the random initialization, uniformly for any vector* $\boldsymbol{v}(t)\in\mathbb{A}_0$ *and time* $t\lesssim\sqrt{m}$*, the following holds*

$$\|\boldsymbol{v}(t)\|_2\leq c\sqrt{m}, \tag{A.18}$$

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

Proof. We shall start our analysis on the whole expressions in set $\mathbb{A}_0$. For any vector $\boldsymbol{v}(t)\in\mathbb{A}_0$, we can write $\boldsymbol{v}(t)=\boldsymbol{e}_s\boldsymbol{e}_{s-1}\cdots\boldsymbol{e}_1\boldsymbol{e}_0$ with $0\leq s\leq 4L$.

We start with the estimate on $\boldsymbol{e}_0$, since $\boldsymbol{e}_0$ is chosen following the rules:

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

- (a). If $\boldsymbol{e}_0=\boldsymbol{a}_t$, then by Lemma A.2, for $t\lesssim\sqrt{m}$,

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

- (b). If $\boldsymbol{e}_0=\sqrt{m}\boldsymbol{x}_\beta^{[l]}$ where $1\leq l\leq L$, then based on Proposition A.4, 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 other terms in the expression $\boldsymbol{e}_j$ where $j\geq 1$.

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

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

  Since $\|\boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\beta)\|_{2\to2}\leq C_{\mathrm{Lip}}\leq 1$, thus for all $j\geq1$ with $\boldsymbol{e}_j=\boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\beta)$

$$\|\boldsymbol{v}(t)\|_2\leq\left(C_{\mathrm{Lip}}\right)^{4L}c\sqrt{m}\leq c\sqrt{m}.$$

- (ii). If $\boldsymbol{e}_j=\boldsymbol{E}_{t,\beta}^{[l]}$ or $\boldsymbol{e}_j=\left(\boldsymbol{E}_{t,\beta}^{[l]}\right)^\top$, then based on Proposition A.3

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

  Since

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

  thus for all $j\geq1$ with $\boldsymbol{e}_j=\boldsymbol{E}_{t,\beta}^{[l]}$ or $\boldsymbol{e}_j=(\boldsymbol{E}_{t,\beta}^{[l]})^\top$,

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

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

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

Combining these two observations, we finish the proof. □

Thus, if we define the quantity $\xi_{\infty,0}(t)$ as follows,

$$\xi_{\infty,0}(t)=\sup_{0\leq t'\leq t}\left\{\left\|\boldsymbol{v}(t')\right\|_2 \colon \boldsymbol{v}(t')\in\mathbb{A}_0\right\}. \tag{A.19}$$

Then directly from Proposition A.5, for time $0\leq t\leq\sqrt{m}/(\ln m)^{C'}$, the following holds

$$\xi_{\infty,0}(t)\leq c\sqrt{m}. \tag{A.20}$$

A.3 Apriori L^∞ bounds for expressions in $\mathbb{A}_0$

In this part, we shall make estimate on the quantity $\eta_{\infty,0}(t)$ defined below

$$\eta_{\infty,0}(t)=\sup_{0\leq t'\leq t}\left\{\left\|\boldsymbol{v}(t')\right\|_\infty \colon \boldsymbol{v}(t')\in\mathbb{A}_0\right\}. \tag{A.21}$$

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

Lemma A.3. *For any i.i.d. normal distribution $X_1,X_2,\cdots,X_m\sim\mathcal{N}(0,1)$, it holds with high probability that the L^∞-norm of the gaussian vector $\boldsymbol{X}=(X_1,X_2,\cdots,X_m)^\intercal$ is upper bounded by*

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

for some really large constant $C>0$.

Proof. For any $X_i\sim\mathcal{N}(0,1)$, we have that for some $\varepsilon,\lambda>0$

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

We optimize over λ,

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

By taking absolute value

$$\mathbb{P}(|X_i|\geq\varepsilon)\leq 2\exp\left(-\frac{\varepsilon^2}{2}\right).$$

Hence if we take over m unions

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

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

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

Note that when $C>0$ is really large, $(\ln m)^C \approx m^\varepsilon$ for some small $\varepsilon>0$. □

We now state a lemma on the matrix two to infinity norm.

Lemma A.4. *Given a matrix $\boldsymbol{W}\in\mathbb{R}^{m\times m}$ with each entry $W_{i,j}\sim\mathcal{N}(0,1)$, then with high probability, the following holds*

$$\|\boldsymbol{W}\|_{2\to\infty} = \sup_{\|\boldsymbol{x}\|_2=1} \|\boldsymbol{W}\boldsymbol{x}\|_\infty \leq (\ln m)^C. \tag{A.22}$$

Proof. Note that $\boldsymbol{W}\boldsymbol{x}$ shares the same distribution as the Gaussian vector $\boldsymbol{X}$ in Lemma A.3, i.e., $\boldsymbol{W}\boldsymbol{x}\sim\boldsymbol{X}$. Then apply Lemma A.3 directly, we obtain the result. □

Finally, to evaluate $\eta_{\infty,0}(t)$, we need to state a lemma.

Lemma A.5. *Under Assumptions* 4.1 *and* 4.2, *for any vector $\boldsymbol{v}(t)\in\mathbb{A}_0$, we can write*

$$\boldsymbol{v}(t) = \boldsymbol{e}_s\boldsymbol{e}_{s-1}\cdots\boldsymbol{e}_1\boldsymbol{e}_0, \quad 0\leq s\leq 4L, \quad t\geq 0.$$

*For some vectors in $\mathbb{A}_0$ with **length** q, we define $\eta_{q,0}(t)$ as*

$$\eta_{q,0}(t) := \sup_{0\leq t'\leq t}\left\{\left\|\boldsymbol{v}_q(t')\right\|_\infty \,|\, \boldsymbol{v}_q(t') = \boldsymbol{e}_q\boldsymbol{e}_{q-1}\cdots\boldsymbol{e}_1\boldsymbol{e}_0,\ \boldsymbol{v}_q(t')\in\mathbb{A}_0\right\}. \tag{A.23}$$

Moreover, we define $\omega(t)$ as

$$\omega(t) := \sup_{0\leq t'\leq t} \max\left\{\left\|\boldsymbol{W}_{t'}^{[2]}\right\|_{2\to\infty}, \left\|\left(\boldsymbol{W}_{t'}^{[2]}\right)^\mathsf{T}\right\|_{2\to\infty}, \cdots, \left\|\boldsymbol{W}_{t'}^{[L]}\right\|_{2\to\infty}, \left\|\left(\boldsymbol{W}_{t'}^{[L]}\right)^\mathsf{T}\right\|_{2\to\infty}\right\},$$

then with high probability with respect to the random initialization, for $t\lesssim\sqrt{m}$,

$$\eta_{q,0}(t) \leq \eta_{0,0}(t) + c\ \omega(t)\left(1+\frac{c_{\text{res}}c_{w,t}}{L}\right)^q, \tag{A.24}$$

where constant $c>0$ is independent of depth L, and $c_{w,t}$ has been defined in Proposition A.3.

Proof. Since for any vector $\boldsymbol{v}_q(t)\in\mathbb{A}_0$ of length q, $0\le q\le 4L$, we can write $\boldsymbol{v}_q(t)$ into

$$\boldsymbol{v}_q(t)=\boldsymbol{e}_q\boldsymbol{e}_{q-1}\cdots\boldsymbol{e}_1\boldsymbol{e}_0,$$

then we shall prove (A.24) by performing induction on q. Firstly, for $q=0$, (A.24) is trivial. While for $q\ge 1$, we shall investigate on the terms $\boldsymbol{e}_j$ in the expression $\boldsymbol{v}_q(t)$, where $j\ge 1$.

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

$$\begin{aligned}\left\|\boldsymbol{v}_q(t)\right\|_\infty&=\left\|\boldsymbol{e}_q\boldsymbol{e}_{q-1}\cdots\boldsymbol{e}_1\boldsymbol{e}_0\right\|_\infty\\&=\left\|\boldsymbol{e}_q\right\|_{\infty\to\infty}\left\|\boldsymbol{e}_{q-1}\right\|_{\infty\to\infty}\cdots\|\boldsymbol{e}_1\|_{\infty\to\infty}\|\boldsymbol{e}_0\|_\infty,\end{aligned}$$

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

$$\left\|\boldsymbol{v}_q(t)\right\|_\infty\le\left(C_{\mathrm{Lip}}\right)^q c(\ln m)^C\le c(\ln m)^C.$$

- (ii). If $\boldsymbol{e}_j=\boldsymbol{E}_{t,\beta}^{[l]}$ or $\boldsymbol{e}_j=\left(\boldsymbol{E}_{t,\beta}^{[l]}\right)^{\mathsf{T}}$ where $2\le l\le L$, $\left\|\boldsymbol{e}_j\right\|_{\infty\to\infty}\ge 1$, so we need to tackle it differently

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

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

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

$$\left\|\boldsymbol{v}_q(t)\right\|_\infty\le\left\|\boldsymbol{v}_{q-1}(t)\right\|_\infty+\frac{c_{\mathrm{res}}}{L\sqrt{m}}\omega(t)\left\|\boldsymbol{v}_{q-1}(t)\right\|_2.$$

  Based on Proposition A.3

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

then

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

Inductively we have

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

where we use the property of a geometric sum. By taking supreme on both sides, we have

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

This completes the proof. □

Based on these lemmas, recall definition (A.21), we are able to make a proposition on the quantity $\eta_{\infty,0}(t)$ at $t=0$.

Proposition A.6. *Under Assumptions* 4.1 *and* 4.2*, with high probability with respect to the random initialization*

$$\eta_{\infty,0}(0) \le c(\ln m)^C, \tag{A.25}$$

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

Proof. As always, for any vector $\boldsymbol{v}(t) \in \mathbb{A}_0$, we can write $\boldsymbol{v}(t)$ as

$$\boldsymbol{v}(t) = \boldsymbol{e}_s\boldsymbol{e}_{s-1}\cdots\boldsymbol{e}_1\boldsymbol{e}_0, \quad 0 \le s \le 4L.$$

We start with the estimate on $\eta_{0,0}(0)$, since $\boldsymbol{e}_0$ is chosen following the rules:

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

- (a). If $\boldsymbol{e}_0 = \boldsymbol{a}_t$, then at $t=0$, by Lemma A.3,

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

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

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

moreover, for $l \geq 1$, based on Proposition A.4,

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

inductively for $1 \leq l \leq L$,

$$\left\|\sqrt{m}\boldsymbol{x}_\beta^{[l]}\right\|_\infty \leq c\left(1 + \frac{l}{L}\right)(\ln m)^C \leq c(\ln m)^C, \tag{A.26}$$

where c is independent of the depth L.

Hence we have

$$\eta_{0,0}(0) \leq c(\ln m)^C. \tag{A.27}$$

Directly from Lemma A.5

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

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

Our next proposition is on $\eta_{\infty,0}(t)$ for time $0 \leq t \leq \sqrt{m}/(\ln m)^{C'}$.

Proposition A.7. *Under Assumptions* 4.1 *and* 4.2, *with high probability with respect to the random initialization, for time* $0 \leq t \leq \sqrt{m}/(\ln m)^{C'}$, *the following holds*

$$\eta_{\infty,0}(t) \leq c(\ln m)^C, \tag{A.28}$$

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

Proof. We shall start with the estimate on $\eta_{0,0}(t)$, since $\boldsymbol{e}_0$ is chosen following the rules:

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

We observe that from the replacement rules given in Section 5.1,

$$\begin{aligned}
\partial_t \boldsymbol{a}_t &= -\frac{1}{n}\sum_{\beta=1}^{n}\frac{1}{\sqrt{m}}\sqrt{m}\boldsymbol{x}_\beta^{[L]}(f_\beta(t)-y_\beta),\\
\partial_t \sqrt{m}\boldsymbol{x}_\alpha^{[l]} &= -\frac{1}{n}\sum_{\beta=1}^{n}\frac{c_\sigma}{\sqrt{m}}\boldsymbol{E}_{t,\alpha}^{[2:l]}\boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\alpha)\boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[2:L]}\right)^{\mathsf{T}}\boldsymbol{a}_t\langle \boldsymbol{x}_\alpha,\boldsymbol{x}_\beta\rangle(f_\beta(t)-y_\beta)\\
&\quad +-\frac{1}{n}\sum_{\beta=1}^{n}\sum_{k=2}^{l}\frac{c_{\mathrm{res}}^2}{L^2\sqrt{m}}\boldsymbol{E}_{t,\alpha}^{[(k+1):l]}\boldsymbol{\sigma}_{[k]}^{(1)}(\boldsymbol{x}_\alpha)\boldsymbol{\sigma}_{[k]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[(k+1):L]}\right)^{\mathsf{T}}\\
&\qquad \boldsymbol{a}_t\left\langle \boldsymbol{x}_\alpha^{[k-1]},\boldsymbol{x}_\beta^{[k-1]}\right\rangle(f_\beta(t)-y_\beta).
\end{aligned}$$

Since for $0\le t\le \sqrt{m}/(\ln m)^{C'}$, then by Proposition A.4

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

By taking supreme on time $0\le t\le \sqrt{m}/(\ln m)^{C'}$, we have

$$\eta_{0,0}(t)\le c(\ln m)^C+\frac{C}{\sqrt{m}}\int_0^t\eta_{\infty,0}(s)\mathrm{d}s. \tag{A.29}$$

For the auxiliary term $\omega(t)$, from the replacement rules again, for $2\le l\le L$

$$\begin{aligned}
\partial_t\boldsymbol{W}_t^{[l]} &= -\frac{1}{n}\sum_{\beta=1}^{n}\frac{c_{\mathrm{res}}}{L\sqrt{m}}\boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[(l+1):L]}\right)^{\mathsf{T}}\boldsymbol{a}_t\otimes(\boldsymbol{x}_\beta^{[l-1]})^{\mathsf{T}}(f_\beta(t)-y_\beta),\\
\partial_t\left(\boldsymbol{W}_t^{[l]}\right)^{\mathsf{T}} &= -\frac{1}{n}\sum_{\beta=1}^{n}\frac{c_{\mathrm{res}}}{L\sqrt{m}}\boldsymbol{x}_\beta^{[l-1]}\otimes\left(\boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[(l+1):L]}\right)^{\mathsf{T}}\boldsymbol{a}_t\right)^{\mathsf{T}}(f_\beta(t)-y_\beta),
\end{aligned}$$

then by Proposition A.5

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

Hence by taking supreme on time $0\le t\le \sqrt{m}/(\ln m)^{C'}$, we have

$$\omega(t)\le(\ln m)^C+\frac{C}{\sqrt{m}}\int_0^t\eta_{\infty,0}(s)\mathrm{d}s. \tag{A.30}$$

Directly from Lemma A.5

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

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

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

This gives us a Gronwall-type inequality, we have that

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

To sum up, for $t \le \sqrt{m}/(\ln m)^{C'}$, the following holds

$$\eta_{\infty,0}(t) \le c(\ln m)^C, \tag{A.31}$$

which finishes the proof. □

A.4 Apriori L^2 and L^∞ bounds for expression in $\mathbb{A}_r$, $r \ge 1$

In this part, we shall make estimates for $\|\cdot\|_\infty$ and $\|\cdot\|_2$ of vectors belonging to higher order sets, i.e., $\mathbb{A}_r$, $r \ge 1$. Then it is natural for us to define several quantities for some vectors in $\mathbb{A}_r$ with length q

$$\xi_{q,r}(t) := \sup_{0\le t'\le t}\left\{\left\|\boldsymbol{v}_q(t')\right\|_2 :\, \boldsymbol{v}_q(t') = \boldsymbol{e}_q\boldsymbol{e}_{q-1}\cdots\boldsymbol{e}_1\boldsymbol{e}_0,\, \boldsymbol{v}_q(t') \in \mathbb{A}_r\right\}. \tag{A.32}$$

Note that from Propositions A.3 and A.5,

$$\xi_{q,0}(t) \le c\left(1+\frac{c_{\text{res}}c_{w,t}}{L}\right)^q \sqrt{m}. \tag{A.33}$$

Moreover, we define that

$$\xi_{\infty,r}(t) = \sup_{0\le q\le 4L}\left\{\xi_{q,r}(t)\right\}. \tag{A.34}$$

Then by taking supreme on $0 \le q \le 4L$ in (A.33)

$$\xi_{\infty,0}(t) \le c\sqrt{m}, \tag{A.35}$$

and recalling the definition we made in Section A.3, we similarly define

$$\eta_{q,r}(t) := \sup_{0\le t'\le t}\left\{\left\|\boldsymbol{v}_q(t')\right\|_\infty \,|\, \boldsymbol{v}_q(t') = \boldsymbol{e}_q\boldsymbol{e}_{q-1}\cdots\boldsymbol{e}_1\boldsymbol{e}_0,\, \boldsymbol{v}_q(t') \in \mathbb{A}_r\right\}. \tag{A.36}$$

Moreover, we also define that

$$\eta_{\infty,r}(t)=\sup_{0\le q\le 4L}\{\eta_{q,r}(t)\}. \tag{A.37}$$

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

$$\boldsymbol{v}(t)=\boldsymbol{e}_s\boldsymbol{e}_{s-1}\cdots\boldsymbol{e}_1\boldsymbol{e}_0,\ 0\le s\le 4L.$$

We shall start with the estimate on $\boldsymbol{e}_0$. Since $\boldsymbol{e}_0$ is chosen following the rules:

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

$\|\mathbf{1}\|_\infty=1,\|\mathbf{1}\|_2=\sqrt{m}$, then for time $0\le t\le\sqrt{m}/(\ln m)^{C'}$, by Propositions A.3 and A.7,

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

Now we proceed to other terms in the expression $\boldsymbol{e}_j$ where $j\ge 1$. For each $\boldsymbol{e}_j$, there are several cases:

- (i) $\boldsymbol{e}_j=\boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\beta)$, $\boldsymbol{e}_j=\boldsymbol{E}_{t,\beta}^{[l]}$ or $\boldsymbol{e}_j=\left(\boldsymbol{E}_{t,\beta}^{[l]}\right)^{\mathsf{T}}$, $2\le l\le L$.
- (ii) $\boldsymbol{e}_j=\mathrm{diag}(\boldsymbol{g})$.
- (iii)

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

  or

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

By our observation, the total number of diag operations in $\boldsymbol{v}(t)\in\mathbb{A}_r$ is r, and that is how we characterize a vector belonging to different hierarchical sets. Especially if for one of those $\boldsymbol{e}_j$ belongs to case (iii), there are two scenarios:

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

- $Q_{u+1}=1$, since diagonal matrices commute, $\boldsymbol{e}_j$ writes into

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

or

$$\boldsymbol{e}_j=\left(\frac{c_{\text{res}}}{L}\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{\mathsf{T}}\boldsymbol{\sigma}_{[l]}^{(u+1)}(\boldsymbol{x}_\beta)\operatorname{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_1}\boldsymbol{g}_1\right)\cdots\operatorname{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right)^{Q_u}\boldsymbol{g}_u\right).$$

We shall take advantage of the special structure of $\boldsymbol{e}_j$. Define a new type of skip-connection matrix, $\widetilde{\boldsymbol{E}}_{t,\beta}^{[l,r]}$, for $r\geq 2$:

$$\widetilde{\boldsymbol{E}}_{t,\beta}^{[l,r]}:=\left(\boldsymbol{I}_m+\frac{c_{\text{res}}}{L}\boldsymbol{\sigma}_{[l]}^{(r)}(\boldsymbol{x}_\beta)\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\right). \tag{A.38}$$

Then we can write $\boldsymbol{e}_j$ into

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

or

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

To illustrate such relation, if some vector $\bar{\boldsymbol{v}}(t)$ contains $\boldsymbol{e}_j$ belonging to case (iii), we write it as

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

From the above analysis, we are able to characterize an element in set $\mathbb{A}_r$. If $\boldsymbol{v}(t)\in\mathbb{A}_r$, then as always, we write it as

$$\boldsymbol{v}(t)=\boldsymbol{e}_s\boldsymbol{e}_{s-1}\cdots\boldsymbol{e}_1\boldsymbol{e}_0,\quad 0\le s\le 4L,$$

and there exists $\boldsymbol{e}_{j_1},\boldsymbol{e}_{j_2},\cdots,\boldsymbol{e}_{j_k}$, such that

$$\begin{aligned}\boldsymbol{e}_{j_1}=&\mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l_1]}}{\sqrt{m}}\right)^{Q_1}\boldsymbol{g}_1\right),\quad \boldsymbol{g}_1\in\mathbb{A}_{r_1-1},\\\boldsymbol{e}_{j_2}=&\mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l_2]}}{\sqrt{m}}\right)^{Q_2}\boldsymbol{g}_2\right),\quad \boldsymbol{g}_2\in\mathbb{A}_{r_2-1},\\&\vdots\\\boldsymbol{e}_{j_k}=&\mathrm{diag}\left(\left(\frac{\boldsymbol{W}_t^{[l_k]}}{\sqrt{m}}\right)^{Q_k}\boldsymbol{g}_k\right),\quad \boldsymbol{g}_k\in\mathbb{A}_{r_k-1},\end{aligned}$$

with

$$r_1+r_2+\cdots+r_k=r,\quad r_1,r_2,\cdots,r_k\in\mathbb{N}^+. \tag{A.39}$$

Eq. (A.39) serves as the counting of the number of diag operations contained in $\boldsymbol{v}(t)$, while for other $\boldsymbol{e}_j(j\notin\{j_1,j_2,\cdots,j_k,0\})$, chosen from the following sets

$$\left\{\boldsymbol{E}_{t,\beta}^{[l]},\left(\boldsymbol{E}_{t,\beta}^{[l]}\right)^{\mathsf{T}}:\,2\le l\le L\right\}_{1\le\beta\le n}, \tag{A.40}$$

$$\left\{\boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_\beta):\,1\le l\le L\right\}_{1\le\beta\le n}, \tag{A.41}$$

$$\left\{\widetilde{\boldsymbol{E}}_{t,\beta}^{[l,p]},\left(\widetilde{\boldsymbol{E}}_{t,\beta}^{[l,p]}\right)^{\mathsf{T}}:\,2\le l\le L,\,p\ge 2\right\}_{1\le\beta\le n}, \tag{A.42}$$

note that the elements in set (A.40) and set (A.42) share the same matrix properties, thanks to Assumption 4.1 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 contained in each vector.

Proposition A.8. *Under Assumptions* 4.1 *and* 4.2*, with high probability with respect to the random initialization, for some finite $r\geq 1$ and time $0\leq t\leq \sqrt{m}/(\ln m)^{C'}$, the following holds*

$$\xi_{\infty,r}(t)\leq c(\ln m)^{C}\sqrt{m}, \tag{A.43}$$

$$\eta_{\infty,r}(t)\leq c(\ln m)^{C}, \tag{A.44}$$

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

Proof. Recalling the definition of $\omega(t)$, $\eta_{\infty,0}(t)$ and $\xi_{\infty,0}(t)$, for time $0\leq t\leq \sqrt{m}/(\ln m)^{C'}$, the following holds with high probability,

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

Let's start with $r=1$. For any $\boldsymbol{v}(t)\in\mathbb{A}_1$, since there is only one solution to equation (A.39), there exists one and only one index i, such that $\boldsymbol{e}_i=\mathrm{diag}(\boldsymbol{g})$ or $\boldsymbol{e}_i=\mathrm{diag}(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\boldsymbol{g})$, with $\boldsymbol{g}\in\mathbb{A}_0$. Then we have

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

for $q>i$,

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

Inductively, we have

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

By taking supreme on q and i, this leads to

$$\xi_{\infty,1}(t)\leq \exp(4c_{\text{res}}c_{w,t})c(\ln m)^C\xi_{\infty,0}(t)\leq c(\ln m)^C\sqrt{m}. \tag{A.45}$$

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

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

and for $q>i$, inductively

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

Then by taking supreme on q and i, combined with (A.45), we obtain

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

In the following we assume that (A.43) and (A.44) holds for $1,2,\cdots,r-1$ and prove it for r.

If $v(t)\in\mathbb{A}_r$, then as always, we write it as

$$v(t)=e_se_{s-1}\cdots e_1e_0,\ 0\leq s\leq 4L,$$

and there exists $e_{j_1},e_{j_2},\cdots,e_{j_k}$, such that

$$\begin{aligned}e_{j_1}&=\text{diag}\left(\left(\frac{W_t^{[l_1]}}{\sqrt{m}}\right)^{Q_1}g_1\right),\quad g_1\in\mathbb{A}_{r_1-1},\\e_{j_2}&=\text{diag}\left(\left(\frac{W_t^{[l_2]}}{\sqrt{m}}\right)^{Q_2}g_2\right),\quad g_2\in\mathbb{A}_{r_2-1},\\&\vdots\\e_{j_k}&=\text{diag}\left(\left(\frac{W_t^{[l_k]}}{\sqrt{m}}\right)^{Q_k}g_k\right),\quad g_k\in\mathbb{A}_{r_k-1},\end{aligned}$$

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 $j_1,j_2,\cdots,j_k$, i.e.,

$$i=\max\{j_1,j_2,\cdots,j_k\},$$

and without loss of generality, let $i=j_1$. We have $\boldsymbol{e}_i=\mathrm{diag}(\boldsymbol{g}_1)$ or $\boldsymbol{e}_i=\mathrm{diag}(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\boldsymbol{g}_1)$ with $\boldsymbol{g}_1\in\mathbb{A}_{r_1-1}$, then

$$\begin{aligned}\xi_{i,r}(t)&\le\sup_{\boldsymbol{g}\in\mathbb{A}_{r_1-1}}\|\mathrm{diag}(\boldsymbol{g})\|_{2\to2}\xi_{i-1,r-r_1}(t)\\&\le\sup_{\boldsymbol{g}\in\mathbb{A}_{r_1-1}}\|\boldsymbol{g}\|_\infty\xi_{i-1,r-r_1}(t)\le\eta_{\infty,r_1-1}(t)\xi_{i-1,r-r_1}(t)\le c(\ln m)^C\xi_{i-1,r-r_1}(t),\\ \text{or}\quad\xi_{i,r}(t)&\le\sup_{\boldsymbol{g}\in\mathbb{A}_{r_1-1}}\left\|\mathrm{diag}\left(\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\boldsymbol{g}\right)\right\|_{2\to2}\xi_{i-1,r-r_1}(t)\le\sup_{\boldsymbol{g}\in\mathbb{A}_{r_1-1}}\left\|\frac{\boldsymbol{W}_t^{[l]}}{\sqrt{m}}\boldsymbol{g}\right\|_\infty\xi_{i-1,r-r_1}(t)\\&\le\frac{\omega(t)}{\sqrt{m}}\xi_{\infty,r_1-1}(t)\xi_{i-1,r-r_1}(t)\le c(\ln m)^C\xi_{i-1,r-r_1}(t),\end{aligned}$$

inductively

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

Then by taking supreme on q and i, we obtain

$$\xi_{\infty,r}(t)\le\exp(4c_{\mathrm{res}}c_{w,t})c(\ln m)^C\xi_{\infty,r-r_1}(t)\le c(\ln m)^C\sqrt{m}.\tag{A.46}$$

For $\eta_{i,r}(t)$, we have

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

and for $q>i$

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

then by taking supreme on q and i,

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

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

B Least eigenvalue of Gram matrices

We shall recall the Gram matrices defined in Section 4.2. We first define a series of matrices $\{\widetilde{\boldsymbol{K}}^{[l]}\}_{l=1}^{L}$, $\{\widetilde{\boldsymbol{A}}^{[l]}\}_{l=1}^{L+1}$, and a series of vectors $\{\widetilde{\boldsymbol{b}}^{[l]}\}_{l=1}^{L}$. Given the input samples $\mathcal{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$, $\|\boldsymbol{x}_i\|_2=1$, for $1\le i\le n$, and $\boldsymbol{x}_i \nparallel \boldsymbol{x}_j$, for any $i\ne j$,

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

Given these definitions, we define that for $2\le l\le L-1$,

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

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

$$\boldsymbol{K}_{ij}^{[l]}=\frac{c_{\text{res}}^2}{L^2}\widetilde{\boldsymbol{K}}_{ij}^{[l-1]}\lim_{m\to\infty}\frac{1}{m}\left\langle\boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_i)\left(\boldsymbol{E}_{0,i}^{[(l+1):L]}\right)^{\intercal}\boldsymbol{a}_0,\boldsymbol{\sigma}_{[l]}^{(1)}(\boldsymbol{x}_j)\left(\boldsymbol{E}_{0,j}^{[(l+1):L]}\right)^{\intercal}\boldsymbol{a}_0\right\rangle, \tag{B.3}$$

$$\boldsymbol{K}_{ij}^{[1]}=c_\sigma\widetilde{\boldsymbol{K}}_{ij}^{[0]}\lim_{m\to\infty}\frac{1}{m}\left\langle\boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_i)\left(\boldsymbol{E}_{0,i}^{[2:L]}\right)^{\intercal}\boldsymbol{a}_0,\boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_j)\left(\boldsymbol{E}_{0,j}^{[2:L]}\right)^{\intercal}\boldsymbol{a}_0\right\rangle. \tag{B.4}$$

We shall state two lemmas concerning full rankness of the Gram matrices, which have been stated as Lemma F.1. and Lemma F.2. in Du et al. [12].

Lemma B.1. *Assume $\sigma(\cdot)$ is analytic and not a polynomial function. Consider input data set as $\mathcal{V}=\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_n\}$, and non-parallel with each other, i.e., $\boldsymbol{v}_j\notin\text{span}(\boldsymbol{v}_k)$ for any $j\neq k$, we define*

$$\boldsymbol{G}(\mathcal{V})_{ij}:=\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}\left[\sigma(\boldsymbol{w}^{\intercal}\boldsymbol{v}_i)\sigma(\boldsymbol{w}^{\intercal}\boldsymbol{v}_j)\right], \tag{B.5}$$

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

Similar to Lemma B.1, we have Lemma B.2

Lemma B.2. *Assume $\sigma(\cdot)$ is analytic and not a polynomial function. Consider input data set as $\mathcal{V}=\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_n\}$, and non-parallel with each other, i.e., $\boldsymbol{v}_j\notin\text{span}(\boldsymbol{v}_k)$ for any $j\neq k$, we define*

$$\boldsymbol{G}(\mathcal{V})_{ij}:=\mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}\left[\sigma^{(1)}(\boldsymbol{w}^{\intercal}\boldsymbol{v}_i)\sigma^{(1)}(\boldsymbol{w}^{\intercal}\boldsymbol{v}_j)\left(\boldsymbol{v}_i^{\intercal}\boldsymbol{v}_j\right)\right], \tag{B.6}$$

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

Now we proceed to quantify the least eigenvalues of these Gram matrices.

B.1 Full rankness for $(L+1)$-th Gram matrix

We begin this part by a lemma on the estimate of the entry of Gram matrices,

Lemma B.3. *Given the input samples $\mathcal{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$, $\|\boldsymbol{x}_i\|_2=1$, $1\leq i\leq n$, and $\boldsymbol{x}_i\nparallel\boldsymbol{x}_j$, for any $i\neq j$, then for every fixed l, where $1\leq l\leq L$, each diagonal entry of $\widetilde{\boldsymbol{K}}^{[l]}$ is the same with each other. Also for every fixed l, where $1\leq l\leq L$, each element of the vector $\widetilde{\boldsymbol{b}}^{[l]}$ is the same with each other, i.e.,*

$$\widetilde{\boldsymbol{K}}_{ii}^{[l_1]}=\widetilde{\boldsymbol{K}}_{jj}^{[l_1]},\quad \widetilde{\boldsymbol{b}}_i^{[l_2]}=\widetilde{\boldsymbol{b}}_j^{[l_2]},\quad i\neq j.$$

Moreover

$$\left(1-\frac{l}{L}\frac{c}{\sqrt{c_\sigma}}\right)^2\leq\widetilde{\boldsymbol{K}}_{ii}^{[l]}\leq\left(1+\frac{l}{L}\frac{c}{\sqrt{c_\sigma}}\right)^2, \tag{B.7}$$

and

$$\left(\widetilde{\boldsymbol{b}}_i^{[l]}\right)^2<\widetilde{\boldsymbol{K}}_{ii}^{[l]}, \tag{B.8}$$

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

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

Since based on definition, recall that $c_\sigma=\left(\mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[\sigma(x)^2\right]\right)^{-1}$,

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

and

$$\widetilde{\boldsymbol{b}}_i^{[1]}=\sqrt{c_\sigma}\mathbb{E}_{u\sim\mathcal{N}(0,\widetilde{\boldsymbol{K}}_{ii}^{[0]})}\left[\sigma(u)\right]=\sqrt{c_\sigma}\mathbb{E}_{u\sim\mathcal{N}(0,1)}\left[\sigma(u)\right],$$

then

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

the last inequality 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 our choice of i, $\widetilde{\boldsymbol{K}}_{ii}^{[1]}=\widetilde{\boldsymbol{K}}_{jj}^{[1]}$, $\widetilde{\boldsymbol{b}}_i^{[1]}=\widetilde{\boldsymbol{b}}_j^{[1]}$, for any $i\neq j$.

Now we assume that it holds for superscript being $1,2,\cdots,l-1$ and want to show that it holds for l. Hence based on definition

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

such quantities are also independent of our choice of i.

Moreover we would like to show that (B.7) and (B.8) hold for all l.

Firstly, for $\widetilde{\boldsymbol{b}}_i^{[l]}$, assume (B.8) holds for $1,2,\cdots,l-1$, then we have

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

showing that (B.8) holds for l.

For $\widetilde{\boldsymbol{K}}_{ii}^{[l]}$, we have

$$\begin{aligned}&\left(\sqrt{\widetilde{\boldsymbol{K}}_{ii}^{[l-1]}}-\frac{c_{\text{res}}}{L}\sqrt{\mathbb{E}_{u\sim\mathcal{N}(0,\widetilde{\boldsymbol{K}}_{ii}^{[l-1]})}[\sigma(u)^2]}\right)^2\\ \leq&\widetilde{\boldsymbol{K}}_{ii}^{[l]}\leq\left(\sqrt{\widetilde{\boldsymbol{K}}_{ii}^{[l-1]}}+\frac{c_{\text{res}}}{L}\sqrt{\mathbb{E}_{u\sim\mathcal{N}(0,\widetilde{\boldsymbol{K}}_{ii}^{[l-1]})}[\sigma(u)^2]}\right)^2.\end{aligned}\tag{B.9}$$

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

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

Then

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

By induction

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

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

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

Then if we choose c wisely, for example, letting

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

by our choice of c, combined with (B.9)

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

and hence

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

which finishes our proof. □

Our next lemma is crucial in that it revels a 'covariance-type' structure for the Gram matrices. We need to introduce a standard notation related to matrices. We denote that $A \succeq B$ if and only if $A-B$ is a semi-positive definite matrix, and $A \succ B$ if and only if $A-B$ is a strictly positive definite matrix.

Proposition B.1. *Given the input samples $\mathcal{X}=\{x_1,x_2,\cdots,x_n\}$, $\|x_i\|_2=1$, for $1\le i\le n$, and $x_i \nparallel x_j$, $i\neq j$, then we have for every fixed l, where $1\le l\le L$,*

$$\widetilde{K}^{[l]}-\widetilde{b}^{[l]}\otimes\left(\widetilde{b}^{[l]}\right)^{\intercal}\succ\widetilde{K}^{[l-1]}-\widetilde{b}^{[l-1]}\otimes\left(\widetilde{b}^{[l-1]}\right)^{\intercal}. \tag{B.10}$$

Moreover, since

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

we denote that

$$\lambda_{\min}\left(\widetilde{K}^{[1]}-\widetilde{b}^{[1]}\otimes\left(\widetilde{b}^{[1]}\right)^{\intercal}\right)=\lambda_0, \tag{B.11}$$

then we can conclude that for $2\le l\le L$,

$$\lambda_{\min}\left(\widetilde{K}^{[l]}\right)>\lambda_0, \tag{B.12}$$

where λ_0 only depends on the activation function and input data and independent of depth L.

Proof. We only need to show that for $1\le i,j\le n$ and $1\le l\le L$

$$\begin{aligned}
&\widetilde{K}_{ij}^{[l]}-\widetilde{b}_i^{[l]}\widetilde{b}_j^{[l]}\\
=&\widetilde{K}_{ij}^{[l-1]}+\mathbb{E}_{(u,v)^{\intercal}\sim\mathcal{N}\left(0,\begin{pmatrix}\widetilde{K}_{ii}^{[l-1]} & \widetilde{K}_{ij}^{[l-1]}\\ \widetilde{K}_{ji}^{[l-1]} & \widetilde{K}_{jj}^{[l-1]}\end{pmatrix}\right)}\left[\frac{c_{\text{res}}\widetilde{b}_i^{[l-1]}\sigma(v)}{L}+\frac{c_{\text{res}}\widetilde{b}_j^{[l-1]}\sigma(u)}{L}+\frac{c_{\text{res}}^2\sigma(u)\sigma(v)}{L^2}\right]\\
&-\left(\widetilde{b}_i^{[l-1]}+\frac{c_{\text{res}}}{L}\mathbb{E}_{u\sim\mathcal{N}(0,\widetilde{K}_{ii}^{[l-1]})}[\sigma(u)]\right)\left(\widetilde{b}_j^{[l-1]}+\frac{c_{\text{res}}}{L}\mathbb{E}_{v\sim\mathcal{N}(0,\widetilde{K}_{jj}^{[l-1]})}[\sigma(v)]\right)\\
=&\widetilde{K}_{ij}^{[l-1]}-\widetilde{b}_i^{[l-1]}\widetilde{b}_j^{[l-1]}+\mathbb{E}_{(u,v)^{\intercal}\sim\mathcal{N}\left(0,\begin{pmatrix}\widetilde{K}_{ii}^{[l-1]} & \widetilde{K}_{ij}^{[l-1]}\\ \widetilde{K}_{ji}^{[l-1]} & \widetilde{K}_{jj}^{[l-1]}\end{pmatrix}\right)}\left[\frac{c_{\text{res}}^2\sigma(u)\sigma(v)}{L^2}\right]\\
&-\frac{c_{\text{res}}}{L}\mathbb{E}_{u\sim\mathcal{N}(0,\widetilde{K}_{ii}^{[l-1]})}[\sigma(u)]\frac{c_{\text{res}}}{L}\mathbb{E}_{v\sim\mathcal{N}(0,\widetilde{K}_{jj}^{[l-1]})}[\sigma(v)]\\
=&\widetilde{K}_{ij}^{[l-1]}-\widetilde{b}_i^{[l-1]}\widetilde{b}_j^{[l-1]}+\frac{c_{\text{res}}^2}{L^2}\text{Cov}_{(u,v)^{\intercal}\sim\mathcal{N}\left(0,\begin{pmatrix}\widetilde{K}_{ii}^{[l-1]} & \widetilde{K}_{ij}^{[l-1]}\\ \widetilde{K}_{ji}^{[l-1]} & \widetilde{K}_{jj}^{[l-1]}\end{pmatrix}\right)}[\sigma(u)\sigma(v)],
\end{aligned}$$

which brings us to the definition of a series of covariance matrices $\left\{\boldsymbol{P}^{[s]}:1\leq s\leq L\right\}$, where

$$\boldsymbol{P}_{ij}^{[s]}:=\frac{c_{\text{res}}^2}{L^2}\text{Cov}_{(u,v)^\intercal\sim\mathcal{N}\left(\boldsymbol{0},\begin{pmatrix}\widetilde{\boldsymbol{K}}_{ii}^{[s]} & \widetilde{\boldsymbol{K}}_{ij}^{[s]}\\ \widetilde{\boldsymbol{K}}_{ji}^{[s]} & \widetilde{\boldsymbol{K}}_{jj}^{[s]}\end{pmatrix}\right)}[\sigma(u)\sigma(v)],\quad 1\leq s\leq L,$$

$\boldsymbol{P}^{[s]}$ are covariance matrices, naturally we have $\boldsymbol{P}^{[s]}\succeq 0$, and $\boldsymbol{P}^{[s]}\succ 0$ except that one sample is an exact linear function of the others. Apply Lemma B.1 directly, we can guarantee that $\boldsymbol{P}^{[s]}$ is positive definite for every s. Hence, inductively we have

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

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

$$\left(\widetilde{\boldsymbol{K}}^{[1]}-\widetilde{\boldsymbol{b}}^{[1]}\otimes\left(\widetilde{\boldsymbol{b}}^{[1]}\right)^\intercal\right)_{ij}=c_\sigma\text{Cov}_{(u,v)^\intercal\sim\mathcal{N}\left(\boldsymbol{0},\begin{pmatrix}\widetilde{\boldsymbol{K}}_{ii}^{[0]} & \widetilde{\boldsymbol{K}}_{ij}^{[0]}\\ \widetilde{\boldsymbol{K}}_{ji}^{[0]} & \widetilde{\boldsymbol{K}}_{jj}^{[0]}\end{pmatrix}\right)}[\sigma(u)\sigma(v)].$$

Then apply Lemma B.1 again

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

and λ_0 only depends on the input data and activation function. □

Corollary B.1. *Given the input samples $\mathcal{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$, $\|\boldsymbol{x}_i\|_2=1$, for $1\leq i\leq n$, and $\boldsymbol{x}_i\nparallel\boldsymbol{x}_j$, $i\neq j$, then we have*

$$\lambda_{\min}\left(\boldsymbol{K}^{[L+1]}\right)>\lambda_0,\tag{B.13}$$

where λ_0 has been defined in (B.11).

Proof. The proof is quite similar to the proof of Proposition B.1. Recall that

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

and we define that

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

Then

$$\boldsymbol{K}_{ij}^{[L+1]} - \boldsymbol{b}_i^{[L+1]}\boldsymbol{b}_j^{[L+1]} = \widetilde{\boldsymbol{K}}_{ij}^{[L]} - \widetilde{\boldsymbol{b}}_i^{[L]}\widetilde{\boldsymbol{b}}_j^{[L]} + \frac{c_{\text{res}}^2}{L^2}\text{Cov}_{(u,v)^\intercal\sim\mathcal{N}\left(\boldsymbol{0},\begin{pmatrix}\widetilde{\boldsymbol{K}}_{ii}^{[L]} & \widetilde{\boldsymbol{K}}_{ij}^{[L]}\\ \widetilde{\boldsymbol{K}}_{ji}^{[L]} & \widetilde{\boldsymbol{K}}_{jj}^{[L]}\end{pmatrix}\right)}[\sigma(u)\sigma(v)].$$

Hence

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

Applying Proposition B.1 directly, we are able to finish the proof. □

By Corollary B.1, we see that $\lambda_{\min}\left(\boldsymbol{K}^{[L+1]}\right)\sim\Omega(1)$.

B.2 Full rankness for L-th Gram matrix

Our next Proposition is related to the eigenvalue of the L-th Gram matrix, whose entries concerning the derivative of the activation function. This proposition has been stated as Proposition F.2 in Du et al. [12], and we will mimic its proof.

Proposition B.2. *Given the input samples $\mathcal{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$, $\|\boldsymbol{x}_i\|_2=1$, for $1\le i\le n$, and $\boldsymbol{x}_i\nparallel\boldsymbol{x}_j$, $i\neq j$, then for $2\le l\le L$*

$$\lambda_{\min}\left(\boldsymbol{K}^{[l]}\right)\ge\frac{c_{\text{res}}^2}{L^2}\kappa, \tag{B.14}$$

where κ is a constant that only depends on $\sigma(\cdot)$ and input samples, independent of depth L.

Proof. Based on Lemma B.3, uniformly for any $1\le l\le L$,

$$1/c\le\widetilde{\boldsymbol{K}}_{ii}^{[l]}\le c,$$

then we can define a function $\boldsymbol{G}:\mathbb{R}^{n\times n}\to\mathbb{R}^{n\times n}$, such that

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

Consequently, a scalar function $g(\lambda)$ can be defined as follows:

$$g(\lambda) := \min_{\boldsymbol{K}:\boldsymbol{K}\succ 0,1/c\le\boldsymbol{K}_{ii}\le c,\lambda_{\min}(\boldsymbol{K})\ge\lambda}\lambda_{\min}(\boldsymbol{G}(\boldsymbol{K})).$$

Then Lemma B.2 guarantees that

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

Moreover, based on Proposition B.1

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

we have

$$\lambda_{\min}\left(\boldsymbol{K}^{[L]}\right)\geq\frac{c_{\text{res}}^2}{L^2}g(\lambda_0). \tag{B.15}$$

Let $\kappa=g(\lambda_0)$. Since κ is independent of depth L, we finish our proof. □

By Proposition B.2, we see that $\lambda_{\min}\left(\boldsymbol{K}^{[L]}\right)\sim\Omega(\frac{1}{L^2})$.

C Random initialization of Gram matrices

In this part, we are going to show that with high probability with respect to the random initialization,

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

where λ_0 is defined in (B.11).

Let's get started with a lemma concerning the Gaussian concentrations.

Lemma C.1. *Let $\boldsymbol{X}=(X_1,\cdots X_p)\in\mathbb{R}^p$, $X_1,\cdots X_p$ be a vector of i.i.d. Gaussian variables from $\mathcal{N}(0,\sigma^2)$, and let $f(\cdot):\mathbb{R}^p\to\mathbb{R}$ be L-Lipschitz function, i.e., $|f(\boldsymbol{x})-f(\boldsymbol{y})|\leq L\|\boldsymbol{x}-\boldsymbol{y}\|_2$ for all $\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^p$, then for any $t\geq 0$*

$$\mathbb{P}\left(|f(\boldsymbol{X})-\mathbb{E}f(\boldsymbol{X})|\geq t\right)\leq 2\exp\Big(-\frac{t^2}{2L^2\sigma^2}\Big). \tag{C.1}$$

Before we proceed to the stability of the randomly initialized Gram matrix of higher order, we need to state two lemmas. The first lemma has been stated as Lemma G.3. in Du et al. [12],

Lemma C.2. *If $\sigma(\cdot)$ is C_{Lip}-Lipschitz, then for $a,b\in\mathbb{R}^+$, with $1/c\leq\min(a,b),\max(a,b)\leq c$ for some $c>0$, then we have*

$$\Big|\mathbb{E}_{z\sim\mathcal{N}(0,1)}[\sigma(az)]-\mathbb{E}_{z\sim\mathcal{N}(0,1)}[\sigma(bz)]\Big|\leq C|a-b|, \tag{C.2}$$

where $C>0$ only depends on c and Lipschitz constant C_{Lip}.

Next lemma has been stated as Lemma G.4. in Du et al. [12]:

Lemma C.3. *If $\sigma(\cdot)$ is C_{Lip}-Lipschitz, define a scalar function $F(\boldsymbol{K})$ as follows:*

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

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

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

and their entries satisfying

$$1/c \le \min(a_1,b_1), \min(a_2,b_2), \max(a_1,b_1), \max(a_2,b_2) \le c,$$

and

$$-1 < \rho_1, \rho_2 < 1$$

for some $c > 0$, then we have

$$|F(\boldsymbol{A}) - F(\boldsymbol{B})| \le C\|\boldsymbol{A}-\boldsymbol{B}\|_{\text{F}} \le 2C\|\boldsymbol{A}-\boldsymbol{B}\|_{\infty},$$

where the constant $C > 0$ only relies on c and the Lipschitz constant C_{Lip}.

We shall begin with a proposition on the initial estimate of the output of each layer $\boldsymbol{x}_j^{[l]}(0)$,

Proposition C.1. *Under Assumptions* 4.1 *and* 4.2*, we have that for some $t > 0$, $1 \le i \le n$, $1 \le l \le L$,*

$$\mathbb{P}\left(\left|\left\|\boldsymbol{x}_i^{[l]}(0)\right\|_2 - \sqrt{\widetilde{\boldsymbol{K}}_{ii}^{[l]}}\right| \ge t\right) \le \exp\left(-cmt^2\right), \tag{C.3}$$

$$\mathbb{P}\left(\left|\left\langle \frac{\boldsymbol{x}_i^{[l]}(0)}{\sqrt{m}}, \boldsymbol{1}\right\rangle - \widetilde{\boldsymbol{b}}_i^{[l]}\right| \ge t\right) \le \exp\left(-cmt^2\right), \tag{C.4}$$

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

Proof. For $l = 1$, we have

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

then

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

Since $\left(\boldsymbol{W}^{[1]}(0)\boldsymbol{x}_i\right)_j$ are i.i.d standard Gaussian variables, and $\sigma(\cdot)$ is 1-Lipschitz, then $\left(\sigma(\boldsymbol{W}^{[1]}(0)\boldsymbol{x}_i)_j\right)$ are sub-exponential variables, and we have for $\lambda>0$,

$$\mathbb{E}\left[\exp\lambda\left(m\left\|\boldsymbol{x}_i^{[1]}(0)\right\|_2^2-m\right)\right]\leq\exp\left(cm\lambda^2\right).$$

Hence applying Markov inequality directly

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

and

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

we obtain

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

We should note that $\boldsymbol{x}_i^{[1]}(0)$ writes into

$$\boldsymbol{x}_i^{[1]}(0)=\sqrt{\frac{c_\sigma}{m}}\sigma(\boldsymbol{X}),$$

with $\boldsymbol{X}$ being a standard normal Gaussian vector. We shall focus on the inner product function $g^{[1]}(\cdot):\mathbb{R}^m\to\mathbb{R}$, with

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

We have 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, then apply Lemma C.1

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

Then we have

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

Our next step is to prove that (C.3) and (C.4) hold for $l\geq 2$, and we will prove it by induction.

Assume that (C.3) and (C.4) hold for 1,2,3,$\cdots$,l and want to show that they hold for $l+1$.

$$\mathbb{P}\left(\left|\left\|\boldsymbol{x}_i^{[l+1]}(0)\right\|_2-\sqrt{\widetilde{\boldsymbol{K}}_{ii}^{[l+1]}}\right|\geq t\right)\leq \exp\left(-cmt^2\right), \tag{C.5}$$

$$\mathbb{P}\left(\left|\left\langle \frac{\boldsymbol{x}_i^{[l+1]}(0)}{\sqrt{m}},\mathbf{1}\right\rangle-\widetilde{\boldsymbol{b}}_i^{[l+1]}\right|\geq t\right)\leq \exp\left(-cmt^2\right). \tag{C.6}$$

We recall that,

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

and the definition of $\widetilde{\boldsymbol{K}}_{ii}^{[l]}$ and $\widetilde{\boldsymbol{b}}_i^{[l]}$

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

$$\widetilde{\boldsymbol{b}}_i^{[l+1]}=\widetilde{\boldsymbol{b}}_i^{[l]}+\frac{c_{\text{res}}}{L}\mathbb{E}_{u\sim\mathcal{N}(0,\widetilde{\boldsymbol{K}}_{ii}^{[l]})}[\sigma(u)].$$

Then we have

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

Then we need to focus on the terms I and II. Note that for term I there is a $\frac{1}{\sqrt{m}}$ scaling factor contained in $\boldsymbol{x}_i^{[l]}(0)$, and $\sigma\left(\boldsymbol{W}^{[l+1]}(0)\boldsymbol{x}_i^{[l]}(0)\right)$ has distribution

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

with $\boldsymbol{Y}$ being a standard normal Gaussian vector. Then we have

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

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

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

we have for any $\boldsymbol{X}_1,\boldsymbol{X}_2\in\mathbb{R}^m$,

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

based on our induction hypothesis, $\|\boldsymbol{x}_i^{[l]}(0)\|_2\le C$ with high probability. Hence $g^{[l]}(\cdot)$ is $\frac{C}{\sqrt{m}}$-Lipschitz. Applying Lemma C.1 again

$$\begin{aligned}&\mathbb{P}\left(\left|\frac{1}{\sqrt{m}}\left\langle \boldsymbol{x}_i^{[l]}(0),\sigma\left(\boldsymbol{W}^{[l+1]}(0)\boldsymbol{x}_i^{[l]}(0)\right)\right\rangle-\frac{1}{\sqrt{m}}\left\langle \boldsymbol{x}_i^{[l]}(0),\mathbb{E}\left[\sigma\left(\left\|\boldsymbol{x}_i^{[l]}(0)\right\|_2\boldsymbol{Y}\right)\right]\right\rangle\right|\ge t\right)\\&\le\exp(-cmt^2),\end{aligned}\tag{C.7}$$

and based on our induction hypothesis,

$$\begin{aligned}&\mathbb{P}\left(\left|\frac{1}{\sqrt{m}}\left\langle \boldsymbol{x}_i^{[l]}(0),\mathbb{E}\left[\sigma\left(\left\|\boldsymbol{x}_i^{[l]}(0)\right\|_2\boldsymbol{Y}\right)\right]\right\rangle-\widetilde{\boldsymbol{b}}_i^{[l]}\mathbb{E}\left[\sigma\left(\left\|\boldsymbol{x}_i^{[l]}(0)\right\|_2\boldsymbol{Y}\right)\right]\right|\ge t\right)\\&\le\exp(-cmt^2),\end{aligned}\tag{C.8}$$

from Lemma C.2

$$\left|\mathbb{E}\left[\sigma\left(\left\|\boldsymbol{x}_i^{[l]}(0)\right\|_2\boldsymbol{Y}\right)\right]-\mathbb{E}\left[\sigma\left(\sqrt{\widetilde{\boldsymbol{K}}_{ii}^{[l]}}\boldsymbol{Y}\right)\right]\right|\le C\left|\left\|\boldsymbol{x}_i^{[l]}(0)\right\|_2-\sqrt{\widetilde{\boldsymbol{K}}_{ii}^{[l]}}\right|,$$

altogether we have

$$\mathbb{P}\left(\left|\widetilde{\boldsymbol{b}}_i^{[l]}\mathbb{E}\left[\sigma\left(\left\|\boldsymbol{x}_i^{[l]}(0)\right\|_2\boldsymbol{Y}\right)\right]-\widetilde{\boldsymbol{b}}_i^{[l]}\mathbb{E}\left[\sigma\left(\sqrt{\widetilde{\boldsymbol{K}}_{ii}^{[l]}}\boldsymbol{Y}\right)\right]\right|\ge t\right)\le\exp(-cmt^2).\tag{C.9}$$

Combining (C.7), (C.8) and (C.9), we have

$$\mathbb{P}\left(\left|\frac{1}{\sqrt{m}}\left\langle \boldsymbol{x}_i^{[l]}(0),\sigma\left(\boldsymbol{W}^{[l+1]}(0)\boldsymbol{x}_i^{[l]}(0)\right)\right\rangle-\widetilde{\boldsymbol{b}}_i^{[l]}\mathbb{E}\left[\sigma\left(\sqrt{\widetilde{\boldsymbol{K}}_{ii}^{[l]}}\boldsymbol{Y}\right)\right]\right|\ge t\right)\le\exp(-cmt^2).\tag{C.10}$$

Finally for term II

$$\mathbb{E}\left[\frac{1}{m}\left\langle \sigma\left(\boldsymbol{W}^{[l+1]}(0)\boldsymbol{x}_i^{[l]}(0)\right),\sigma\left(\boldsymbol{W}^{[l+1]}(0)\boldsymbol{x}_i^{[l]}(0)\right)\right\rangle\right]=\mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[\sigma\left(\left\|\boldsymbol{x}_i^{[l]}(0)\right\|_2x\right)^2\right],$$

since $\left(\boldsymbol{W}^{[l+1]}(0)\boldsymbol{x}_i^{[l]}(0)\right)$ are i.i.d standard Gaussian variables, and $\sigma(\cdot)$ is 1-Lipschitz, then $\left(\boldsymbol{W}^{[l+1]}(0)\boldsymbol{x}_i^{[l]}(0)\right)$ are sub-exponential variables and we have

$$\begin{aligned}&\mathbb{P}\left(\left|\frac{1}{m}\left\langle\sigma\left(\boldsymbol{W}^{[l+1]}(0)\boldsymbol{x}_i^{[l]}(0)\right),\sigma\left(\boldsymbol{W}^{[l+1]}(0)\boldsymbol{x}_i^{[l]}(0)\right)\right\rangle-\mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[\sigma\left(\left\|\boldsymbol{x}_i^{[l]}(0)\right\|_2 x\right)^2\right]\right|\geq t\right)\\ \leq&\exp\left(-cmt^2\right).\end{aligned}\tag{C.11}$$

Applying Lemma C.3

$$\left|\mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[\sigma\left(\left\|\boldsymbol{x}_i^{[l]}(0)\right\|_2 x\right)^2\right]-\mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[\sigma\left(\sqrt{\widetilde{\boldsymbol{K}_{ii}}^{[l]}}x\right)^2\right]\right|\leq C\left|\left\|\boldsymbol{x}_i^{[l]}(0)\right\|_2-\sqrt{\widetilde{\boldsymbol{K}_{ii}}^{[l]}}\right|,$$

then based on our induction hypothesis, we have

$$\mathbb{P}\left(\left|\mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[\sigma\left(\left\|\boldsymbol{x}_i^{[l]}(0)\right\|_2 x\right)^2\right]-\mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[\sigma\left(\sqrt{\widetilde{\boldsymbol{K}_{ii}}^{[l]}}x\right)^2\right]\right|\geq t\right)\leq\exp\left(-cmt^2\right).\tag{C.12}$$

Combining (C.11) and (C.12)

$$\begin{aligned}&\mathbb{P}\left(\left|\frac{1}{m}\left\langle\sigma\left(\boldsymbol{W}^{[l+1]}(0)\boldsymbol{x}_i^{[l]}(0)\right),\sigma\left(\boldsymbol{W}^{[l+1]}(0)\boldsymbol{x}_i^{[l]}(0)\right)\right\rangle-\mathbb{E}_{x\sim\mathcal{N}(0,1)}\left[\sigma\left(\sqrt{\widetilde{\boldsymbol{K}_{ii}}^{[l]}}x\right)^2\right]\right|\geq t\right)\\ \leq&\exp\left(-cmt^2\right).\end{aligned}\tag{C.13}$$

Since we have

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

then

$$\begin{aligned}&\mathbb{P}\left(\left|\left\|\boldsymbol{x}_i^{[l+1]}(0)\right\|_2^2-\widetilde{\boldsymbol{K}_{ii}}^{[l+1]}\right|\geq t\left(1+\frac{c_{\text{res}}}{L}\right)^2\right)\\ \leq&\mathbb{P}\left(\left|\left\|\boldsymbol{x}_i^{[l]}(0)\right\|_2^2-\widetilde{\boldsymbol{K}_{ii}}^{[l]}\right|\geq t\right)+\mathbb{P}\left(|\text{II}|\geq t\right)+\mathbb{P}\left(|\text{III}|\geq t\right)\\ \leq&\exp\left(-cmt^2\right).\end{aligned}\tag{C.14}$$

We shall see that thanks to the $\frac{c_{\text{res}}}{L}$ structure, with high probability the difference of $\left|\|\boldsymbol{x}_i^{[l]}(0)\|_2^2-\widetilde{\boldsymbol{K}_{ii}}^{[l]}\right|$ does not explode exponentially layer by layer.

For $\widetilde{\boldsymbol{b}}_i^{[l+1]}$, applying Lemma C.1, we have

$$\mathbb{P}\left(\left|\left\langle\frac{\sigma\left(\boldsymbol{W}^{[l+1]}(0)\boldsymbol{x}_i^{[l]}(0)\right)}{m},\boldsymbol{1}\right\rangle-\mathbb{E}\left[\sigma\left(\left\|\boldsymbol{x}_i^{[l]}(0)\right\|_2\boldsymbol{Y}\right)\right]\right|\geq t\right)\leq\exp\left(-cmt^2\right). \tag{C.15}$$

And applying Lemma C.2, we have

$$\mathbb{P}\left(\left|\mathbb{E}\left[\sigma\left(\left\|\boldsymbol{x}_i^{[l]}(0)\right\|_2\boldsymbol{Y}\right)\right]-\mathbb{E}\left[\sigma\left(\sqrt{\widetilde{\boldsymbol{K}_{ii}}^{[l]}}\boldsymbol{Y}\right)\right]\right|\geq t\right)\leq\exp\left(-cmt^2\right). \tag{C.16}$$

Combining (C.15) and (C.16),

$$\mathbb{P}\left(\left|\left\langle\frac{\sigma\left(\boldsymbol{W}^{[l+1]}(0)\boldsymbol{x}_i^{[l]}(0)\right)}{m},\boldsymbol{1}\right\rangle-\mathbb{E}\left[\sigma\left(\sqrt{\widetilde{\boldsymbol{K}_{ii}}^{[l]}}\boldsymbol{Y}\right)\right]\right|\geq t\right)\leq\exp\left(-cmt^2\right). \tag{C.17}$$

Then

$$\begin{aligned}
&\mathbb{P}\left(\left|\left\langle\frac{\boldsymbol{x}_i^{[l+1]}(0)}{\sqrt{m}},\boldsymbol{1}\right\rangle-\widetilde{\boldsymbol{b}}_i^{[l+1]}\right|\geq t\left(1+\frac{c_{\text{res}}}{L}\right)\right)\\
\leq&\mathbb{P}\left(\left|\left\langle\frac{\boldsymbol{x}_i^{[l]}(0)}{\sqrt{m}},\boldsymbol{1}\right\rangle-\widetilde{\boldsymbol{b}}_i^{[l]}\right|\geq t\right)+\mathbb{P}\left(\left|\left\langle\frac{\sigma\left(\boldsymbol{W}^{[l+1]}(0)\boldsymbol{x}_i^{[l]}(0)\right)}{m},\boldsymbol{1}\right\rangle-\mathbb{E}\left[\sigma\left(\sqrt{\widetilde{\boldsymbol{K}_{ii}}^{[l]}}\boldsymbol{Y}\right)\right]\right|\geq t\right)\\
\leq&\exp\left(-cmt^2\right).
\end{aligned} \tag{C.18}$$

We shall see again that thanks to the $\frac{c_{\text{res}}}{L}$ structure, with high probability the difference of $\left|<\frac{\boldsymbol{x}_i^{[l]}(0)}{\sqrt{m}},\boldsymbol{1}>-\widetilde{\boldsymbol{b}}_i^{[l]}\right|$ only has slight increment with respect to each layer l. □

Our next Proposition is on the least eigenvalue of the randomly initialized Gram matrix $\boldsymbol{G}^{[1]}(0)$.

Proposition C.2. *Under Assumptions* 4.1 *and* 4.2, *if* $m=\Omega\left(\left(\frac{n}{\lambda_0}\right)^{2+\varepsilon}\right)$, *then with high probability*

$$\lambda_{\min}\left(\boldsymbol{G}^{[1]}(0)\right)\geq\frac{3\lambda_0}{4}, \tag{C.19}$$

where λ_0 *has been defined in* (B.11).

Proof. We have that

$$\begin{aligned}
&\boldsymbol{G}_{ij}^{[1]}(0)=\left\langle \boldsymbol{x}_i^{[1]}(0),\boldsymbol{x}_j^{[1]}(0)\right\rangle,\\
&\widetilde{\boldsymbol{K}}_{ij}^{[0]}=\left\langle \boldsymbol{x}_i,\boldsymbol{x}_j\right\rangle,\\
&\widetilde{\boldsymbol{K}}_{ij}^{[1]}=c_\sigma\mathbb{E}_{(u,v)^\intercal\sim\mathcal{N}\left(\boldsymbol{0},\begin{pmatrix}\widetilde{\boldsymbol{K}}_{ii}^{[0]} & \widetilde{\boldsymbol{K}}_{ij}^{[0]}\\ \widetilde{\boldsymbol{K}}_{ji}^{[0]} & \widetilde{\boldsymbol{K}}_{jj}^{[0]}\end{pmatrix}\right)}[\sigma(u)\sigma(v)].
\end{aligned}$$

Now we need to apply Lemma C.1 again, except that this time we are going to apply it to the inner product function $h^{[1]}(\cdot):\mathbb{R}^{2m}\to\mathbb{R}$, with

$$h^{[1]}(\boldsymbol{Z})=\frac{c_\sigma}{m}\left\langle \sigma(\boldsymbol{X}),\sigma(\rho\boldsymbol{X}+\sqrt{1-\rho^2}\boldsymbol{Y})\right\rangle,$$

where $-1\le\rho\le 1$.

Specifically with $\boldsymbol{Z}^\intercal=(\boldsymbol{X}^\intercal,\boldsymbol{Y}^\intercal)$, we have for any $\boldsymbol{Z}_1,\boldsymbol{Z}_2\in\mathbb{R}^m$,

$$\begin{aligned}
\left|h^{[1]}(\boldsymbol{Z}_1)-h^{[1]}(\boldsymbol{Z}_2)\right|\le&\sqrt{\frac{c_\sigma}{m}}\left\|\sigma(\rho\boldsymbol{X}_1+\sqrt{1-\rho^2}\boldsymbol{Y}_1)\right\|_2\sqrt{\frac{c_\sigma}{m}}\|\boldsymbol{X}_1-\boldsymbol{X}_2\|_2\\
&+\sqrt{\frac{c_\sigma}{m}}\|\sigma(\boldsymbol{X}_2)\|_2\sqrt{\frac{c_\sigma}{m}}\left(|\rho|\,\|\boldsymbol{X}_1-\boldsymbol{X}_2\|_2+\sqrt{1-\rho^2}\|\boldsymbol{Y}_1-\boldsymbol{Y}_2\|_2\right),
\end{aligned}$$

combined with Proposition C.1, with probability $1-\exp(-cm)$,

$$\sqrt{\frac{c_\sigma}{m}}\left\|\sigma(\rho\boldsymbol{X}_1+\sqrt{1-\rho^2}\boldsymbol{Y}_1)\right\|_2,\sqrt{\frac{c_\sigma}{m}}\|\sigma(\boldsymbol{X}_2)\|_2\le 2.$$

So we have

$$\left|h^{[1]}(\boldsymbol{Z}_1)-h^{[1]}(\boldsymbol{Z}_2)\right|\le 8\sqrt{\frac{c_\sigma}{m}}\|\boldsymbol{Z}_1-\boldsymbol{Z}_2\|_2,$$

hence $h^{[1]}(\boldsymbol{Z})$ is $8\sqrt{\frac{c_\sigma}{m}}$-Lipschitz, then we shall set $\rho=\widetilde{\boldsymbol{K}}_{ij}^{[0]}$,

$$\mathbb{P}\left(\left|\boldsymbol{G}_{ij}^{[1]}(0)-\widetilde{\boldsymbol{K}}_{ij}^{[1]}\right|\ge t\right)\le\exp(-cmt^2). \tag{C.20}$$

Noting that

$$\left\|\boldsymbol{G}^{[1]}(0)-\widetilde{\boldsymbol{K}}^{[1]}\right\|_{2\to 2}\le\left\|\boldsymbol{G}^{[1]}(0)-\widetilde{\boldsymbol{K}}^{[1]}\right\|_{\mathrm{F}}\le n\left\|\boldsymbol{G}^{[1]}(0)-\widetilde{\boldsymbol{K}}^{[1]}\right\|_{\infty},$$

based on Proposition B.1, $\lambda_{\min}(\widetilde{\boldsymbol{K}}^{[1]})\ge\lambda_0$, then if we choose $t=\frac{\lambda_0}{4n}$ and with a union m^2 such events, we have with probability $1-m^2\exp\left(-cm\lambda_0^2/n^2\right)$

$$\left\|\boldsymbol{G}^{[1]}(0)-\widetilde{\boldsymbol{K}}^{[1]}\right\|_{2\to 2}\le\frac{\lambda_0}{4}. \tag{C.21}$$

Hence if $m=\Omega\big((\frac{n}{\lambda_0})^{2+\varepsilon}\big)$, we have with probability $1-\exp(-m^{\varepsilon})$

$$\lambda_{\min}(\boldsymbol{G}^{[1]}(0))\geq\lambda_{\min}(\widetilde{\boldsymbol{K}}^{[1]})-\left\|\boldsymbol{G}^{[1]}(0)-\widetilde{\boldsymbol{K}}^{[1]}\right\|_{2\to 2}\geq\frac{3\lambda_0}{4}. \tag{C.22}$$

This completes the proof. □

Our next proposition on the stability of the randomly initialized Gram matrix $\boldsymbol{G}^{[l]}(0)$ for $2\leq l\leq L+1$.

Proposition C.3. *Under Assumptions* 4.1 *and* 4.2, *if* $m=\Omega\big((\frac{n}{\lambda_0})^{2+\varepsilon}\big)$, *then with high probability*

$$\lambda_{\min}\left(\boldsymbol{G}^{[l]}(0)\right)\geq\frac{3\lambda_0}{4},\quad 2\leq l\leq L+1, \tag{C.23}$$

where λ_0 *has been defined in* (B.11).

Proof. For $l=2$, we shall make estimate on the norm, $\|\boldsymbol{G}^{[2]}(0)-\widetilde{\boldsymbol{K}}^{[2]}\|_{\infty}$, since by definition

$$\begin{aligned}
\boldsymbol{G}_{ij}^{[2]}(0)&=\left\langle\boldsymbol{x}_i^{[2]}(0),\boldsymbol{x}_j^{[2]}(0)\right\rangle=\boldsymbol{G}_{ij}^{[1]}(0)+\frac{c_{\rm res}}{L}\underbrace{\frac{1}{\sqrt{m}}\left\langle\boldsymbol{x}_i^{[1]}(0),\sigma\left(\boldsymbol{W}^{[2]}(0)\boldsymbol{x}_j^{[1]}(0)\right)\right\rangle}_{\rm I}\\
&\quad+\frac{c_{\rm res}}{L}\underbrace{\frac{1}{\sqrt{m}}\left\langle\boldsymbol{x}_j^{[1]}(0),\sigma\left(\boldsymbol{W}^{[2]}(0)\boldsymbol{x}_i^{[1]}(0)\right)\right\rangle}_{\rm II}\\
&\quad+\frac{c_{\rm res}^2}{L^2}\underbrace{\frac{1}{m}\left\langle\sigma\left(\boldsymbol{W}^{[2]}(0)\boldsymbol{x}_i^{[1]}(0)\right),\sigma\left(\boldsymbol{W}^{[2]}(0)\boldsymbol{x}_j^{[1]}(0)\right)\right\rangle}_{\rm III},\\
\widetilde{\boldsymbol{K}}_{ij}^{[1]}&=c_\sigma\mathbb{E}_{(u,v)^{\intercal}\sim\mathcal{N}\left(\boldsymbol{0},\begin{pmatrix}\widetilde{\boldsymbol{K}}_{ii}^{[0]}&\widetilde{\boldsymbol{K}}_{ij}^{[0]}\\\widetilde{\boldsymbol{K}}_{ji}^{[0]}&\widetilde{\boldsymbol{K}}_{jj}^{[0]}\end{pmatrix}\right)}[\sigma(u)\sigma(v)],\\
\widetilde{\boldsymbol{b}}_i^{[1]}&=\sqrt{c_\sigma}\mathbb{E}_{u\sim\mathcal{N}(0,\widetilde{\boldsymbol{K}}_{ii}^{[0]})}[\sigma(u)],\\
\widetilde{\boldsymbol{A}}_{ij}^{[2]}&=\begin{pmatrix}\widetilde{\boldsymbol{K}}_{ii}^{[1]}&\widetilde{\boldsymbol{K}}_{ij}^{[1]}\\\widetilde{\boldsymbol{K}}_{ji}^{[1]}&\widetilde{\boldsymbol{K}}_{jj}^{[1]}\end{pmatrix},\\
\widetilde{\boldsymbol{K}}_{ij}^{[2]}&=\widetilde{\boldsymbol{K}}_{ij}^{[1]}+\mathbb{E}_{(u,v)^{\intercal}\sim\mathcal{N}\left(\boldsymbol{0},\widetilde{\boldsymbol{A}}_{ij}^{[2]}\right)}\left[\frac{c_{\rm res}}{L}\underbrace{\widetilde{\boldsymbol{b}}_i^{[1]}\sigma(v)}_{\rm I'}+\frac{c_{\rm res}}{L}\underbrace{\widetilde{\boldsymbol{b}}_j^{[1]}\sigma(u)}_{\rm II'}+\frac{c_{\rm res}^2}{L^2}\underbrace{\sigma(u)\sigma(v)}_{\rm III'}\right].
\end{aligned}$$

We need to tackle the difference between I and I′. In order for that, we need to write

the difference into

$$\begin{aligned}
&\left|\frac{1}{\sqrt{m}}\left\langle \boldsymbol{x}_i^{[1]}(0),\sigma\left(\boldsymbol{W}^{[2]}(0)\boldsymbol{x}_j^{[1]}(0)\right)\right\rangle-\widetilde{\boldsymbol{b}}_i^{[1]}\sigma(v)\right| \\
\leq&\left|\frac{1}{\sqrt{m}}\left\langle \boldsymbol{x}_i^{[1]}(0),\sigma\left(\boldsymbol{W}^{[2]}(0)\boldsymbol{x}_j^{[1]}(0)\right)\right\rangle-\frac{1}{\sqrt{m}}\left\langle \boldsymbol{x}_i^{[1]}(0),\mathbb{E}\left[\sigma\left(\left\|\boldsymbol{x}_j^{[1]}(0)\right\|_2\boldsymbol{Y}\right)\right]\right\rangle\right| \\
&+\left|\frac{1}{\sqrt{m}}\left\langle \boldsymbol{x}_i^{[1]}(0),\mathbb{E}\left[\sigma\left(\left\|\boldsymbol{x}_j^{[1]}(0)\right\|_2\boldsymbol{Y}\right)\right]\right\rangle-\widetilde{\boldsymbol{b}}_i^{[1]}\mathbb{E}\left[\sigma\left(\left\|\boldsymbol{x}_j^{[1]}(0)\right\|_2\boldsymbol{Y}\right)\right]\right| \\
&+\left|\widetilde{\boldsymbol{b}}_i^{[1]}\mathbb{E}\left[\sigma\left(\left\|\boldsymbol{x}_j^{[1]}(0)\right\|_2\boldsymbol{Y}\right)\right]-\widetilde{\boldsymbol{b}}_i^{[1]}\mathbb{E}\left[\sigma\left(\sqrt{\widetilde{\boldsymbol{K}_{jj}}^{[1]}}\boldsymbol{Y}\right)\right]\right|,
\end{aligned}$$

similar to the proof in Proposition C.1 with $\boldsymbol{Y}$ being a standard normal Gaussian vector

$$\mathbb{P}\left(\left|\mathrm{I}-\mathrm{I}'\right|\geq t\right)\leq\exp\left(-cmt^2\right). \tag{C.24}$$

Similarly, we have

$$\mathbb{P}\left(\left|\mathrm{II}-\mathrm{II}'\right|\geq t\right)\leq\exp\left(-cmt^2\right). \tag{C.25}$$

For the difference between III and III', we need to define another inner product function $h^{[2]}(\cdot):\mathbb{R}^{2m}\to\mathbb{R}$, being

$$h^{[2]}(\boldsymbol{Z})=\frac{1}{m}\left\langle\sigma(C_2\boldsymbol{X}),\sigma\left(D_2\left(\rho\boldsymbol{X}+\sqrt{1-\rho^2}\boldsymbol{Y}\right)\right)\right\rangle,$$

with $C_2,D_2>0$ being constants and $-1\leq\rho\leq1$.

Note that the form $\boldsymbol{Z}^\intercal=(\boldsymbol{X}^\intercal,\boldsymbol{Y}^\intercal)$, similar to $h^{[1]}(\cdot)$ defined in the proof of Proposition C.2, $h^{[2]}(\cdot)$ is $\frac{C}{\sqrt{m}}$-Lipschitz. Then we have

$$\mathbb{P}\left(\left|h^{[2]}(\boldsymbol{Z})-\mathbb{E}h^{[2]}(\boldsymbol{Z})\right|\geq t\right)\leq\exp(-cmt^2).$$

Hence we have

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{m}\left\langle\sigma\left(\boldsymbol{W}^{[2]}(0)\boldsymbol{x}_i^{[1]}(0)\right),\sigma\left(\boldsymbol{W}^{[2]}(0)\boldsymbol{x}_j^{[1]}(0)\right)\right\rangle-\mathbb{E}_{(u,v)^\intercal\sim\mathcal{N}\left(\boldsymbol{0},\boldsymbol{A}_{ij}^{[2]}\right)}\left[\sigma(u)\sigma(v)\right]\right|\geq t\right) \\
\leq&\exp(-cmt^2),
\end{aligned} \tag{C.26}$$

with

$$\boldsymbol{A}_{ij}^{[2]}=\begin{pmatrix}\left\langle\boldsymbol{x}_i^{[1]}(0),\boldsymbol{x}_i^{[1]}(0)\right\rangle & \left\langle\boldsymbol{x}_i^{[1]}(0),\boldsymbol{x}_j^{[1]}(0)\right\rangle \\ \left\langle\boldsymbol{x}_j^{[1]}(0),\boldsymbol{x}_i^{[1]}(0)\right\rangle & \left\langle\boldsymbol{x}_j^{[1]}(0),\boldsymbol{x}_j^{[1]}(0)\right\rangle\end{pmatrix},$$

combined with Lemma C.3 and Proposition C.1

$$\mathbb{P}\left(\left|\mathbb{E}_{(u,v)^\intercal\sim\mathcal{N}\left(\mathbf{0},\boldsymbol{A}_{ij}^{[2]}\right)}\left[\sigma(u)\sigma(v)\right]-\mathbb{E}_{(u,v)^\intercal\sim\mathcal{N}\left(\mathbf{0},\widetilde{\boldsymbol{A}}_{ij}^{[2]}\right)}\left[\sigma(u)\sigma(v)\right]\right|\geq t\right)\leq\exp(-cmt^2). \quad \text{(C.27)}$$

Combining (C.26) and (C.27)

$$\mathbb{P}\left(\left|\mathrm{III}-\mathrm{III}'\right|\geq t\right)\leq\exp\left(-cmt^2\right), \quad \text{(C.28)}$$

we have that

$$\begin{aligned}&\mathbb{P}\left(\left|\boldsymbol{G}_{ij}^{[2]}(0)-\widetilde{\boldsymbol{K}}_{ij}^{[2]}\right|\geq t\left(1+\frac{c_{\mathrm{res}}}{L}\right)^2\right)\\&\leq\mathbb{P}\left(\left|\boldsymbol{G}_{ij}^{[1]}(0)-\widetilde{\boldsymbol{K}}_{ij}^{[1]}\right|\geq t\right)+\mathbb{P}\left(\left|\mathrm{I}-\mathrm{I}'\right|\geq t\right)+\mathbb{P}\left(\left|\mathrm{II}-\mathrm{II}'\right|\geq t\right)+\mathbb{P}\left(\left|\mathrm{III}-\mathrm{III}'\right|\geq t\right)\\&\leq\exp\left(-cmt^2\right). \quad \text{(C.29)}\end{aligned}$$

Hence inductively, for $2\leq l\leq L$

$$\mathbb{P}\left(\left|\boldsymbol{G}_{ij}^{[l]}(0)-\widetilde{\boldsymbol{K}}_{ij}^{[l]}\right|\geq t\left(1+\frac{c_{\mathrm{res}}}{L}\right)^{2l-2}\right)\leq\exp\left(-cmt^2\right). \quad \text{(C.30)}$$

Moreover,

$$\mathbb{P}\left(\left|\boldsymbol{G}_{ij}^{[L+1]}(0)-\boldsymbol{K}_{ij}^{[L+1]}\right|\geq t\left(1+\frac{c_{\mathrm{res}}}{L}\right)^{2L}\right)\leq\exp\left(-cmt^2\right), \quad \text{(C.31)}$$

note that we have

$$\left\|\boldsymbol{G}^{[L+1]}(0)-\boldsymbol{K}^{[L+1]}\right\|_{2\to2}\leq\left\|\boldsymbol{G}^{[L+1]}(0)-\boldsymbol{K}^{[L+1]}\right\|_{\mathrm{F}}\leq n\left\|\boldsymbol{G}^{[L+1]}(0)-\boldsymbol{K}^{[L+1]}\right\|_{\infty},$$

based on Proposition B.1, $\lambda_{\min}(\boldsymbol{K}^{[L+1]})>\lambda_0$, then if we choose $t=\frac{\lambda_0}{4n\exp(2c_{\mathrm{res}})}$, for $2\leq l\leq L$, with probability $1-\exp\left(-cm\lambda_0^2/n^2\right)$,

$$\left\|\boldsymbol{G}^{[l]}(0)-\widetilde{\boldsymbol{K}}^{[l]}\right\|_{2\to2}\leq\frac{\lambda_0}{4}, \quad \text{(C.32)}$$

hence if $m=\Omega\left((\frac{n}{\lambda_0})^{2+\varepsilon}\right)$, we have with probability $1-\exp(-m^\varepsilon)$

$$\lambda_{\min}(\boldsymbol{G}^{[l]}(0))\geq\lambda_{\min}(\widetilde{\boldsymbol{K}}^{[l]})-\left\|\boldsymbol{G}^{[l]}(0)-\widetilde{\boldsymbol{K}}^{[l]}\right\|_{2\to2}>\frac{3\lambda_0}{4}. \quad \text{(C.33)}$$

In particular, we have that with probability $1-\exp\left(-cm\lambda_0^2/n^2\right)$,

$$\left\|\boldsymbol{G}^{[L+1]}(0)-\boldsymbol{K}^{[L+1]}\right\|_{2\to2}\leq\frac{\lambda_0}{4}, \quad \text{(C.34)}$$

hence if $m=\Omega\left((\frac{n}{\lambda_0})^{2+\varepsilon}\right)$, we have with probability $1-\exp(-m^\varepsilon)$

$$\lambda_{\min}(\boldsymbol{G}^{[L+1]}(0))\geq\lambda_{\min}(\boldsymbol{K}^{[L+1]})-\left\|\boldsymbol{G}^{[L+1]}(0)-\boldsymbol{K}^{[L+1]}\right\|_{2\to2}>\frac{3\lambda_0}{4}. \quad \text{(C.35)}$$

This completes the proof. □

D Proof of Theorem 4.2 and Corollary 4.1

We shall begin with the detailed proof of Theorem 4.2.

Proof of Theorem 4.2. We are only going to use $\mathcal{G}_t^{[L+1]}(\cdot)$ instead of the whole NTK $\mathcal{K}_t^{(2)}(\cdot)$, thanks to the simple structure of $\mathcal{G}_t^{[L+1]}(\cdot)$, we are able to bring about a more concrete proof.

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 and we are going to show that (4.11) holds. Firstly, we need to denote $\mathcal{G}_t^{[L+1]}(\cdot)$ by $\mathcal{G}_t^{[2]}(\cdot)$, i.e.,

$$\mathcal{G}_t^{(2)}(\boldsymbol{x}_{\alpha_1},\boldsymbol{x}_{\alpha_2}) := \mathcal{G}_t^{[L+1]}(\boldsymbol{x}_{\alpha_1},\boldsymbol{x}_{\alpha_2}) = \left\langle \boldsymbol{x}_{\alpha_1}^{[L]},\boldsymbol{x}_{\alpha_2}^{[L]} \right\rangle,$$

then it's natural for us to get that $C(2,L)=\mathcal{O}(1)$, since there is only one term.

Secondly, by the replacement rule, all the possible terms generated from $\mathcal{G}_t^{(2)}(\cdot)$ are

$$\begin{aligned}
&\mathcal{G}_t^{(2)}(\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^{(3)}(\boldsymbol{x}_{\alpha_1},\boldsymbol{x}_{\alpha_2},\boldsymbol{x}_\beta),\\
&\mathcal{G}_t^{(3)}(\boldsymbol{x}_{\alpha_1},\boldsymbol{x}_{\alpha_2},\boldsymbol{x}_\beta)\\
&\quad = \frac{c_\sigma}{m}\underbrace{\left\langle \operatorname{diag}\left(\boldsymbol{E}_{t,\alpha_1}^{[2:L]}\dot{\boldsymbol{\sigma}}_{[1]}^{(1)}(\boldsymbol{x}_{\alpha_1})\dot{\boldsymbol{\sigma}}_{[1]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[2:L]}\right)^\top \boldsymbol{a}_t\right)\mathbf{1},\boldsymbol{x}_{\alpha_2}^{[L]}\right\rangle \langle \boldsymbol{x}_{\alpha_1},\boldsymbol{x}_\beta\rangle}_{\mathrm{I}}\\
&\qquad + \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]}\dot{\boldsymbol{\sigma}}_{[k]}^{(1)}(\boldsymbol{x}_{\alpha_1})\dot{\boldsymbol{\sigma}}_{[k]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[(k+1):L]}\right)^\top \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}}\\
&\qquad + \frac{c_\sigma}{m}\left\langle \operatorname{diag}\left(\boldsymbol{E}_{t,\alpha_2}^{[2:L]}\dot{\boldsymbol{\sigma}}_{[1]}^{(1)}(\boldsymbol{x}_{\alpha_2})\dot{\boldsymbol{\sigma}}_{[1]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[2:L]}\right)^\top \boldsymbol{a}_t\right)\mathbf{1},\boldsymbol{x}_{\alpha_1}^{[L]}\right\rangle \langle \boldsymbol{x}_{\alpha_2},\boldsymbol{x}_\beta\rangle\\
&\qquad + \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]}\dot{\boldsymbol{\sigma}}_{[k]}^{(1)}(\boldsymbol{x}_{\alpha_2})\dot{\boldsymbol{\sigma}}_{[k]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[(k+1):L]}\right)^\top \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

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

Finally for $\mathcal{G}_t^{(4)}(\cdot)$, by symmetry, we are only going to analyze terms I and II. Since there are at most $(2L+2)$ symbols in term I to be replaced, and by the replacement rules, each replacement 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 we have

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

Using (4.9) in Theorem 4.1, it holds that for time $0\le t\le \sqrt{m}/(\ln m)^{C'}$

$$\left\|\mathcal{G}_t^{(4)}(\cdot)\right\|_\infty \le C(4,L)\frac{(\ln m)^C}{m},$$

based on (4.7)

$$\begin{aligned}\left|\partial_t \mathcal{G}_t^{(3)}(\boldsymbol{x}_{\alpha_1},\boldsymbol{x}_{\alpha_2},\boldsymbol{x}_{\alpha_3})\right| &\le \sup_{1\le\beta\le n}\left|\mathcal{G}_t^{(4)}(\boldsymbol{x}_{\alpha_1},\boldsymbol{x}_{\alpha_2},\boldsymbol{x}_{\alpha_3},\boldsymbol{x}_\beta)\right|\sqrt{\frac{\sum_{\beta=1}^n |f_\beta(t)-y_\beta|^2}{n}}\\ &\le \left\|\mathcal{G}_t^{(4)}(\cdot)\right\|_\infty \sqrt{R_S(\boldsymbol{\theta}_0)}\\ &\le C(4,L)\frac{(\ln m)^C}{m},\end{aligned}$$

then for any $1\le\alpha_1,\alpha_2,\alpha_3\le n$, with time $0\le t\le\sqrt{m}/(\ln m)^{C'}$

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

Finally, we need to make estimate on $\|\mathcal{G}_0^{(3)}(\cdot)\|_\infty$. We shall take advantage of the $\mathrm{diag}(\cdot)\mathbf{1}$ structure and rewrite $\mathcal{G}_t^{(3)}(\cdot)$ into

$$\begin{aligned}&\mathcal{G}_t^{(3)}(\boldsymbol{x}_{\alpha_1},\boldsymbol{x}_{\alpha_2},\boldsymbol{x}_\beta)\\ =&\frac{c_\sigma}{m}\left\langle \boldsymbol{E}_{t,\alpha_1}^{[2:L]}\boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_{\alpha_1})\boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[2:L]}\right)^\mathsf{T}\boldsymbol{a}_t,\boldsymbol{x}_{\alpha_2}^{[L]}\right\rangle\langle\boldsymbol{x}_{\alpha_1},\boldsymbol{x}_\beta\rangle\\ &+\sum_{k=2}^L\frac{c_{\mathrm{res}}^2}{L^2m}\left\langle \boldsymbol{E}_{t,\alpha_1}^{[(k+1):L]}\boldsymbol{\sigma}_{[k]}^{(1)}(\boldsymbol{x}_{\alpha_1})\boldsymbol{\sigma}_{[k]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[(k+1):L]}\right)^\mathsf{T}\boldsymbol{a}_t,\boldsymbol{x}_{\alpha_2}^{[L]}\right\rangle\left\langle\boldsymbol{x}_{\alpha_1}^{[k-1]},\boldsymbol{x}_\beta^{[k-1]}\right\rangle\\ &+\frac{c_\sigma}{m}\left\langle \boldsymbol{E}_{t,\alpha_2}^{[2:L]}\boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_{\alpha_2})\boldsymbol{\sigma}_{[1]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[2:L]}\right)^\mathsf{T}\boldsymbol{a}_t,\boldsymbol{x}_{\alpha_1}^{[L]}\right\rangle\langle\boldsymbol{x}_{\alpha_2},\boldsymbol{x}_\beta\rangle\\ &+\sum_{k=2}^L\frac{c_{\mathrm{res}}^2}{L^2m}\left\langle \boldsymbol{E}_{t,\alpha_2}^{[(k+1):L]}\boldsymbol{\sigma}_{[k]}^{(1)}(\boldsymbol{x}_{\alpha_2})\boldsymbol{\sigma}_{[k]}^{(1)}(\boldsymbol{x}_\beta)\left(\boldsymbol{E}_{t,\beta}^{[(k+1):L]}\right)^\mathsf{T}\boldsymbol{a}_t,\boldsymbol{x}_{\alpha_1}^{[L]}\right\rangle\left\langle\boldsymbol{x}_{\alpha_2}^{[k-1]},\boldsymbol{x}_\beta^{[k-1]}\right\rangle,\end{aligned}$$

then at time $t=0$, without loss of generality, each term in $\mathcal{G}_0^{(3)}(\cdot)$ is of the form

$$\frac{c}{m}\left\langle \boldsymbol{B}\boldsymbol{a}_0,\boldsymbol{x}_{\alpha_1}^{[L]}\right\rangle\left\langle\boldsymbol{x}_{\alpha_2}^{[l]},\boldsymbol{x}_\beta^{[l]}\right\rangle,\quad 0\le l\le L-1, \tag{D.1}$$

where $\boldsymbol{B}$ is some specific matrix that changes from term to term, then we can rewrite the inner product into:

$$\frac{c}{m}\left\langle \boldsymbol{a}_0,\boldsymbol{B}^\mathsf{T}\boldsymbol{x}_{\alpha_1}^{[L]}\right\rangle\left\langle\boldsymbol{x}_{\alpha_2}^{[l]},\boldsymbol{x}_\beta^{[l]}\right\rangle, \tag{D.2}$$

we shall focus on the term

$$\left\langle \boldsymbol{a}_0, \boldsymbol{B}^\mathsf{T} \boldsymbol{x}_{\alpha_1}^{[L]} \right\rangle \left\langle \boldsymbol{x}_{\alpha_2}^{[l]}, \boldsymbol{x}_\beta^{[l]} \right\rangle, \tag{D.3}$$

note that each entry of $\boldsymbol{a}_0$ is i.i.d $\mathcal{N}(0,1)$, also based on Propositions A.3 and A.4, with high probability with respect to random initialization, for time $0 \le t \le (\ln m)^{C'}$

$$\|\boldsymbol{B}^\mathsf{T}\|_{2\to 2}, \boldsymbol{x}_{\alpha_1}^{[L]}, \boldsymbol{x}_{\alpha_2}^{[l]}, \boldsymbol{x}_\beta^{[l]} \le c,$$

then after taking conditional expectation except for the random variable $\boldsymbol{a}_0$

$$\left\langle \boldsymbol{a}_0, \boldsymbol{B}^\mathsf{T} \boldsymbol{x}_{\alpha_1}^{[L]} \right\rangle \left\langle \boldsymbol{x}_{\alpha_2}^{[l]}, \boldsymbol{x}_\beta^{[l]} \right\rangle \sim \mathcal{N}(0,c), \tag{D.4}$$

apply Lemma A.3 directly, with high probability

$$\frac{c}{m} \left\langle \boldsymbol{a}_0, \boldsymbol{B}^\mathsf{T} \boldsymbol{x}_{\alpha_1}^{[L]} \right\rangle \left\langle \boldsymbol{x}_{\alpha_2}^{[l]}, \boldsymbol{x}_\beta^{[l]} \right\rangle \le c \frac{(\ln m)^C}{m}, \tag{D.5}$$

consequently

$$\left\| \mathcal{G}_0^{(3)}(\cdot) \right\|_\infty \le C(3,L) \frac{(\ln m)^C}{m}, \tag{D.6}$$

then for any $1 \le \alpha_1, \alpha_2, \alpha_3 \le n$, with time $0 \le t \le \sqrt{m}/(\ln m)^{C'}$

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

Similarly, based on (4.7), for time $0 \le t \le \sqrt{m}/(\ln m)^{C'}$

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

set $\boldsymbol{x}_\beta = \boldsymbol{x}_{\alpha_3}$

$$\begin{aligned} \left| \partial_t \mathcal{G}_t^{(2)}(\boldsymbol{x}_{\alpha_1}, \boldsymbol{x}_{\alpha_2}) \right| &\le \left(C(3,L) \frac{(\ln m)^C}{m} + tC(4,L) \frac{(\ln m)^C}{m} \right) \sqrt{\frac{\sum_{\beta=1}^n \left| f_\beta(t) - y_\beta \right|^2}{n}} \\ &\le (C(3,L) + tC(4,L)) \frac{(\ln m)^C}{m}, \end{aligned} \tag{D.7}$$

and (D.7) finishes the proof of Theorem 4.2. □

Proof of Corollary 4.1. Firstly, based on Proposition C.3, if $m=\Omega\big((\frac{n}{\lambda_0})^{2+\varepsilon}\big)$, we have with high probability with respect to random initialization,

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

set $\lambda=\frac{3\lambda_0}{4}$, which finishes the proof of (4.12).

We shall move on to the change of the least eigenvalue of the NTK. Recall (D.7) in the proof of Theorem 4.2, for time $0\le t\le\sqrt{m}/(\ln m)^{C'}$,

$$\left|\partial_t\mathcal{G}_t^{(2)}\left(\boldsymbol{x}_{\alpha_1},\boldsymbol{x}_{\alpha_2}\right)\right|\le\left(C(3,L)+tC(4,L)\right)\frac{(\ln m)^C}{m},$$

consequently

$$\left|\mathcal{G}_t^{(2)}\left(\boldsymbol{x}_{\alpha_1},\boldsymbol{x}_{\alpha_2}\right)-\mathcal{G}_0^{(2)}\left(\boldsymbol{x}_{\alpha_1},\boldsymbol{x}_{\alpha_2}\right)\right|\le t\left(C(3,L)+tC(4,L)\right)\frac{(\ln m)^C}{m}.$$

The above inequality can be used to derive a bound of the change of the least eigenvalue of the $\mathcal{G}_t^{(2)}(\cdot)$

$$\begin{aligned}\left\|\mathcal{G}_t^{(2)}-\mathcal{G}_0^{(2)}\right\|_{2\to2}\le\left\|\mathcal{G}_t^{(2)}-\mathcal{G}_0^{(2)}\right\|_{\mathrm{F}}&\le n\left\|\mathcal{G}_t^{(2)}-\mathcal{G}_0^{(2)}\right\|_\infty\\&\le nt\left(C(3,L)+tC(4,L)\right)\frac{(\ln m)^C}{m},\end{aligned}$$

we set t^* satisfying

$$nt^*\left(C(3,L)+t^*C(4,L)\right)\frac{(\ln m)^C}{m}=\frac{\lambda}{2},$$

rewrite the equation above, we have

$$C(4,L)(t^*)^2+C(3,L)t^*=\frac{\lambda m}{2(\ln m)^C n},\tag{D.8}$$

solve (D.8), we obtain that

$$t^*=\frac{-C(3,L)+\sqrt{\left(C(3,L)\right)^2+2C(4,L)\frac{\lambda m}{(\ln m)^C n}}}{2C(4,L)},\tag{D.9}$$

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

$$t^*\ge\frac{1}{2}\sqrt{\frac{\frac{\lambda m}{(\ln m)^C n}}{C(4,L)}}=\frac{1}{2}\sqrt{\frac{\lambda m}{C(4,L)(\ln m)^C n}}.\tag{D.10}$$

Moreover

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

then let

$$\bar{t} := \inf\left\{t:\ \lambda_{\min}\left[\mathcal{K}_t^{(2)}\left(\boldsymbol{x}_\alpha,\boldsymbol{x}_\beta\right)\right]_{1\le\alpha,\beta\le n}\ge \lambda/2\right\},$$

naturally

$$t^*\le \bar{t}, \tag{D.11}$$

using (4.6), we have for any $0\le t\le\bar{t}$,

$$\begin{aligned}\partial_t\sum_{\alpha=1}^n\|f_\alpha(t)-y_\alpha\|_2^2 &\le \sum_{\alpha,\beta=1}^n -\frac{2}{n}K_t^{(2)}(\boldsymbol{x}_\alpha,\boldsymbol{x}_\beta)(f_\alpha(t)-y_\alpha)(f_\beta(t)-y_\beta)\\ &\le -\frac{\lambda}{n}\sum_{\alpha=1}^n\|f_\alpha(t)-y_\alpha\|_2^2,\end{aligned} \tag{D.12}$$

then

$$\sum_{\alpha=1}^n\|f_\alpha(t)-y_\alpha\|_2^2\le\exp\left(-\frac{\lambda t}{n}\right)\sum_{\alpha=1}^n\|f_\alpha(0)-y_\alpha\|_2^2, \tag{D.13}$$

we can rewrite (D.13) into

$$R_S(\boldsymbol{\theta}_t)\le\exp\left(-\frac{\lambda t}{n}\right)R_S(\boldsymbol{\theta}_0), \tag{D.14}$$

set $R_S(\boldsymbol{\theta}_t)=\varepsilon$, it takes time $t\le\frac{n}{\lambda}\ln(\frac{C'}{\varepsilon})$ for loss $R_S(\boldsymbol{\theta}_t)$ to reach accuracy ε, hence if the following holds

$$t\le\frac{n}{\lambda}\ln\left(\frac{C'}{\varepsilon}\right)\le t^*\le\bar{t}, \tag{D.15}$$

then the width m is required to yield the lower bound for t^* derived in (D.10),

$$\frac{n}{\lambda}\ln\left(\frac{C'}{\varepsilon}\right)\le\frac{1}{2}\sqrt{\frac{\lambda m}{C(4,L)(\ln m)^C n}}. \tag{D.16}$$

Then we have

$$m\ge C(4,L)\left(\frac{n}{\lambda}\right)^3(\ln m)^C\ln\left(\frac{C'}{\varepsilon}\right)^2,$$

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

$$m=\Omega\left(\left(\frac{n}{\lambda}\right)^3L^2(\ln m)^C\ln\left(\frac{C'}{\varepsilon}\right)\right), \tag{D.17}$$

where ε is the desired training accuracy. □

References

[1] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 242–252. PMLR, 09–15 Jun 2019.

[2] Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. In *International Conference on Learning Representations*, 2019.

[3] Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 322–332. PMLR, 09–15 Jun 2019.

[4] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On Exact Computation with an Infinitely Wide Neural Net. In *Advances in Neural Information Processing Systems*, pages 8139–8148, 2019.

[5] Peter L Bartlett, David P Helmbold, and Philip M Long. Gradient Descent with Identity Initialization Efficiently Learns Positive-definite Linear Transformations by Deep Residual Networks. *Neural Comput.*, 31(3):477–502, 2019.

[6] Alon Brutzkus and Amir Globerson. Globally optimal gradient descent for a convnet with gaussian inputs. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 605–614. JMLR, 2017.

[7] Lenaic Chizat and Francis Bach. On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport. In *Advances in Neural Information Processing Systems*, pages 3036–3046, 2018.

[8] Ronan Collobert and Jason Weston. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. In *Proceedings of the 25th International Conference on Machine Learning*, pages 160–167, 2008.

[9] George E Dahl, Dong Yu, Li Deng, and Alex Acero. Context-dependent Pre-trained Deep Neural Networks for Large-vocabulary Speech Recognition. *IEEE Audio, Speech, Language Process*, 20(1):30–42, 2011.

[10] Amit Daniely, Roy Frostig, and Yoram Singer. Toward Deeper Understanding of Neural Networks: The Power of Initialization and a Dual View on Expressivity. In *Advances In Neural Information Processing Systems*, pages 2253–2261, 2016.

[11] Simon Du and Jason Lee. On the power of over-parametrization in neural networks with quadratic activation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1329–1338. PMLR, 10–15 Jul 2018.

[12] Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1675–1685. PMLR, 09–15 Jun 2019.

[13] Simon Du, Jason Lee, Yuandong Tian, Aarti Singh, and Barnabas Poczos. Gradient descent learns one-hidden-layer CNN: Don't be afraid of spurious local minima. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1339–1348. PMLR, 10–15 Jul

2018.

[14] Simon S Du, Chi Jin, Jason D Lee, Michael I Jordan, Aarti Singh, and Barnabas Poczos. Gradient Descent can take Exponential Time to Escape Saddle Points. In *Advances in Neural Information Processing Systems*, pages 1067–1077, 2017.

[15] Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019.

[16] Weinan E, Chao Ma, and Lei Wu. A Comparative Analysis of the Optimization and Generalization Property of Two-layer Neural Network and Random Feature Models under Gradient Descent Dynamics. *Science China Mathematics*, pages 1–24, 2020.

[17] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from Saddle PointsOnline Stochastic Gradient for Tensor Decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

[18] Rong Ge, Jason D. Lee, and Tengyu Ma. Learning one-hidden-layer neural networks with landscape design. In *International Conference on Learning Representations*, 2018.

[19] Xavier Glorot and Yoshua Bengio. Understanding the Difficulty of Training Deep Feedforward Neural Networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256, 2010.

[20] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[21] Moritz Hardt and Tengyu Ma. Identity matters in deep learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[23] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely Connected Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.

[24] Jiaoyang Huang and Horng-Tzer Yau. Dynamics of deep neural networks and neural tangent hierarchy. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4542–4551. PMLR, 13–18 Jul 2020.

[25] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In *Advances in Neural Information Processing Systems*, pages 8571–8580, 2018.

[26] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to Escape Saddle Points Efficiently. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1724–1732. JMLR. org, 2017.

[27] Kenji Kawaguchi. Deep learning without Poor Local Minima. In *Advances in Neural Information Processing Systems*, pages 586–594, 2016.

[28] Kenji Kawaguchi and Jiaoyang Huang. Gradient Descent Finds Global Minima for Generalizable Deep Neural Networks of Practical Sizes. In *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 92–99. IEEE, 2019.

[29] Beatrice Laurent and Pascal Massart. Adaptive Estimation of a Quadratic Functional by Model Selection. *Annals of Statistics*, pages 1302–1338, 2000.

[30] Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear mod-

els under gradient descent. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[31] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient Descent only Converges to Minimizers. In *Conference on Learning Theory*, pages 1246–1257, 2016.

[32] Yuanzhi Li and Yingyu Liang. Learning Overparameterized Neural Networks via Stochastic Gradient Descent on Structured Data. In *Advances in Neural Information Processing Systems*, pages 8157–8166, 2018.

[33] Yuanzhi Li and Yang Yuan. Convergence Analysis of Two-layer Neural Networks with Relu Activation. In *Advances in Neural Information Processing Systems*, pages 597–607, 2017.

[34] Yuqing Li, Tao Luo, and Chao Ma. Nonlinear weighted directed acyclic graph and a priori estimates for neural networks. *arXiv preprint arXiv:2103.16355*, 2021.

[35] Chao Ma, Qingcan Wang, and Weinan E. A Priori Estimates of the Population Risk for Residual Networks. *arXiv preprint arXiv:1903.02154*, 2019.

[36] Chao Ma, Qingcan Wang, Lei Wu, et al. Analysis of the Gradient Descent Algorithm for a Deep Neural Network Model with Skip-connections. *arXiv preprint arXiv:1904.05263*, 2019.

[37] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A Mean Field View of the Landscape of Two-layer Neural Networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

[38] Quynh Nguyen and Matthias Hein. The Loss Surface of Deep and Wide Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2603–2612. JMLR, 2017.

[39] Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. Xnor-net: Imagenet Classification Using Binary Convolutional Neural Networks. In *European Conference on Computer Vision*, pages 525–542. Springer, 2016.

[40] Grant Rotskoff and Eric Vanden-Eijnden. Parameters as Interacting Particles: Long Time Convergence and Asymptotic Error Scaling of Neural Networks. In *Advances in Neural Information Processing Systems*, pages 7146–7155, 2018.

[41] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the Game of Go with Deep Neural Networks and Tree Search. *Nature*, 529(7587):484, 2016.

[42] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the Game of Go without Human Knowledge. *Nature*, 550(7676):354–359, 2017.

[43] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020.

[44] Zhao Song and Xin Yang. Quadratic Suffices for Over-parametrization via Matrix Chernoff Bound. *arXiv preprint arXiv:1906.03593*, 2019.

[45] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. Inception-v4, Inception-resnet and the Impact of Residual Connections on Learning (2016). *arXiv preprint arXiv:1602.07261*, 2016.

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

[47] Greg 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.

[48] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 87.1–87.12. BMVA Press, September 2016.

[49] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding Deep Learning Requires Rethinking Generalization, 2018.

[50] Guodong Zhang, James Martens, and Roger B Grosse. Fast Convergence of Natural Gradient Descent for Over-parameterized Neural Networks. In *Advances in Neural Information Processing Systems*, pages 8080–8091, 2019.

[51] Yi Zhou and Yingbin Liang. Critical Points of Neural Networks: Analytical Forms and Landscape Properties. *arXiv preprint arXiv:1710.11205*, 2017.

[52] Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep ReLU networks. *Machine Learning*, 109(3):467–492, 2020.