Neural Tangent Kernel Analysis of Deep Narrow Neural Networks

Jongmin Lee [1] Joo Young Choi [1] Ernest K. Ryu [1] Albert No [2]

Abstract

The tremendous recent progress in analyzing the training dynamics of overparameterized neural networks has primarily focused on wide networks and therefore does not sufficiently address the role of depth in deep learning. In this work, we present the first trainability guarantee of infinitely deep but narrow neural networks. We study the infinite-depth limit of a multilayer perceptron (MLP) with a specific initialization and establish a trainability guarantee using the NTK theory. We then extend the analysis to an infinitely deep convolutional neural network (CNN) and perform brief experiments.

1. Introduction

Despite the remarkable experimental advancements of deep learning in many domains, a theoretical understanding behind this success remains elusive. Recently, significant progress has been made by analyzing limits of infinitely large neural networks to obtain provable guarantees. The neural tangent kernel (NTK) and the mean-field (MF) theory are the most prominent results.

However, these prior analyses primarily focus on the infinite width limit and therefore do not sufficiently address the role of depth in deep learning. After all, substantial experimental evidence indicates that depth is indeed an essential component to the success of modern deep neural networks. Analyses directly addressing the limit of infinite depth may lead to an understanding of the role of depth.

In this work, we present the first trainability guarantee of infinitely deep but narrow neural networks. We study the infinite-depth limit of a multilayer perceptron (MLP) with a very specific initialization and establish a trainability guarantee using the NTK theory. The MLP uses ReLU activation functions and has width on the order of input dimension + output dimension. Furthermore, we extend the analysis to an infinitely deep convolutional neural network (CNN) and perform brief experiments.

1.1. Prior works

The classical universal approximation theorem establishes that wide 2-layer neural networks can approximate any continuous function (Cybenko, 1989; Funahashi, 1989). Extensions and generalizations (Hornik, 1991; Leshno et al., 1993; Jones, 1992; Barron, 1993; Pinkus, 1999) and random feature learning (Rahimi & Recht, 2007; 2008a;b), a constructive version of the universal approximation theorem, use large width in their analyses. As overparameterization got recognized as a key component in understanding the performance of deep learning (Zhang et al., 2017), analyses of large neural networks started to appear in the literature (Soltanolkotabi et al., 2019; Allen-Zhu et al., 2019; Du et al., 2019a;b; Zou et al., 2020; Li & Liang, 2018), and their infinite-width limits such as neural network as Gaussian process (NNGP) (Neal, 1994; 1996; Williams, 1997; Lee et al., 2018; Matthews et al., 2018; Novak et al., 2019), neural tangent kernel (NTK) (Jacot et al., 2018), and mean-field (MF) (Mei et al., 2018; Chizat & Bach, 2018; Rotskoff & Vanden-Eijnden, 2018; Rotskoff et al., 2019; Sirignano & Spiliopoulos, 2020a;b; Nguyen & Pham, 2021; Pham & Nguyen, 2021) were formulated. NNGP characterizes the neural network at initialization, while NTK and MF analyses provide guarantees of trainability with SGD. This line of research naturally raises the question of whether very deep neural networks also enjoy similar properties as wide neural networks, especially given the importance of depth in modern deep learning.

The analogous line of research for very deep neural networks has a shorter history. Universality of deep narrow neural networks (Lu et al., 2017; Hanin & Sellke, 2017; Lin & Jegelka, 2018; Hanin, 2019; Kidger & Lyons, 2020; Park et al., 2021; Tabuada & Gharesifard, 2021), lower bounds on the minimum width necessary for universality (Lu et al., 2017; Hanin & Sellke, 2017; Johnson, 2019; Park et al., 2021), and quantitative analyses showing the benefit of depth over width in approximating certain functions (Telgarsky, 2016) are all very recent developments. On the other

[1]Department of Mathematical Sciences, Seoul National University, Seoul, Korea [2]Department of Electronic and Electrical Engineering, Hongik University, Seoul, Korea. Correspondence to: Albert No <albertno@hongik.ac.kr>, Ernest K. Ryu <ernestryu@snu.ac.kr>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

hand, the neural ODE (Chen et al., 2018) is a continuous-depth model that can be considered an infinite-depth limit of a neural network with residual connections. Also stochastic extensions of neural ODE, viewing infinitely deep ResNets as diffusion processes, have been considered in (Peluchetti & Favaro, 2020; 2021). However, these continuous-depth limits do not come with any trainability or generalization guarantees. In the infinite width *and* depth limit, a quantitative universal approximation result has been established (Lu et al., 2021) and a trainability guarantee was obtained in setups combining the MF limit with the continuous-depth limit inspired by the neural ODE (Lu et al., 2020; Ding et al., 2021; 2022).

Efforts to understand the trainability of deep non-wide neural networks have been made. Arora et al. (2019a); Shamir (2019) establishes trainability guarantees for linear deep networks (no activation functions). Pennington et al. (2017) studied the so-called dynamic isometry property, Hanin (2018) studied the exploding and vanishing gradient problem for ReLU MLPs, and Huang et al. (2020) studied the NTK of deep MLPs and ResNets at initialization, but these results are limited to the state of the neural network at initialization and therefore do not directly establish guarantees on the training dynamics. To the best of our knowledge, no prior work has yet established a trainability guarantee on (non-linear) deep narrow neural networks.

Many extensions and variations of the NTK have been studied: NTK analysis with convolutional layers (Arora et al., 2019b; Yang, 2019), further refined analyses and experiments (Lee et al., 2019), NTK analysis with regularizers and noisy gradients (Chen et al., 2020), finite-width NTK analysis (Hanin & Nica, 2020), quantitative universality result of the NTK (Ji et al., 2020), generalization properties of overparameterized neural networks (Bietti & Mairal, 2019; Woodworth et al., 2020), unified analysis of NTK and MF (Geiger et al., 2020), notion of lazy training generalizing linearization effect of the NTK regime (Chizat et al., 2019), closed-form evaluations of NTK kernel values (Cho & Saul, 2009; Arora et al., 2019b), analyses of distribution shift and meta learning in the infinite-width regime (Adlam et al., 2021; Nguyen et al., 2021), library implementations of infinitely wide neural networks as kernel methods with NTK (Novak et al., 2020), empirical evaluation of finite vs. infinite neural networks (Lee et al., 2020), and constructing better-performing kernels with improved extrapolation of the standard initialization (Sohl-Dickstein et al., 2020). Such results have mostly focused on the analysis and application of wide, rather than deep, neural networks.

1.2. Contribution

In this work, we analyze the training dynamics of a deep narrow MLP and CNN without residual connections and establish trainability guarantees. To the best of our knowledge, this is the first trainability guarantee of deep narrow neural networks, with or without residual connections; prior trainability guarantees consider the infinite width or the infinite depth and width limits. Essential to our analysis is the particular choice of initialization, as very deep networks with the usual initialization schemes are known to not be trainable (He & Sun, 2015, Section 4.4), (Srivastava et al., 2015, Section 3.1), (Huang et al., 2020, Section 4.1). By considering the limit in which only the depth is infinite, we demonstrate that infinite depth is sufficient to obtain a neural network with trainability guarantees.

The key technical challenge of our work arises from bounding the variation of the scaled NTK in the infinite-depth limit via a delicate Lyapunov analysis. Our infinite-depth analysis requires that we control an infinite composition of layers and an infinite product of matrices, which is significantly more technical than the prior infinite-width analyses that can more directly apply the central limit theorem and law of large numbers.

The MLP we analyze is narrow in the sense that it is within a factor of 2 of the width lower bound for universal approximation (Park et al., 2021). In particular, the width does not depend on the number of data points.

2. Preliminaries and notation

In this section, we review the necessary background and set up the notation. We largely follow the notions of Jacot et al. (2018), although our notation has some minor differences.

Given a function $f_\theta(x) \in \mathbb{R}^n$ with $\theta \in \mathbb{R}^p$ and $x \in \mathbb{R}^d_+$, write $\partial_\theta f_\theta(x) \in \mathbb{R}^{n\times p}$ to denote the Jacobian matrix with respect to θ. If $f_\theta(x) \in \mathbb{R}$ is scalar-valued, write $\nabla_\theta f_\theta(x) \in \mathbb{R}^{p\times 1}$ for the gradient. The gradient and Jacobian matrices are, by convention, transposes of each other, i.e., $\nabla_\theta f_\theta(x) = (\partial_\theta f_\theta(x))^\intercal$. Write $\|\cdot\|$ to denote the standard Euclidean norm for vectors and the standard operator norm for matrices. Write $\langle \cdot, \cdot \rangle$ for the vector and Frobenius inner products. Write $\mathbb{R}^d_+ \subset \mathbb{R}^d$ for the strict positive orthant, i.e., $\mathbb{R}^d_+$ is the set of vectors with element-wise positive entries. Write $\xrightarrow{p}$ to denote convergence in probability. Write $g \sim \mathcal{GP}(\mu, \Sigma)$ to denote that g is a Gaussian process with mean $\mu(x)$ and covariance kernel $\Sigma(x, x')$ (Rasmussen, 2004; Rasmussen & Williams, 2005).

Let p^{in} be the empirical distribution on a training dataset $x_1, x_2, \ldots, x_N \in \mathbb{R}^{d_{in}}_+$, which we assume are element-wise positive. Let $\mathcal{F} = \{f : \mathbb{R}^{d_{in}}_+ \to \mathbb{R}^{d_{out}}\}$ be the function space with a seminorm $\|\cdot\|_{p^{in}}$ induced by the bilinear map

$$\langle f, g \rangle_{p^{in}} = \mathbb{E}_{x\sim p^{in}}[f(x)^\intercal g(x)].$$

If W is a matrix-valued function, write

$$\|W\|_{p^{in}}^2 = \mathbb{E}_{x\sim p^{in}}\|W(x)\|^2,$$

where $\|\cdot\|$ here is the operator norm.

Kernel gradient flow. Let $\mathcal{L}\colon \mathcal{F} \to \mathbb{R}$ be a functional loss. We primarily consider $\mathcal{L}(f) = \frac{1}{2}\|f - f^\star\|_{p^{in}}^2$ with a given target function $f^\star$. We train a neural network f_θ by solving

$$\underset{\theta}{\text{minimize}} \quad \mathcal{L}(f_\theta).$$

Let $\mathcal{F}^*$ denote the dual of $\mathcal{F}$ with respect to p^{in}. So $\mathcal{F}^*$ consists of linear maps $\langle \delta, \cdot\rangle_{p^{in}}$ for some $\delta \in \mathcal{F}$. Let $\partial_f \mathcal{L}|_{f_0}$ denote the functional derivative of the loss at f_0. Since $\partial_f \mathcal{L}|_{f_0} \in \mathcal{F}^*$, there exists a corresponding dual element $\delta|_{f_0} \in \mathcal{F}$, where $\partial_f \mathcal{L}|_{f_0} = \langle \delta|_{f_0}, \cdot\rangle_{p^{in}}$. To clarify, $\delta|_{f_0}(x)$ is defined to be a length d_{out} column vector.

A multi-dimensional kernel $K\colon \mathbb{R}_+^{d_{\text{in}}} \times \mathbb{R}_+^{d_{\text{in}}} \to \mathbb{R}^{d_{\text{out}}\times d_{\text{out}}}$ is a function such that $K(x, x') = K(x', x)^\intercal$ for all $x, x' \in \mathbb{R}_+^{d_{\text{in}}}$. A multi-dimensional kernel is positive semidefinite if

$$\mathbb{E}_{x,x'\sim p^{in}}\left[f(x)^\intercal K(x, x') f(x')\right] \geq 0$$

for all $f \in \mathcal{F}$ and (strictly) positive definite if the inequality holds strictly when $\|f\|_{p^{in}} \neq 0$. To clarify, $\mathbb{E}_{x,x'\sim p^{in}}$ denotes the expectation with respect to x and x' sampled independently from p^{in}. The kernel gradient of $\mathcal{L}$ at f_0 with respect to the kernel K is defined as

$$\nabla_K \mathcal{L}|_{f_0}(x) = \mathbb{E}_{x'\sim p^{in}}\left[K(x, x')\delta|_{f_0}(x')\right]$$

for all $x \in \mathbb{R}_+^{d_{\text{in}}}$. We say a time-dependent function f_t follows the kernel gradient flow with respect to K if

$$\partial_t f_t(x) = -\nabla_K \mathcal{L}|_{f_t}(x)$$

for all $t > 0$ and $x \in \mathbb{R}_+^{d_{\text{in}}}$. During the kernel gradient flow, the loss $\mathcal{L}(f_t)$ evolves as

$$\partial_t \mathcal{L}|_{f_t} = -\,\mathbb{E}_{x,x'\sim p^{in}}\left[\delta|_{f_t}(x)^\intercal K(x, x')\delta|_{f_t}(x')\right].$$

If K is positive definite and if certain regularity conditions hold, then kernel gradient flow converges a critical point and it converges to a global minimum if $\mathcal{L}$ is convex and bounded from below.

Neural tangent kernel. Given a neural network $f_{\theta(t)}$, Jacot et al. (2018) defines the neural tangent kernel (NTK) at time t as

$$\Theta_t(x, x') = \partial_\theta f_{\theta(t)}(x)\left(\partial_\theta f_{\theta(t)}(x')\right)^\intercal,$$

which, by definition, is a positive semidefinite kernel. Jacot et al. (2018) pointed out that a neural network trained with gradient flow, which we define and discuss in Section 3.2, can be viewed as kernel gradient flow with respect to Θ_t, i.e.,

$$\partial_t f_{\theta(t)}(x) = -\nabla_{\Theta_t}\mathcal{L}|_{f_t}(x).$$

However, even though Θ_t is always positive semidefinite, the time-dependence of Θ_t makes the training dynamics non-convex and prevents one from establishing trainability guarantees in general. The contribution of Jacot et al. (2018) is showing that $\Theta_t \xrightarrow{p} \Theta$ in an appropriate infinite-width limit, where Θ is a fixed limit that does not depend on time. Then, since Θ is fixed, kernel gradient flow generically converges provided that the loss $\mathcal{L}$ is convex.

3. NTK analysis of infinitely deep MLP

Consider an L-layer multilayer perceptron (MLP) $f_\theta^L\colon \mathbb{R}_+^{d_{\text{in}}} \to \mathbb{R}^{d_{\text{out}}}$ parameterized by θ, where the input $x \in \mathbb{R}_+^{d_{\text{in}}}$ is a d_{in}-dimensional vector with positive entries and the output is a d_{out}-dimensional vector. The network consists of $L-1$ fully connected hidden layers with uniform width $d_{\text{in}} + d_{\text{out}} + 1$, each followed by the ReLU activation function. The final output layer has width d_{out} and is not followed by an activation function.

Let us set up specific notation. Define the pre-activation values as

$$\begin{aligned} f^1_{\theta^{(1)}}(x) &= W^1 x + b^1 \\ f^l_{\theta^{(l)}}(x) &= W^l \sigma(f^{l-1}_{\theta^{(l-1)}}(x)) + b^l, \qquad 2 \leq l \leq L. \end{aligned}$$

We use ReLU for the activation σ. The weight matrices have dimension $W^1 \in \mathbb{R}^{(d_{\text{in}}+d_{\text{out}}+1)\times d_{\text{in}}}$, $W^2, \dots, W^{L-1} \in \mathbb{R}^{(d_{\text{in}}+d_{\text{out}}+1)\times(d_{\text{in}}+d_{\text{out}}+1)}$, and $W^L \in \mathbb{R}^{d_{\text{out}}\times(d_{\text{in}}+d_{\text{out}}+1)}$. The bias vectors have dimension $b^1, \dots, b^{L-1} \in \mathbb{R}^{(d_{\text{in}}+d_{\text{out}}+1)\times 1}$ and $b^L \in \mathbb{R}^{d_{\text{out}}\times 1}$. For $1 \leq l \leq L$, write $\theta^{(l)} = \{W^i, b^i : i \leq l\}$ to denote the collection of parameters up to l-th layer. Let $f_\theta^L = f^L_{\theta^{(L)}}$.

3.1. Initialization

Motivated by Kidger & Lyons (2020), initialize the weights of our L-layer MLP f_θ^L as follows:

$$\begin{aligned} W^1 &= \begin{bmatrix} C_L I_{d_{\text{in}}} \\ u^1 \\ 0_{d_{\text{out}}\times d_{\text{in}}} \end{bmatrix} \\ W^l &= \begin{bmatrix} I_{d_{\text{in}}} & 0_{d_{\text{in}}\times 1} & 0_{d_{\text{in}}\times d_{\text{out}}} \\ u^l & 0_{1\times 1} & 0_{1\times d_{\text{out}}} \\ 0_{d_{\text{out}}\times d_{\text{in}}} & 0_{d_{\text{out}}\times 1} & I_{d_{\text{out}}} \end{bmatrix} \\ W^L &= \begin{bmatrix} 0_{d_{\text{out}}\times d_{\text{in}}} & 0_{d_{\text{out}}\times 1} & I_{d_{\text{out}}} \end{bmatrix} \end{aligned}$$

for $2 \leq l \leq L-1$, where $u^l \in \mathbb{R}^{1\times d_{\text{in}}}$ is randomly sampled $u^l_i \overset{iid}{\sim} \mathcal{N}(0, \frac{1}{d_{\text{in}}}\rho^2)$ for $1 \leq l \leq L-1$. Here, I_d is the $d \times d$ identity matrix and $0_{m\times n}$ is the m by n matrix with all zero entries, $C_L > 0$ is a scalar growing as a function of L at a rate satisfying $L^2/C_L \to 0$, and $\rho > 0$ is a fixed variance parameter. Initializes the biases as follows:

$$b^1 = \begin{bmatrix} 0_{d_{\text{in}}\times 1} \\ v^1 \\ C_L \mathbb{1}_{d_{\text{out}}} \end{bmatrix}$$
$$b^l = \begin{bmatrix} 0_{d_{\text{in}}\times 1} \\ v^l \\ 0_{d_{\text{out}}\times 1} \end{bmatrix}$$
$$b^L = \begin{bmatrix} -C_L \mathbb{1}_{d_{\text{out}}} \end{bmatrix}$$

for $2 \le l \le L-1$, where $v^l \in \mathbb{R}$ is randomly sampled $v^l \overset{iid}{\sim} \mathcal{N}(0, C_L^2\beta^2)$ for $1 \le l \le L-1$. Here, $\beta > 0$ is a fixed variance parameter. Note that $\mathbb{1}_k \in \mathbb{R}^{k\times 1}$ is the vector whose entries are all 1. Figure 1 illustrates this initialization scheme.

We clarify that while we use many specific non-random initializations, those parameters are not fixed throughout training. In other words, all parameters are trainable, just as one would expect from a standard MLP.

Kidger & Lyons (2020) used a similar construction to establish a universal approximation result for deep MLPs by showing that their deep MLP mimics a 2-layer wide MLP. However, their main concern is the existence of a weight configuration that approximates a given function, which does not guarantee that such a configuration can be found through training. On the other hand, we propose an explicit initialization and establish a trainability guarantee; our network outputs 0 at initialization and converges to the desired configuration through training.

3.2. Gradient flow and neural tangent kernel

We are now ready to describe the training of our neural network f_θ^L via gradient flow, a continuous-time model of gradient descent. Since our initialization scales the input by C_L at the first layer, we scale the learning rate accordingly, both in the continuous-time analysis of Section 4 and in the discrete-time experiments of Section 5, so that we get meaningful limits as $L \to \infty$.

Train f_θ^L with

$$\begin{aligned}\partial_t\theta(t) &\overset{\text{(a)}}{=} -\frac{1}{LC_L^2}\nabla_\theta \text{Loss}(\theta)\Big|_{\theta=\theta(t)} \\ &\overset{\text{(b)}}{=} -\frac{1}{LC_L^2}\left(\partial_\theta \mathcal{L}(f_\theta^L)\right)^\intercal\Big|_{\theta=\theta(t)} \\ &\overset{\text{(c)}}{=} -\frac{1}{LC_L^2}\mathbb{E}_{x\sim p^{in}}\left[(\partial_\theta f_\theta^L(x))^\intercal \delta|_{f_\theta^L}(x)\right]\Big|_{\theta=\theta(t)},\end{aligned}$$

where (a) defines the θ-update to be gradient flow with learning rate $1/(LC_L^2)$, (b) plugs in our notation, and (c) follows from the chain rule.

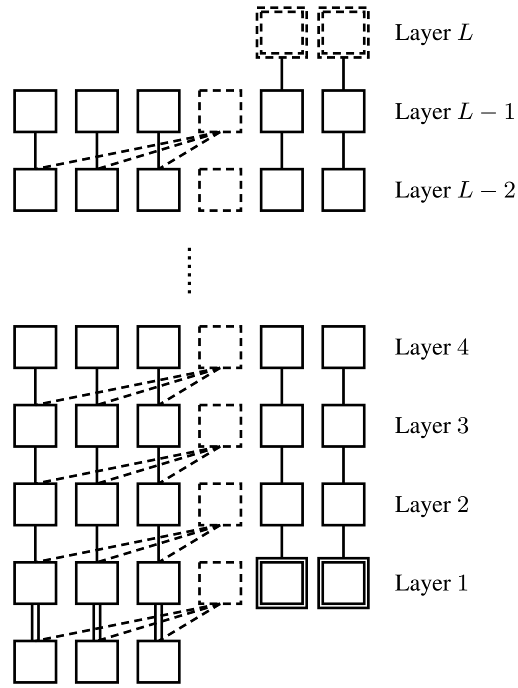


Figure 1. Initialization of deep MLP with $d_{in} = 3$ and $d_{out} = 2$. Intermediate layers have width $d_{in}+1+d_{out}$. Line styles indicate types of weight initializations (solid:1, double:C_L, dash:Gaussian, none:0). Box styles indicate types of bias initializations (solid:0, dash:Gaussian, double:C_L, double-dash:$-C_L$).

This gradient flow defines $\theta(t)$ to be a function of time, but we will often write θ rather than $\theta(t)$ for notational conciseness. This gradient flow and the chain rule induces the functional dynamics of the neural network:

$$\begin{aligned}\partial_t f_\theta^L(x) &= \partial_\theta f_\theta^L(x)\partial_t\theta \\ &= -\frac{1}{LC_L^2}\mathbb{E}_{x'\sim p^{in}}\left[\partial_\theta f_\theta^L(x)\left(\partial_\theta f_\theta^L(x')\right)^\intercal \delta|_{f_\theta^L}(x')\right].\end{aligned}$$

Since we use a scaling factor in our gradient flow, we define the *scaled NTK* at time t as

$$\tilde{\Theta}_t^L(x,x') = \frac{1}{LC_L^2}\partial_\theta f_\theta^L(x)\left(\partial_\theta f_\theta^L(x')\right)^\intercal.$$

Then,

$$\begin{aligned}\partial_t f_\theta^L(x) &= -\mathbb{E}_{x'\sim p^{in}}\left[\tilde{\Theta}_t^L(x,x')\delta|_{f_\theta^L}(x')\right] \\ &= -\nabla_{\tilde{\Theta}_t^L}\mathcal{L}|_{f_\theta^L}(x).\end{aligned}$$

3.3. Convergence in infinite-depth limit

We now analyze the convergence of the infinitely deep MLP.

Theorem 1 establishes that the scaled NTK at initialization (before training) of the randomly initialized MLP converges to a deterministic limit with a closed-form expression as the depth L becomes infinite.

Theorem 1 (Scaled NTK at initialization). *Suppose f_θ^L is initialized as in Section 3.1. For any $x, x' \in \mathbb{R}_+^{d_{in}}$,*

$$\tilde{\Theta}_0^L(x, x') \xrightarrow{p} \tilde{\Theta}^\infty(x, x')$$

as $L \to \infty$, where

$$\tilde{\Theta}^\infty(x, x') = (x^\intercal x' + 1 + \mathbb{E}_g[\sigma(g(x))\sigma(g(x'))])\, I_{d_{\text{out}}},$$

and $g \sim \mathcal{GP}(0, \frac{\rho^2}{d_{\text{in}}} x^\intercal x' + \beta^2)$.

When $L < \infty$, the scaled NTK $\tilde{\Theta}_t^L(x, x')$ depends on time through its dependence on $\theta(t)$. Theorem 2 establishes that $\tilde{\Theta}_t^L(x, x')$ becomes independent of t as $L \to \infty$.

Theorem 2 (Invariance of scaled NTK). *Let $T > 0$. Suppose $\int_0^T \left\| \delta|_{f_\theta^L} \right\|_{p^{in}} dt$ is stochastically bounded as $L \to \infty$. Then, for any $x, x' \in \mathbb{R}_+^{d_{in}}$,*

$$\tilde{\Theta}_t^L(x, x') \xrightarrow{p} \tilde{\Theta}^\infty(x, x')$$

uniformly for $t \in [0, T]$ as $L \to \infty$.

Let $\mathcal{L}(f) = \frac{1}{2}\|f - f^\star\|_{p^{in}}^2$ be the quadratic loss. In this case, the stochastic boundedness assumption of Theorem 2 holds, as we show in Lemma 16 of the appendix, and we characterize the training dynamics explicitly. For the sake of notational simplicity, assume the MLP's prediction is a scalar, i.e., assume $d_{out} = 1$. The generalization to multi-dimensional outputs is straightforward, following the arguments of (Jacot et al., 2018, Section 5).

Theorem 3 concludes the analysis by characterizing the trained MLP as $L \to \infty$. Define the kernel regression predictor as

$$f_{\text{ntk}}(x) = \left(\tilde{\Theta}^\infty(x, x_1), \dots, \tilde{\Theta}^\infty(x, x_N)\right) K^{-1} f^\star(X),$$

where $K_{i,j} = \tilde{\Theta}^\infty(x_i, x_j)$ and $[f^\star(X)]_i = f^\star(x_i)$ for $i, j \in \{1, \dots, N\}$. Let f_t be trained with the limiting kernel gradient flow of $f_{\theta(t)}^L$ as $L \to \infty$, i.e., $f_0 = 0$ and f_t follows

$$\partial_t f_t = -\nabla_{\tilde{\Theta}^\infty} \mathcal{L}|_{f_t}(x).$$

Then the infinitely deep training dynamics converge to f_{ntk} in the following sense.

Theorem 3 (Equivalence between deep MLP and kernel regression). *Let $\mathcal{L}(f) = \frac{1}{2}\|f - f^\star\|_{p^{in}}^2$. Let $\tilde{\Theta}^\infty$ be positive definite. If f_t follows the kernel gradient flow with respect to $\tilde{\Theta}^\infty$, then for any $x \in \mathbb{R}_+^{d_{in}}$,*

$$\lim_{t \to \infty} f_t(x) = f_{\text{ntk}}(x).$$

3.4. Proof outline

At a high level, our analysis follows the same line of argument as that of the original NTK paper (Jacot et al., 2018): Theorem 1 characterizes the limiting NTK at initialization, Theorem 2 establishes that the NTK remains invariant throughout training, and Theorem 3 establishes convergence in the case of quadratic loss functions. The proof of Theorem 3 follows from arguments similar to those of (Jacot et al., 2018, Theorem 3). The proof of Theorem 1 follows from identifying the recursive structure and noticing that the initialization is designed to simplify this recursion.

The key technical challenge of this work is in Theorem 2. The analysis is based on defining the Lyapunov function

$$\Gamma_L(t) = \Psi_{L,2}(t) + \Psi_{L,8}(t) + \sum_{j=1}^{8} \Phi_{L,j}(t)$$

(the individual terms will be defined soon), establishing

$$\Gamma_L(t) \le \Gamma_L(0) + \int_0^t \mathcal{O}(L/C_L)\Gamma_L(s)^4\, ds,$$

and appealing to Grönwall's lemma to show that $\Gamma_L(t)$ is invariant, i.e., $\Gamma_L(t) \to \Gamma_\infty(0)$ as $L \to \infty$ uniformly in $t \in [0, T]$. For $j = 1, \dots, 8$, the first term is defined as

$$\Phi_{L,j} = \frac{1}{(L-1)C_L^j} \sum_{l=1}^{L-1} \left(\left\| f_{\theta^{(l)}(0)}^l \right\|_{p^{in}} + \left\| f_{\theta^{(l)}(t)}^l - f_{\theta^{(l)}(0)}^l \right\|_{p^{in}} \right)^j$$

and its invariance implies that the average variation of the layers vanish for inputs $x \in \{x_1, \dots, x_N\}$ to the MLP.

Because we study the infinite-depth regime, our Lyapunov analysis is significantly more technical compared to the prior NTK analyses studying the infinite-width regime. Prior work dealt with sums of infinitely many terms, which were analyzed with the central limit theorem and law of large numbers. In contrast, the infinite depth of our setup leads to an infinite composition of layers and an infinite product of matrices, which must be controlled through a more delicate Lyapunov analysis.

For $L \ge j \ge k \ge 2$, define

$$\mathfrak{W}_j^k(x, t) = W^j \text{diag}(\dot{\sigma}(f_{\theta^{(j-1)}}^{j-1})) \cdots W^k \text{diag}(\dot{\sigma}(f_{\theta^{(k-1)}}^{k-1}))$$

(where W^j and $\theta^{(j)}$ depend on t and $f_{\theta^{(l)}}^l$ depends on x and t) and we bound its change by incorporating the following terms into the Lyapunov function:

$$\Psi_{L,2} = \tfrac{1}{(L-1)^2} \textstyle\sum_{l=2}^{L} \tfrac{L-1}{l-1} \sum_{i=2}^{l} \left(\left\| \mathfrak{W}_l^i(\cdot, 0) \right\|_{p^{in}} + \left\| \mathfrak{W}_l^i(\cdot, t) - \mathfrak{W}_l^i(\cdot, 0) \right\|_{p^{in}} \right)^2$$

$$\Psi_{L,8} = \tfrac{1}{L-1} \textstyle\sum_{l=2}^{L} \left(\left\| \mathfrak{W}_L^l(\cdot, 0) \right\|_{p^{in}} + \left\| \mathfrak{W}_L^l(\cdot, t) - \mathfrak{W}_L^l(\cdot, 0) \right\|_{p^{in}} \right)^8.$$

Establishing the invariance of $\Gamma_L(t)$ as $L \to \infty$ is the first step of the analysis, but, by itself, it does not characterize the limiting MLP for inputs $x \notin \{x_1, \dots, x_N\}$ and it only bounds the average variation of the layers, rather than the layer-wise variation. Hence, we generalize the invariance results with two additional Lyapunov analyses and combine these results to establish the scaled NTK's invariance.

Our Lyapunov analysis significantly is simplified by setting $\ddot{\sigma}(s) = 0$ and thereby removing terms involving $\ddot{\sigma}$. However, while $\ddot{\sigma}(s) = 0$ for $s \neq 0$, we cannot ignore the fact that $\ddot{\sigma}(0) \neq 0$ (in fact undefined). We resolve this issue by showing that all instances of $\sigma(s)$ never encounter the input $s = 0$ throughout training with probability approaching 1 as $L \to \infty$. We outline this argument below.

Due to the randomness of the initialization, the pre-activation values $f^l_{\theta^{(l)}(0)}(x)$ are element-wise nonzero for all $x \in \{x_1, \dots, x_N\}$ with probability 1. If $(f^l_{\theta^{(l)}(t)}(x))_r = 0$ for some l, r, and t, i.e., if a *zero-crossing* happens for a neuron at time t, then the analysis must somehow deal with the behavior of $\sigma(s)$ at $s = 0$. However, if zero-crossing happens for no neurons for all $t \in [0, T]$, then we can safely set $\ddot{\sigma} = 0$ in our analysis. We prove that the probability of a zero crossing (over all neurons of all layers and all $t \in [0, T]$) vanishes as $L \to \infty$. We specifically establish this claim by showing that as $L \to \infty$,

$$\sup_{1 \leq l \leq L-1, t \in [0,T]} \left\| f^l_{\theta^{(l)}(t)}(x) - f^l_{\theta^{(l)}(0)}(x) \right\| \leq KL$$

with high probability and

$$\Pr\left[\inf_{l,r} \left| \left(f^l_{\theta^{(l)}(0)}(x) \right)_r \right| > (K+1)L \right] \to 1.$$

for some constant $K > 0$. These two results establish that the zero-crossing probability vanishes as $L \to \infty$. We provide the details in Section C.

4. NTK analysis of infinitely deep CNN

Consider an L-layer convolutional neural network (CNN) $f^{L+1}_\theta \colon \mathbb{R}^{d \times d}_+ \to \mathbb{R}$ parameterized by θ, where the input $x \in \mathbb{R}^{d \times d}_+$ is a $d \times d$ image with positive entries and the output is a scalar. The network consists of $L-1$ convolutional layers using 3×3 filters and zero-padding of 1 with 3 output channels, each followed by the ReLU activation function. The L-th convolutional layer uses a 3×3 filter with zero-padding of 1 and has a single output channel. This is followed by a global average pool with no activation function applied before or after the average pool. All convolutional layers use stride of 1.

Let us set up specific notation. For a convolutional filter $w \in \mathbb{R}^{3 \times 3}$ and a (single-channel) image $x \in \mathbb{R}^{d \times d}_+$, denote the convolution operation with zero padding as

$$[w * x]_{i,j} = \langle w, \phi_{i,j}(x) \rangle$$

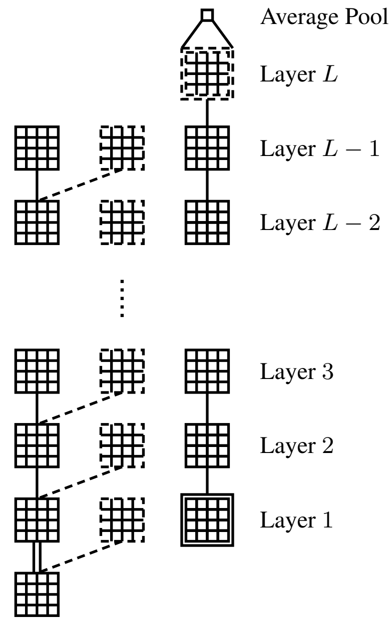


Figure 2. Initialization of deep CNN with 4×4 input. The 3 grids per row represent the 3 channels per layer, and the box at the top represents the scalar output of the final average pool. Line styles indicate types of weight initializations (solid:diag$(0,1,0)$, double:diag$(0, C_L, 0)$, dash:Gaussian, none:$0_{3\times3}$). Box styles indicate types of bias initializations (solid:0, dash:Gaussian, double:C_L, double-dash:$-C_L$).

for $1 \leq i, j \leq d$, where $\phi_{i,j}(x) = [x]_{i-1:i+1, j-1:j+1}$ and $x_{pq} = 0$ if p or q is less than 1 or greater than d, i.e., $x_{pq} = 0$ if the index is out of bounds. Define $\boldsymbol{\iota}_{3\times3} = \text{diag}(0,1,0)$ to be the 3×3 filter serving as the identity map. So $\boldsymbol{\iota}_{3\times3} * x = x$. Define the pre-activation values as

$$\left(f^1_{\theta^{(1)}}\right)_{r,:,:} = w^1_{r,1,:,:} * x + \mathbb{1}_{d\times d} b^1_r,$$

$$\left(f^l_{\theta^{(l)}}\right)_{r,:,:} = \sum_{s=1}^{n_{l-1}} w^l_{r,s,:,:} * \sigma\left(\left(f^{l-1}_{\theta^{(l-1)}}\right)_{s,:,:} \right) + \mathbb{1}_{d\times d} b^l_r,$$

for $2 \leq l \leq L$, where $\mathbb{1}_{d\times d}$ is the d by d matrix with all unit entries, n_l is the number of channels of the l-th layer, and $r = 1, \dots, n_l$. Our notation indexing the 3D and 4D tensors resembles the PyTorch convention and is defined precisely in Section I.1. We use ReLU for the activation σ. The number of channels are $3 = n_1 = \cdots = n_{L-1}$ and $n_0 = n_L = 1$. So, $f^l_{\theta^{(l)}}(x) \in \mathbb{R}^{n_l \times d \times d}$, $w^l \in \mathbb{R}^{n_l \times n_{l-1} \times 3 \times 3}$, and $b^l \in \mathbb{R}^{n_l}$ for $1 \leq l \leq L$. Write average pool as $S(A) = \frac{1}{d^2} \sum_{i=1}^d \sum_{j=1}^d A_{i,j}$ for $A \in \mathbb{R}^{d\times d}$. The CNN outputs

$$f^{L+1}_\theta = S(f^L_{\theta^{(L)}}).$$

For $1 \leq l \leq L$, write $\theta^{(l)} = \{w^i, b^i : i \leq l\}$ to denote the collection of parameters up to l-th layer.

4.1. Initialization

Initialize the filters of our L-layer CNN f_θ^{L+1} as follows:

$$w^1_{1,1,:,:} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & C_L & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad w^1_{2,1,:,:} = \begin{bmatrix} u^1_{1,1} & u^1_{1,2} & u^1_{1,3} \\ u^1_{2,1} & u^1_{2,2} & u^1_{2,3} \\ u^1_{3,1} & u^1_{3,2} & u^1_{3,3} \end{bmatrix},$$

$$w^1_{3,1,:,:} = 0_{3\times 3}$$
$$w^l_{1,:,:,:} = \boldsymbol{\iota}_{3\times 3}, 0_{3\times 3}, 0_{3\times 3}$$
$$w^l_{2,:,:,:} = \begin{bmatrix} u^l_{1,1} & u^l_{1,2} & u^l_{1,3} \\ u^l_{2,1} & u^l_{2,2} & u^l_{2,3} \\ u^l_{3,1} & u^l_{3,2} & u^l_{3,3} \end{bmatrix}, 0_{3\times 3}, 0_{3\times 3}$$
$$w^l_{3,:,:,:} = 0_{3\times 3}, 0_{3\times 3}, \boldsymbol{\iota}_{3\times 3}$$
$$w^L_{1,:,:,:} = 0_{3\times 3}, 0_{3\times 3}, \boldsymbol{\iota}_{3\times 3}$$

for $2 \le l \le L-1$, where $u^l_{i,j} \in \mathbb{R}^{3\times 3}$ is randomly sampled $u^l_{i,j} \overset{iid}{\sim} \mathcal{N}(0, \rho^2)$ for $1 \le l \le L-1$. Here, $0_{3\times 3}$ is the 3 by 3 matrix with all zero entries, $C_L > 0$ is a scalar growing as a function of L at a rate satisfying $L^2/C_L \to 0$, and $\rho > 0$ is a fixed variance parameter. Initialize the biases as follows:

$$(b^1_1, b^1_2, b^1_3) = (0, v^1, C_L)$$
$$(b^l_1, b^l_2, b^l_3) = (0, v^l, 0)$$
$$b^L = -C_L$$

for $2 \le l \le L-1$, where $v^l \in \mathbb{R}$ is randomly sampled $v^l \overset{iid}{\sim} \mathcal{N}(0, C_L^2\beta^2)$ for $1 \le l \le L-1$. Here, $\beta > 0$ is a fixed variance parameter.

4.2. Convergence in infinite-depth limit

We now analyze the convergence of the infinitely deep CNN.

Theorem 4 (Scaled NTK at initialization)**.** *Suppose f_θ^{L+1} is initialized as in Section 4.1. For any $x, x' \in \mathbb{R}_+^{d\times d}$,*

$$S\left(\tilde{\Theta}^L_0(x, x')\right) \xrightarrow{p} \tilde{\Theta}^\infty(x, x')$$

as $L \to \infty$, where

$$\tilde{\Theta}^\infty(x,x') = \frac{1}{d^2}\sum_{s=1}^{3}\sum_{u=1}^{3}\Bigg(p_d(s,u) + d^2 S(x_{\psi_{s,u}})S(x'_{\psi_{s,u}}) + \sum_{i,j\in\psi_{s,u}}\sum_{i',j'\in\psi_{s,u}} \mathbb{E}[\sigma(g(\phi_{i,j}(x)))\sigma(g(\phi_{i',j'}(x')))]\Bigg),$$

$g \sim \mathcal{GP}(0, \rho^2\langle x, x'\rangle + \beta^2)$, $\psi_{s,u}$ is the set of coordinate which satisfy $x_{\psi_{s,u}} = [x]_{s-1:d+s-2,u-1:d+u-2}$ and $x_{pq} = 0, \phi_{mn}(x) = 0$ if the index is out of bounds, and

$$p_d(s,u) = \begin{cases} d^4 & (s,u) = (2,2), \\ d^2(d-1)^2 & |s-u| = 1, \\ (d-1)^4 & \textit{else}. \end{cases}$$

Theorem 5 (Invariance of scaled NTK)**.** *Let $T > 0$. Suppose $\int_0^T \left\|\delta|_{f_\theta^L}\right\|_{p^{in}} dt$ is stochastically bounded as $L \to \infty$. Then, for any $x, x' \in \mathbb{R}_+^{d\times d}$,*

$$S\left(\tilde{\Theta}^L_t(x, x')\right) \xrightarrow{p} \tilde{\Theta}^\infty(x, x')$$

uniformly for $t \in [0, T]$ as $L \to \infty$.

Again, define the kernel regression predictor as

$$f_{\text{ntk}}(x) = \left(\tilde{\Theta}^\infty(x, x_1), \ldots, \tilde{\Theta}^\infty(x, x_N)\right) K^{-1} f^\star(X),$$

where $K_{i,j} = \tilde{\Theta}^\infty(x_i, x_j)$ and $[f^\star(X)]_i = f^\star(x_i)$ for $i, j \in \{1, \ldots, N\}$.

Theorem 6 (Equivalence between deep CNN and kernel regression)**.** *Let $\mathcal{L}(f) = \frac{1}{2}\|f - f^\star\|^2_{p^{in}}$. Let $\tilde{\Theta}^\infty$ be positive definite. If f_t follows the kernel gradient flow with respect to $\tilde{\Theta}^\infty$, then for any $x \in \mathbb{R}_+^{d\times d}$,*

$$\lim_{t\to\infty} f_t(x) = f_{\text{ntk}}(x).$$

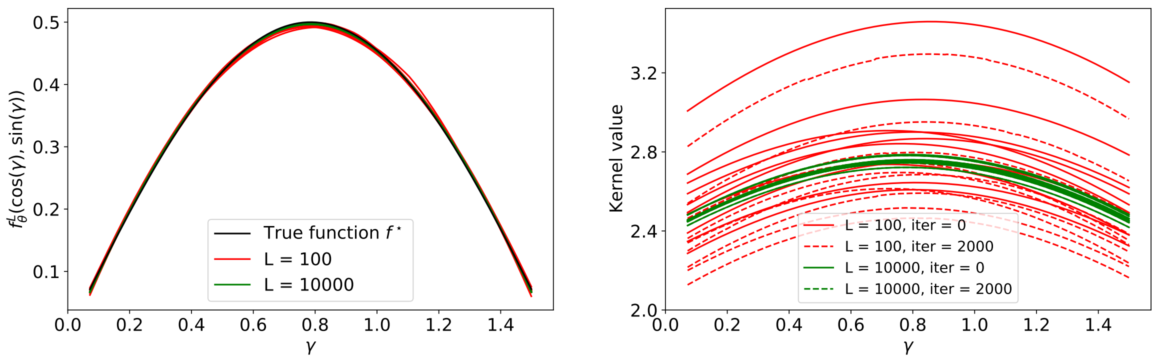
Generalizations. At the expense of slight notational complications, we can generalize our results as follows. We assumed the convolutional filter size is 3×3, but we can use larger filters by assigning a symbol for the filter size and managing the summation indices with care. We assumed the number of input channels and output scalar dimension are 1, but we can have c input channels and k outputs, i.e., $f_\theta^{L+1} \colon \mathbb{R}_+^{c\times d\times d} \to \mathbb{R}^k$, by letting the intermediate layers have $c + 1 + k$ channels. In fact, Section 5.2 presents a 10-class classification of MNIST with a deep CNN using $1 + 1 + 10 = 12$ channels in intermediate layers.

5. Experiments

In this section, we experimentally demonstrate the invariance of the scaled NTK and the trainability of deep neural networks. The code is provided as supplementary material.

5.1. Convergence of the scaled NTK

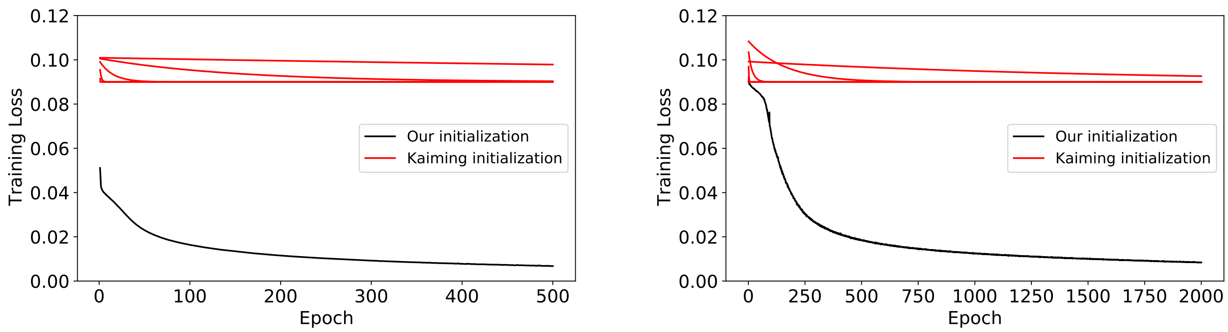
Our first experiment, inspired by Jacot et al. (2018), trains L-layer MLPs on 2-dimensional inputs and shows that the network and its scaled NTK converges as the depth L increases. For $L = 100$ and $L = 10000$, we initialize 10 independent MLP instances and train them to approximate $f^\star(x_1, x_2) = x_1 x_2$ using the quadratic loss. To verify that the networks are indeed successfully trained, we compare the trained MLP against the true function $f^\star$ in Figure 3(a). We then plot the scaled NTK $\tilde{\Theta}^L(x_0, x)$ for fixed $x_0 = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$ and $x = (\cos(\gamma), \sin(\gamma))$ for $0 < \gamma < \pi/2$ in Figure 3(b). The kernels are plotted at initialization ($t = 0$), and after 2000 iterations of gradient descent.



(a) True function and trained MLP after 2000 iterations.

(b) Scaled NTK values

Figure 3. Depth L MLPs learning a toy function with 2D inputs as described in Section 5.1, (Left) Trained deep MLP approximates the true function well, i.e., training succeeds. (Right) Kernel values evaluated at initialization and after training with 10 independent initialization-training trials each for $L = 100$ and $L = 10000$. As L grows, initialization becomes less random, as Theorem 1 predicts, and the kernel changes less throughout training, as Theorem 2 predicts.



(a) Depth 1000 MLP

(b) Depth 1000 CNN

Figure 4. Depth 1000 MLPs and CNNs with MNIST are trainable with our proposed initialization but not with the standard Kaiming He initialization. For Kaiming initialization, we show trials with learning rates 1×10^{-5}, 1×10^{-4}, 1×10^{-3}, 0.01, 0.1, and 1.

Task	Architecture	Depth	C_L	ρ	β	Learning rate	Epochs	Training loss	Test accuracy
10-class	MLP	4000	4	1	1	1×10^{-5}	1500	0.0055	97.48%
10-class	CNN	4000	4	$1/\sqrt{3}$	1	1×10^{-5}	2000	0.014578	94.02%
Binary	CNN	20000	20	$1/\sqrt{3}$	1	1×10^{-8}	1000	0.031257	98.87%

Table 1. Very deep MLP and CNNs trained with MNIST. As the theory predicts, the very deep networks are trainable.

5.2. Trainability of the deep narrow neural network

Next, we demonstrate the empirical trainability of the deep narrow networks on the MNIST dataset.

Very deep MLP. We train L-layer MLPs with $d_{\text{in}} = 784$ and $d_{\text{out}} = 10$ using the quadratic loss with one-hot vectors as targets. To establish a point of comparison, we attempt to train a 1000-layer MLP with the typical Kaiming He uniform initialization (He et al., 2015). We tuned the learning rate via a grid search from 0.00001 to 1.0, but the network was untrainable, as one would expect based on the prior findings of (He & Sun, 2015; Srivastava et al., 2015; Huang et al., 2020). In contrast, when we use the initialization defined in Section 3.1, the deep MLP was trainable. Figure 4(a) reports these results. To push the depth, we also trained a 4000-layer MLP and report the results in Table 1. To the best of our knowledge, this 4000-layer MLP holds the record for the deepest trained MLP.

Very deep CNN. We train L-layer CNNs using the quadratic loss with one-hot probability vectors as targets. To reduce the computational cost, we insert a 4×4 average pool before the first layer to reduce the MNIST input size to 7×7. As in the MLP experiment, we attempt to train a 1000-layer CNN with the typical Kaiming He uniform initialization, but the network was untrainable. In contrast, when we use the initialization defined in Section 4.1, the

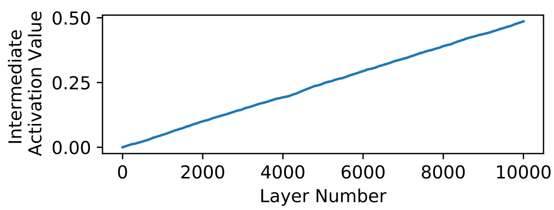


Figure 5. Intermediate activation values of the rightmost neurons of each layer after training. (The effect of the bias C_L is compensated for to help visualize the change.)

deep CNN was trainable. Figure 4(b) reports these results. To push the depth, we also trained a 4000-layer CNN and report the results in Table 1. To further push the depth, we simplify the problem to binary classification between digits 0 and 1 and use values 0 and 1 as targets. We then trained a 20000-layer CNN and report the results in Table 1. This CNN surpasses the 10000-depth CNN of (Xiao et al., 2018) and, to the best of our knowledge, holds the record of the deepest trained CNN.

5.3. Accumulation of the layer-wise effect

To observe the accumulation of the output value throughout the depth, we plot the intermediate activation values of the rightmost neurons of each layer from the 10000-layer MLP trained in Section 5.1. Precisely, we plot $(\sigma(f^{(l)}(x)))_{d_{\text{in}}+d_{\text{out}}+1} - C_L$ for $1 \leq l \leq L-1$ and $(f^{(L)}(x))_{d_{\text{in}}+d_{\text{out}}+1}$. Figure 5 shows that the target output value is achieved through the accumulation of the effect of 10000-layers.

6. Conclusion

This work presents an NTK analysis of a deep narrow MLP and CNN in the infinite-depth limit and establishes the first trainability guarantee on deep narrow neural networks. Our results serve as a demonstration that infinitely deep neural networks can be made provably trainable using the right initialization, just as the infinitely wide counterparts are. However, our results do have the following limitations. First, while our proposed initialization is straightforwardly implementable, it is far from the initializations used in practice. Second, our results do not indicate any benefit of using deep neural networks compared to wide neural networks. Further investigating the trainability of overparameterized deep neural networks to address these questions would be an interesting direction of future work.

Acknowledgements

JL, JYC, and EKR were supported by the National Research Foundation of Korea (NRF) Grant funded by the Korean Government (MSIP) [No. 2022R1C1C1010010], the National Research Foundation of Korea (NRF) Grant funded by the Korean Government (MSIP) [No. 2022R1A5A6000840], and the Samsung Science and Technology Foundation (Project Number SSTF-BA2101-02). AN was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2021R1F1A1059567). We thank Jisun Park and TaeHo Yoon for providing valuable feedback. We also thank anonymous reviewers for giving thoughtful comments.

References

Adlam, B., Lee, J., Xiao, L., Pennington, J., and Snoek, J. Exploring the uncertainty properties of neural networks' implicit priors in the infinite-width limit. *International Conference on Learning Representations*, 2021.

Allen-Zhu, Z., Li, Y., and Song, Z. A convergence theory for deep learning via over-parameterization. *International Conference on Machine Learning*, 2019.

Arora, S., Cohen, N., Golowich, N., and Hu, W. A convergence analysis of gradient descent for deep linear neural networks. *International Conference on Learning Representations*, 2019a.

Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R. R., and Wang, R. On exact computation with an infinitely wide neural net. *Neural Information Processing Systems*, 2019b.

Barron, A. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993.

Bietti, A. and Mairal, J. On the inductive bias of neural tangent kernels. *Neural Information Processing Systems*, 2019.

Chen, R. T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural ordinary differential equations. *Neural Information Processing Systems*, 2018.

Chen, Z., Cao, Y., Gu, Q., and Zhang, T. A generalized neural tangent kernel analysis for two-layer neural networks. *Neural Information Processing Systems*, 2020.

Chizat, L. and Bach, F. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Neural Information Processing Systems*, 2018.

Chizat, L., Oyallon, E., and Bach, F. R. On lazy training in differentiable programming. *Neural Information Processing Systems*, 2019.

Cho, Y. and Saul, L. Kernel methods for deep learning. *Neural Information Processing Systems*, 2009.

Cybenko, G. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.

Ding, Z., Chen, S., Li, Q., and Wright, S. On the global convergence of gradient descent for multi-layer ResNets in the mean-field regime. *arXiv:2110.02926*, 2021.

Ding, Z., Chen, S., Li, Q., and Wright, S. J. Overparameterization of deep resnet: Zero loss and mean-field analysis. *Journal of Machine Learning Research*, 23(48):1–65, 2022.

Dragomir, S. *Some Gronwall Type Inequalities and Applications*. Nova Science Publishers, 2003.

Du, S. S., Lee, J., Li, H., Wang, L., and Zhai, X. Gradient descent finds global minima of deep neural networks. *International Conference on Machine Learning*, 2019a.

Du, S. S., Zhai, X., Póczos, B., and Singh, A. Gradient descent provably optimizes over-parameterized neural networks. *International Conference on Learning Representations*, 2019b.

Funahashi, K.-I. On the approximate realization of continuous mappings by neural networks. *Neural Networks*, 2 (3):183–192, 1989.

Geiger, M., Spigler, S., Jacot, A., and Wyart, M. Disentangling feature and lazy training in deep neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(11):113301, 2020.

Hanin, B. Which neural net architectures give rise to exploding and vanishing gradients? *Neural Information Processing Systems*, 2018.

Hanin, B. Universal function approximation by deep neural nets with bounded width and ReLU activations. *Mathematics*, 7(10):992, 2019.

Hanin, B. and Nica, M. Finite depth and width corrections to the neural tangent kernel. *International Conference on Learning Representations*, 2020.

Hanin, B. and Sellke, M. Approximating continuous functions by ReLU nets of minimal width. *arXiv:1710.11278v2*, 2017.

He, K. and Sun, J. Convolutional neural networks at constrained time cost. *Computer Vision and Pattern Recognition*, 2015.

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. *International Conference on Computer Vision*, 2015.

Hornik, K. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, 1991.

Huang, K., Wang, Y., Tao, M., and Zhao, T. Why do deep residual networks generalize better than deep feedforward networks? — A neural tangent kernel perspective. *Neural Information Processing Systems*, 2020.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *Neural Information Processing Systems*, 2018.

Ji, Z., Telgarsky, M., and Xian, R. Neural tangent kernels, transportation mappings, and universal approximation. *International Conference on Learning Representations*, 2020.

Johnson, J. Deep, skinny neural networks are not universal approximators. *International Conference on Learning Representations*, 2019.

Jones, L. K. A simple lemma on greedy approximation in hilbert space and convergence rates for projection pursuit regression and neural network training. *The Annals of Statistics*, 20(1):608–613, 1992.

Kidger, P. and Lyons, T. Universal approximation with deep narrow networks. *Conference on Learning Theory*, 2020.

Lee, J., Bahri, Y., Novak, R., Schoenholz, S. S., Pennington, J., and Sohl-Dickstein, J. Deep neural networks as gaussian processes. *International Conference on Learning Representations*, 2018.

Lee, J., Xiao, L., Schoenholz, S., Bahri, Y., Novak, R., Sohl-Dickstein, J., and Pennington, J. Wide neural networks of any depth evolve as linear models under gradient descent. *Neural Information Processing Systems*, 2019.

Lee, J., Schoenholz, S., Pennington, J., Adlam, B., Xiao, L., Novak, R., and Sohl-Dickstein, J. Finite versus infinite neural networks: An empirical study. *Neural Information Processing Systems*, 2020.

Leshno, M., Lin, V. Y., Pinkus, A., and Schocken, S. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, 1993.

Li, Y. and Liang, Y. Learning overparameterized neural networks via stochastic gradient descent on structured data. *Neural Information Processing Systems*, 2018.

Lin, H. and Jegelka, S. ResNet with one-neuron hidden layers is a universal approximator. *Neural Information Processing Systems*, 2018.

Lu, J., Shen, Z., Yang, H., and Zhang, S. Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*, 53(5):5465–5506, 2021.

Lu, Y., Ma, C., Lu, Y., Lu, J., and Ying, L. A mean field analysis of deep resnet and beyond: Towards provably optimization via overparameterization from depth. *International Conference on Machine Learning*, 2020.

Lu, Z., Pu, H., Wang, F., Hu, Z., and Wang, L. The expressive power of neural networks: A view from the width. *Neural Information Processing Systems*, 2017.

Matthews, A. G. d. G., Hron, J., Rowland, M., Turner, R. E., and Ghahramani, Z. Gaussian process behaviour in wide deep neural networks. *International Conference on Learning Representations*, 2018.

Mei, S., Montanari, A., and Nguyen, P.-M. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33): E7665–E7671, 2018.

Neal, R. M. Priors for infinite networks. Technical Report crg-tr-94-1, University of Toronto, 1994.

Neal, R. M. *Bayesian Learning for Neural Networks*. University of Toronto, 1996.

Nguyen, P.-M. and Pham, H. T. A rigorous framework for the mean field limit of multilayer neural networks. *arXiv:2001.11443*, 2021.

Nguyen, T., Novak, R., Xiao, L., and Lee, J. Dataset distillation with infinitely wide convolutional networks. *Neural Information Processing Systems*, 2021.

Novak, R., Xiao, L., Bahri, Y., Lee, J., Yang, G., Hron, J., Abolafia, D. A., Pennington, J., and Sohl-dickstein, J. Bayesian deep convolutional networks with many channels are gaussian processes. *International Conference on Learning Representations*, 2019.

Novak, R., Xiao, L., Hron, J., Lee, J., Alemi, A. A., Sohl-Dickstein, J., and Schoenholz, S. S. Neural tangents: Fast and easy infinite neural networks in python. *International Conference on Learning Representations*, 2020.

Park, S., Yun, C., Lee, J., and Shin, J. Minimum width for universal approximation. *International Conference on Learning Representations*, 2021.

Peluchetti, S. and Favaro, S. Infinitely deep neural networks as diffusion processes. *International Conference on Artificial Intelligence and Statistics*, 2020.

Peluchetti, S. and Favaro, S. Doubly infinite residual neural networks: a diffusion process approach. *Journal of Machine Learning Research*, 2021.

Pennington, J., Schoenholz, S., and Ganguli, S. Resurrecting the sigmoid in deep learning through dynamical isometry: Theory and practice. *Neural Information Processing Systems*, 2017.

Pham, H. T. and Nguyen, P.-M. Global convergence of three-layer neural networks in the nean field regime. *International Conference on Learning Representations*, 2021.

Pinkus, A. Approximation theory of the MLP model in neural networks. *Acta Numerica*, 8:143–195, 1999.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. *Neural Information Processing Systems*, 2007.

Rahimi, A. and Recht, B. Uniform approximation of functions with random bases. *Allerton Conference on Communication, Control, and Computing*, 2008a.

Rahimi, A. and Recht, B. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. *Neural Information Processing Systems*, 2008b.

Rasmussen, C. E. Gaussian Processes in Machine Learning. In Bousquet, O., von Luxburg, U., and Rätsch, G. (eds.), *Advanced Lectures on Machine Learning*, pp. 63–71. Springer, 2004.

Rasmussen, C. E. and Williams, C. K. I. *Gaussian Processes for Machine Learning*. The MIT Press, 2005.

Rotskoff, G. and Vanden-Eijnden, E. Parameters as interacting particles: Long time convergence and asymptotic error scaling of neural networks. *Neural Information Processing Systems*, 2018.

Rotskoff, G., Jelassi, S., Bruna, J., and Vanden-Eijnden, E. Neuron birth-death dynamics accelerates gradient descent and converges asymptotically. *International Conference on Machine Learning*, 2019.

Shamir, O. Exponential convergence time of gradient descent for one-dimensional deep linear neural networks. *Conference on Learning Theory*, 2019.

Sirignano, J. and Spiliopoulos, K. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020a.

Sirignano, J. and Spiliopoulos, K. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 130(3):1820–1852, 2020b.

Sohl-Dickstein, J., Novak, R., Schoenholz, S. S., and Lee, J. On the infinite width limit of neural networks with a standard parameterization. *arXiv:2001.07301*, 2020.

Soltanolkotabi, M., Javanmard, A., and Lee, J. D. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2019.

Srivastava, R. K., Greff, K., and Schmidhuber, J. Highway networks. *arXiv:1505.00387*, 2015.

Tabuada, P. and Gharesifard, B. Universal approximation power of deep residual neural networks via nonlinear control theory. *International Conference on Learning Representations*, 2021.

Telgarsky, M. Benefits of depth in neural networks. *Conference on Learning Theory*, 2016.

Williams, C. Computing with infinite networks. *Neural Information Processing Systems*, 1997.

Woodworth, B., Gunasekar, S., Lee, J. D., Moroshko, E., Savarese, P., Golan, I., Soudry, D., and Srebro, N. Kernel and rich regimes in overparametrized models. *Conference on Learning Theory*, 2020.

Xiao, L., Bahri, Y., Sohl-Dickstein, J., Schoenholz, S., and Pennington, J. Dynamical isometry and a mean field theory of CNNs: How to train 10,000-layer vanilla convolutional neural networks. *International Conference on Machine Learning*, 2018.

Yang, G. Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation. *arXiv:1902.04760v3*, 2019.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. *International Conference on Learning Representations*, 2017.

Zou, D., Cao, Y., Zhou, D., and Gu, Q. Gradient descent optimizes over-parameterized deep ReLU networks. *Machine Learning*, 109(3):467–492, 2020.

A. Preliminaries

A.1. Grönwall's lemma

The following is a generalized version of Grönwall's lemma, which is essential to prove the invariance of kernel.

Lemma 7 (Dragomir 2003)**.** *Let the x, a, b and k be continuous and nonnegative functions on $J = [\alpha, \beta]$. Let $n \geq 2$ be a positive integer $\frac{a}{b}$ be a nondecreasing function. If*

$$z(t) \leq a(t) + b(t) \int_{\alpha}^{t} k(s) z^n(s)\, ds,\ \ t \in J$$

then

$$z(t) \leq a(t) \left\{ 1 - (n-1) \int_{\alpha}^{t} k(s) b(s) a^{n-1}(s)\, ds \right\}^{\frac{1}{1-n}},$$

where $\alpha \leq t \leq \beta_n$, for β_n defined by

$$\beta_n = \sup \left\{ t \in J : (n-1) \int_{\alpha}^{t} kba^{n-1}\, ds < 1 \right\}.$$

B. Proof of Theorems of NTK for MLP

B.1. Proof of Theorem 1

The following lemma shows the recursive relation of the neural tangent kernel (without the scaling factor).

Lemma 8. *For $t \geq 0$, $l \geq 2$ and $x, x' \in \mathbb{R}_+^{d_{in}}$,*

$$\begin{aligned}\Theta_t^l(x, x') &= W^l(t) diag(\dot\sigma(f^{l-1}_{\theta^{(l-1)}}(x))) \Theta_t^{l-1}(x, x') diag(\dot\sigma(f^{l-1}_{\theta^{(l-1)}(t)}(x'))) (W^l(t))^\intercal \\ &\quad + (\sigma(f^{l-1}_{\theta^{(l-1)}(t)}(x'))^\intercal \sigma(f^{l-1}_{\theta^{(l-1)}(t)}(x)) + 1) I_{d_{\text{in}}+d_{\text{out}}+1},\end{aligned}$$

where $\Theta_t^1(x, x') = (x^\intercal x' + 1) I_{d_{\text{in}}+d_{\text{out}}+1}$ if $l = 1$.

Proof of Lemma 8.

$$\begin{aligned}\Theta_t^l(x, x') &= \partial_{\theta^{(l)}} f^l_{\theta^{(l)}(t)}(x) \left(\partial_{\theta^{(l)}} f^l_{\theta^{(l)}(t)}(x')\right)^\intercal \\ &= \sum_{\theta_i \in \theta^{(l)}} \partial_{\theta_i} f^l_{\theta^{(l)}(t)}(x) \left(\partial_{\theta_i} f^l_{\theta^{(l)}(t)}(x')\right)^\intercal \\ &= \sum_{\theta_i \in \theta^{(l-1)}} \partial_{\theta_i} f^l_{\theta^{(l)}(t)}(x) \left(\partial_{\theta_i} f^l_{\theta^{(l)}(t)}(x')\right)^\intercal \\ &\quad + \sum_{W^l_{ij}} \partial_{W^l_{ij}} f^l_{\theta^{(l)}(t)}(x) \left(\partial_{W^l_{ij}} f^l_{\theta^{(l)}(t)}(x')\right)^\intercal + \sum_{b^l_i} \partial_{b^l_i} f^l_{\theta^{(l)}(t)}(x) \left(\partial_{b^l_i} f^l_{\theta^{(l)}(t)}(x')\right)^\intercal.\end{aligned}$$

The last two terms of RHS is

$$\begin{aligned}&\sum_{W^l_{ij}} \left(\partial_{W^l_{ij}} f^l_{\theta^{(l)}(t)}(x)\right)_k \left(\partial_{W^l_{ij}} f^l_{\theta^{(l)}(t)}(x')\right)_{k'} + \sum_{b^l_i} \left(\partial_{b^l_i} f^l_{\theta^{(l)}(t)}(x)\right)_k \left(\partial_{b^l_i} f^l_{\theta^{(l)}(t)}(x')\right)_{k'} \\ &= (\sigma(f^{l-1}_{\theta^{(l-1)}(t)}(x))^\intercal \sigma(f^{l-1}_{\theta^{(l-1)}(t)}(x')) + 1)\delta_{k,k'},\end{aligned}$$

where $(\cdot)_k$ denote the k-th component of the vector, and $\delta_{k,k'} = 1$ if $k = k'$ and 0 otherwise. Also, since

$$\partial_{\theta^{(l-1)}} f^l_{\theta^{(l)}(t)}(x) = W^l(t) \text{diag}(\dot\sigma(f^{l-1}_{\theta^{(l-1)}(t)}(x))) \partial_{\theta^{(l-1)}} f^{l-1}_{\theta^{(l-1)}(t)}(x),$$

the first term of RHS is

$$\sum_{\theta_i \in \theta^{(l-1)}} \partial_{\theta_i} f^l_{\theta^{(l)}(t)}(x) \left(\partial_{\theta_i} f^l_{\theta^{(l)}(t)}(x')\right)^\intercal = \partial_{\theta^{(l-1)}} f^l_{\theta^{(l)}(t)}(x) \left(\partial_{\theta^{(l-1)}} f^l_{\theta^{(l)}(t)}(x')\right)^\intercal$$
$$= W^l(t)\text{diag}(\dot\sigma(f^{l-1}_{\theta^{(l-1)}(t)}(x)))\Theta^{l-1}_t(x,x')\text{diag}(\dot\sigma(f^{l-1}_{\theta^{(l-1)}(t)}(x')))(W^l(t))^\intercal.$$

Thus, we have a recursive relation for kernel

$$\Theta^l_t(x,x') = W^l(t)\text{diag}(\dot\sigma(f^{l-1}_{\theta^{(l-1)}}(x)))\Theta^{l-1}_t(x,x')\text{diag}(\dot\sigma(f^{l-1}_{\theta^{(l-1)}(t)}(x')))(W^l(t))^\intercal$$
$$+ (\sigma(f^{l-1}_{\theta^{(l-1)}(t)}(x))^\intercal \sigma(f^{l-1}_{\theta^{(l-1)}(t)}(x')) + 1)I_{d_{\text{in}}+d_{\text{out}}+1}.$$

for $l \geq 2$, where $\Theta^1_t(x,x') = (x^\intercal x' + 1)I_{d_{\text{in}}+d_{\text{out}}+1}$ if $l = 1$. □

For notational simplicity, we let

$$\mathfrak{W}^k_j(x,t) = W^j(t)\text{diag}\left(\dot\sigma\left(f^{j-1}_{\theta^{(j-1)}(t)}(x)\right)\right)W^{j-1}(t)\text{diag}\left(\dot\sigma\left(f^{j-2}_{\theta^{(j-2)}(t)}(x)\right)\right)\cdots W^k(t)\text{diag}\left(\dot\sigma\left(f^{k-1}_{\theta^{(k-1)}(t)}(x)\right)\right)$$

for $2 \leq k, j \leq L$, where $\mathfrak{W}^k_j(x,t) = 1$ if $k > j$. By applying the above recursive relation inductively, we get the following corollary.

Corollary 9. *For any $x, x' \in \mathbb{R}^{d_{in}}_+$, and $t \geq 0$,*

$$\tilde\Theta^L_t(x,x') = \frac{1}{L}\sum_{l=0}^{L-1}\left[\left(\frac{1}{C_L}\sigma\left(f^l_{\theta^{(l)}(t)}(x)\right)^\intercal \frac{1}{C_L}\sigma\left(f^l_{\theta^{(l)}(t)}(x')\right)\right)\mathfrak{W}^{l+2}_L(x,t)\mathfrak{W}^{l+2}_L(x',t)^\intercal\right]$$
$$+ \frac{1}{LC_L^2}\sum_{l=0}^{L-1}\left[\mathfrak{W}^{l+2}_L(x,t)\mathfrak{W}^{l+2}_L(x',t)^\intercal\right],$$

where we define $\sigma(f^0(x)) = x$ for convenience.

Now, we are ready to prove the Theorem 1.

Proof of Theorem 1. At initialization, we can simplify the kernel further. Since we assumed positive input, we have

$$W^l(0)\text{diag}(\dot\sigma(f^{l-1}_{\theta^{(l-1)}(0)}(x))) = \begin{bmatrix} I_{d_{\text{in}}} & 0_{d_{\text{in}}\times 1} & 0_{d_{\text{in}}\times d_{\text{out}}} \\ u^l & 0_{1\times 1} & 0_{1\times d_{\text{out}}} \\ 0_{d_{\text{out}}\times d_{\text{in}}} & 0_{d_{\text{out}}\times 1} & I_{d_{\text{out}}} \end{bmatrix} \quad \text{for } 1 \leq l \leq L-1,$$
$$W^L(0)\text{diag}(\dot\sigma(f^{L-1}_{\theta^{(L-1)}(0)}(x))) = \begin{bmatrix} 0_{d_{\text{out}}\times d_{\text{in}}} & 0_{d_{\text{out}}\times 1} & I_{d_{\text{out}}} \end{bmatrix}.$$

This implies

$$\mathfrak{W}^k_l(x,0) = W^l(0)\text{diag}(\dot\sigma(f^{l-1}_{\theta^{(l-1)}(0)})) = W^l(0) \tag{1}$$

for $l \geq k$. Also, our initialization implies

$$f^1_{\theta^{(1)}(0)}(x) = \begin{bmatrix} C_L x \\ (u^1)^\intercal x + v^1 \\ C_L \mathbb{1}_{d_{\text{out}}} \end{bmatrix},\ f^l_{\theta^{(l)}(0)}(x) = \begin{bmatrix} C_L x \\ C_L(u^l)^\intercal x + v^l \\ C_L \mathbb{1}_{d_{\text{out}}} \end{bmatrix},\ f^L_{\theta^{(L)}(0)} = 0_{d_{\text{in}}\times 1} \tag{2}$$

for $2 \leq l \leq L-1$. Thus, the scaled NTK is given by

$$\tilde\Theta^L_0(x,x') = \frac{1}{LC_L^2}\left(x^\intercal x' + 1 + \sum_{j=0}^{L-1}(C_L^2 x^\intercal x' + \sigma(C_L(u^j)^\intercal x + v^j)\sigma(C_L(u^j)^\intercal x' + v^j) + C_L^2 + 1)\right)I_{d_{\text{out}}}.$$

Let $g^j(x) = u^j x + v^j/C_L$, then $\{g^j(x)\}_{j=0}^{L-1}$ are i.i.d. Gaussian process, where $g^j \sim \mathcal{GP}(0, \frac{\rho^2}{d_{in}} x^\intercal x' + \beta^2)$. This is because

$$\mathbb{E}_{u^j,v^j}[g^j] = 0$$

and

$$\begin{aligned}
\mathbb{E}_{u^j,v^j}[g^j(x)g^j(x')] &= \mathbb{E}_{u^j,v^j}\left[\left((u^j)^\intercal x + v^j/C_L\right)\left((u^j)^\intercal x' + v^j/C_L\right)\right] \\
&= \mathbb{E}_{u^j}\left[\left((u^j)^\intercal x\right)\left((u^j)^\intercal x'\right)\right] + \beta^2 \\
&= \frac{\rho^2}{d} x^\intercal x' + \beta^2.
\end{aligned}$$

Hence, we have

$$\tilde{\Theta}_0^L(x,x') \xrightarrow{p} \left(x^\intercal x' + 1 + \mathbb{E}_{g\sim\mathcal{GP}(0,\frac{\rho^2}{d_{\text{in}}}x^\intercal x' + \beta^2)}[\sigma(g(x))\sigma(g(x'))]\right) I_{d_{\text{out}}}$$

as $L \to \infty$ by the law of large number, which concludes the proof. □

B.2. Proof of Theorem 2

For proving our scaled NTK stays constant during training, we need to show that intermediate weight and pre-activation values are effectively unchanged. In this paper, we also consider Lyapunov functions similar to (Jacot et al., 2018); however, we need more delicate Lyapunov function to handle infinite depth network.

For the sake of simplicity, we set the scaling factor of gradient flow by $\frac{1}{(L-1)C_L^2}$ throughout the proof. Also, without loss of generality, we assume the norms of data points $\|x_i\|$ are bounded by 1 for all $1 \le i \le N$, and we further assume that all other inputs $x \notin \{x_1, \ldots, x_N\}$ also have bounded norm, i.e., $\|x\| \le 1$.

We define Lyapunov functions to track intermediate weights and pre-activation values of the network. For $1 \le j \le 8$ and $C_L > 0$ satisfying $L^2/C_L \to 0$,

$$\begin{aligned}
\Phi_{L,j}(t) &= \frac{1}{(L-1)C_L^j}\sum_{l=1}^{L-1}\left(\left\|f^l_{\theta^{(l)}(0)}(x)\right\|_{p^{in}} + \left\|f^l_{\theta^{(l)}(t)}(x) - f^l_{\theta^{(l)}(0)}(x)\right\|_{p^{in}}\right)^j \\
\Psi_{L,2}(t) &= \frac{1}{(L-1)^2}\sum_{l=2}^{L}\frac{L-1}{l-1}\sum_{i=2}^{l}\left(\left\|\mathfrak{W}^i_l(x,0)\right\|_{p^{in}} + \left\|\mathfrak{W}^i_l(x,t) - \mathfrak{W}^i_l(x,0)\right\|_{p^{in}}\right)^2 \\
\Psi_{L,8}(t) &= \frac{1}{L-1}\sum_{l=2}^{L}\left(\left\|\mathfrak{W}^l_L(x,0)\right\|_{p^{in}} + \left\|\mathfrak{W}^l_L(x,t) - \mathfrak{W}^l_L(x,0)\right\|_{p^{in}}\right)^8,
\end{aligned}$$

where $\Phi_{L,0} = 1$.

At initialization ($t = 0$), we have

$$\begin{aligned}
\Phi_{L,j}(0) &= \frac{1}{(L-1)C_L^j}\sum_{l=1}^{L-1}\left(\left\|f^l_{\theta^{(l)}(0)}(x)\right\|_{p^{in}}\right)^j \\
\Psi_{L,2}(0) &= \frac{1}{(L-1)^2}\sum_{l=2}^{L}\frac{L-1}{l-1}\sum_{i=2}^{l}\left(\left\|\mathfrak{W}^i_l(x,0)\right\|_{p^{in}}\right)^2 \\
\Psi_{L,8}(0) &= \frac{1}{L-1}\sum_{l=2}^{L}\left(\left\|\mathfrak{W}^l_L(x,0)\right\|_{p^{in}}\right)^8.
\end{aligned}$$

From (1) and (2), imply that $\Phi_{L,j}(0)$ and $\Psi_{L,2}(0)$ converges (in probability) to constant values $\Phi_{\infty,j}(0)$ and $\Psi_{\infty,2}(0)$ by the law of large number, respectively. Furthermore, $\Psi_{L,8}(0)$ also converges (in probability) to constant value $\Psi_{\infty,8}(0)$ since $\mathfrak{W}^l_L(x,0) = W^L(0)$. Then, the following proposition implies Lyapunov function remains constant during training.

Proposition 1. *For $1 \leq j \leq 8$,*

$$\Phi_{L,j}(t) \xrightarrow{p} \Phi_{\infty,j}(0), \quad \Psi_{L,2}(t) \xrightarrow{p} \Psi_{\infty,2}(0), \quad \Psi_{L,8}(t) \xrightarrow{p} \Psi_{\infty,8}(0)$$

uniformly for $t \in [0, T]$ as $L \to \infty$.

The following lemma is a key step to prove the proposition which allows us to apply Grönwall's Lemma. For the sake of simplicity, we let $\delta_t^L = \partial_f C = \delta|_{f_\theta^L}$.

Lemma 10. *For $t \geq 0$, if $f^l_{\theta^{(l)}(t)}(x)$ is element-wise nonzero for $1 \leq l \leq L-1$ and $x \in \{x_1, \ldots, x_N\}$, then*

$$\begin{aligned}
\partial_t \Phi_{L,j}(t) &\leq \frac{2jN^{3/2}(L-1)}{C_L} \Phi_{L,j-1}(t)\Phi_{L,8}(t)\Psi_{L,2}(t)\Psi_{L,8}(t) \left\|\delta_t^L\right\|_{p^{in}} \\
\partial_t \Psi_{L,2}(t) &\leq \frac{2N(L-1)}{C_L} \Phi_{L,4}(t) \left(\Psi_{L,2}(t)\right)^2 (t) \Psi_{L,8}(t) \left\|\delta_t^L\right\|_{p^{in}} \\
\partial_t \Psi_{L,8}(t) &\leq \frac{8N(L-1)}{C_L} \Phi_{L,4}(t)\Psi_{L,2}(t) \left(\Psi_{L,8}(t)\right)^2 \left\|\delta_t^L\right\|_{p^{in}},
\end{aligned}$$

for $1 \leq j \leq 8$.

The proof of Lemma 10 is given in Section F.

Proof of Proposition 1. In order to apply Lemma 10 for all $t \in [0, T]$, we need to show that $f^l_{\theta^{(l)}(t)}(x)$ is element-wise nonzero for all $t \in [0, T]$, $1 \leq l \leq L-1$, and $x \in \{x_1, \ldots, x_N\}$. This is indeed true when L is large enough, which we proved in Section C. From Lemma 10, we get

$$\partial_t \left(\sum_{j=1}^{8} \Phi_{L,j}(t) + \Psi_{L,2}(t) + \Psi_{L,8}(t) \right) \leq \frac{16N^{3/2}(L-1)}{C_L} \left(\sum_{j=1}^{8} \Phi_{L,j}(t) + \Psi_{L,2}(t) + \Psi_{L,8}(t) \right)^4 \left\|\delta_t^L\right\|_{p^{in}},$$

which implies

$$\Gamma_L(t) \leq \Gamma_L(0) + \frac{16N^{3/2}(L-1)}{C_L} \int_0^t \Gamma_L(s)^4 \left\|\delta_s^L\right\|_{p^{in}} ds,$$

where $\Gamma_L(t) = \sum_{j=1}^8 \Phi_{L,j}(t) + \Psi_{L,2}(t) + \Psi_{L,8}(t)$. By Grönwall's lemma with $z(t) = \Gamma_L(t)$, $a(t) = \Gamma_L(0)$, $b(t) = \frac{16N^{3/2}(L-1)}{C_L}$, $k(s) = \left\|\delta_s^L\right\|_{p^{in}}$,

$$\Gamma_L(t) \leq \Gamma_L(0) \left\{ 1 - \frac{48N^{3/2}(L-1)}{C_L} \Gamma_L(0)^3 \int_0^t \left\|\delta_s^L\right\|_{p^{in}} ds \right\}^{-\frac{1}{3}}$$

for $0 \leq t \leq \beta_L$. Recall that $\int_0^T \left\|\delta_s^L\right\|_{p^{in}} ds$ is stochastically bounded and $\Gamma_L(0)$ converges to a constant. Thus, for $L^2/C_L \to 0$, we get $\frac{48N^{3/2}(L-1)}{C_L}\Gamma_L(0)^3 \int_0^t \left\|\delta_s^L\right\|_{p^{in}} ds \xrightarrow{p} 0$. This implies the upper bound of $\Gamma_\infty(t)$ converges to $\Gamma_\infty(0)$ in probability. On the other hand, $\Gamma_L(t) \geq \Gamma_L(0)$ by construction and $\beta_L \to \infty$ as $L \to \infty$. Thus, $\Gamma_L(t) \xrightarrow{p} \Gamma_\infty(0)$ uniformly for $t \in [0, T]$. This concludes the proof. □

Next, we consider similar Lyapunov functions but under ℓ_2 norm. This is because scaled NTK is not restricted to the dataset. For $1 \leq j \leq 8$,

$$\begin{aligned}
\tilde{\Phi}_{L,j}(x,t) &= \frac{1}{(L-1)C_L^j} \sum_{l=1}^{L-1} \left(\left\| f^l_{\theta^{(l)}(0)}(x) \right\| + \left\| f^l_{\theta^{(l)}(t)}(x) - f^l_{\theta^{(l)}(0)}(x) \right\| \right)^j \\
\tilde{\Psi}_{L,2}(x,t) &= \frac{1}{(L-1)^2} \sum_{l=2}^{L} \frac{L-1}{l-1} \sum_{i=2}^{l} \left(\left\| \mathfrak{W}_l^i(x,0) \right\| + \left\| \mathfrak{W}_l^i(x,t) - \mathfrak{W}_l^i(x,0) \right\| \right)^2 \\
\tilde{\Psi}_{L,8}(x,t) &= \frac{1}{L-1} \sum_{l=2}^{L} \left(\left\| \mathfrak{W}_L^l(x,0) \right\| + \left\| \mathfrak{W}_L^l(x,t) - \mathfrak{W}_L^l(x,0) \right\| \right)^8,
\end{aligned}$$

where $\tilde{\Phi}_{L,0} = 1$.

Note that Lyapunov functions under ℓ_2-norm are functions of (x, t). Similar to Lyapunov functions under p^{in}-norm, at initialization ($t = 0$), we have $\tilde{\Phi}_{L,j}(x, 0) \xrightarrow{p} \tilde{\Phi}_{\infty,j}(x, 0)$, $\tilde{\Psi}_{L,2}(x, 0) \xrightarrow{p} \tilde{\Psi}_{\infty,2}(x, 0)$, $\tilde{\Psi}_{L,8}(x, 0) \xrightarrow{p} \tilde{\Psi}_{\infty,8}(x, 0)$, where $\tilde{\Phi}_{L,\infty}(x, 0), \tilde{\Psi}_{\infty,2}(x, 0), \tilde{\Psi}_{\infty,8}(x, 0)$ are constant by the law of large number. The similar proposition holds, which implies the intermediate weights and pre-activation values are invariant during training in ℓ_2-norm sense.

Proposition 2. *For $1 \le j \le 8$ and any $T > 0$,*

$$\tilde{\Phi}_{L,j}(x, t) \xrightarrow{p} \tilde{\Phi}_{\infty,j}(x, 0), \quad \tilde{\Psi}_{L,2}(x, t) \xrightarrow{p} \tilde{\Psi}_{\infty,2}(x, 0), \quad \tilde{\Psi}_{L,8}(x, t) \xrightarrow{p} \tilde{\Psi}_{\infty,8}(x, 0)$$

uniformly for $t \in [0, T]$ as $L \to \infty$.

The following lemma, which corresponds to Lemma 10, allows us to apply Grönwall's lemma.

Lemma 11. *For $t \ge 0$ and $x \in \mathbb{R}_+^{d_{in}}$, if $f^l_{\theta^{(l)}(t)}(x)$ is element-wise nonzero for $1 \le l \le L - 1$, then*

$$\partial_t \tilde{\Psi}_{L,2}(x, t) \le \frac{2(L-1)}{C_L} \left(\tilde{\Psi}_{L,2}(x, t)\right)^2 \Phi_{L,4}(t)\Psi_{L,8}(t) \left\|\delta_t^L\right\|_{p^{in}}$$
$$\partial_t \tilde{\Phi}_{L,j}(x, t) \le \frac{2jN^{1/2}(L-1)}{C_L} \tilde{\Phi}_{L,j-1}(t)\tilde{\Phi}_{L,8}(x, t)\tilde{\Psi}_{L,2}(x, t)\Phi_{L,8}(t)\Psi_{L,8}(t) \left\|\delta_t^L\right\|_{p^{in}}$$
$$\partial_t \tilde{\Psi}_{L,8}(x, t) \le \frac{8(L-1)}{C_L} \tilde{\Phi}_{L,4}(x, t)\tilde{\Psi}_{L,2}(x, t) \left(\tilde{\Psi}_{L,8}(x, t)\right)^2 \Psi_{L,8}(t) \left\|\delta_t^L\right\|_{p^{in}},$$

for $1 \le j \le 8$.

The proof of Lemma 11 is given in Section G.

Proof of Proposition 2. Similar to the proof of Proposition 1, we need to show that $f^l_{\theta^{(l)}(t)}(x)$ is element-wise nonzero for all $t \in [0, T]$, $1 \le l \le L - 1$, and $x \in \mathbb{R}_+^{d_{in}}$. Again, the element-wise nonzero assumption also holds if L is large enough, and the proof is given in Section C. From Lemma 11,

$$\tilde{\Psi}_{L,2}(x, t) \le \tilde{\Psi}_{L,2}(x, 0) + \frac{2(L-1)}{C_L} \int_0^t \left(\tilde{\Psi}_{L,2}(x, s)\right)^2 \Phi_{L,4}(s)\Psi_{L,8}(s) \left\|\delta_s^L\right\|_{p^{in}} ds.$$

Then using Grönwall's lemma where $z(t) = \tilde{\Psi}_{L,2}(x, t)$, $a(t) = \tilde{\Psi}_{L,2}(x, 0)$, $b(t) = \frac{2(L-1)}{C_L}$, $k(s) = \Phi_{L,4}(s)\Psi_{L,8}(s) \left\|\delta_s^L\right\|_{p^{in}}$, we get

$$\tilde{\Psi}_{L,2}(x, t) \le \tilde{\Psi}_{L,2}(x, 0) \left\{1 - \frac{2(L-1)}{C_L} \tilde{\Psi}_{L,2}(x, 0) \int_0^t \Phi_{L,4}(s)\Psi_{L,8}(s) \left\|\delta_s^L\right\|_{p^{in}} ds\right\}^{-1}$$

for all $0 \le t \le \beta_L$. Similar to the proof of Proposition 1, if $L^2/C_L \to 0$, then we get $\tilde{\Psi}_{L,2}(t) \xrightarrow{p} \tilde{\Psi}_{\infty,2}(0)$ uniformly for $t \in [0, T]$.

On the other hand, Lemma 11 also implies

$$\sum_{j=1}^{8} \tilde{\Phi}_{L,j}(x, t) \le \sum_{j=1}^{8} \tilde{\Phi}_{L,j}(x, 0) + \int_0^t \frac{16N^{1/2}(L-1)}{C_L} \left(\sum_{j=1}^{8} \tilde{\Phi}_{L,j-1}(x, s)\right) \tilde{\Phi}_{L,8}(x, s)\tilde{\Psi}_{L,2}(x, s)\Phi_{L,8}(s)\Psi_{L,8}(s) \left\|\delta_s^L\right\|_{p^{in}} ds$$
$$\le \sum_{j=1}^{8} \tilde{\Phi}_{L,j}(x, 0) + \int_0^t \frac{16N^{1/2}(L-1)}{C_L} \left(\sum_{j=1}^{8} \tilde{\Phi}_{L,j}(x, s)\right)^2 \tilde{\Psi}_{L,2}(x, s)\Phi_{L,8}(s)\Psi_{L,8}(s) \left\|\delta_s^L\right\|_{p^{in}} ds$$
$$\tilde{\Psi}_{L,8}(x, t) \le \tilde{\Psi}_{L,8}(x, 0) + \int_0^t \frac{8(L-1)}{C_L} \left(\tilde{\Psi}_{L,8}(x, s)\right)^2 \tilde{\Phi}_{L,4}(x, s)\tilde{\Psi}_{L,2}(x, s)\Psi_{L,8}(s) \left\|\delta_s^L\right\|_{p^{in}} ds.$$

We can similarly apply Grönwall's lemma to show that $\sum_{j=1}^{8} \tilde{\Phi}_{L,j}(x, t)$ and $\tilde{\Psi}_{L,8}(x, t)$ converge to constant values. $\square$

Now, we consider our last and core Lyapunov function which contains intermediate weight and pre-activation values from a single layer. Define

$$\tilde{\Psi}^k_L(x,t) = \left\|\mathfrak{W}^k_L(x,0)\right\| + \left\|\mathfrak{W}^k_L(x,t) - \mathfrak{W}^k_L(x,0)\right\|$$
$$\tilde{\Phi}^l_L(x,t) = \frac{1}{C_L}\left(\left\|f^l_{\theta^{(l)}(0)}(x)\right\| + \left\|f^l_{\theta^{(l)}(t)}(x) - f^l_{\theta^{(l)}(0)}(x)\right\|\right)$$

for $1 \le l \le L$ and $2 \le k \le L$.

At initialization, $\tilde{\Psi}^k_L(x,0)$ and $\tilde{\Phi}^l_L(x,0)$ are stochastically bounded. Then the following proposition implies the invariance of Lyapunov functions during training.

Proposition 3. *For $1 \le l \le L$, $2 \le k \le L$ and any $T > 0$,*

$$\tilde{\Psi}^k_L(x,t) \xrightarrow{p} \tilde{\Psi}^k_\infty(x,0) \quad \tilde{\Phi}^l_L(x,t) \xrightarrow{p} \tilde{\Phi}^l_\infty(x,0).$$

uniformly for $t \in [0,T]$ as $L \to \infty$.

This implies individual intermediate weights and pre-activation values are effectively unchanged at each layer. Recall that this layer-wise invariance is straightforward in an infinite width network with finite depth. However, in our setting, an infinitely deep network has composition of infinitely many weights, which requires much careful analysis including the following lemma as well as previous propositions.

Lemma 12. *For $t \ge 0$ and $x \in \mathbb{R}^{d_{in}}_+$, if $f^l_{\theta^{(l)}(t)}(x)$ is element-wise nonzero for $1 \le l \le L-1$, then*

$$\partial_t \tilde{\Psi}^k_L(x,t) \le \frac{L-1}{C_L}\Psi_{L,8}(t)\tilde{\Phi}_{L,4}(x,t)\tilde{\Psi}_{L,2}(x,t)\tilde{\Psi}_{L,8}(x,t)\left\|\delta^L_t\right\|_{p^{in}}$$
$$\partial_t \tilde{\Phi}^l_L(x,t) \le \frac{2N^{1/2}(L-1)}{C_L}\Phi_{L,8}(t)\Psi_{L,8}(t)\tilde{\Phi}_{L,8}(x,t)\tilde{\Psi}_{L,2}(x,t)\left\|\delta^L_t\right\|_{p^{in}},$$

for $1 \le l \le L$ and $2 \le k \le L$.

The proof of Lemma 12 is given in Section H.

Proof of Proposition 3. As we discussed in the proof of Proposition 2, if L is large enough, $f^l_{\theta^{(l)}(t)}(x)$ is element-wise nonzero for all $t \in [0,T]$, $1 \le l \le L-1$, and $x \in \mathbb{R}^{d_{in}}_+$. Lemma 12 implies

$$\tilde{\Psi}^k_L(x,t) \le \tilde{\Psi}^k_L(x,0) + \int_0^t \frac{L-1}{C_L}\Psi_{L,8}(t)\tilde{\Phi}_{L,4}(x,t)\tilde{\Psi}_{L,2}(x,t)\tilde{\Psi}_{L,8}(x,t)\left\|\delta^L_t\right\|_{p^{in}}\, ds$$
$$\tilde{\Phi}^l_L(x,t) \le \tilde{\Phi}^l_L(x,0) + \int_0^t \frac{2N^{1/2}(L-1)}{C_L}\Phi_{L,8}(t)\Psi_{L,8}(t)\tilde{\Phi}_{L,8}(x,t)\tilde{\Psi}_{L,2}(x,t)\left\|\delta^L_t\right\|_{p^{in}}\, ds.$$

Unlike previous proofs, we do not need Grönwall's Lemma. From Proposition 1 and Proposition 2, all terms in RHS (such as $\Psi_{L,8}(t)$, $\tilde{\Phi}_{L,4}(t)$, $\tilde{\Psi}_{L,2}(t)$, etc.) converge to constants in probability. Since $L^2/C_L \to 0$, integral terms converge to zero as $L \to \infty$. On the other hand, it is clear that $\tilde{\Psi}^k_L(x,t) \ge \tilde{\Psi}^k_L(x,0)$ and $\tilde{\Phi}^l_L(x,t) \ge \tilde{\Phi}^l_L(x,0)$ by construction. Thus, $\tilde{\Psi}^k_L(x,t)$ and $\tilde{\Phi}^l_L(x,t)$ converges to $\tilde{\Psi}^k_L(x,0)$ and $\tilde{\Phi}^l_L(x,0)$ in probability, respectively. Since the integral terms are independent from the choice of k and l, we get uniform convergence. □

Proposition 3 implies the following desired result which implies that the variation of intermediate pre-activation values and weights must be negligible.

Corollary 13. *For any x, as $L \to \infty$, we have*

$$\sup_{1\le l\le L, t\in[0,T]} \frac{1}{C_L}\left\|f^l_{\theta^{(l)}(t)}(x) - f^l_{\theta^{(l)}(0)}(x)\right\| \xrightarrow{p} 0$$
$$\sup_{2\le k\le L, t\in[0,T]} \left\|\mathfrak{W}^k_L(x,t) - \mathfrak{W}^k_L(x,0)\right\| \xrightarrow{p} 0.$$

With this corollary, we are now ready to prove our main theorem, the invariance of scaled NTK.

Theorem 2 (Invariance of scaled NTK). *Let $T > 0$. Suppose $\int_0^T \left\| \delta|_{f_\theta^L} \right\|_{p^{in}} dt$ is stochastically bounded as $L \to \infty$. Then, for any $x, x' \in \mathbb{R}_+^d$,*

$$\tilde{\Theta}_t^L(x, x') \xrightarrow{p} \tilde{\Theta}^\infty(x, x')$$

uniformly for $t \in [0, T]$ as $L \to \infty$.

Proof of Theorem 2. By definition,

$$\begin{aligned}\tilde{\Theta}_t^L(x, x') = \frac{1}{L} \sum_{l=0}^{L-1} \left[\left(\frac{1}{C_L} \sigma \left(f_{\theta^{(l)}(t)}^l (x) \right)^\intercal \frac{1}{C_L} \sigma \left(f_{\theta^{(l)}(t)}^l (x') \right) \right) \mathfrak{W}_L^{l+2}(x, t) \mathfrak{W}_L^{l+2}(x', t)^\intercal \right] \\ + \frac{1}{LC_L^2} \sum_{l=0}^{L-1} \left[\mathfrak{W}_L^{l+2}(x, t) \mathfrak{W}_L^{l+2}(x', t)^\intercal \right].\end{aligned}$$

Informally, Proposition 3 implies that $f_{\theta^{(l)}(t)}^l(x)$ and $\mathfrak{W}_L^{l+2}(x, t)$ are effectively invariant, and therefore the kernel $\tilde{\Theta}_t^L(x, x')$ is invariant. An additional effort is required to handle the summation of L terms since L also increases. More formal proof is given in the following.

Since ReLU is 1-Lipshitz, Corollary 13 implies

$$\begin{aligned}\sup_{1 \le l \le L, t \in [0,T]} \left\| \frac{1}{C_L} \sigma \left(f_{\theta^{(l)}(t)}^l (x) \right)^\intercal \frac{1}{C_L} \sigma \left(f_{\theta^{(l)}(t)}^l (x') \right) - \frac{1}{C_L} \sigma \left(f_{\theta^{(l)}(0)}^l (x) \right)^\intercal \frac{1}{C_L} \sigma \left(f_{\theta^{(l)}(0)}^l (x') \right) \right\| \xrightarrow{p} 0 \\ \sup_{2 \le k \le L, t \in [0,T]} \left\| \mathfrak{W}_L^k(t, x) \mathfrak{W}_L^k(t, x') - \mathfrak{W}_L^k(0, x) \mathfrak{W}_L^k(0, x') \right\| \xrightarrow{p} 0\end{aligned}$$

as $L \to \infty$ for all x and x'. On the other hand, the norm of the difference between kernels is bounded by

$$\begin{aligned}\left\| \tilde{\Theta}_t^L(x, x') - \tilde{\Theta}_0^L(x, x') \right\| \le \frac{1}{L} \sum_{l=0}^{L} \Bigg\| \left(\frac{1}{C_L} \sigma \left(f_{\theta^{(l)}(t)}^l (x) \right)^\intercal \frac{1}{C_L} \sigma \left(f_{\theta^{(l)}(t)}^l (x') \right) \right) \mathfrak{W}_L^{l+2}(x, t) \mathfrak{W}_L^{l+2}(x', t)^\intercal \\ - \left(\frac{1}{C_L} \sigma \left(f_{\theta^{(l)}(0)}^l (x) \right)^\intercal \frac{1}{C_L} \sigma \left(f_{\theta^{(l)}(0)}^l (x') \right) \right) \mathfrak{W}_L^{l+2}(x, 0) \mathfrak{W}_L^{l+2}(x', 0)^\intercal \Bigg\| \\ + \frac{1}{LC_L^2} \sum_{l=0}^{L-1} \left\| \left[\mathfrak{W}_L^{l+2}(x, t) \mathfrak{W}_L^{l+2}(x', t)^\intercal \right] - \left[\mathfrak{W}_L^{l+2}(x, 0) \mathfrak{W}_L^{l+2}(x', 0)^\intercal \right] \right\| \\ \le \max_l \Bigg\| \left(\frac{1}{C_L} \sigma \left(f_{\theta^{(l)}(t)}^l (x) \right)^\intercal \frac{1}{C_L} \sigma \left(f_{\theta^{(l)}(t)}^l (x') \right) \right) \mathfrak{W}_L^{l+2}(x, t) \mathfrak{W}_L^{l+2}(x', t)^\intercal \\ - \left(\frac{1}{C_L} \sigma \left(f_{\theta^{(l)}(0)}^l (x) \right)^\intercal \frac{1}{C_L} \sigma \left(f_{\theta^{(l)}(0)}^l (x') \right) \right) \mathfrak{W}_L^{l+2}(x, 0) \mathfrak{W}_L^{l+2}(x', 0)^\intercal \Bigg\| \\ + \frac{1}{C_L^2} \max_l \left\| \left[\mathfrak{W}_L^{l+2}(x, t) \mathfrak{W}_L^{l+2}(x', t)^\intercal \right] - \left[\mathfrak{W}_L^{l+2}(x, 0) \mathfrak{W}_L^{l+2}(x', 0)^\intercal \right] \right\|.\end{aligned}$$

Thus, as $L \to \infty$, we have

$$\sup_{t \in [0,T]} \left\| \tilde{\Theta}_t^L(x, x') - \tilde{\Theta}_0^L(x, x') \right\| \xrightarrow{p} 0.$$

Finally, Theorem 1 implies $\tilde{\Theta}_0^L$ converges to $\tilde{\Theta}^\infty$ in probability, and therefore $\tilde{\Theta}_t^L$ converges to $\tilde{\Theta}^\infty$ in probability. □

B.3. Proof of Theorem 3

Theorem 3 implies that the solution, a fully-trained inifinite MLP, matches the kernel regression predictor f_{ntk}.

Theorem 3 (Equivalence between deep MLP and kernel regression)**.** *Let $\mathcal{L}(f) = \frac{1}{2}\|f - f^\star\|^2_{p^{in}}$. Let $\tilde{\Theta}^\infty$ be positive definite. If f_t follows the kernel gradient flow with respect to $\tilde{\Theta}^\infty$, then for any $x \in \mathbb{R}^{d_{in}}_+$,*

$$\lim_{t\to\infty} f_t(x) = f_{\text{ntk}}(x).$$

Proof of Theorem 3. Throughout the proof, we implicitly assume $d_{out} = 1$ for the sake of simplicity. Our proof conveys all important idea of the general proof, and the extension to general d_{out} is straightforward.

The gradient flow is

$$\partial_t f_t(x') = -\mathbb{E}_{x\sim p^{in}}\left[\tilde{\Theta}^\infty(x', x)\left(f_t(x) - f^\star(x)\right)\right].$$

Define an linear operator $D : \mathcal{F} \to \mathcal{F}$ such that $D(g)(x') = \mathbb{E}_{x\sim p^{in}}\left[\tilde{\Theta}^\infty(x', x)g(x)\right]$. Then, we have

$$\partial_t f_t(x') = -D\left(f_t(x) - f^\star(x)\right)$$

which has an explicit solution:

$$f_t(x') = f^\star(x') + e^{-tD}(f_0(x) - f^\star(x)),$$

where $e^{-tD} = \sum_{k=0}^\infty \frac{(-t)^k}{k!}D^k$. Then following lemma holds.

Lemma 14. *Any function $f : \mathbb{R}^d_+ \to \mathbb{R}$ is a linear combination of eigenfunctions of D. More precisely, if we let $f(X) = (f(x_1), \ldots, f(x_N)) \in \mathbb{R}^N$, then f can be expressed as*

$$f(x) = (\tilde{\Theta}^\infty(x, x_1), \cdots, \tilde{\Theta}^\infty(x, x_N))K^{-1}f(X) + h(x),$$

where $(\tilde{\Theta}^\infty_0(x, x_1), \cdots, \tilde{\Theta}^\infty(x, x_N))K^{-1}f(X)$ is a linear combination of eigenfunctions of D for nonzero eigenvalues, and $h(x)$ is an eigenfunction for zero eigenvalue.

Proof. Let K be an N by N matrix, where $K_{i,j} = \tilde{\Theta}^\infty(x_i, x_j)$. Then, K is positive definite, and K has N eigenvectors with nonzero eigenvalues. Suppose $a = (a_1, \ldots, a_N) \in \mathbb{R}^N$ be an eigenvector of K for eigenvalue $\lambda > 0$, then

$$\begin{aligned}
D\left(\sum_{i=1}^N a_i\tilde{\Theta}^\infty(\cdot, x_i)\right)(x') =& \frac{1}{N}\sum_{j=1}^N\left(\tilde{\Theta}^\infty(x', x_j)\sum_{i=1}^N\tilde{\Theta}^\infty(x_j, x_i)a_i\right)\\
=& \frac{1}{N}\sum_{j=1}^N\left(\tilde{\Theta}^\infty(x', x_j)\lambda a_j\right)\\
=& \frac{\lambda}{N}\sum_{i=1}^N a_i\tilde{\Theta}^\infty(x', x_i).
\end{aligned}$$

In other words, $\sum_{i=1}^N a_i\tilde{\Theta}^\infty(\cdot, x_i)$ is an eigenvector of D. Since K is positive definite, any linear combination of $\{\tilde{\Theta}^\infty(\cdot, x_i)\}_{i=1}^N$ is a linear combination of eigenfunctions of D for nonzero eigenvalues, including $(\tilde{\Theta}^\infty(x, x_1), \cdots, \tilde{\Theta}^\infty(x, x_N))K^{-1}f(X)$ for any $f : \mathbb{R}^d_+ \to \mathbb{R}$.

On the other hand, for a given function f, let $h(x) = f(x) - (\tilde{\Theta}^\infty(x, x_1), \cdots, \tilde{\Theta}^\infty(x, x_N))K^{-1}f(X)$. Since $h(x_i) = 0$ for all i,

$$D(h)(x') = \frac{1}{N}\sum_{i=1}^N\tilde{\Theta}^\infty(x', x_i)h(x_i) = 0,$$

and therefore h is an eigenfunction of D for zero eigenvalue. Thus, we have

$$f(x) = (\tilde{\Theta}^{\infty}(x', x_1), \cdots, \tilde{\Theta}^{\infty}(x', x_N))K^{-1}f(X) + h(x),$$

which is a linear combination of eigenfunctions of D. □

From Lemma 14, we have

$$f_0(x') - f^\star(x') = \sum a_i g_i(x') + h(x'),$$

where g_i are eigenfunctions of D for nonzero eigenvalues and h is an eigenfunction for zero eigenvalue. Then,

$$f_t(x') = f^\star(x') + h(x') + \sum_i e^{-t\lambda_i} a_i g_i(x').$$

At initialization $t = 0$,

$$f_0(x') = f^\star(x') + h(x') + \sum_i a_i g_i(x').$$

On the other hand, as $t \to \infty$, we have

$$\lim_{t\to\infty} f_t(x') = f^\star(x') + h(x').$$

By combining two equations, we have

$$\lim_{t\to\infty} f_t(x') = f_0(x') - \sum_i a_i g_i(x').$$

Finally, we have

$$\begin{aligned}\lim_{t\to\infty} f_t(x') =& f_0(x') - (\tilde{\Theta}^{\infty}(x', x_1), \cdots, \tilde{\Theta}^{\infty}(x', x_N))K^{-1}(f_0(X) - f^\star(X))\\ =&(\tilde{\Theta}^{\infty}(x', x_1), \cdots, \tilde{\Theta}^{\infty}(x', x_N))K^{-1}f^\star(X),\end{aligned}$$

where the last equality is from $f_0(x') = 0$ by initialization. This concludes the proof. □

C. Zero-crossing does not happen during training MLP

While proving key statements in the proof of Theorem 2 (Proposition 1, Proposition 2, and Proposition 3), we assumed that all elements of $f^l_{\theta^{(l)}(t)}(x)$ are nonzero for $1 \le l \le L-1$, $x \in \{x_1, \dots, x_N\}$, and any input $x \in \mathbb{R}^{d_{in}}_+$ while training. In this section, for any $x', x'' \in \mathbb{R}^{d_{in}}_+$, we show that all entries of $\{f^l_{\theta^{(l)}(t)}(x_1), \dots f^l_{\theta^{(l)}(t)}(x_N), f^l_{\theta^{(l)}(t)}(x'), f^l_{\theta^{(l)}(t)}(x'')\}$ are nonzero with high probability while training. For the sake of notational simplicity, we set $x_{N+1} = x'$ and $x_{N+2} = x''$ in the following lemma.

Lemma 15. *Suppose $\int_0^T \left\|\delta|_{f^L_\theta}\right\|_{p^{in}} dt$ is stochastically bounded as $L \to \infty$. For any $T > 0$, let*

$$T_L = \bigwedge_{i=1}^{N+2} \inf_t \left\{ \left(f^l_{\theta^{(l)}(t)}(x_i)\right)_r = 0 \textit{ for some } 1 \le l \le L-1 \textit{ and } r \right\} \bigwedge T,$$

where $x_{N+1}, x_{N+2} \in \mathbb{R}^{d_{in}}_+$ are arbitrary inputs. Then, for any $\epsilon > 0$, there exists $L_{\max}$ such that Pr $[T_L = T] \ge 1 - \epsilon$ for all $L > L_{\max}$.

Proof of Lemma 15. For $\epsilon' > 0$, let $M = 1 + \epsilon'$. Since $\int_0^T \left\|\delta|_{f^L_\theta}\right\|_{p^{in}} dt$ is stochastically bounded as $L \to \infty$, there exists $K_0 > 0$ and large enough $L_0 > 0$ such that

$$\Pr\left[\int_0^T \left\|\delta|_{f^L_\theta}\right\|_{p^{in}} dt > K_0\right] < \frac{\epsilon}{3}$$

for all $L \geq L_0$. We define a complement of such event by

$$E_0 = \left\{ \int_0^T \left\| \delta|_{f_\theta^L} \right\|_{p^{in}} \, dt \leq K_0 \right\},$$

where $\Pr[E_0] \geq 1 - \frac{\epsilon}{3}$.

On the other hand, recall that $\tilde{\Psi}_{\infty,2}(x_i, 0), \tilde{\Psi}_{\infty,8}(x, 0), \tilde{\Phi}_{\infty,j}(x_i, 0), \Gamma_\infty(0)$ are all (non-random) constants for all $1 \leq j \leq 8$ and $1 \leq i \leq N + 2$. Since $L^2/C_L \to 0$ as $L \to \infty$, we have

$$\begin{aligned}
\frac{48N^{3/2}M^3(L-1)}{C_L}\Gamma_\infty(0)^3 K_0 &\to 0, \\
\frac{2M^5(L-1)}{C_L}\tilde{\Psi}_{\infty,2}(x_i,0)\Gamma_\infty^2(0)K_0 &\to 0, \\
\frac{M^7 16N^{1/2}(L-1)}{C_L}\left(\sum_{j=1}^{8}\tilde{\Phi}_{\infty,j}(x_i,0)\right)\tilde{\Psi}_{\infty,2}(0)\Gamma_\infty^2(0)K_0 &\to 0, \\
\frac{M^7 8(L-1)}{C_L}\tilde{\Psi}_{\infty,8}(x_i,0)\left(\sum_{j=1}^{8}\tilde{\Phi}_{\infty,j}(x_i,0)\right)\tilde{\Psi}_{\infty,2}(x_i,0)\Gamma_\infty(0)K_0 &\to 0,
\end{aligned}$$

as $L \to \infty$. Thus, there exists L_1 such that

$$\begin{aligned}
\frac{48N^{3/2}M^3(L-1)}{C_L}\Gamma_\infty(0)^3 K_0 &< 1 - M^{-3}, \\
\frac{2M^5(L-1)}{C_L}\tilde{\Psi}_{\infty,2}(x_i,0)\Gamma_\infty^2(0)K_0 &< 1 - M^{-1}, \\
\frac{M^7 16N^{1/2}(L-1)}{C_L}\left(\sum_{j=1}^{8}\tilde{\Phi}_{\infty,j}(x_i,0)\right)\tilde{\Psi}_{\infty,2}(0)\Gamma_\infty^2(0)K_0 &< 1 - M^{-1}, \\
\frac{M^7 8(L-1)}{C_L}\tilde{\Psi}_{\infty,8}(x_i,0)\left(\sum_{j=1}^{8}\tilde{\Phi}_{\infty,j}(x_i,0)\right)\tilde{\Psi}_{\infty,2}(x_i,0)\Gamma_\infty(0)K_0 &< 1 - M^{-1},
\end{aligned}$$

for all $L \geq L_1$.

Now, we define following events which indicate Lyapunov functions converge to constant values by the law of large numbers.

$$\begin{aligned}
E_1^{(L)} &= \{\Gamma_L(0) \leq M\Gamma_\infty(0)\} \\
E_2^{(L)} &= \bigcup_{i=1}^{N+2} \left\{ \tilde{\Psi}_{L,8}(x_i,0) \leq M\tilde{\Psi}_{\infty,8}(x_i,0) \right\} \\
E_3^{(L)} &= \bigcup_{i=1}^{N+2} \left\{ \tilde{\Psi}_{L,2}(x_i,0) \leq M\tilde{\Psi}_{\infty,2}(x_i,0) \right\} \\
E_4^{(L)} &= \bigcup_{i=1}^{N+2} \left\{ \sum_{j=1}^{8}\tilde{\Phi}_{L,j}(x_i,0) \leq M\sum_{j=1}^{8}\tilde{\Phi}_{\infty,j}(x_i,0) \right\}.
\end{aligned}$$

By the law of large numbers, all events $E_1^{(L)}, E_2^{(L)}, E_3^{(L)}, E_4^{(L)}$ have probabilities converge to 1 as $L \to \infty$. Thus, there exists L_2 such that

$$\Pr\left[E_1^{(L)} \cap E_2^{(L)} \cap E_3^{(L)} \cap E_4^{(L)}\right] \geq 1 - \frac{\epsilon}{3}$$

for all $L \geq L_2$.

Finally, we are interested in the event

$$E_5 = \bigcup_{i=1}^{N+2} \left\{ \left| \left(f^l_{\theta(0)}(x_i) \right)_r \right| \geq \frac{L}{C_L}(K_1 + 1) \text{ for all } 1 \leq l \leq L-1 \text{ and } r \right\},$$

where we specify $K_1 > 0$ later. Consider a random element of intermediate pre-activation values

$$\begin{aligned} g^1(x) &= \left(f^1_{\theta(0)}(x) \right)_{d_{in}+1} = \frac{1}{C_L}(u^1)^\intercal x + \frac{1}{C_L} v^1 \\ g^l(x) &= \left(f^l_{\theta^{(l)}(0)}(x) \right)_{d_{in}+1} = (u^l)^\intercal x + \frac{1}{C_L} v^l \text{ for } 2 \leq l \leq L-1. \end{aligned}$$

Then, $g^1 \sim \mathcal{N}(0, \sigma_1^2)$ and $g^l \sim \mathcal{N}(0, \sigma_g^2)$ for $2 \leq l \leq L-1$ are independent Gaussian random variables with finite variances. Since

$$\Pr\left[|g^l| < \tfrac{L}{C_L}(K+1) \text{ for some } 1 \leq l \leq L-1 \right] \leq 2(K+1)\frac{L(L-2)}{C_L}\frac{1}{\sqrt{2\pi\sigma_g^2}} + 2(K+1)\frac{L}{C_L}\frac{1}{\sqrt{2\pi\sigma_1^2}}$$

by union bound, the probability converges to 0 as $L \to \infty$. Thus, there exists L_3 such that $\Pr[E_5] \geq 1 - \frac{\epsilon}{3}$ holds for all $L \geq L_3$.

Then, we would like to show the following claim.

Claim 1. *If $L \geq L_1$, and all events $E_0, E_1^{(L)}, E_2^{(L)}, E_3^{(L)}, E_4^{(L)}, E_5$ are given, then all elements of intermediate pre-activation values never cross zeros while training for $t \in [0, T]$.*

If the claim holds, then for $L \geq L_{\max} = \max\{L_0, L_1, L_2, L_3\}$,

$$\Pr\left[E_0 \cap E_1^{(L)} \cap E_2^{(L)} \cap E_3^{(L)} \cap E_4^{(L)} \cap E_5 \right] \geq 1 - \epsilon,$$

which concludes the proof of Lemma 15. In the rest of the proof, we will prove the above claim. Suppose $L \geq L_1$ and all events $E_0, E_1^{(L)}, E_2^{(L)}, E_3^{(L)}, E_4^{(L)}, E_5$ are given. Since all intermediate pre-activation values at initialization

$$f^1_{\theta^{(1)}(0)}(x_i) = \begin{bmatrix} C_L x_i \\ (u^1)^\intercal x_i + v^1 \\ C_L \mathbb{1}_{d_{\text{out}}} \end{bmatrix}, \quad f^l_{\theta^{(l)}(0)}(x_i) = \begin{bmatrix} C_L x_i \\ C_L (u^l)^\intercal x_i + v^l \\ C_L \mathbb{1}_{d_{\text{out}}} \end{bmatrix}$$

are nonzero for $2 \leq l \leq L-1$ and $1 \leq i \leq N+2$ with probability 1. Then, for all $t \in [0, T_L]$, Lemma 10 holds which is essential in the proof of Proposition 1. In the proof, using Grönwall's lemma, we showed

$$\Gamma_L(t) \leq \Gamma_L(0) \left\{ 1 - \frac{48N^{3/2}(L-1)}{C_L} \Gamma_L(0)^3 \int_0^{\beta_L} \left\| \delta_s^L \right\|_{p^{in}} ds \right\}^{-\frac{1}{3}}$$

for $t \in \left[0, \beta_L^{(1)}\right]$, where

$$\beta_L^{(1)} = \sup\left\{ t : \frac{48N^{3/2}(L-1)}{C_L}\Gamma_L(0)^3 \int_0^t \left\| \delta_s^L \right\|_{p^{in}} ds < 1 \right\}.$$

Since $E_1^{(L)}$ is given, $\Gamma_L(0) \leq M\Gamma_\infty(0)$. Then, since E_0 is given,

$$\begin{aligned} \frac{48N^{3/2}(L-1)}{C_L}\Gamma_L(0)^3 \int_0^t \left\| \delta_s^L \right\|_{p^{in}} ds \leq & \frac{48M^3N^{3/2}(L-1)}{C_L}\Gamma_\infty(0)^3 K_0 \\ < & 1 - M^{-3} \end{aligned}$$

by definition of L_1. This implies that $\beta_L^{(1)} = T_L$ and $\Gamma_L(t) \leq M^2\Gamma_\infty(0)$ if E_0 and $E_1^{(L)}$ are given and $L \geq L_1$. Recall that $\Gamma_L(t) = \sum_{j=1}^8 \Phi_{L,j}(t) + \Psi_{L,2}(t) + \Psi_{L,8}(t)$, and therefore $\Gamma_L(t) \leq M^2\Gamma_\infty(0)$ also implies $\Psi_{L,8}(t) \leq M^2\Gamma_\infty(0)$, $\Psi_{L,2}(t) \leq M^2\Gamma_\infty(0)$, and $\Phi_{L,j}(t) \leq M^2\Gamma_\infty(0)$ for all $1 \leq j \leq 8$.

Similarly, in the proof of Proposition 2, we proved

$$\tilde{\Psi}_{L,2}(x_i, t) \leq \tilde{\Psi}_{L,2}(x_i, 0) \left\{1 - \frac{2(L-1)}{C_L}\tilde{\Psi}_{L,2}(x_i, 0)\int_0^t \Phi_{L,4}(s)\Psi_{L,4}(s)\left\|\delta_s^L\right\|_{p^{in}} ds\right\}^{-1}$$

for all $t \in \left[0, \beta_L^{(2)}\right]$, where

$$\beta_L^{(2)} = \sup\left\{t : \frac{2(L-1)}{C_L}\tilde{\Psi}_{L,2}(x_i, 0)\int_0^t \Phi_{L,4}(s)\Psi_{L,8}(s)\left\|\delta_s^L\right\|_{p^{in}} ds < 1\right\}.$$

We have $\tilde{\Psi}_{L,8}(x_i, 0) < M\tilde{\Psi}_{\infty,2}(x_i, 0)$ if $E_3^{(L)}$ is given. In this case, with a definition of L_1,

$$\begin{aligned}\frac{2(L-1)}{C_L}\tilde{\Psi}_{L,2}(x_i, 0)\int_0^t \Phi_{L,4}(s)\Psi_{L,8}(s)\left\|\delta_s^L\right\|_{p^{in}} ds &\leq \frac{2M^5(L-1)}{C_L}\tilde{\Psi}_{\infty,2}(x_i, 0)\Gamma_\infty^2(0)K_0 \\ &< 1 - M^{-1}\end{aligned}$$

if E_0 is also given. This implies that $\beta_L^{(2)} = T_L$ and $\tilde{\Psi}_{L,2}(x_i, t) \leq M^2\tilde{\Psi}_{\infty,2}(x_i, 0)$, if $E_3^{(L)}$ and E_0 are given and $L \geq L_1$. In similar way, we obtain $\sum_{j=1}^8 \tilde{\Phi}_{\infty,j}(x_i, t) \leq M^2\sum_{j=1}^8 \tilde{\Phi}_{\infty,j}(x_i, 0)$ for $t \in [0, T_L]$ if $E_4^{(L)}$ and E_0 are given and $L \geq L_1$. Also, we have $\tilde{\Psi}_{L,8}(x_i, t) \leq M^2\tilde{\Psi}_{\infty,8}(x_i, 0)$ for $t \in [0, T_L]$ if $E_2^{(L)}$ and E_0 are given and $L \geq L_1$.

Finally, in the proof of Proposition 3, we proved

$$\tilde{\Phi}_L^l(x, t) \leq \tilde{\Phi}_L^l(x, 0) + \int_0^t \frac{2N^{1/2}(L-1)}{C_L}\Phi_{L,8}(t)\Psi_{L,8}(t)\tilde{\Phi}_{L,8}(x, t)\tilde{\Psi}_{L,2}(x, t)\left\|\delta_t^L\right\|_{p^{in}} ds.$$

Then, if $E_0, E_1^{(L)}, E_2^{(L)}, E_3^{(L)}, E_4^{(L)}$ are given and $L \geq L_1$,

$$\begin{aligned}&\int_0^{T_L} \frac{2N^{1/2}(L-1)}{C_L}\Phi_{L,8}(t)\Psi_{L,8}(t)\tilde{\Phi}_{L,8}(x, t)\tilde{\Psi}_{L,2}(x, t)\left\|\delta_t^L\right\|_{p^{in}} ds \\ &\quad\leq \frac{2N^{1/2}M^8(L-1)}{C_L}\Gamma_\infty^2(0)\left(\sum_{j=1}^8 \tilde{\Phi}_{\infty,j}(x_i, 0)\right)\tilde{\Psi}_{\infty,2}(x_i, 0)K_0.\end{aligned}$$

If we let $K_1 = \sup_{1\leq i\leq N+2} 2N^{1/2}M^8\Gamma_\infty^2(0)\left(\sum_{j=1}^8 \tilde{\Phi}_{\infty,j}(x_i, 0)\right)\tilde{\Psi}_{\infty,2}(x_i, 0)K_0$, then this inequality implies

$$\sup_{1\leq l\leq L-1, t\in[0,T_L]} \frac{1}{C_L}\left\|f_{\theta^{(l)}(t)}^l(x_i) - f_{\theta^{(l)}(0)}^l(x_i)\right\| < \frac{L}{C_L}K_1$$

for all i. On the other hand, if E_5 is given,

$$\frac{1}{C_L}\left|\left(f_{\theta(0)}^l(x_i)\right)_r\right| > \frac{L}{C_L}(K_1 + 1)$$

for all i, l and r, and therefore

$$\inf_{1\leq l\leq L-1, t\in[0,T_L]} \left|\left(f_{\theta^{(l)}(t)}^l(x_i)\right)_r\right| > L$$

for all i and r. This implies $T_L = T$, which concludes the proof of Claim 1. □

Equipped with Lemma 15, we provide a brief outline of the rigorous proof of Theorem 2. For any $T > 0$, let

$$T_L = \bigwedge_{i=1}^{N+2} \inf_t \left\{ \left(f^l_{\theta^{(l)}(t)}(x_i) \right)_r = 0 \text{ for some } 1 \le l \le L-1 \text{ and } r \right\} \bigwedge T,$$

where $x_{N+1}, x_{N+2} \in \mathbb{R}_+^{d_{in}}$ are arbitrary inputs. Also, for $\epsilon, \epsilon' > 0$, where $M = 1 + \epsilon'$, we can define events $E_0, E_1^{(L)}, E_2^{(L)}, E_3^{(L)}, E_4^{(L)}, E_5$ and constants K_0, L_0, L_1, L_2, L_3 as defined in the proof of Lemma 15. We further let

$$K_2 = \max \left\{ K_1, \sup_{1 \le i \le N+2} M^8 \Gamma_\infty(0) \left(\sum_{j=1}^{8} \tilde{\Phi}_{\infty,j}(x_i, 0) \right) \tilde{\Psi}_{\infty,2}(x_i, 0) \tilde{\Psi}_{\infty,8}(x_i, 0) K_0 \right\}.$$

and

$$\begin{aligned} E_6 = & \Bigg\{ \frac{1}{L} \sum_l \left(1 + \frac{L}{C_L} K_2 + \left\| \frac{1}{C_L} \sigma\left(f^l_{\theta^{(l)}(0)}(x) \right) \right\| \right) \left(1 + \frac{L}{C_L} K_2 + \left\| \frac{1}{C_L} \sigma\left(f^l_{\theta^{(l)}(0)}(x') \right) \right\| \right) \\ & \times \left(1 + \frac{L}{C_L} K_2 + \left\| \mathfrak{W}_L^{l+2}(x, 0) \right\| \right) \left(1 + \frac{L}{C_L} K_2 + \left\| \mathfrak{W}_L^{l+2}(x', 0) \right\| \right) \times \frac{L}{C_L} K_2 \\ & + \frac{1}{L C_L^2} \sum_l \left(1 + \frac{L}{C_L} K_2 + \left\| \mathfrak{W}_L^{l+2}(x, 0) \right\| \right) \left(1 + \frac{L}{C_L} K_2 + \left\| \mathfrak{W}_L^{l+2}(x', 0) \right\| \right) \times \frac{L}{C_L} K_2 < \epsilon'' \Bigg\} \end{aligned}$$

for $\epsilon'' > 0$. Since $L^2/C_L \to 0$ as $L \to \infty$, the probability of $E_6^{(L)}$ converges to 1 by the law of large numbers, and there exists L_4 such that

$$\Pr\left[E_6^{(L)} \right] > 1 - \epsilon$$

for all $L \ge L_4$. Suppose all events $E_0, E_1^{(L)}, E_2^{(L)}, E_3^{(L)}, E_4^{(L)}, E_5, E_6^{(L)}$ are given, and $L \ge L_{\max} = \max\{L_0, L_1, L_2, L_3, L_4\}$. Then, instead of Proposition 1, the following inequalities hold without having $L \to \infty$:

$$\begin{aligned} \Phi_{L,j}(t) &\le M^2 \Gamma_\infty(0) \\ \Psi_{L,2}(t) &\le M^2 \Gamma_\infty(0) \\ \Psi_{L,8}(t) &\le M^2 \Gamma_\infty(0) \end{aligned}$$

for all $1 \le j \le 8$ and $t \in [0, T_L]$ since $\beta_L^{(1)} = T_L$. Similarly, we can rewrite Proposition 2 without having $L \to \infty$:

$$\begin{aligned} \sum_{j=1}^{8} \tilde{\Phi}_{\infty,j}(x_i, 0) \le & \sum_{j=1}^{8} \tilde{\Phi}_{L,j}(x_i, t) \le M^2 \sum_{j=1}^{8} \tilde{\Phi}_{\infty,j}(x_i, 0) \\ \tilde{\Psi}_{\infty,2}(x_i, 0) \le & \tilde{\Psi}_{L,2}(x_i, t) \le M^2 \tilde{\Psi}_{\infty,2}(x_i, 0) \\ \tilde{\Psi}_{\infty,8}(x_i, 0) \le & \tilde{\Psi}_{L,8}(x_i, t) \le M^2 \tilde{\Psi}_{\infty,8}(x_i, 0) \end{aligned}$$

for all $1 \le i \le N+2$ and $t \in [0, T_L]$. Finally, Corollary 13 can be replaced by

$$\sup_{1 \le l \le L-1, t \in [0, T_L]} \frac{1}{C_L} \left\| f^l_{\theta^{(l)}(t)}(x_i) - f^l_{\theta^{(l)}(0)}(x_i) \right\| < \frac{L}{C_L} K_1$$

for all i, where K_1 is as defined in the proof of Lemma 15. This already implies $T_L = T$.

In addition, by Lemma 12,

$$\sup_{1 \le l \le L-1, t \in [0, T]} \left\| \mathfrak{W}_L^k(x_i, t) - \mathfrak{W}_L^k(x_i, 0) \right\| < \frac{L}{C_L} K_2.$$

Thus, for $x, x' \in \{x_1, \dots, x_{N+2}\}$, we can bound the deviation of $\tilde{\Theta}^L_t(x, x')$ by

$$\begin{aligned}
\left\|\tilde{\Theta}^L_t(x,x') - \tilde{\Theta}^L_0(x,x')\right\| &\le \frac{1}{L}\sum_l \Bigg\| \left(\frac{1}{C_L}\sigma\left(f^l_{\theta^{(l)}(t)}(x)\right)^\intercal \frac{1}{C_L}\sigma\left(f^l_{\theta^{(l)}(t)}(x')\right)\right)\mathfrak{W}^{l+2}_L(x,t)\mathfrak{W}^{l+2}_L(x',t)^\intercal \\
&\qquad - \left(\frac{1}{C_L}\sigma\left(f^l_{\theta^{(l)}(0)}(x)\right)^\intercal \frac{1}{C_L}\sigma\left(f^l_{\theta^{(l)}(0)}(x')\right)\right)\mathfrak{W}^{l+2}_L(x,0)\mathfrak{W}^{l+2}_L(x',0)^\intercal \Bigg\| \\
&\quad + \frac{1}{LC_L^2}\sum_l \left\|\left[\mathfrak{W}^{l+2}_L(x,t)\mathfrak{W}^{l+2}_L(x',t)^\intercal\right] - \left[\mathfrak{W}^{l+2}_L(x,0)\mathfrak{W}^{l+2}_L(x',0)^\intercal\right]\right\| \\
&\le \frac{1}{L}\sum_l \Bigg[\left(\frac{L}{C_L}K_2 + \left\|\frac{1}{C_L}\sigma\left(f^l_{\theta^{(l)}(0)}(x)\right)\right\|\right)\left(\frac{L}{C_L}K_2 + \left\|\frac{1}{C_L}\sigma\left(f^l_{\theta^{(l)}(0)}(x')\right)\right\|\right) \\
&\qquad \times \left(\frac{L}{C_L}K_2 + \left\|\mathfrak{W}^{l+2}_L(x,0)\right\|\right)\left(\frac{L}{C_L}K_2 + \left\|\mathfrak{W}^{l+2}_L(x',0)\right\|\right) \\
&\qquad - \left\|\frac{1}{C_L}\sigma\left(f^l_{\theta^{(l)}(0)}(x)\right)\right\| \left\|\frac{1}{C_L}\sigma\left(f^l_{\theta^{(l)}(0)}(x')\right)\right\| \left\|\mathfrak{W}^{l+2}_L(x,0)\right\| \left\|\mathfrak{W}^{l+2}_L(x',0)\right\| \Bigg] \\
&\quad + \frac{1}{LC_L^2}\sum_l \Bigg[\left(\frac{L}{C_L}K_2 + \left\|\mathfrak{W}^{l+2}_L(x,0)\right\|\right)\left(\frac{L}{C_L}K_2 + \left\|\mathfrak{W}^{l+2}_L(x',0)\right\|\right) \\
&\qquad - \left\|\mathfrak{W}^{l+2}_L(x,0)\right\| \left\|\mathfrak{W}^{l+2}_L(x',0)\right\| \Bigg] \\
&\le \frac{1}{L}\sum_l \Bigg[\left(1 + \frac{L}{C_L}K_2 + \left\|\frac{1}{C_L}\sigma\left(f^l_{\theta^{(l)}(0)}(x)\right)\right\|\right)\left(1 + \frac{L}{C_L}K_2 + \left\|\frac{1}{C_L}\sigma\left(f^l_{\theta^{(l)}(0)}(x')\right)\right\|\right) \\
&\qquad \times \left(1 + \frac{L}{C_L}K_2 + \left\|\mathfrak{W}^{l+2}_L(x,0)\right\|\right)\left(1 + \frac{L}{C_L}K_2 + \left\|\mathfrak{W}^{l+2}_L(x',0)\right\|\right)\Bigg] \times \frac{L}{C_L}K_2 \\
&\quad + \frac{1}{LC_L^2}\sum_l \left(1 + \frac{L}{C_L}K_2 + \left\|\mathfrak{W}^{l+2}_L(x,0)\right\|\right)\left(1 + \frac{L}{C_L}K_2 + \left\|\mathfrak{W}^{l+2}_L(x',0)\right\|\right) \times \frac{L}{C_L}K_2 \\
&\le \epsilon''
\end{aligned}$$

Finally,

$$\Pr\left[\left\|\tilde{\Theta}^L_t(x,x') - \tilde{\Theta}^L_0(x,x')\right\| < \epsilon''\right] \ge 1 - 2\epsilon$$

for all $L \ge L'_{\max}$. Since ϵ, ϵ', and ϵ'' are arbitrary, we can conclude $\tilde{\Theta}^L_t(x,x') \xrightarrow{p} \tilde{\Theta}^\infty(x,x')$.

D. Stochastically bounded assumption of Theorem 2

Since the scaled NTK stays constant during training by Theorem 2, we can find an explicit solution to the differential equation when the loss function is quadratic loss. However, Theorem 2 requires stochastically bounded assumption of $\int_0^T \left\|\delta^L_t\right\|_{p^{in}} dt$, which is provided by the following lemma.

Lemma 16. *If $\mathcal{L}(f) = \|f - f^\star\|^2_{p^{in}}$, then $\int_0^T \left\|\delta^L_t\right\|_{p^{in}} dt$ is stochastically bounded for all $T > 0$.*

Proof of Lemma 16. Since $\Theta^L(x,x')$ is semidefinite by definition,

$$\begin{aligned}
\partial_t \mathcal{L}|_{f_t} &= -\mathbb{E}_{x\sim p^{in}}\left[\delta|_{f_t}(x)\partial_t f_t(x)\right] \\
&= -\mathbb{E}_{x,x'\sim p^{in}}\left[\delta|_{f_t}(x)^\intercal \Theta^L(x,x')\delta|_{f_t}(x')\right] \le 0.
\end{aligned}$$

This implies $\|f - f^\star\|^2_{p^{in}}$ is non increasing during training. Then $\left\|\delta^L_t\right\|_{p^{in}} = \|f - f^\star\|_{p^{in}}$ is also non increasing, and therefore $\int_0^T \left\|\delta^L_t\right\|_{p^{in}} dt$ is stochastically bounded for all $T > 0$. $\square$

E. Equalites and inequalities of sub layers for MLP

In this section, we introduce some useful equalities and inequalities in our problem setting.

First, we prove some property of $\|\cdot\|_{p^{in}}$.

Lemma 17. *$\|\cdot\|_{p^{in}}$ is seminorm for both matrix and vector.*

Proof. Since it is clear that $\|\cdot\|_{p^{in}}$ satisfies absolute homogeneity and nonnegativity, it is enough to show the triangle inequality (which suffices to prove for matrix only).

$$\begin{aligned}\|A(x)+B(x)\|_{p^{in}} &= \sqrt{\frac{1}{N}\sum_i \|A(x_i)+B(x_i)\|^2} \\ &\le \sqrt{\frac{1}{N}\sum_i (\|A(x_i)\|+\|B(x_i)\|)^2} \\ &\le \sqrt{\frac{1}{N}\sum_i \|A(x_i)\|^2} + \sqrt{\frac{1}{N}\sum_i \|B(x_i)\|^2} \\ &= \|A(x)\|_{p^{in}} + \|B(x)\|_{p^{in}}\end{aligned}$$

by Cauchy inequality. □

Lemma 18.

$$\|A(x)B(x)\|_{p^{in}} \le \sqrt{N}\,\|A(x)\|_{p^{in}}\,\|B(x)\|_{p^{in}}$$

Proof.

$$\begin{aligned}\|A(x)B(x)\|_{p^{in}} &= \sqrt{\frac{1}{N}\sum_i \|A(x_i)B(x_i)\|^2} \\ &\le \sqrt{\frac{1}{N}\sum_i \|A(x_i)\|^2 \max_i \|B(x_i)\|^2} \\ &\le \sqrt{N}\sqrt{\frac{1}{N}\sum_i \|A(x_i)\|^2}\sqrt{\frac{1}{N}\sum_i \|B(x_i)\|^2} \\ &= \sqrt{N}\,\|A(x)\|_{p^{in}}\,\|B(x)\|_{p^{in}}\end{aligned}$$

by the property of ℓ_2 norm. □

Note that if W does not depend on x, then $\|W\|_{p^{in}} = \|W\|$.

For simplicity, we use $f^l \equiv f^l_{\theta^{(l)}}$ if it is clear from the context. As previously mentioned, we set the scaling factor of gradient flow by $\frac{1}{(L-1)C_L^2}$ throughout the proof. Hence the gradient flow is given by

$$\partial_t \theta^l = -\frac{1}{(L-1)C_L^2}\mathbb{E}_{x\sim p^{in}}[(\partial_{\theta^l} f^l_{\theta^{(l)}(t)}(x))^\intercal \delta^l_t],$$

where $\delta^l_t = \left(\mathfrak{W}^{l+1}_L(x,t)\right)^\intercal \delta^L_t$. This is because

$$\partial_{\theta^l} f^L = \left(\prod_{j=L}^{l+1} W^j \operatorname{diag}(\dot\sigma(f^{j-1}))\right)\partial_{\theta^l} f^l \quad \text{for } L \ge l.$$

We use inverted indexing for $\prod$ to emphasize the order of product, i.e.,

$$\prod_{j=L}^{l+1} W^j \mathrm{diag}(\dot{\sigma}(f^{j-1})) = W^L \mathrm{diag}(\dot{\sigma}(f^{L-1})) \cdots W^{l+1}\mathrm{diag}(\dot{\sigma}(f^l))$$

Then, by the chain rule, we obtain the following lemma.

Lemma 19. *For $t \geq 0$ and $1 \leq l \leq L$,*

$$\partial_t W^l = -\frac{1}{(L-1)C_L^2}\mathbb{E}_{x\sim p^{in}}\left[\delta_t^l\left(\sigma(f^{l-1}_{\theta^{(l-1)}}(x))^\intercal\right)\right], \quad \partial_t b^l = -\frac{1}{(L-1)C_L^2}\mathbb{E}_{x\sim p^{in}}\left[\delta_t^l\right].$$

Then, the following lemma provides a bound of derivatives.

Lemma 20. *For any $t \geq 0$ and $1 \leq l \leq L$,*

$$\left\|\partial_t W^l\right\|_{p^{in}} \leq \frac{1}{(L-1)C_L^2}\left\|\delta_t^l\right\|_{p^{in}}\left\|f^{l-1}\right\|_{p^{in}}, \quad \left\|\partial_t b^l\right\|_{p^{in}} \leq \frac{1}{(L-1)C_L^2}\left\|\delta_t^l\right\|_{p^{in}}.$$

Proof of Lemma 20. From Lemma 19,

$$\begin{aligned}
\left\|\partial_t W^l\right\|_{p^{in}} &= \|\partial_t W^l\| \\
&= \left\|\frac{1}{(L-1)C_L^2}\mathbb{E}_{x\sim p^{in}}\left[\delta_t^l \sigma(f^{l-1})^\intercal\right]\right\| \\
&\leq \frac{1}{(L-1)C_L^2}\mathbb{E}_{x\sim p^{in}}\left[\|(\delta_t^l \sigma(f^{l-1})^\intercal\|\right] \\
&\leq \frac{1}{(L-1)C_L^2}\mathbb{E}_{x\sim p^{in}}\left[\|\delta_t^l\|\|\sigma(f^{l-1})\|\right] \\
&\leq \frac{1}{(L-1)C_L^2}\sqrt{\mathbb{E}_{x\sim p^{in}}\left[\|\delta_t^l\|^2\right]}\sqrt{\mathbb{E}_{x\sim p^{in}}\left[\|\sigma(f^{l-1})\|^2\right]} \\
&= \frac{1}{(L-1)C_L^2}\left\|\delta_t^l\right\|_{p^{in}}\left\|\sigma(f^{l-1})\right\|_{p^{in}} \\
&\leq \frac{1}{(L-1)C_L^2}\left\|\delta_t^l\right\|_{p^{in}}\left\|f^{l-1}\right\|_{p^{in}},
\end{aligned}$$

where the first inequality is from Jensen's inequality and third inequality comes from Cauchy's inequality. Finally, we can bound $\left\|\partial_t b^l\right\|_{p^{in}}$ similarly. □

The gradient flow also implies

$$\partial_t f^l_{\theta^{(l)}(t)}(x') = -\frac{1}{(L-1)C_L^2}\mathbb{E}_{x\sim p^{in}}[\partial_{\theta^l} f^l_{\theta^{(l)}(t)}(x')(\partial_{\theta^l} f^l_{\theta^{(l)}(t)}(x))^\intercal \delta_t^l(x)],$$

where the following lemma provides a bound of the norm.

Lemma 21. *For $1 \leq l \leq L$ and $t \geq 0$,*

$$\left\|\partial_t f^l(x')\right\|_{p^{in}} \leq \frac{N}{(L-1)C_L^2}\mathbb{E}_{x,x'\sim p^{in}}\left[\left\|\partial_\theta f^l_\theta(x')\left(\partial_\theta f^l_\theta(x)\right)^\intercal \delta_t^l(x)\right\|\right].$$

Proof of Lemma 21. From previous gradient flow,

$$
\begin{aligned}
\left\|\partial_t f^l(x')\right\|_{p^{in}} &= \sqrt{\mathbb{E}_{x'\sim p^{in}}\left[\|\partial_{\theta^l} f^l_{\theta^l}(x')\frac{1}{(L-1)C_L^2}\mathbb{E}_{x\sim p^{in}}\left[(\partial_{\theta^l} f^l_{\theta^l}(x))^\intercal \delta^l_t(x)\right]\|^2\right]} \\
&= \frac{1}{(L-1)C_L^2}\sqrt{\mathbb{E}_{x'\sim p^{in}}\left[\left\|\mathbb{E}_{x\sim p^{in}}\left[\partial_{\theta^l} f^l_{\theta^l}(x')(\partial_{\theta^l} f^l_{\theta^l}(x))^\intercal \delta^l_t(x)\right]\right\|^2\right]} \\
&\le \frac{1}{(L-1)C_L^2}\sqrt{\frac{1}{N}\mathbb{E}_{x'\sim p^{in}}\left[\sum_i \left\|\partial f^l_{\theta^l}(x')(\partial_{\theta^l} f^l(x_i))^\intercal \delta^l_t(x)\right\|^2\right]} \\
&\le \frac{1}{(L-1)C_L^2}\sqrt{\frac{1}{N^2}\sum_j\sum_i \left\|\partial f^l_{\theta^l}(x'_j)(\partial_{\theta^l} f^l(x_i))^\intercal \delta^l_t(x_i)\right\|^2} \\
&\le \frac{1}{(L-1)C_L^2}\frac{1}{N}\sum_j\sum_i \left\|\partial f^l_{\theta^l}(x'_j)(\partial_{\theta^l} f^l(x_i))^\intercal \delta^l_t(x_i)\right\| \\
&\le \frac{N}{(L-1)C_L^2}\mathbb{E}_{x,x'\sim p^{in}}\left[\left\|\partial_{\theta^l} f^l(x')(\partial_{\theta^l} f^l(x))^\intercal \delta^l_t(x)\right\|\right].
\end{aligned}
$$

□

The following is the key lemma to prove invariances of Lyapunov functions.

Lemma 22. *Suppose $f^i_{\theta^{(i)}(t)}(x')$ is element-wise nonzero for $1 \le i \le L-1$ and x' in dataset. Then, for $t \in [0,T]$, $1 \le k \le j \le L$, and $1 \le i \le L$,*

$$
\partial_t \left\|\mathfrak{W}^k_j(x,t)\right\|_{p^{in}} \le \frac{N}{(L-1)C_L^2}\sum_{i=k}^{j}\left\|\mathfrak{W}^{i+1}_j(x,t)\right\|_{p^{in}}\left\|\mathfrak{W}^k_{i-1}(x,t)\right\|_{p^{in}}\left\|\delta^i_t\right\|_{p^{in}}\left\|\sigma\left(f^{i-1}_{\theta^{(i-1)}(t)}(x)\right)\right\|_{p^{in}}
$$

$$
\begin{aligned}
\partial_t \left\|f^i_{\theta^{(i)}(t)}(x')\right\|_{p^{in}} \le{}& \frac{N^{3/2}}{(L-1)C_L^2}\sum_{l=0}^{i-1}\left\|f^l_{\theta^{(l)}(t)}(x)\right\|_{p^{in}}\left\|f^l_{\theta^{(l)}(t)}(x')\right\|_{p^{in}}\left\|\mathfrak{W}^{l+2}_i\right\|_{p^{in}}\left\|(\mathfrak{W}^{l+2}_L)^\intercal\right\|_{p^{in}}\left\|\delta^L_t\right\|_{p^{in}} \\
&+\frac{N}{(L-1)C_L^2}\sum_{l=0}^{i-1}\left\|\mathfrak{W}^{l+2}_i\right\|_{p^{in}}\left\|(\mathfrak{W}^{l+2}_L)^\intercal\right\|_{p^{in}}\left\|\delta^L_t\right\|_{p^{in}}.
\end{aligned}
$$

Proof of Lemma 22. By the chain rule,

$$
\begin{aligned}
\partial_t \left\|\mathfrak{W}^k_j(x,t)\right\|_{p^{in}} &\le \left\|\partial_t \prod_{l=j}^{k} W^l(t)\mathrm{diag}\left(\dot\sigma\left(f^{l-1}_{\theta^{(l-1)}(t)}(x)\right)\right)\right\|_{p^{in}} \\
&\le \sum_{i=k}^{j}\left\|\mathfrak{W}^{i+1}_j \partial_t W^i(t)\mathrm{diag}(\dot\sigma(f^{i-1}_{\theta^{(i-1)}(t)}(x)))\mathfrak{W}^k_{i-1}\right\|_{p^{in}},
\end{aligned}
$$

where $\mathfrak{W}^{i+1}_j$ denotes $\mathfrak{W}^{i+1}_j(x,t)$ if it is clear from the context. Then,

$$
\begin{aligned}
\partial_t \left\|\mathfrak{W}^k_j(x,t)\right\|_{p^{in}} &\le \sum_{i=k}^{j}\left\|\mathfrak{W}^{i+1}_j \frac{1}{(L-1)C_L^2}\mathbb{E}_{x\sim p^{in}}\left[\delta^i_t \sigma(f^{i-1}_{\theta^{(i-1)}(t)}(x))^\intercal\right]\mathrm{diag}(\dot\sigma(f^{i-1}_{\theta^{(i-1)}(t)}(x)))\mathfrak{W}^k_{i-1}\right\|_{p^{in}} \\
&\le \frac{N}{(L-1)C_L^2}\sum_{i=k}^{j}\left\|\mathfrak{W}^{i+1}_j\right\|_{p^{in}}\left\|\mathbb{E}_{x\sim p^{in}}\left[\delta^i_t \sigma(f^{i-1}_{\theta^{(i-1)}(t)}(x))^\intercal\right]\right\|\left\|\mathrm{diag}(\dot\sigma(f^{i-1}_{\theta^{(i-1)}(t)}(x)))\right\|_{p^{in}}\left\|\mathfrak{W}^k_{i-1}\right\|_{p^{in}} \\
&\le \frac{N}{(L-1)C_L^2}\sum_{i=k}^{j}\left\|\mathfrak{W}^{i+1}_j(x,t)\right\|_{p^{in}}\left\|\mathfrak{W}^k_{i-1}(x,t)\right\|_{p^{in}}\left\|\delta^i_t\right\|_{p^{in}}\left\|\sigma(f^{i-1}_{\theta^{(i-1)}(t)}(x))\right\|_{p^{in}},
\end{aligned}
$$

where first inequality comes from Lemma 20.

On the other hand, from Lemma 21 and Corollary 9,

$$\begin{aligned}
&\left\|\partial_t f^l_{\theta^{(l)}(t)}(x)\right\|_{p^{in}} \\
&\leq \frac{N}{(L-1)C_L^2}\mathbb{E}_{x,x'\sim p^{in}}\left[\left\|\partial_{\theta^l} f^l_{\theta^{(l)}(t)}(x')\left(\partial_{\theta^l} f^l_{\theta^{(l)}(t)}(x)\right)^\intercal \left(\mathfrak{W}_L^{l+1}(x)\right)^\intercal \delta_t^L(x)\right\|\right] \\
&\leq \frac{N}{(L-1)C_L^2}\mathbb{E}_{x,x'\sim p^{in}}\left[\sum_{i=0}^{l-1}\left\|(\sigma(f^i_{\theta^{(i)}(t)}(x'))^\intercal\sigma(f^i_{\theta^{(i)}(t)}(x))+1)\mathfrak{W}_l^{i+2}(x')(\mathfrak{W}_l^{i+2}(x))^\intercal(\mathfrak{W}_L^{l+1}(x))^\intercal\delta_t^L(x)\right\|\right] \\
&\leq \frac{N}{(L-1)C_L^2}\sum_{i=0}^{l-1}\mathbb{E}_{x,x'\sim p^{in}}\left\|\sigma(f^i(x'))^\intercal\sigma(f^i(x))\mathfrak{W}_l^{i+2}(x')(\mathfrak{W}_L^{i+2}(x))^\intercal\delta_t^L(x)\right\| \\
&\quad+\frac{N}{(L-1)C_L^2}\sum_{i=0}^{l-1}\mathbb{E}_{x,x'\sim p^{in}}\left[\left\|\mathfrak{W}_l^{i+2}(x')(\mathfrak{W}_L^{i+2}(x))^\intercal\delta_t^L(x)\right\|\right] \\
&\leq \frac{N}{(L-1)C_L^2}\sum_{i=0}^{l-1}\mathbb{E}_{x,x'\sim p^{in}}\left[\|(\sigma(f^i(x))\|\|\sigma(f^i(x'))\|\|\mathfrak{W}_l^{i+2}(x')\|\|(\mathfrak{W}_L^{i+2}(x))^\intercal\|\|\delta_t^L(x)\|\right] \\
&\quad+\frac{N}{(L-1)C_L^2}\sum_{i=0}^{l-1}\mathbb{E}_{x,x'\sim p^{in}}\left[\|\mathfrak{W}_l^{i+2}(x')\|\|(\mathfrak{W}_L^{i+2}(x))^\intercal\|\|\delta_t^L(x)\|\right] \\
&\leq \frac{N^{3/2}}{(L-1)C_L^2}\sum_{i=0}^{l-1}\left\|\sigma(f^i(x))\right\|_{p^{in}}\left\|\sigma(f^i(x'))\right\|_{p^{in}}\left\|\mathfrak{W}_l^{i+2}\right\|_{p^{in}}\left\|(\mathfrak{W}_L^{i+2})^\intercal\right\|_{p^{in}}\left\|\delta_t^L(x)\right\|_{p^{in}} \\
&\quad+\frac{N}{(L-1)C_L^2}\sum_{i=0}^{l-1}\left\|\mathfrak{W}_l^{i+2}\right\|_{p^{in}}\left\|(\mathfrak{W}_L^{i+2})^\intercal\right\|_{p^{in}}\left\|\delta_t^L(x)\right\|_{p^{in}},
\end{aligned}$$

where the last inequality holds due to the property of p^{in}-norm

$$\mathbb{E}_{x\sim p^{in}}\left[\|(\sigma(f^i(x))\|\|(\mathfrak{W}_L^{i+2})^\intercal\|\|\delta_t^L(x)\|\right]\leq\sqrt{N}\left\|(\sigma(f^i(x))\right\|_{p^{in}}\left\|(\mathfrak{W}_L^{i+2}(x))^\intercal\right\|_{p^{in}}\left\|\delta_t^L(x)\right\|_{p^{in}}.$$

□

We have a similar bounds for derivatives of ℓ_2-norms.

Lemma 23. *For $x'\in\mathbb{R}_+^{d_{in}}$, if all elements of $f^i_{\theta^{(i)}(t)}(x')$ are nonzero for $1\leq i\leq L-1$, then, for $t\in[0,T]$, $1\leq l\leq j\leq L$, and $1\leq i\leq L$*

$$\begin{aligned}
\partial_t\left\|\mathfrak{W}_l^k(x',t)\right\| &\leq \frac{1}{(L-1)C_L^2}\sum_{l\geq i\geq k}\left\|\mathfrak{W}_l^{i+1}(x')\right\|\left\|\mathfrak{W}_{i-1}^k(x')\right\|\left\|\delta_t^i\right\|_{p^{in}}\left\|\sigma(f^{i-1})^\intercal\right\|_{p^{in}} \\
\partial_t\left\|f^i_{\theta^{(l)}(t)}(x')\right\| &\leq \frac{N^{1/2}}{(L-1)C_L^2}\sum_{i=0}^{l-1}\left\|\sigma(f^i(x))\right\|_{p^{in}}\left\|\sigma(f^i(x'))\right\|\left\|\mathfrak{W}_l^{i+2}(x')\right\|\left\|(\mathfrak{W}_L^{i+2})^\intercal\right\|_{p^{in}}\left\|\delta_t^L(x)\right\|_{p^{in}} \\
&\quad+\frac{1}{(L-1)C_L^2}\sum_{i=0}^{l-1}\left\|\mathfrak{W}_l^{i+2}(x')\right\|\left\|(\mathfrak{W}_L^{i+2})^\intercal\right\|_{p^{in}}\left\|\delta_t^L(x)\right\|_{p^{in}}.
\end{aligned}$$

Proof of Lemma 23. The proof is similar to that of Lemma 22.

$$
\begin{aligned}
\partial_t \left\|\mathfrak{W}_l^k\right\| &\le \left\|\partial_t \mathfrak{W}_l^k\right\| \\
&\le \sum_{l\ge i\ge k} \left\| \mathfrak{W}_l^{i+1} \frac{1}{(L-1)C_L^2} \mathbb{E}_{x\sim p^{in}} \left[\delta_t^i \sigma(f^{i-1})^\intercal\right] \mathrm{diag}(\dot\sigma(f^{i-1}(t)))\mathfrak{W}_{i-1}^k \right\| \\
&\le \frac{1}{(L-1)C_L^2} \sum_{l\ge i\ge k} \left\|\mathfrak{W}_l^{i+1}\right\| \left\|\mathbb{E}_{x\sim p^{in}} \left[\delta_t^i \sigma(f^{i-1})^\intercal\right]\right\| \left\|\mathrm{diag}(\dot\sigma(f^{i-1}(t)))\right\| \left\|\mathfrak{W}_{i-1}^k\right\| \\
&\le \frac{1}{(L-1)C_L^2} \sum_{l\ge i\ge k} \left\|\mathfrak{W}_l^{i+1}\right\| \left\|\mathfrak{W}_{i-1}^k\right\| \left\|(\delta_t^i)\right\|_{p^{in}} \left\|\sigma(f^{i-1})^\intercal\right\|_{p^{in}} .
\end{aligned}
$$

On the other hand,

$$
\begin{aligned}
\left\|\partial_t f^l(t)\right\| &\le \frac{1}{(L-1)C_L^2} \left[\left\| \partial_{\theta^l} f^l(x') \mathbb{E}_{x\sim p^{in}} \left[\left(\partial f_{\theta^l}^l(x)\right)^\intercal \left(\mathfrak{W}_L^{l+1}(x)\right)^\intercal \delta_t^L(x)\right]\right\|\right] \\
&\le \frac{1}{(L-1)C_L^2} \mathbb{E}_{x\sim p^{in}} \left[\sum_{i=0}^{l-1} \left\| \left(\left(\sigma(f^i(x'))\right)^\intercal \sigma(f^i(x)) + 1\right) \mathfrak{W}_l^{i+2}(x') (\mathfrak{W}_L^{i+2}(x))^\intercal \delta_t^L(x) \right\| \right] \\
&\le \frac{1}{(L-1)C_L^2} \sum_{i=0}^{l-1} \mathbb{E}_{x\sim p^{in}} \left\|\sigma(f^i(x'))^\intercal\right\| \left\|\sigma(f^i(x))\right\| \left\|\mathfrak{W}_l^{i+2}(x')\right\| \left\|(\mathfrak{W}_L^{i+2}(x))^\intercal\right\| \left\|\delta_t^L(x)\right\| \\
&\quad + \frac{1}{(L-1)C_L^2} \sum_{i=0}^{l-1} \mathbb{E}_{x\sim p^{in}} \left[\left\|\mathfrak{W}_l^{i+2}(x')\right\| \left\|(\mathfrak{W}_L^{i+2}(x))^\intercal\right\| \left\|\delta_t^L(x)\right\|\right] \\
&\le \frac{N^{1/2}}{(L-1)C_L^2} \sum_{i=0}^{l-1} \left\|\sigma(f^i(x)\right\|_{p^{in}} \left\|\sigma(f^i(x')\right\| \left\|\mathfrak{W}_l^{i+2}(x')\right\| \left\|(\mathfrak{W}_L^{i+2}(x))^\intercal\right\|_{p^{in}} \left\|\delta_t^L(x)\right\|_{p^{in}} \\
&\quad + \frac{1}{(L-1)C_L^2} \sum_{i=0}^{l-1} \left\|\mathfrak{W}_l^{i+2}(x')\right\| \left\|(\mathfrak{W}_L^{i+2})^\intercal\right\|_{p^{in}} \left\|\delta_t^L(x)\right\|_{p^{in}} .
\end{aligned}
$$

□

F. Proof of Lemma 10 for Proposition 1

Lemma 10. *For $t \ge 0$, if $f^l_{\theta^{(l)}(t)}(x)$ is element-wise nonzero for $1 \le l \le L-1$ and $x \in \{x_1, \ldots, x_N\}$, then*

$$
\begin{aligned}
\partial_t \Phi_{L,j}(t) &\le \frac{2jN^{3/2}(L-1)}{C_L} \Phi_{L,j-1}(t)\Phi_{L,8}(t)\Psi_{L,2}(t)\Psi_{L,8}(t) \left\|\delta_t^L\right\|_{p^{in}} \\
\partial_t \Psi_{L,2}(t) &\le \frac{2N(L-1)}{C_L} \Phi_{L,4}(t) \left(\Psi_{L,2}(t)\right)^2 (t)\Psi_{L,8}(t) \left\|\delta_t^L\right\|_{p^{in}} \\
\partial_t \Psi_{L,8}(t) &\le \frac{8N(L-1)}{C_L} \Phi_{L,4}(t)\Psi_{L,2}(t) \left(\Psi_{L,8}(t)\right)^2 \left\|\delta_t^L\right\|_{p^{in}} ,
\end{aligned}
$$

for $1 \le j \le 8$.

Proof of Lemma 10. Consider the derivative of $\Psi_{L,8}$. In Section C, we show that the intermediate pre-activation values never reach zero while training, and therefore we can apply Lemma 22.

$$
\begin{aligned}
&\partial_t \Psi_{L,8}(t) \\
&\le \frac{8N}{(L-1)^2 C_L^2} \sum_{l=2}^{L} \sum_{l=i}^{L} \left\|\mathfrak{W}_L^{l+1}(t)\right\|_{p^{in}}^2 \left\|\mathfrak{W}_{l-1}^i(t)\right\|_{p^{in}} \left(\left\|\mathfrak{W}_L^i(0)\right\|_{p^{in}} + \left\|\mathfrak{W}_L^i(t) - \mathfrak{W}_L^i(0)\right\|_{p^{in}}\right)^7 \\
&\quad \times \left\| f^{l-1}_{\theta^{(l-1)}(t)}(x)\right\|_{p^{in}} \left\|\delta_t^L\right\|_{p^{in}} .
\end{aligned}
$$

First, the following rough bound holds.

$$\begin{aligned}
&\sum_{l=2}^{L}\sum_{l=i}^{L}\left\|\mathfrak{W}_L^{l+1}(t)\right\|_{p^{in}}^2\left\|\mathfrak{W}_{l-1}^i(t)\right\|_{p^{in}}\left(\left\|\mathfrak{W}_L^i(0)\right\|_{p^{in}}+\left\|\mathfrak{W}_L^i(t)-\mathfrak{W}_L^i(0)\right\|_{p^{in}}\right)^7\left\|f_{\theta^{(l-1)}(t)}^{l-1}(x)\right\|_{p^{in}}\\
&\le\left(\sum_{i=2}^{L}\sum_{l=i}^{L}\left\|\mathfrak{W}_{l-1}^i(t)\right\|_{p^{in}}\left(\left\|f_{\theta^{(l-1)}(0)}^{l-1}\right\|_{p^{in}}+\left\|f_{\theta^{(l-1)}(t)}^{l-1}-f_{\theta^{(l-1)}(0)}^{l-1}\right\|_{p^{in}}\right)\right.\\
&\quad\left.\times\left(\left\|\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}+\left\|\mathfrak{W}_L^{l+1}(t)-\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}\right)^2\right)\sum_{i=2}^{L}\left(\left\|\mathfrak{W}_L^i(0)\right\|_{p^{in}}+\left\|\mathfrak{W}_L^i(t)-\mathfrak{W}_L^i(0)\right\|_{p^{in}}\right)^8\\
&\le(L-1)\Psi_{L,8}(L-1)^2C_L\left(\Psi_{L,2}\right)^{(1/2)}\left(\Phi_{L,4}\right)^{(1/4)}\left(\Psi_{L,8}\right)^{(1/4)},
\end{aligned}$$

where the last inequality is from

$$\begin{aligned}
&\left(\sum_{i=2}^{L}\sum_{l=i}^{L}\left\|\mathfrak{W}_{l-1}^i\right\|_{p^{in}}\left(\left\|f_{\theta^{(l-1)}(0)}^{l-1}\right\|_{p^{in}}+\left\|f_{\theta^{(l-1)}(t)}^{l-1}-f_{\theta^{(l-1)}(0)}^{l-1}\right\|_{p^{in}}\right)\left(\left\|\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}+\left\|\mathfrak{W}_L^{l+1}(t)-\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}\right)^2\right)^2\\
&\le\left(\sum_{i=2}^{L}\sum_{l=i}^{L}\frac{L-1}{l-1}\left\|\mathfrak{W}_l^i\right\|_{p^{in}}^2\right)\\
&\quad\times\left((L-1)\sum_{l=2}^{L}\left(\left\|f_{\theta^{(l-1)}(0)}^{l-1}\right\|_{p^{in}}+\left\|f_{\theta^{(l-1)}(t)}^{l-1}-f_{\theta^{(l-1)}(0)}^{l-1}\right\|_{p^{in}}\right)^2\left(\left\|\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}+\left\|\mathfrak{W}_L^{l+1}(t)-\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}\right)^4\right)\\
&\le(L-1)^2\Psi_{L,2}(L-1)\\
&\quad\times\sqrt{\left(\sum_{l=1}^{L-1}\left(\left\|f_{\theta^{(l)}(0)}^{l}\right\|_{p^{in}}+\left\|f_{\theta^{(l)}(t)}^{l}-f_{\theta^{(l)}(0)}^{l}\right\|_{p^{in}}\right)^4\right)\left(\sum_{l=2}^{L}\left(\left\|\mathfrak{W}_L^l(0)\right\|_{p^{in}}+\left\|\mathfrak{W}_L^l(t)-\mathfrak{W}_L^l(0)\right\|_{p^{in}}\right)^8\right)}\\
&\le(L-1)^2\Psi_{L,2}(L-1)\sqrt{(L-1)^2C_L^4\Phi_{L,4}\Psi_{L,8}}.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\partial_t\Psi_{L,8}(t)\\
&\le\frac{8N}{(L-1)^2C_L^2}\sum_{l=2}^{L}\sum_{l=i}^{L}\left\|\mathfrak{W}_L^{l+1}(t)\right\|_{p^{in}}^2\left\|\mathfrak{W}_{l-1}^i(t)\right\|_{p^{in}}\left(\left\|\mathfrak{W}_L^i(0)\right\|_{p^{in}}+\left\|\mathfrak{W}_L^i(t)-\mathfrak{W}_L^i(0)\right\|_{p^{in}}\right)^7\\
&\quad\times\left\|f_{\theta^{(l-1)}(t)}^{l-1}(x)\right\|_{p^{in}}\left\|\delta_t^L\right\|_{p^{in}}\\
&\le\frac{8N(L-1)}{C_L}\left(\Psi_{L,2}\right)^{1/2}\left(\Phi_{L,4}\right)^{1/4}\left(\Psi_{L,8}\right)^{5/4}\left\|\delta_t^L\right\|_{p^{in}}\\
&\le\frac{8N(L-1)}{C_L}\Psi_{L,2}\Phi_{L,4}\left(\Psi_{L,8}\right)^2\left\|\delta_t^L\right\|_{p^{in}}.
\end{aligned}$$

Then, let consider the derivative of $\Psi_{L,2}$. From Lemma 22,

$$
\begin{aligned}
&\partial_t \Psi_{L,2}(t) \\
&\leq \frac{2N}{(L-1)^3 C_L^2} \sum_{i=2}^{L} \sum_{k=2}^{i} \sum_{l=k}^{i} \frac{L-1}{i-1} \left\|\mathfrak{W}_L^{l+1}\right\|_{p^{in}} \left\|\mathfrak{W}_i^{l+1}\right\|_{p^{in}} \left\|\mathfrak{W}_{l-1}^{k}\right\|_{p^{in}} \left(\left\|\mathfrak{W}_i^{k}(0)\right\|_{p^{in}} + \left\|\mathfrak{W}_i^{k}(t) - \mathfrak{W}_i^{k}(0)\right\|_{p^{in}}\right) \\
&\quad \times \left\|f^{l-1}_{\theta^{(l-1)}(t)}\right\|_{p^{in}} \left\|\delta_t^L\right\|_{p^{in}} \\
&\leq \frac{N}{(L-1)^3 C_L^2} \sum_{i=2}^{L} \sum_{k=2}^{i} \sum_{l=k}^{i} \frac{L-1}{i-1} \left\|\mathfrak{W}_L^{l+1}\right\|_{p^{in}} \left\|\mathfrak{W}_{l-1}^{k}\right\|_{p^{in}} \left(\left\|\mathfrak{W}_i^{l+1}\right\|_{p^{in}}^2 + \left(\left\|\mathfrak{W}_i^{k}(0)\right\|_{p^{in}} + \left\|\mathfrak{W}_i^{k}(t) - \mathfrak{W}_i^{k}(0)\right\|_{p^{in}}\right)^2\right) \\
&\quad \times \left\|f^{l-1}_{\theta^{(l-1)}(t)}\right\|_{p^{in}} \left\|\delta_t^L\right\|_{p^{in}} \\
&= \frac{N}{(L-1)^3 C_L^2} \sum_{i=2}^{L} \sum_{k=2}^{i} \sum_{l=k}^{i} \frac{L-1}{i-1} \left\|\mathfrak{W}_L^{l+1}\right\|_{p^{in}} \left\|\mathfrak{W}_{l-1}^{k}\right\|_{p^{in}} \left\|\mathfrak{W}_i^{l+1}\right\|_{p^{in}}^2 \left\|f^{l-1}_{\theta^{(l-1)}(t)}\right\|_{p^{in}} \left\|\delta_t^L\right\|_{p^{in}} \\
&\quad + \frac{N}{(L-1)^3 C_L^2} \sum_{i=2}^{L} \sum_{k=2}^{i} \sum_{l=k}^{i} \frac{L-1}{i-1} \left\|\mathfrak{W}_L^{l+1}\right\|_{p^{in}} \left\|\mathfrak{W}_{l-1}^{k}\right\|_{p^{in}} \left(\left\|\mathfrak{W}_i^{k}(0)\right\|_{p^{in}} + \left\|\mathfrak{W}_i^{k}(t) - \mathfrak{W}_i^{k}(0)\right\|_{p^{in}}\right)^2 \\
&\quad \times \left\|f^{l-1}_{\theta^{(l-1)}(t)}\right\|_{p^{in}} \left\|\delta_t^L\right\|_{p^{in}}.
\end{aligned}
$$

From Cauchy-Schwartz inequality,

$$
\begin{aligned}
&\left(\sum_{k=2}^{L} \sum_{l=k}^{L} \left\|\mathfrak{W}_{l-1}^{k}\right\|_{p^{in}} \left(\left\|f^{l-1}_{\theta^{(l-1)}(0)}\right\|_{p^{in}} + \left\|f^{l-1}_{\theta^{(l-1)}(t)} - f^{l-1}_{\theta^{(l-1)}(0)}\right\|_{p^{in}}\right) \left(\left\|\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}} + \left\|\mathfrak{W}_L^{l+1}(t) - \mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}\right)\right)^2 \\
&\leq \left(\sum_{k=2}^{L} \sum_{l=k}^{L} \frac{L-1}{l-1} \left\|\mathfrak{W}_l^{k}\right\|_{p^{in}}^2\right) \\
&\quad \times \left((L-1) \sum_{l=2}^{L} \left(\left\|f^{l-1}_{\theta^{(l-1)}(0)}\right\|_{p^{in}} + \left\|f^{l-1}_{\theta^{(l-1)}(t)} - f^{l-1}_{\theta^{(l-1)}(0)}\right\|_{p^{in}}\right)^2 \left(\left\|\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}} + \left\|\mathfrak{W}_L^{l+1}(t) - \mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}\right)^2\right) \\
&\leq (L-1)^2 \Psi_{L,2} (L-1) \\
&\quad \times \sqrt{\left(\sum_{l=1}^{L-1} \left(\left\|f^{l}_{\theta^{(l)}(0)}\right\|_{p^{in}} + \left\|f^{l}_{\theta^{(l)}(t)} - f^{l}_{\theta^{(l)}(0)}\right\|_{p^{in}}\right)^4\right) \left(\sum_{l=2}^{L} \left(\left\|\mathfrak{W}_L^{l}(0)\right\|_{p^{in}} + \left\|\mathfrak{W}_L^{l}(t) - \mathfrak{W}_L^{l}(0)\right\|_{p^{in}}\right)^4\right)} \\
&\leq (L-1)^2 \Psi_{L,2} (L-1) \\
&\quad \times \sqrt{\left(\sum_{l=1}^{L-1} \left(\left\|f^{l}_{\theta^{(l)}(0)}\right\|_{p^{in}} + \left\|f^{l}_{\theta^{(l)}(t)} - f^{l}_{\theta^{(l)}(0)}\right\|_{p^{in}}\right)^4\right) \left(\sum_{l=2}^{L} \left(\left\|\mathfrak{W}_L^{l}(0)\right\|_{p^{in}} + \left\|\mathfrak{W}_L^{l}(t) - \mathfrak{W}_L^{l}(0)\right\|_{p^{in}}\right)^8\right)} \\
&\leq (L-1)^2 \Psi_{L,2} (L-1) \sqrt{(L-1)^2 C_L^4 \Phi_{L,4} \Psi_{L,8}} \\
&= (L-1)^4 C_L^2 \Psi_{L,2} \left(\Phi_{L,4}\right)^{(1/2)} \left(\Psi_{L,8}\right)^{(1/2)}.
\end{aligned}
$$

Using the above inequality, we have the following upper bounds:

$$\begin{aligned}
&\sum_{i=2}^{L}\sum_{k=2}^{i}\sum_{l=k}^{i}\frac{L-1}{i-1}\left\|\mathfrak{W}_L^{l+1}\right\|_{p^{in}}\left\|\mathfrak{W}_{l-1}^{k}\right\|_{p^{in}}\left\|\mathfrak{W}_i^{l+1}\right\|_{p^{in}}^2\left\|f_{\theta^{(l-1)}(t)}^{l-1}\right\|_{p^{in}}\\
&\leq\left(\sum_{k=2}^{L}\sum_{l=k}^{L}\left\|\mathfrak{W}_{l-1}^{k}\right\|_{p^{in}}\left(\left\|f_{\theta^{(l-1)}(0)}^{l-1}\right\|_{p^{in}}+\left\|f_{\theta^{(l-1)}(t)}^{l-1}-f_{\theta^{(l-1)}(0)}^{l-1}\right\|_{p^{in}}\right)\right.\\
&\quad\left.\times\left(\left\|\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}+\left\|\mathfrak{W}_L^{l+1}(t)-\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}\right)\right)\left(\sum_{i=2}^{L}\sum_{l=2}^{i}\frac{L-1}{i-1}\left\|\mathfrak{W}_i^{l+1}\right\|_{p^{in}}^2\right)\\
&\leq(L-1)^2\Psi_{L,2}(L-1)^2C_L\left(\Psi_{L,2}\right)^{(1/2)}\left(\Phi_{L,4}\right)^{(1/4)}\left(\Psi_{L,8}\right)^{(1/4)}\\
&=(L-1)^4C_L\left(\Psi_{L,2}\right)^{(3/2)}\left(\Phi_{L,4}\right)^{(1/4)}\left(\Psi_{L,8}\right)^{(1/4)}.
\end{aligned}$$

and

$$\begin{aligned}
&\sum_{i=2}^{L}\sum_{k=2}^{i}\sum_{l=k}^{i}\frac{L-1}{i-1}\left\|\mathfrak{W}_L^{l+1}\right\|_{p^{in}}\left\|\mathfrak{W}_{l-1}^{k}\right\|_{p^{in}}\left(\left\|\mathfrak{W}_i^{k}(0)\right\|_{p^{in}}+\left\|\mathfrak{W}_i^{k}(t)-\mathfrak{W}_i^{k}(0)\right\|_{p^{in}}\right)^2\left\|f_{\theta^{(l-1)}(t)}^{l-1}\right\|_{p^{in}}\\
&\leq\left(\sum_{k=2}^{L}\sum_{l=k}^{L}\left\|\mathfrak{W}_{l-1}^{k}\right\|_{p^{in}}\left(\left\|f_{\theta^{(l-1)}(0)}^{l-1}\right\|_{p^{in}}+\left\|f_{\theta^{(l-1)}(t)}^{l-1}-f_{\theta^{(l-1)}(0)}^{l-1}\right\|_{p^{in}}\right)\right.\\
&\quad\left.\times\left(\left\|\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}+\left\|\mathfrak{W}_L^{l+1}(t)-\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}\right)\right)\left(\sum_{i=2}^{L}\sum_{k=2}^{i}\frac{L-1}{i-1}\left(\left\|\mathfrak{W}_i^{k}(0)\right\|_{p^{in}}+\left\|\mathfrak{W}_i^{k}(t)-\mathfrak{W}_i^{k}(0)\right\|_{p^{in}}\right)^2\right)\\
&\leq(L-1)^2\Psi_{L,2}(L-1)^2C_L\left(\Psi_{L,2}\right)^{(1/2)}\left(\Phi_{L,4}\right)^{(1/4)}\left(\Psi_{L,8}\right)^{(1/4)}\\
&=(L-1)^4C_L\left(\Psi_{L,2}\right)^{(3/2)}\left(\Phi_{L,4}\right)^{(1/4)}\left(\Psi_{L,8}\right)^{(1/4)}.
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&\partial_t\Psi_L^{(2)}(t)\\
&\leq\frac{N}{(L-1)^3C_L^2}\sum_{i=2}^{L}\sum_{k=2}^{i}\sum_{l=k}^{i}\frac{L-1}{i-1}\left\|\mathfrak{W}_L^{l+1}\right\|_{p^{in}}\left\|\mathfrak{W}_{l-1}^{k}\right\|_{p^{in}}\left(\left\|\mathfrak{W}_i^{l+1}\right\|_{p^{in}}^2+\left(\left\|\mathfrak{W}_i^{k}(0)\right\|_{p^{in}}+\left\|\mathfrak{W}_i^{k}(t)-\mathfrak{W}_i^{k}(0)\right\|_{p^{in}}\right)^2\right)\\
&\quad\times\left\|f_{\theta^{(l-1)}(t)}^{l-1}\right\|_{p^{in}}\left\|\delta_t^L\right\|_{p^{in}}\\
&\leq\frac{N2(L-1)}{C_L}\left(\Psi_{L,2}\right)^{(3/2)}\left(\Phi_{L,4}\right)^{(1/4)}\left(\Psi_{L,8}\right)^{(1/4)}\left\|\delta_t^L\right\|_{p^{in}}\\
&\leq\frac{N2(L-1)}{C_L}\left(\Psi_{L,2}\right)^2\Phi_{L,4}\Psi_{L,8}\left\|\delta_t^L\right\|_{p^{in}}.
\end{aligned}$$

Finally, let consider the derivatives of $\Phi_{L,j}$ for $1\leq j\leq 8$. From Lemma 22,

$$\begin{aligned}
\partial_t\Phi_{L,j}(t)\leq{}&\frac{jN^2}{(L-1)^2C_L^{2+j}}\sum_{i=1}^{L-1}\sum_{l=0}^{i-1}\left(\left\|f_{\theta^{(i)}(0)}^{i}\right\|_{p^{in}}+\left\|f_{\theta^{(i)}(t)}^{i}-f_{\theta^{(i)}(0)}^{i}\right\|_{p^{in}}\right)^{j-1}\left\|f_{\theta^{(l)}(t)}^{l}(x)\right\|_{p^{in}}\left\|f_{\theta^{(l)}(t)}^{l}(x')\right\|_{p^{in}}\\
&\times\left\|\mathfrak{W}_i^{l+2}\right\|_{p^{in}}\left\|(\mathfrak{W}_L^{l+2})^\intercal\right\|_{p^{in}}\left\|\delta_t^L\right\|_{p^{in}}\\
&+\frac{jN^2}{(L-1)^2C_L^{2+j}}\sum_{i=1}^{L-1}\sum_{l=0}^{i-1}\left(\left\|f_{\theta^{(i)}(0)}^{i}\right\|_{p^{in}}+\left\|f_{\theta^{(i)}(t)}^{i}-f_{\theta^{(i)}(0)}^{i}\right\|_{p^{in}}\right)^{j-1}\left\|\mathfrak{W}_i^{l+2}\right\|_{p^{in}}\left\|(\mathfrak{W}_L^{l+2})^\intercal\right\|_{p^{in}}\left\|\delta_t^L\right\|_{p^{in}}.
\end{aligned}$$

Similar to the previous inequalities, we have

$$
\begin{aligned}
&\left(\sum_{i=1}^{L-1}\sum_{l=0}^{i-1}\left\|\mathfrak{W}_i^{l+2}\right\|_{p^{in}}\left(\left\|f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}+\left\|f^l_{\theta^{(l)}(t)}-f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}\right)^2\left(\left\|\mathfrak{W}_L^{l+2}(0)\right\|_{p^{in}}+\left\|\mathfrak{W}_L^{l+2}(t)-\mathfrak{W}_L^{l+2}(0)\right\|_{p^{in}}\right)\right)^2\\
&\le\left(\sum_{i=1}^{L-1}\sum_{l=0}^{i-1}\frac{L-1}{i-1}\left\|\mathfrak{W}_i^{l+2}\right\|^2_{p^{in}}\right)\\
&\quad\times\left((L-1)\sum_{l=0}^{L-2}\left(\left\|f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}+\left\|f^l_{\theta^{(l)}(t)}-f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}\right)^4\left(\left\|\mathfrak{W}_L^{l+2}(0)\right\|_{p^{in}}+\left\|\mathfrak{W}_L^{l+2}(t)-\mathfrak{W}_L^{l+2}(0)\right\|_{p^{in}}\right)^2\right)\\
&\le(L-1)^2\Psi_{L,2}(t)(L-1)\\
&\quad\times\sqrt{\left(\sum_{l=0}^{L-2}\left(\left\|f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}+\left\|f^l_{\theta^{(l)}(t)}-f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}\right)^8\right)\left(\sum_{l=2}^{L}\left(\left\|\mathfrak{W}_L^{l}(0)\right\|_{p^{in}}+\left\|\mathfrak{W}_L^{l}(t)-\mathfrak{W}_L^{l}(0)\right\|_{p^{in}}\right)^4\right)}\\
&\le(L-1)^2\Psi_{L,2}(t)(L-1)\\
&\quad\times\sqrt{\left(\sum_{l=0}^{L-2}\left(\left\|f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}+\left\|f^l_{\theta^{(l)}(t)}-f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}\right)^8\right)\left(\sum_{l=2}^{L}\left(\left\|\mathfrak{W}_L^{l}(0)\right\|_{p^{in}}+\left\|\mathfrak{W}_L^{l}(t)-\mathfrak{W}_L^{l}(0)\right\|_{p^{in}}\right)^8\right)}\\
&\le(L-1)^2\Psi_{L,2}(t)(L-1)\times(L-1)C_L^4\left(\Phi_{L,8}(t)\right)^{(1/2)}\left(\Psi_{L,8}(t)\right)^{(1/2)}\\
&=(L-1)^4C_L^4\Psi_{L,2}(t)\left(\Phi_{L,8}(t)\right)^{(1/2)}\left(\Psi_{L,8}(t)\right)^{(1/2)}.
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
&\sum_{i=1}^{L-1}\sum_{l=0}^{i-1}\left(\left\|f^i_{\theta^{(i)}(0)}\right\|_{p^{in}}+\left\|f^i_{\theta^{(i)}(t)}-f^i_{\theta^{(i)}(0)}\right\|_{p^{in}}\right)^{j-1}\left\|f^l_{\theta^{(l)}(t)}(x)\right\|_{p^{in}}\left\|f^l_{\theta^{(l)}(t)}(x')\right\|_{p^{in}}\left\|\mathfrak{W}_i^{l+2}\right\|_{p^{in}}\left\|(\mathfrak{W}_L^{l+2})^\intercal\right\|_{p^{in}}\\
&\le\sum_{i=1}^{L-1}\sum_{l=0}^{i-1}\left\|\mathfrak{W}_i^{l+2}\right\|_{p^{in}}\left(\left\|f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}+\left\|f^l_{\theta^{(l)}(t)}-f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}\right)^2\left(\left\|\mathfrak{W}_L^{l+2}(0)\right\|_{p^{in}}+\left\|\mathfrak{W}_L^{l+2}(t)-\mathfrak{W}_L^{l+2}(0)\right\|_{p^{in}}\right)\\
&\quad\times\sum_{i=1}^{L-1}\left(\left\|f^i_{\theta^{(i)}(0)}\right\|_{p^{in}}+\left\|f^i_{\theta^{(i)}(t)}-f^i_{\theta^{(i)}(0)}\right\|_{p^{in}}\right)^{j-1}\\
&\le(L-1)C_L^{j-1}\Phi_{L,j-1}(t)(L-1)^2C_L^2\left(\Psi_{L,2}(t)\right)^{1/2}\left(\Phi_{L,8}(t)\right)^{(1/4)}\left(\Psi_{L,8}(t)\right)^{(1/4)}\\
&=(L-1)^3C_L^{j+1}\Phi_{L,j-1}(t)\left(\Psi_{L,2}(t)\right)^{1/2}\left(\Phi_{L,8}(t)\right)^{(1/4)}\left(\Psi_{L,8}(t)\right)^{(1/4)}
\end{aligned}
$$

and

$$
\begin{aligned}
&\frac{jN^{3/2}}{(L-1)^2C_L^{2+j}}\sum_{i=1}^{L-1}\sum_{l=0}^{i-1}\left(\left\|f^i_{\theta^{(i)}(0)}\right\|_{p^{in}}+\left\|f^i_{\theta^{(i)}(t)}-f^i_{\theta^{(i)}(0)}\right\|_{p^{in}}\right)^{j-1}\left\|\mathfrak{W}_i^{l+2}\right\|_{p^{in}}\left\|(\mathfrak{W}_L^{l+2})^\intercal\right\|_{p^{in}}\\
&\le\frac{jN^{3/2}}{(L-1)^2C_L^{2+j}}\sum_{i=1}^{L-1}\sum_{l=0}^{i-1}\left(\left\|f^i_{\theta^{(i)}(0)}\right\|_{p^{in}}+\left\|f^i_{\theta^{(i)}(t)}-f^i_{\theta^{(i)}(0)}\right\|_{p^{in}}\right)^{j-1}\left\|f^l_{\theta^{(l)}(t)}(x)\right\|_{p^{in}}\left\|f^l_{\theta^{(l)}(t)}(x')\right\|_{p^{in}}\\
&\quad\times\left\|\mathfrak{W}_i^{l+2}\right\|_{p^{in}}\left\|(\mathfrak{W}_L^{l+2})^\intercal\right\|_{p^{in}}.
\end{aligned}
$$

By combining the above inequalities, we have

$$\begin{aligned}
\partial_t \Phi_{L,j}(t) \leq& \frac{jN^{3/2}}{(L-1)^2 C_L^{2+j}} \sum_{i=1}^{L-1}\sum_{l=0}^{i-1} \left(\left\| f^i_{\theta^{(i)}(0)} \right\|_{p^{in}} + \left\| f^i_{\theta^{(i)}(t)} - f^i_{\theta^{(i)}(0)} \right\|_{p^{in}} \right)^{j-1} \left\| f^l_{\theta^{(l)}(t)}(x) \right\|_{p^{in}} \left\| f^l_{\theta^{(l)}(t)}(x') \right\|_{p^{in}} \\
&\times \left\| \mathfrak{W}_i^{l+2} \right\|_{p^{in}} \left\| (\mathfrak{W}_L^{l+2})^\intercal \right\|_{p^{in}} \left\| \delta_t^L \right\|_{p^{in}} \\
&+ \frac{jN^{3/2}}{(L-1)^2 C_L^{2+j}} \sum_{i=1}^{L-1}\sum_{l=0}^{i-1} \left(\left\| f^i_{\theta^{(i)}(0)} \right\|_{p^{in}} + \left\| f^i_{\theta^{(i)}(t)} - f^i_{\theta^{(i)}(0)} \right\|_{p^{in}} \right)^{j-1} \left\| \mathfrak{W}_i^{l+2} \right\|_{p^{in}} \left\| (\mathfrak{W}_L^{l+2})^\intercal \right\|_{p^{in}} \left\| \delta_t^L \right\|_{p^{in}} \\
\leq& \frac{2jN^{3/2}(L-1)}{C_L} \Phi_{L,j-1}(t) (\Psi_{L,2}(t))^{1/2} (\Phi_{L,8}(t))^{(1/4)} (\Psi_{L,8}(t))^{(1/4)} \left\| \delta_t^L \right\|_{p^{in}} \\
\leq& \frac{2jN^{3/2}(L-1)}{C_L} \Phi_{L,j-1}(t) \Psi_{L,2}(t) \Phi_{L,8}(t) \Psi_{L,8}(t) \left\| \delta_t^L \right\|_{p^{in}} .
\end{aligned}$$

□

G. Proof of Lemma 11 for Proposition 2

Lemma 11. *For $t \geq 0$ and $x \in \mathbb{R}_+^{d_{in}}$, if $f^l_{\theta^{(l)}(t)}(x)$ is element-wise nonzero for $1 \leq l \leq L-1$, then*

$$\begin{aligned}
\partial_t \tilde{\Psi}_{L,2}(x,t) &\leq \frac{2(L-1)}{C_L} \left(\tilde{\Psi}_{L,2}(x,t) \right)^2 \Phi_{L,4}(t) \Psi_{L,8}(t) \left\| \delta_t^L \right\|_{p^{in}} \\
\partial_t \tilde{\Phi}_{L,j}(x,t) &\leq \frac{2jN^{1/2}(L-1)}{C_L} \tilde{\Phi}_{L,j-1}(t) \tilde{\Phi}_{L,8}(x,t) \tilde{\Psi}_{L,2}(x,t) \Phi_{L,8}(t) \Psi_{L,8}(t) \left\| \delta_t^L \right\|_{p^{in}} \\
\partial_t \tilde{\Psi}_{L,8}(x,t) &\leq \frac{8(L-1)}{C_L} \tilde{\Phi}_{L,4}(x,t) \tilde{\Psi}_{L,2}(x,t) \left(\tilde{\Psi}_{L,8}(x,t) \right)^2 \Psi_{L,8}(t) \left\| \delta_t^L \right\|_{p^{in}} ,
\end{aligned}$$

for $1 \leq j \leq 8$.

Proof of Lemma 11. Consider the derivative of $\tilde{\Psi}_{L,8}$ first. In Section C, we show that the intermediate pre-activation values never reach zero while training, and therefore we can apply Lemma 23.

$$\begin{aligned}
&\partial_t \tilde{\Psi}_{L,8}(t) \\
&\leq \frac{8}{(L-1)^2 C_L^2} \sum_{l=2}^{L}\sum_{i=l}^{L} \left\| \mathfrak{W}_L^{l+1}(t) \right\|_{p^{in}} \left\| \mathfrak{W}_L^{l+1}(t) \right\| \left\| \mathfrak{W}_{l-1}^{i}(t) \right\| \left(\left\| \mathfrak{W}_L^{i}(0) \right\| + \left\| \mathfrak{W}_L^{i}(t) - \mathfrak{W}_L^{i}(0) \right\| \right)^7 \\
&\quad \times \left\| f^{l-1}_{\theta^{(l-1)}(t)}(x) \right\|_{p^{in}} \left\| \delta_t^L \right\|_{p^{in}} .
\end{aligned}$$

Then,

$$\begin{aligned}
&\sum_{l=2}^{L}\sum_{i=l}^{L}\left\|\mathfrak{W}_L^{l+1}(t)\right\|_{p^{in}}\left\|\mathfrak{W}_L^{l+1}(t)\right\|\left\|\mathfrak{W}_{l-1}^{i}(t)\right\|\left(\left\|\mathfrak{W}_L^{i}(0)\right\|+\left\|\mathfrak{W}_L^{i}(t)-\mathfrak{W}_L^{i}(0)\right\|\right)^7\left\|f_{\theta^{(l-1)}(t)}^{l-1}(x)\right\|_{p^{in}}\\
&\leq\left(\sum_{i=2}^{L}\sum_{i=l}^{L}\left\|\mathfrak{W}_{l-1}^{i}\right\|\left(\left\|f_{\theta^{(l-1)}(0)}^{l-1}\right\|_{p^{in}}+\left\|f_{\theta^{(l-1)}(t)}^{l-1}-f_{\theta^{(l-1)}(0)}^{l-1}\right\|_{p^{in}}\right)\left(\left\|\mathfrak{W}_L^{l+1}(0)\right\|+\left\|\mathfrak{W}_L^{l+1}(t)-\mathfrak{W}_L^{l+1}(0)\right\|\right)\right.\\
&\quad\left.\times\left(\left\|\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}+\left\|\mathfrak{W}_L^{l+1}(t)-\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}\right)\right)\left(\sum_{i=2}^{L}\left(\left\|\mathfrak{W}_L^{i}(0)\right\|_{p^{in}}+\left\|\mathfrak{W}_L^{i}(t)-\mathfrak{W}_L^{i}(0)\right\|_{p^{in}}\right)^8\right)\\
&\leq\left(\sum_{l=2}^{L}\sum_{i=2}^{l}\frac{L-1}{l-1}\left\|\mathfrak{W}_l^{i}\right\|^2\right)^{1/2}\left((L-1)\sum_{l=2}^{L}\left(\left\|f_{\theta^{(l-1)}(0)}^{l-1}\right\|_{p^{in}}+\left\|f_{\theta^{(l-1)}(t)}^{l-1}-f_{\theta^{(l-1)}(0)}^{l-1}\right\|_{p^{in}}\right)^2\right.\\
&\quad\left.\times\left(\left\|\mathfrak{W}_L^{l+1}(0)\right\|+\left\|\mathfrak{W}_L^{l+1}(t)-\mathfrak{W}_L^{l+1}(0)\right\|\right)^2\left(\left\|\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}+\left\|\mathfrak{W}_L^{l+1}(t)-\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}\right)^2\right)^{1/2}(L-1)\tilde{\Psi}_{L,8}\\
&\leq\left((L-1)^2\tilde{\Psi}_{L,2}\right)^{1/2}\sqrt{L-1}\left(\sum_{l=2}^{L}\left(\left\|f_{\theta^{(l-1)}(0)}^{l-1}\right\|_{p^{in}}+\left\|f_{\theta^{(l-1)}(t)}^{l-1}-f_{\theta^{(l-1)}(0)}^{l-1}\right\|_{p^{in}}\right)^4\right)^{1/4}\\
&\quad\times\left(\sum_{l=2}^{L}\left(\left\|\mathfrak{W}_L^{l+1}(0)\right\|+\left\|\mathfrak{W}_L^{l+1}(t)-\mathfrak{W}_L^{l+1}(0)\right\|\right)^4\left(\left\|\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}+\left\|\mathfrak{W}_L^{l+1}(t)-\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}\right)^4\right)^{1/4}(L-1)\tilde{\Psi}_{L,8}\\
&\leq\left((L-1)^2\tilde{\Psi}_{L,2}\right)^{1/2}\sqrt{L-1}(LC_L^4)^{1/4}(\Phi_{L,4})^{1/4}(L-1)\tilde{\Psi}_{L,8}\\
&\quad\times\left(\sum_{l=2}^{L}\left(\left\|\mathfrak{W}_L^{l+1}(0)\right\|+\left\|\mathfrak{W}_L^{l+1}(t)-\mathfrak{W}_L^{l+1}(0)\right\|\right)^8\right)^{1/8}\left(\sum_{l=2}^{L}\left(\left\|\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}+\left\|\mathfrak{W}_L^{l+1}(t)-\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}\right)^8\right)^{1/8}\\
&=(L-1)\tilde{\Psi}_{L,8}(L-1)^2C_L\left(\tilde{\Psi}_{L,2}\right)^{1/2}\left(\tilde{\Phi}_{L,4}\right)^{(1/4)}\left(\Psi_{L,8}\right)^{(1/8)}\left(\tilde{\Psi}_{L,8}\right)^{(1/8)}
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&\partial_t\tilde{\Psi}_{L,8}(t)\\
&\leq\frac{8}{(L-1)^2C_L^2}\sum_{l=2}^{L}\sum_{i=l}^{L}\left\|\mathfrak{W}_L^{l+1}(t)\right\|_{p^{in}}\left\|\mathfrak{W}_L^{l+1}(t)\right\|\left\|\mathfrak{W}_{l-1}^{i}(t)\right\|\left(\left\|\mathfrak{W}_L^{i}(0)\right\|+\left\|\mathfrak{W}_L^{i}(t)-\mathfrak{W}_L^{i}(0)\right\|\right)^7\\
&\quad\times\left\|f_{\theta^{(l-1)}(t)}^{l-1}(x)\right\|_{p^{in}}\left\|\delta_t^L\right\|_{p^{in}}\\
&\leq\frac{8(L-1)}{C_L}\left(\tilde{\Psi}_{L,2}\right)^{1/2}\left(\tilde{\Phi}_{L,4}\right)^{1/4}\left(\Psi_{L,8}\right)^{1/8}\left(\tilde{\Psi}_{L,8}\right)^{9/8}\left\|\delta_t^L\right\|_{p^{in}}\\
&\leq\frac{8(L-1)}{C_L}\tilde{\Psi}_{L,2}\tilde{\Phi}_{L,4}\Psi_{L,8}\left(\tilde{\Psi}_{L,8}\right)^2\left\|\delta_t^L\right\|_{p^{in}}.
\end{aligned}$$

Now, let consider the derivative of $\tilde{\Psi}_{L,2}$. From Lemma 23,

$$\begin{aligned}
&\partial_t \tilde{\Psi}_{L,2}(t) \\
&\leq \frac{2}{(L-1)^3 C_L^2} \sum_{i=2}^{L}\sum_{k=2}^{i}\sum_{l=k}^{i} \frac{L-1}{i-1} \left\|\mathfrak{W}_L^{l+1}\right\|_{p^{in}} \left\|\mathfrak{W}_i^{l+1}\right\| \left\|\mathfrak{W}_{l-1}^{k}\right\| \left(\left\|\mathfrak{W}_i^{k}(0)\right\| + \left\|\mathfrak{W}_i^{k}(t) - \mathfrak{W}_i^{k}(0)\right\|\right) \\
&\quad \times \left\|f^{l-1}_{\theta^{(l-1)}(t)}\right\|_{p^{in}} \left\|\delta_t^L\right\|_{p^{in}} \\
&\leq \frac{1}{(L-1)^3 C_L^2} \sum_{i=2}^{L}\sum_{k=2}^{i}\sum_{l=k}^{i} \frac{L-1}{i-1} \left\|\mathfrak{W}_L^{l+1}\right\|_{p^{in}} \left\|\mathfrak{W}_{l-1}^{k}\right\| \left(\left\|\mathfrak{W}_i^{l+1}\right\|^2 + \left(\left\|\mathfrak{W}_i^{k}(0)\right\| + \left\|\mathfrak{W}_i^{k}(t) - \mathfrak{W}_i^{k}(0)\right\|\right)^2\right) \\
&\quad \times \left\|f^{l-1}_{\theta^{(l-1)}(t)}\right\|_{p^{in}} \left\|\delta_t^L\right\|_{p^{in}} \\
&= \frac{1}{(L-1)^3 C_L^2} \sum_{i=2}^{L}\sum_{k=2}^{i}\sum_{l=k}^{i} \frac{L-1}{i-1} \left\|\mathfrak{W}_L^{l+1}\right\|_{p^{in}} \left\|\mathfrak{W}_{l-1}^{k}\right\| \left\|\mathfrak{W}_i^{l+1}\right\|^2 \left\|f^{l-1}_{\theta^{(l-1)}(t)}\right\|_{p^{in}} \left\|\delta_t^L\right\|_{p^{in}} \\
&\quad + \frac{1}{(L-1)^3 C_L^2} \sum_{i=2}^{L}\sum_{k=2}^{i}\sum_{l=k}^{i} \frac{L-1}{i-1} \left\|\mathfrak{W}_L^{l+1}\right\|_{p^{in}} \left\|\mathfrak{W}_{l-1}^{k}\right\| \left(\left\|\mathfrak{W}_i^{k}(0)\right\| + \left\|\mathfrak{W}_i^{k}(t) - \mathfrak{W}_i^{k}(0)\right\|\right)^2 \left\|f^{l-1}_{\theta^{(l-1)}(t)}\right\|_{p^{in}} \left\|\delta_t^L\right\|_{p^{in}}.
\end{aligned}$$

By Cauchy-Schwartz inequality,

$$\begin{aligned}
&\left(\sum_{k=2}^{L}\sum_{l=k}^{L} \left\|\mathfrak{W}_{l-1}^{k}\right\| \left(\left\|f^{l-1}_{\theta^{(l-1)}(0)}\right\|_{p^{in}} + \left\|f^{l-1}_{\theta^{(l-1)}(t)} - f^{l-1}_{\theta^{(l-1)}(0)}\right\|_{p^{in}}\right) \left(\left\|\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}} + \left\|\mathfrak{W}_L^{l+1}(t) - \mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}\right)\right)^2 \\
&\leq \left(\sum_{k=2}^{L}\sum_{l=k}^{L} \frac{L-1}{l-1} \left\|\mathfrak{W}_l^{k}\right\|^2\right) \\
&\quad \times \left((L-1)\sum_{l=2}^{L} \left(\left\|f^{l-1}_{\theta^{(l-1)}(0)}\right\|_{p^{in}} + \left\|f^{l-1}_{\theta^{(l-1)}(t)} - f^{l-1}_{\theta^{(l-1)}(0)}\right\|_{p^{in}}\right)^2 \left(\left\|\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}} + \left\|\mathfrak{W}_L^{l+1}(t) - \mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}\right)^2\right) \\
&\leq (L-1)^2 \tilde{\Psi}_{L,2}(L-1)\sqrt{(L-1)C_L^4 \Phi_{L,4} L \Psi_{L,8}} \\
&\leq (L-1)^4 C_L^2 \tilde{\Psi}_{L,2} \left(\Phi_{L,4}\right)^{(1/2)} \left(\Psi_{L,8}\right)^{(1/2)}.
\end{aligned}$$

Then, we have

$$\begin{aligned}
&\sum_{i=2}^{L}\sum_{k=2}^{i}\sum_{l=k}^{i} \frac{L-1}{i-1} \left\|\mathfrak{W}_L^{l+1}\right\|_{p^{in}} \left\|\mathfrak{W}_{l-1}^{k}\right\| \left\|\mathfrak{W}_i^{l+1}\right\|^2 \left\|f^{l-1}_{\theta^{(l-1)}(t)}\right\|_{p^{in}} \\
&\leq \left(\sum_{k=2}^{L}\sum_{l=k}^{L} \left\|\mathfrak{W}_{l-1}^{k}\right\| \left(\left\|f^{l-1}_{\theta^{(l-1)}(0)}\right\|_{p^{in}} + \left\|f^{l-1}_{\theta^{(l-1)}(t)} - f^{l-1}_{\theta^{(l-1)}(0)}\right\|_{p^{in}}\right) \left(\left\|\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}} + \left\|\mathfrak{W}_L^{l+1}(t) - \mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}\right)\right) \\
&\quad \times \left(\sum_{i=2}^{L}\sum_{l=2}^{i} \frac{L-1}{i-1} \left\|\mathfrak{W}_i^{l+1}\right\|^2\right) \\
&\leq (L-1)^2 \tilde{\Psi}_{L,2}(L-1)^2 C_L \left(\tilde{\Psi}_{L,2}\right)^{(1/2)} \left(\Phi_{L,2}\right)^{(1/4)} \left(\Psi_{L,8}\right)^{(1/4)} \\
&= (L-1)^4 C_L \left(\tilde{\Psi}_{L,2}\right)^{(3/2)} \left(\Phi_{L,4}\right)^{(1/2)} \left(\Psi_{L,8}\right)^{(1/4)}
\end{aligned}$$

and

$$
\begin{aligned}
&\sum_{i=2}^{L}\sum_{k=2}^{i}\sum_{l=k}^{i}\frac{L-1}{i-1}\left\|\mathfrak{W}_L^{l+1}\right\|_{p^{in}}\left\|\mathfrak{W}_{l-1}^{k}\right\|\left(\left\|\mathfrak{W}_i^k(0)\right\|+\left\|\mathfrak{W}_i^k(t)-\mathfrak{W}_i^k(0)\right\|\right)^2\left\|f_{\theta^{(l-1)}(t)}^{l-1}\right\|_{p^{in}}\\
&\leq\left(\sum_{k=2}^{L}\sum_{l=k}^{L}\left\|\mathfrak{W}_{l-1}^{k}\right\|\left(\left\|f_{\theta^{(l-1)}(0)}^{l-1}\right\|_{p^{in}}+\left\|f_{\theta^{(l-1)}(t)}^{l-1}-f_{\theta^{(l-1)}(0)}^{l-1}\right\|_{p^{in}}\right)\left(\left\|\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}+\left\|\mathfrak{W}_L^{l+1}(t)-\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}\right)\right)\\
&\quad\times\left(\sum_{i=2}^{L}\sum_{k=2}^{i}\frac{L-1}{i-1}\left(\left\|\mathfrak{W}_i^k(0)\right\|+\left\|\mathfrak{W}_i^k(t)-\mathfrak{W}_i^k(0)\right\|\right)^2\right)\\
&\leq(L-1)^2\tilde{\Psi}_{L,2}(L-1)^2C_L\left(\tilde{\Psi}_{L,2}\right)^{(1/2)}\left(\Phi_{L,4}\right)^{(1/4)}\left(\Psi_{L,8}\right)^{(1/4)}\\
&=(L-1)^4C_L\left(\tilde{\Psi}_{L,2}\right)^{(3/2)}\left(\Phi_{L,4}\right)^{(1/4)}\left(\Psi_{L,8}\right)^{(1/4)}.
\end{aligned}
$$

Thus, by combining the above two inequalities,

$$
\begin{aligned}
&\partial_t\tilde{\Psi}_{L,2}(t)\\
&\leq\frac{1}{(L-1)^3C_L^2}\sum_{i=2}^{L}\sum_{k=2}^{i}\sum_{l=k}^{i}\frac{L-1}{i-1}\left\|\mathfrak{W}_L^{l+1}\right\|_{p^{in}}\left\|\mathfrak{W}_{l-1}^{k}\right\|\left(\left\|\mathfrak{W}_i^{l+1}\right\|^2+\left(\left\|\mathfrak{W}_i^k(0)\right\|+\left\|\mathfrak{W}_i^k(t)-\mathfrak{W}_i^k(0)\right\|\right)^2\right)\\
&\quad\times\left\|f_{\theta^{(l-1)}(t)}^{l-1}\right\|_{p^{in}}\left\|\delta_t^L\right\|_{p^{in}}\\
&\leq\frac{2(L-1)}{C_L}\left(\tilde{\Psi}_{L,2}\right)^{3/2}\left(\Phi_{L,4}\right)^{1/4}\left(\Psi_{L,8}\right)^{1/4}\left\|\delta_t^L\right\|_{p^{in}}\\
&\leq\frac{2(L-1)}{C_L}\left(\tilde{\Psi}_{L,2}\right)^{2}\Phi_{L,4}\Psi_{L,8}\left\|\delta_t^L\right\|_{p^{in}}
\end{aligned}
$$

Finally, let consider the derivatives of $\tilde{\Phi}_{L,j}$ for $1\leq j\leq 8$. From Lemma 23,

$$
\begin{aligned}
&\partial_t\tilde{\Phi}_{L,j}(t)\\
&\leq\frac{jN^{1/2}}{(L-1)^2C_L^{2+j}}\sum_{i=1}^{L-1}\sum_{l=0}^{i-1}\left(\left\|f_{\theta^{(i)}(0)}^{i}\right\|+\left\|f_{\theta^{(i)}(t)}^{i}-f_{\theta^{(i)}(0)}^{i}\right\|\right)^{j-1}\left\|f_{\theta^{(l)}(t)}^{l}(x)\right\|_{p^{in}}\left\|f_{\theta^{(l)}(t)}^{l}(x')\right\|\\
&\quad\times\left\|\mathfrak{W}_i^{l+2}\right\|\left\|(\mathfrak{W}_L^{l+2})^{\intercal}\right\|_{p^{in}}\left\|\delta_t^L\right\|_{p^{in}}\\
&\quad+\frac{j}{(L-1)^2C_L^{2+j}}\sum_{i=1}^{L-1}\sum_{l=0}^{i-1}\left(\left\|f_{\theta^{(i)}(0)}^{i}\right\|+\left\|f_{\theta^{(i)}(t)}^{i}-f_{\theta^{(i)}(0)}^{i}\right\|\right)^{j-1}\left\|\mathfrak{W}_i^{l+2}\right\|\left\|(\mathfrak{W}_L^{l+2})^{\intercal}\right\|_{p^{in}}\left\|\delta_t^L\right\|_{p^{in}}\\
&\leq\frac{2jN^{1/2}}{(L-1)^2C_L^{2+j}}\sum_{i=1}^{L-1}\sum_{l=0}^{i-1}\left(\left\|f_{\theta^{(i)}(0)}^{i}\right\|+\left\|f_{\theta^{(i)}(t)}^{i}-f_{\theta^{(i)}(0)}^{i}\right\|\right)^{j-1}\left\|f_{\theta^{(l)}(t)}^{l}(x)\right\|_{p^{in}}\left\|f_{\theta^{(l)}(t)}^{l}(x')\right\|\\
&\quad\times\left\|\mathfrak{W}_i^{l+2}\right\|\left\|(\mathfrak{W}_L^{l+2})^{\intercal}\right\|_{p^{in}}\left\|\delta_t^L\right\|_{p^{in}}
\end{aligned}
$$

since $\left\|f_{\theta^{(l)}(t)}^{l}(x)\right\|\geq 1$.

By Cauchy-Schwartz inequality,

$$
\begin{aligned}
&\Bigg(\sum_{l=0}^{L-2}\left(\left\|f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}+\left\|f^l_{\theta^{(l)}(t)}-f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}\right)^2\left(\left\|f^l_{\theta^{(l)}(0)}\right\|+\left\|f^l_{\theta^{(l)}(t)}-f^l_{\theta^{(l)}(0)}\right\|\right)^2\\
&\times\left(\left\|\mathfrak{W}^{l+2}_L(0)\right\|_{p^{in}}+\left\|\mathfrak{W}^{l+2}_L(t)-\mathfrak{W}^{l+2}_L(0)\right\|_{p^{in}}\right)^2\Bigg)^2\\
&\le\left(\sum_{l=0}^{L-2}\left(\left\|f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}+\left\|f^l_{\theta^{(l)}(t)}-f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}\right)^4\left(\left\|f^l_{\theta^{(l)}(0)}\right\|+\left\|f^l_{\theta^{(l)}(t)}-f^l_{\theta^{(l)}(0)}\right\|\right)^4\right)\\
&\le\left(\sum_{l=0}^{L-2}\left(\left\|f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}+\left\|f^l_{\theta^{(l)}(t)}-f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}\right)^4\left(\left\|f^l_{\theta^{(l)}(0)}\right\|+\left\|f^l_{\theta^{(l)}(t)}-f^l_{\theta^{(l)}(0)}\right\|\right)^4\right)\\
&\times\left(\sum_{l=2}^{L}\left(\left\|\mathfrak{W}^{l}_L(0)\right\|_{p^{in}}+\left\|\mathfrak{W}^{l}_L(t)-\mathfrak{W}^{l}_L(0)\right\|_{p^{in}}\right)^4\right)\\
&\le\left(\sum_{l=0}^{L-2}\left(\left\|f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}+\left\|f^l_{\theta^{(l)}(t)}-f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}\right)^8\right)^{1/2}\left(\left(\left\|f^l_{\theta^{(l)}(0)}\right\|+\left\|f^l_{\theta^{(l)}(t)}-f^l_{\theta^{(l)}(0)}\right\|\right)^8\right)^{1/2}\\
&\times\left(\sum_{l=2}^{L}\left(\left\|\mathfrak{W}^{l}_L(0)\right\|_{p^{in}}+\left\|\mathfrak{W}^{l}_L(t)-\mathfrak{W}^{l}_L(0)\right\|_{p^{in}}\right)^8\right)\\
&\le(L-1)\Psi_{L,8}(L-1)C_L^8\left(\Phi_{L,8}\right)^{1/2}\left(\tilde{\Phi}_{L,8}\right)^{1/2}.
\end{aligned}
$$

Then, we have

$$
\begin{aligned}
&\sum_{i=1}^{L-1}\sum_{l=0}^{i-1}\left(\left\|f^i_{\theta^{(i)}(0)}\right\|+\left\|f^i_{\theta^{(i)}(t)}-f^i_{\theta^{(i)}(0)}\right\|\right)^{j-1}\left\|f^l_{\theta^{(l)}(t)}(x)\right\|_{p^{in}}\left\|f^l_{\theta^{(l)}(t)}(x')\right\|\left\|\mathfrak{W}^{l+2}_i\right\|\left\|(\mathfrak{W}^{l+2}_L)^\intercal\right\|_{p^{in}}\\
&\le\left(\sum_{i=1}^{L-1}\left(\left\|f^i_{\theta^{(i)}(0)}\right\|+\left\|f^i_{\theta^{(i)}(t)}-f^i_{\theta^{(i)}(0)}\right\|\right)^{j-1}\right)\Bigg(\sum_{i=1}^{L-1}\sum_{l=0}^{i-1}\frac{L-1}{i-1}\left\|\mathfrak{W}^{l+2}_i\right\|\left(\left\|f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}+\left\|f^l_{\theta^{(l)}(t)}-f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}\right)\\
&\quad\times\left(\left\|f^l_{\theta^{(l)}(0)}\right\|+\left\|f^l_{\theta^{(l)}(t)}-f^l_{\theta^{(l)}(0)}\right\|\right)\left(\left\|\mathfrak{W}^{l+2}_L(0)\right\|_{p^{in}}+\left\|\mathfrak{W}^{l+2}_L(t)-\mathfrak{W}^{l+2}_L(0)\right\|_{p^{in}}\right)\Bigg)\\
&\le(L-1)C_L^{j-1}\tilde{\Phi}_{L,j-1}\Bigg(\sum_{i=1}^{L-1}\sum_{l=0}^{i-1}\frac{L-1}{i-1}\left\|\mathfrak{W}^{l+2}_i\right\|\left(\left\|f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}+\left\|f^l_{\theta^{(l)}(t)}-f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}\right)\\
&\quad\times\left(\left\|f^l_{\theta^{(l)}(0)}\right\|+\left\|f^l_{\theta^{(l)}(t)}-f^l_{\theta^{(l)}(0)}\right\|\right)\left(\left\|\mathfrak{W}^{l+2}_L(0)\right\|_{p^{in}}+\left\|\mathfrak{W}^{l+2}_L(t)-\mathfrak{W}^{l+2}_L(0)\right\|_{p^{in}}\right)\Bigg)\\
&\le(L-1)C_L^{j-1}\tilde{\Phi}_{L,j-1}\left(\sum_{i=1}^{L-1}\sum_{l=0}^{i-1}\frac{L-1}{i-1}\left\|\mathfrak{W}^{l+2}_i\right\|^2\right)^{1/2}\Bigg((L-1)\sum_{l=0}^{L-2}\left(\left\|f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}+\left\|f^l_{\theta^{(l)}(t)}-f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}\right)^2\\
&\quad\left(\left\|f^l_{\theta^{(l)}(0)}\right\|+\left\|f^l_{\theta^{(l)}(t)}-f^l_{\theta^{(l)}(0)}\right\|\right)^2\times\left(\left\|\mathfrak{W}^{l+2}_L(0)\right\|_{p^{in}}+\left\|\mathfrak{W}^{l+2}_L(t)-\mathfrak{W}^{l+2}_L(0)\right\|_{p^{in}}\right)^2\Bigg)^{1/2}\\
&\le(L-1)C_L^{j-1}\tilde{\Phi}_{L,j-1}\left((L-1)^2\tilde{\Psi}_{L,2}\right)^{1/2}\left((L-1)^2C_L^4\left(\Phi_{L,8}\right)^{1/4}\left(\Psi_{L,8}\right)^{1/2}\left(\tilde{\Phi}_{L,8}\right)^{1/4}\right)^{1/2}\\
&=(L-1)^3C_L^{j+1}\tilde{\Phi}_{L,j-1}\left(\tilde{\Psi}_{L,2}\right)^{1/2}\left(\Phi_{L,8}\right)^{1/8}\left(\Psi_{L,8}\right)^{1/4}\left(\tilde{\Phi}_{L,8}\right)^{1/8}
\end{aligned}
$$

Thus, by combining the above inqualities,

$$
\begin{aligned}
&\partial_t \tilde{\Phi}_{L,j}(t) \\
&\le 2\frac{jN^{1/2}}{(L-1)^2 C_L^{2+j}} \sum_{i=1}^{L-1}\sum_{l=0}^{i-1} \left(\left\|f^i_{\theta^{(i)}(0)}\right\| + \left\|f^i_{\theta^{(i)}(t)} - f^i_{\theta^{(i)}(0)}\right\|\right)^{j-1} \left\|f^l_{\theta^{(l)}(t)}(x)\right\|_{p^{in}} \left\|f^l_{\theta^{(l)}(t)}(x')\right\| \\
&\quad \times \left\|\mathfrak{W}_i^{l+2}\right\| \left\|(\mathfrak{W}_L^{l+2})^\intercal\right\|_{p^{in}} \left\|\delta_t^L\right\|_{p^{in}} \\
&\le \frac{2jN^{1/2}(L-1)}{C_L} \tilde{\Phi}_{L,j-1} \left(\tilde{\Psi}_{L,2}\right)^{1/2} \left(\Phi_{L,8}\right)^{1/8} \left(\Psi_{L,8}\right)^{1/4} \left(\tilde{\Phi}_{L,8}\right)^{1/8} \left\|\delta_t^L\right\|_{p^{in}} \\
&\le \frac{2jN^{1/2}(L-1)}{C_L} \tilde{\Phi}_{L,j-1} \tilde{\Psi}_{L,2} \Phi_{L,8} \Psi_{L,8} \tilde{\Phi}_{L,8} \left\|\delta_t^L\right\|_{p^{in}}.
\end{aligned}
$$

□

H. Proof of Lemma 12 for Proposition 3

Lemma 12. *For $t \ge 0$ and $x \in \mathbb{R}_+^{d_{in}}$, if $f^l_{\theta^{(l)}(t)}(x)$ is element-wise nonzero for $1 \le l \le L-1$, then*

$$
\begin{aligned}
\partial_t \tilde{\Psi}_L^k(x,t) &\le \frac{L-1}{C_L} \Psi_{L,8}(t) \tilde{\Phi}_{L,4}(x,t) \tilde{\Psi}_{L,2}(x,t) \tilde{\Psi}_{L,8}(x,t) \left\|\delta_t^L\right\|_{p^{in}} \\
\partial_t \tilde{\Phi}_L^l(x,t) &\le \frac{2N^{1/2}(L-1)}{C_L} \Phi_{L,8}(t) \Psi_{L,8}(t) \tilde{\Phi}_{L,8}(x,t) \tilde{\Psi}_{L,2}(x,t) \left\|\delta_t^L\right\|_{p^{in}},
\end{aligned}
$$

for $1 \le l \le L$ and $2 \le k \le L$.

Proof of Lemma 12. Consider the derivative of $\tilde{\Psi}_L^k$ first. In the proof of Lemma 11, we showed

$$
\begin{aligned}
&\left(\sum_{i=2}^{L}\sum_{l=i}^{L} \left\|\mathfrak{W}^i_{l-1}\right\| \left(\left\|f^{l-1}_{\theta^{(l-1)}(0)}\right\|_{p^{in}} + \left\|f^{l-1}_{\theta^{(l-1)}(t)} - f^{l-1}_{\theta^{(l-1)}(0)}\right\|_{p^{in}}\right) \left(\left\|\mathfrak{W}_L^{l+1}(0)\right\| + \left\|\mathfrak{W}_L^{l+1}(t) - \mathfrak{W}_L^{l+1}(0)\right\|\right)\right. \\
&\quad \left. \times \left(\left\|\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}} + \left\|\mathfrak{W}_L^{l+1}(t) - \mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}\right)\right)^2 \\
&\le (L-1)^2 \tilde{\Psi}_{L,2} (L-1) \times (L-1) C_L^2 \left(\Phi_{L,4}\right)^{1/2} \left(\Psi_{L,8}(t)\right)^{1/4} \left(\tilde{\Psi}_{L,8}(t)\right)^{1/4}.
\end{aligned}
$$

Thus, from Lemma 23,

$$
\begin{aligned}
\partial_t \tilde{\Psi}_L^k(t) &\le \frac{1}{(L-1)C_L^2} \sum_{l=k}^{L} \left\|\mathfrak{W}_L^{l+1}(t)\right\|_{p^{in}} \left\|\mathfrak{W}_L^{l+1}(t)\right\| \left\|\mathfrak{W}^k_{l-1}(t)\right\| \left\|f^{l-1}_{\theta^{(l-1)}(t)}(x)\right\|_{p^{in}} \left\|\delta_t^L\right\|_{p^{in}} \\
&\le \frac{L-1}{C_L} \left(\tilde{\Psi}_{L,2}\right)^{1/2} \left(\tilde{\Phi}_{L,4}\right)^{(1/4)} \left(\Psi_{L,8}\right)^{(1/8)} \left(\tilde{\Psi}_{L,8}\right)^{(1/8)} \left\|\delta_t^L\right\|_{p^{in}} \\
&\le \frac{L-1}{C_L} \tilde{\Psi}_{L,2} \tilde{\Phi}_{L,4} \Psi_{L,8} \tilde{\Psi}_{L,8} \left\|\delta_t^L\right\|_{p^{in}}.
\end{aligned}
$$

On the other hand, let consider the derivative of $\tilde{\Phi}_L^l(t)$. We also showed the following inequality from the proof of Lemma 11.

$$
\begin{aligned}
&\sum_{i=1}^{L-1}\sum_{l=0}^{i-1} \frac{L-1}{i-1} \left\|\mathfrak{W}_i^{l+2}\right\| \left(\left\|f^l_{\theta^{(l)}(0)}\right\|_{p^{in}} + \left\|f^l_{\theta^{(l)}(t)} - f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}\right) \left(\left\|f^l_{\theta^{(l)}(0)}\right\| + \left\|f^l_{\theta^{(l)}(t)} - f^l_{\theta^{(l)}(0)}\right\|\right) \\
&\quad \times \left(\left\|\mathfrak{W}_L^{l+2}(0)\right\|_{p^{in}} + \left\|\mathfrak{W}_L^{l+2}(t) - \mathfrak{W}_L^{l+2}(0)\right\|_{p^{in}}\right) \\
&\le (L-1)^2 C_L^2 \left(\tilde{\Psi}_{L,2}\right)^{1/2} \left(\Phi_{L,8}\right)^{1/8} \left(\Psi_{L,8}\right)^{1/4} \left(\tilde{\Phi}_{L,8}\right)^{1/8}.
\end{aligned}
$$

Thus, from Lemma 23,

$$\begin{aligned}\partial_t \tilde{\Phi}_L^l(t) &\leq \frac{N^{1/2}}{(L-1)C_L^3} \sum_{i=0}^{l-1} \left\| f_{\theta^{(i)}(t)}^i(x) \right\|_{p^{in}} \left\| f_{\theta^{(i)}(t)}^i(x') \right\| \left\| \mathfrak{W}_l^{i+2} \right\| \left\| (\mathfrak{W}_L^{i+2})^\intercal \right\|_{p^{in}} \left\| \delta_t^L \right\|_{p^{in}} \\ &\quad + \frac{N^{1/2}}{(L-1)C_L^3} \sum_{i=0}^{l-1} \left\| \mathfrak{W}_l^{i+2} \right\| \left\| (\mathfrak{W}_L^{i+2})^\intercal \right\|_{p^{in}} \left\| \delta_t^L \right\|_{p^{in}} \\ &\leq \frac{2N^{1/2}(L-1)}{C_L} \left(\tilde{\Psi}_{L,2}\right)^{1/2} \left(\Phi_{L,8}\right)^{1/8} \left(\Psi_{L,8}\right)^{1/4} \left(\tilde{\Phi}_{L,8}\right)^{1/8} \left\| \delta_t^L \right\|_{p^{in}} \\ &\leq \frac{2N^{1/2}(L-1)}{C_L} \tilde{\Psi}_{L,2} \Phi_{L,8} \Psi_{L,8} \tilde{\Phi}_{L,8} \left\| \delta_t^L \right\|_{p^{in}}.\end{aligned}$$

□

I. Proof of Theorems of NTK for CNN

I.1. Preliminaries

We define the Jacobian of tensors through appropriate vectorization: if $u \in \mathbb{R}^{P\times Q\times R}$ and $w \in \mathbb{R}^{P\times Q\times R\times S}$, then

$$\partial_w u = \partial_{\mathrm{vec}(w)} \mathrm{vec}\left(u\right),$$

where $\mathrm{vec}\left(\cdot\right)$ that maps a tensor to a vector element-wise. In the following, we use this vectorization and the Jacobian defined through vectorization implicitly.

To index 3D and 4D tensors, we use notation inspired by the slice notation of PyTorch. However, our notation differs in that indices starts from 1 and our slice includes the last element (i.e., $w_{p,q,r_1:r_2,s_1:s_2}$ contains w_{p,q,r_2,s_2}). For a 4D tensor $w \in \mathbb{R}^{P\times Q\times R\times S}$, we define a slice $w_{p,q,r_1:r_2,s_1:s_2} \in \mathbb{R}^{(r_2-r_1+1)\times(s_2-s_1+1)}$, where $\left(w_{p,q,r_1:r_2,s_1:s_2}\right)_{i,j} = w_{p,q,r_1-1+i,s_2-1+j}$. Also define $w_{p,:,:,:} = w_{p,1:Q,1:R,1:S}$. Slicing for 3D tensors is defined analogously. For a function f, where the output is a 3D tensor (i.e., $f(x) \in \mathbb{R}^{P\times Q\times R}$ for an input x), we define a slice of function f similarly, i.e., $f_{p,q_1:q_2,r_1:r_2}(x) = \left(f(x)\right)_{p,q_1:q_2,r_1:r_2} \in \mathbb{R}^{(q_2-q_1+1)\times(r_2-r_1+1)}$ and $f_{p,:,:}(x) = f_{p,1:Q,1:R}(x)$. For the sake of notational simplicity, we often omit commas (i.e., $w_{p,q,r,s} = w_{pqrs}$), and colons (i.e., $f_{p,:,:}(x) = f_p(x)$).

I.2. Proof of Theorem 4

For simplicity, we use $f^l \equiv f^l_{\theta^{(l)}}$ if it is clear from the context.

Proof of Theorem 4. For notational simpicity, we let

$$\mathfrak{W}^k_l(x,t) = \partial_{f^{l-1}} f^l(x) \cdots \partial_{f^{k-1}} f^k(x).$$

for $2 \le k, l \le L$, where $\mathfrak{W}^k_l(x,t) = 1$ if $k > l$. Similar to MLP, we have a recursive relation for kernel

$$\begin{aligned}\tilde{\Theta}^L_t(x,x') =& \frac{1}{LC_L^2}\sum_{l=0}^{L-1}\Bigg[\mathfrak{W}^{l+2}_L(x,t)\Bigg(\sum_{c',c,s,u}\partial_{w^{l+1}_{c'csu}} f^{l+1}_{\theta^{(l+1)}(t)}(x)\left(\partial_{w^{l+1}_{c'csu}} f^{l+1}_{\theta^{(l+1)}(t)}(x')\right)^{\intercal} \\ &+ \sum_i \partial_{b^{l+1}_i} f^{l+1}_{\theta^{(l+1)}(t)}(x)\left(\partial_{b^{l+1}_i} f^{l+1}_{\theta^{(l+1)}(t)}(x')\right)^{\intercal}\Bigg)\mathfrak{W}^{l+2}_L(x',t)^{\intercal}.\Bigg]\end{aligned}$$

By calculation, the following holds:

$$\begin{aligned}\partial_{w^l_{c'csu}} f^l_{r,i,j} &= \partial_{w^l_{c'csu}} \sum_{s=1}^{n_{l-1}}\left(w^l_{r,s} * \sigma(f^{l-1}_s) + \mathbb{1}_{h\times h} b^l_r\right)_{i,j} \\ &= \partial_{w^l_{c',c,s,u}} \sum_{s=1}^{n_{l-1}}\sum_{q=-1}^{1}\sum_{p=-1}^{1} w_{c',s,p+2,q+2}\sigma(f^{l-1}_{s,p+i,q+j})\mathbf{1}_{\{r=c'\}} \\ &= \partial_{w^l_{c',c,s,u}} \sum_{q=-1}^{1}\sum_{p=-1}^{1} w_{c',c,p+2,q+2}\sigma(f^{l-1}_{c,p+i,q+j})\mathbf{1}_{\{r=c'\}} \\ &= \sigma(f^{l-1}_{c,s+i-2,u+j-2})\mathbf{1}_{\{r=c'\}}.\end{aligned}$$

for $1 \le l \le L$, where $f_{pqr} = 0$ if the index is out of bounds. This implies

$$\partial_{w^l_{c'csu}} f^l_r = \sigma(f^{l-1}_c(x))_{\psi_{su}}\mathbf{1}_{\{r=c'\}}, \quad \partial_{w^l_{c'csu}} f^l = P_{c'}\left(\sigma(f^{l-1}_c(x))_{\psi_{su}}\right),$$

where $P_{c'}(\cdot) : \mathbb{R}^{d\times d} \to \mathbb{R}^{n_l\times d\times d}$ is operation such that $\left(P_{c'}(f^{l-1}_c)\right)_q = f^{l-1}_c(x)\mathbf{1}_{\{c'=q\}}$. (As mentioned in I.1 Preliminaries, we denote $P_{c'}(M) = \mathrm{vec}\left(P_{c'}(M)\right)$.) Also

$$\begin{aligned}\partial_{b^l_k} f^l_{r,i,j} &= \partial_{b^l_k} \sum_{s=1}^{n_{l-1}}\left(w^l_{r,s} * \sigma(f^{l-1}_s) + \mathbb{1}_{d\times d} b^l_r\right)_{i,j} \\ &= \mathbf{1}_{\{r=k\}}\end{aligned}$$

for $1 \le l \le L$. Then if we let

$$B^l = \sum_i \partial_{b_i^{l+1}} f^{l+1}(x) \left(\partial_{b_i^{l+1}} f^{l+1}(x')\right)^\intercal = \begin{bmatrix} \mathbb{1}_{d^2\times d^2} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbb{1}_{d^2\times d^2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbb{1}_{d^2\times d^2} \end{bmatrix}, \quad B^{L-1} = \mathbb{1}_{d^2\times d^2}$$

for $0 \le l \le L-2$, we have following lemma.

Lemma 24. *For any x, x', and $t \ge 0$,*

$$\tilde{\Theta}_t^L(x,x') = \frac{1}{LC_L^2}\sum_{l=0}^{L-1}\sum_{c'csu} \mathfrak{W}_L^{l+2}(x,t)P_{c'}\left(\sigma\left(\left(f_{\theta(t)}^{l-1}(x)\right)_c\right)_{\psi_{su}}\right)\left(\frac{1}{C_L}P_{c'}\left(\sigma\left(\left(f_{\theta(t)}^{l-1}(x')\right)_c\right)_{\psi_{su}}\right)\right)^\intercal \mathfrak{W}_L^{l+2}(x',t)^\intercal$$
$$+ \frac{1}{LC_L^2}\sum_{l=0}^{L-1}\mathfrak{W}_L^{l+2}(x,t)B^l\mathfrak{W}_L^{l+2}(x',t)^\intercal,$$

where we define $\sigma(f^0(x)) = x$ for convenience.

Also our initialization ($t=0$) implies

$$\begin{aligned} &f_1^1 = C_L x, \quad f_2^1 = w_{2,1}^1 * x + \mathbb{1}_{d\times d}v^1, \quad f_3^1 = C_L\mathbb{1}_{d\times d} \\ &f_1^l = C_L x, \quad f_2^l = w_{2,1}^l * C_L x + \mathbb{1}_{d\times d}v^l, \quad f_3^l = C_L\mathbb{1}_{d\times d} \\ &f_1^L = 0_{d\times d} \end{aligned} \tag{3}$$

for $2 \le l \le L-1$. Note that $\mathbb{1}_{d\times d} \in \mathbb{R}^{d\times d}$ is a matrix whose entries are all 1.

Since, for $1 \le l \le L$,

$$\begin{aligned} \partial_{f_{c,i,j}^{l-1}} f_{c',s,u}^l &= \partial_{f_{c,i,j}^{l-1}} \sum_{c=1}^{n_{l-1}}\sum_{q=-1}^{1}\sum_{p=-1}^{1} w_{c',c,p+2,q+2}^l\sigma(f_{c,p+s,q+u}^{l-1}(x)) + b_{c'}^l \\ &= w_{c',c,i-s+2,j-u+2}^l\dot{\sigma}(f_{c,i,j}^{l-1})\mathbf{1}_{\{s-1\le i\le s+1\}}\mathbf{1}_{\{u-1\le j\le u+1\}} \end{aligned}$$

is nonzero for $(c',c) = (1,1),(2,1),(3,3)$ at initialization, we have

$$\partial_{f_{\theta^{L-1}(0)}^{L-1}} f_{\theta^L(0)}^L(x) = \begin{bmatrix} 0_{d^2\times d^2} & 0_{d^2\times d^2} & I_{d^2} \end{bmatrix}, \quad \partial_{f_{\theta^{l-1}(0)}^{l-1}} f_{\theta^l(0)}^l(x) = \begin{bmatrix} I_{d^2} & 0_{d^2\times d^2} & 0_{d^2\times d^2} \\ D^l & 0_{d^2\times d^2} & 0_{d^2\times d^2} \\ 0_{d^2\times d^2} & 0_{d^2\times d^2} & I_{d^2} \end{bmatrix}$$

for $2 \le l \le L-1$ and some $D^l \in \mathbb{R}^{d^2\times d^2}$. This implies

$$\mathfrak{W}_l^k(0,x) = \partial_{f_{\theta^{l-1}(0)}^{l-1}} f_{\theta^l(0)}^l(x) \tag{4}$$

for $2 \le l \le L$. Thus, the scaled NTK is given by

$$\tilde{\Theta}_0^L = \frac{1}{LC_L^2}\sum_{l=0}^{L-1}\left(\mathbb{1}_{d^2\times d^2} + \sum_{c,s,u}\sigma(f_c^l(x))_{\psi_{su}}\sigma(f_c^l(x'))_{\psi_{su}}^\intercal\right),$$

where ψ_{su} is set of coordinate which satisfy $x_{\psi_{su}} = [x]_{s-1:d+s-2,u-1:d+u-2}$ and $x_{pq} = 0$ if the index is out of bounds and we define $\sigma(f^0(x)) = x$ for convenience. Since we apply global averaging, we have to cosider $S\left(\tilde{\Theta}^L\right)$.

$$\begin{aligned} S\left(\tilde{\Theta}^L\right) = \frac{1}{LC_L^2}\sum_{l=1}^{L-1}\Bigg(& d^2 + C_L^2\sum_{s,u} S\left((\mathbb{1}_{d\times d})_{\psi_{su}}\right)^2 + C_L^2\sum_{s,u} S\left(x_{\psi_{su}}\right)S\left(x'_{\psi_{su}}\right) \\ &+ \sum_{t,u} S\left(\sigma\left(w_{2,1}^l * C_L x + v^l\mathbb{1}_{d\times d}\right)_{\psi_{su}}\right)S\left(\sigma\left(w_{2,1}^l * C_L x' + v^l\mathbb{1}_{d\times d}\right)_{\psi_{su}}\right)\Bigg) \\ &+ \frac{1}{LC_L^2}\left(d^2 + \sum_{s,u} S\left(x_{\psi_{su}}\right)S\left(x'_{\psi_{su}}\right)\right) \end{aligned}$$

Let $g^l(x) = \langle w^l_{2,1}, x\rangle + v^l/C_L$, where $x \in \mathbb{R}^{3\times 3}$. Then $\{g^l(x)\}_{l=0}^{L-1}$ are i.i.d. Gaussian process, where $g^l_{i,j} \sim \mathcal{GP}(0, \rho^2 \langle x, x'\rangle + \beta^2)$. Thus, as $L \to \infty$, by the law of large number, we have

$$S\left(\tilde{\Theta}^L(x,x')\right) \to \frac{1}{d^2}\sum_{s=1}^{3}\sum_{u=1}^{3}\left(P_d(s,u) + d^2 S(x_{\psi_{su}})S(x'_{\psi_{su}}) + \sum_{ij\in\psi_{su}}\sum_{i'j'\in\psi_{su}} \mathbb{E}[\sigma(g(\phi_{ij}(x))\sigma(g(\phi_{i'j'}(x'))]\right),$$

$g \sim \mathcal{GP}(0, \rho^2 \langle x, x'\rangle + \beta^2))$, $\phi_{mn}(x) = 0$ if the index is out of bounds and

$$p_d(s,u) = \begin{cases} d^4, \ \ (s,u) = (2,2), \\ d^2(d-1)^2, \ \ |s-u| = 1, \\ (d-1)^4, \ \ \text{else.} \end{cases}$$

This concludes the proof.

□

I.3. Proof of Theorem 5

Similar to MLP, we define Lyapunov functions to track intermediate weights and pre-activation values of the network. Again, for the sake of simplicity, we set the scaling factor of gradient flow by $\frac{1}{(L-1)C_L^2}$ throughout the proof. Also, without loss of generality, we assume the norms of vectorized data points $\|\text{vec}\,(x_i)\|$ are bounded by 1 for all $1 \le i \le N$, and we further assume that all other vectorized inputs $x \notin \{x_1, \dots, x_N\}$ also have bounded norm, i.e., $\|\text{vec}\,(x)\| \le 1$.

For $1 \le j \le 8$ and $C_L > 0$ satisfying $L^2/C_L \to 0$,

$$\Phi_{L,j}(t) = \frac{1}{(L-1)C_L^j}\sum_{l=1}^{L-1}\left(\left\|f^l_{\theta^{(l)}(0)}(x)\right\|_{p^{in}} + \left\|f^l_{\theta^{(l)}(t)}(x) - f^l_{\theta^{(l)}(0)}(x)\right\|_{p^{in}}\right)^j$$
$$\Psi_{L,2}(t) = \frac{1}{(L-1)^2}\sum_{l=2}^{L}\frac{L-1}{l-1}\sum_{i=2}^{l}\left(\left\|\mathfrak{W}^i_l(x,0)\right\|_{p^{in}} + \left\|\mathfrak{W}^i_l(x,t) - \mathfrak{W}^i_l(x,0)\right\|_{p^{in}}\right)^2$$
$$\Psi_{L,8}(t) = \frac{1}{(L-1)}\sum_{l=2}^{L}\left(\left\|\mathfrak{W}^l_L(x,0)\right\|_{p^{in}} + \left\|\mathfrak{W}^l_L(x,t) - \mathfrak{W}^l_L(x,0)\right\|_{p^{in}}\right)^8,$$

where $\Phi_{L,0} = 1$ and we define

$$\begin{aligned}&\left(\left\|f^l_{\theta^{(l)}(0)}(x)\right\|_{p^{in}} + \left\|f^l_{\theta^{(l)}(t)}(x) - f^l_{\theta^{(l)}(0)}(x)\right\|_{p^{in}}\right)^j \\ &= \left(\left\|\text{vec}\left(f^l_{\theta^{(l)}(0)}(x)\right)\right\|_{p^{in}} + \left\|\text{vec}\left(f^l_{\theta^{(l)}(t)}(x)\right) - \text{vec}\left(f^l_{\theta^{(l)}(0)}(x)\right)\right\|_{p^{in}}\right)^j.\end{aligned}$$

for all j and l in $\Phi_{L,j}(t)$. At initialization ($t = 0$), we have

$$\Phi_{L,j}(0) = \frac{1}{(L-1)C_L^j}\sum_{l=1}^{L-1}\left(\left\|f^l(x)\right\|_{p^{in}}\right)^j$$
$$\Psi_{L,2}(0) = \frac{1}{(L-1)^2}\sum_{l=2}^{L}\frac{L-1}{l-1}\sum_{i=2}^{l}\left(\left\|\mathfrak{W}^i_l(x,0)\right\|_{p^{in}}\right)^2$$
$$\Psi_{L,8}(0) = \frac{1}{L-1}\sum_{l=2}^{L}\left(\left\|\mathfrak{W}^l_L(x,0)\right\|_{p^{in}}\right)^8.$$

From (4) and (3) and imply that $\Phi_{L,j}(0)$ and $\Psi_{L,2}(0)$ converges (in probability) to constant values $\Phi_{\infty,j}(0)$ and $\Psi_{\infty,2}(0)$ by the law of large number, respectively. Furthermore, $\Psi_{L,8}(0)$ also converges (in probability) to constant value $\Psi_{\infty,8}(0)$ since $\mathfrak{W}^l_L(x,0) = \mathfrak{W}^L_L(x,0)$. Now, we are going to prove following proposition which implies Lyapunov function remains constant during training.

Proposition 4. *For $1 \le j \le 8$ and any $T > 0$,*

$$\Phi_{L,j}(t) \xrightarrow{p} \Phi_{\infty,j}(0), \quad \Psi_{L,2}(t) \xrightarrow{p} \Psi_{\infty,2}(0), \quad \Psi_{L,8}(t) \xrightarrow{p} \Psi_{\infty,8}(0)$$

uniformly for $t \in [0, T]$ as $L \to \infty$.

The following lemma is a key step to prove the proposition which allows us to apply Grönwall's Lemma.

Lemma 25. *For $t \ge 0$, if $f^l_{\theta^{(l)}(t)}(x)$ is element-wise nonzero for $1 \le l \le L-1$ and $x \in \{x_1, \dots, x_N\}$, then*

$$\partial_t \Phi_{L,j}(t) \le \frac{162 \max_l\{n_l^4\} d^4 j N^{3/2}(L-1)}{C_L} \Phi_{L,j-1}(t)\Phi_{L,8}(t)\Psi_{L,2}(t)\Psi_{L,8}(t) \left\|\delta_t^L\right\|_{p^{in}}$$
$$\partial_t \Psi_{L,2}(t) \le \frac{18 d^4 N(L-1)}{C_L} \Phi_{L,4}(t) \left(\Psi_{L,2}(t)\right)^2 (t) \Psi_{L,8}(t) \left\|\delta_t^L\right\|_{p^{in}}$$
$$\partial_t \Psi_{L,8}(t) \le \frac{72 N d^4 (L-1)}{C_L} \Phi_{L,4}(t)\Psi_{L,2}(t) \left(\Psi_{L,8}(t)\right)^2 \left\|\delta_t^L\right\|_{p^{in}}$$

for $1 \le j \le 8$.

The proof of Lemma 25 is given in Section L.

Proof of Proposition 4. Like proof of Theorem 2, we need to show that $f^l_{\theta^{(l)}(t)}(x)$ is element-wise nonzero for large enough L for all $t \in [0, T]$, $1 \le l \le L-1$, and $x \in \{x_1, \dots, x_N\}$. This element-wise nonzero assumption also holds if L is large enough, and the proof is given in Section J. From Lemma L, we get

$$\partial_t \left(\sum_{j=1}^{8} \Phi_{L,j}(t) + \Psi_{L,2}(t) + \Psi_{L,8}(t) \right) \le \frac{1296 \max_l\{n_l^4\} d^4 N^{3/2}(L-1)}{C_L} \left(\sum_{j=1}^{8} \Phi_{L,j}(t) + \Psi_{L,2}(t) + \Psi_{L,8}(t) \right)^4 \left\|\delta_t^L\right\|_{p^{in}},$$

which implies

$$\Gamma_L(t) \le \Gamma_L(0) + \frac{1296 \max_l\{n_l^4\} d^4 j N^{3/2}(L-1)}{C_L} \int_0^t \Gamma_L(s)^4 \left\|\delta_s^L\right\|_{p^{in}} ds,$$

where $\Gamma_L(t) = \sum_{j=1}^8 \Phi_{L,j}(t) + \Psi_{L,2}(t) + \Psi_{L,8}(t)$. By Grönwall's lemma with $z(t) = \Gamma_L(t)$, $a(t) = \Gamma_L(0)$, $b(t) = \frac{1296 \max_l\{n_l^4\} d^4 N^{3/2}(L-1)}{C_L}$, $k(s) = \left\|\delta_s^L\right\|_{p^{in}}$,

$$\Gamma_L(t) \le \Gamma_L(0) \left\{ 1 - \frac{3888 \max_l\{n_l^4\} d^4 N^{3/2}(L-1)}{C_L} \Gamma_L(0)^3 \int_0^t \left\|\delta_s^L\right\|_{p^{in}} ds \right\}^{-\frac{1}{3}}$$

for $0 \le t \le \beta_L$. Recall that $\Gamma_L(0)$ converge to constant and we assumed stochastically boundness of $\int_0^T \left\|\delta_s^L\right\|_{p^{in}} ds$. Thus, if $L^2/C_L \to 0$, then $\frac{3888 \max_l\{n_l^4\} d^4 N^{3/2}(L-1)}{C_L} \Gamma_L(0)^3 \int_0^t \left\|\delta_s^L\right\|_{p^{in}} ds$ converge to 0 in probability, which implies $\Gamma_\infty(t) \le \Gamma_\infty(0)$. On the other hand, $\Gamma_L(t) \ge \Gamma_L(0)$ by construction and $\beta_L \to \infty$ as $L \to \infty$. Thus, $\Gamma_L(t)$ converges to $\Gamma_\infty(0)$ uniformly in probability for $t \in [0, T]$. This concludes the proof. □

Next, we consider similar Lyapunov functions but under ℓ_2 norm. This is because scaled NTK is not restricted to the dataset. For $1 \le j \le 8$,

$$\tilde{\Phi}_{L,j}(x,t) = \frac{1}{(L-1)C_L^j} \sum_{l=1}^{L-1} \left(\left\| f^l_{\theta^{(l)}(0)}(x) \right\| + \left\| f^l_{\theta^{(l)}(t)}(x) f^l_{\theta^{(l)}(0)}(x) \right\| \right)^j$$
$$\tilde{\Psi}_{L,2}(x,t) = \frac{1}{(L-1)^2} \sum_{l=2}^{L} \frac{L-1}{l-1} \sum_{i=2}^{l} \left(\left\| \mathfrak{W}_l^i(x,0) \right\| + \left\| \mathfrak{W}_l^i(x,t) - \mathfrak{W}_l^i(x,0) \right\| \right)^2$$
$$\tilde{\Psi}_{L,8}(x,t) = \frac{1}{L-1} \sum_{l=2}^{L} \left(\left\| \mathfrak{W}_L^l(x,0) \right\| + \left\| \mathfrak{W}_L^l(x,t) - \mathfrak{W}_L^l(x,0) \right\| \right)^8,$$

where $\tilde{\Phi}_{L,0} = 1$ and we define

$$\left(\left\|f^l_{\theta^{(l)}(0)}(x)\right\| + \left\|f^l_{\theta^{(l)}(t)}(x) - f^l_{\theta^{(l)}(0)}(x)\right\|\right)^j = \left(\left\|\mathrm{vec}\left(f^l_{\theta^{(l)}(0)}(x)\right)\right\| + \left\|\mathrm{vec}\left(f^l_{\theta^{(l)}(t)}(x)\right) - \mathrm{vec}\left(f^l_{\theta^{(l)}(0)}(x)\right)\right\|\right)^j.$$

for all j and l in $\tilde{\Phi}_{L,j}(t)$.

Note that Lyapunov functions under ℓ_2-norm are functions of (x,t). Similar to Lyapunov functions under p^{in}-norm, at initialization ($t = 0$), we have $\tilde{\Phi}_{L,j}(x,0) \xrightarrow{p} \tilde{\Phi}_{\infty,j}(x,0)$, $\tilde{\Psi}_{L,2}(x,0) \xrightarrow{p} \tilde{\Psi}_{\infty,2}(x,0)$, $\tilde{\Psi}_{L,8}(x,0) \xrightarrow{p} \tilde{\Psi}_{\infty,8}(x,0)$, where $\tilde{\Phi}_{L,\infty}(x,0), \tilde{\Psi}_{\infty,2}(x,0), \tilde{\Psi}_{\infty,8}(x,0)$ are constant by the law of large number. The similar proposition holds, which implies the intermediate weights and pre-activation values are invariant during training in ℓ_2-norm sense.

Proposition 5. *For $1 \le j \le 8$ and any $T > 0$,*

$$\tilde{\Phi}_{L,j}(x,t) \xrightarrow{p} \tilde{\Phi}_{\infty,j}(x,0), \quad \tilde{\Psi}_{L,2}(x,t) \xrightarrow{p} \tilde{\Psi}_{\infty,2}(x,0), \quad \tilde{\Psi}_{L,8}(x,t) \xrightarrow{p} \tilde{\Psi}_{\infty,8}(x,0)$$

uniformly for $t \in [0,T]$ as $L \to \infty$.

The following lemma, which corresponds to Lemma 25 , allows us to apply Grönwall's lemma.

Lemma 26. *For $t \ge 0$ and $1 \le j \le 8$, and $x \in \mathbb{R}^{d\times d}_+$, if $f^l_{\theta^{(l)}(t)}(x)$ is element-wise nonzero for all $1 \le l \le L-1$, then*

$$\partial_t \tilde{\Psi}_{L,2}(x,t) \le \frac{18d^4(L-1)}{C_L}\left(\tilde{\Psi}_{L,2}(x,t)\right)^2 \Phi_{L,4}(t)\Psi_{L,8}(t)\left\|\delta^L_t\right\|_{p^{in}}$$
$$\partial_t \tilde{\Phi}_{L,j}(x,t) \le \frac{j162d^4\max_l\{n_l^4\}N^{1/2}(L-1)}{C_L}\tilde{\Phi}_{L,j-1}(t)\tilde{\Phi}_{L,8}(x,t)\tilde{\Psi}_{L,2}(x,t)\Phi_{L,8}(t)\Psi_{L,8}(t)\left\|\delta^L_t\right\|_{p^{in}}$$
$$\partial_t \tilde{\Psi}_{L,8}(x,t) \le \frac{72d^4(L-1)}{C_L}\tilde{\Phi}_{L,4}(x,t)\tilde{\Psi}_{L,2}(x,t)\left(\tilde{\Psi}_{L,8}(x,t)\right)^2\Psi_{L,8}(t)\left\|\delta^L_t\right\|_{p^{in}}.$$

The proof of Lemma 26 is given in Section M.

Proof of Proposition 5. As we discussed in the proof of Proposition 4, $f^l_{\theta^{(l)}(t)}(x)$ is element-wise nonzero for large enough L for all $t \in [0,T]$, $1 \le l \le L-1$, and $x \in \mathbb{R}^{d\times d}_+$. From Lemma 26,

$$\tilde{\Psi}_{L,2}(x,t) \le \tilde{\Psi}_{L,2}(x,0) + \frac{18d^4(L-1)}{C_L}\int_0^t \left(\tilde{\Psi}_{L,2}(x,s)\right)^2 \Phi_{L,4}(s)\Psi_{L,8}(s)\left\|\delta^L_s\right\|_{p^{in}} ds.$$

Then using Grönwall's lemma where $z(t) = \tilde{\Psi}_{L,2}(x,t)$, $a(t) = \tilde{\Psi}_{L,2}(x,0)$, $b(t) = \frac{18d^4(L-1)}{C_L}$, $k(s) = \Phi_{L,4}(s)\Psi_{L,8}(s)\left\|\delta^L_s\right\|_{p^{in}}$, we get

$$\tilde{\Psi}_{L,2}(x,t) \le \tilde{\Psi}_{L,2}(x,0)\left\{1 - \frac{18d^4(L-1)}{C_L}\tilde{\Psi}_{L,2}(x,0)\int_0^t \Phi_{L,4}(s)\Psi_{L,4}(s)\left\|\delta^L_s\right\|_{p^{in}} ds\right\}^{-1}$$

for all $0 \le t \le \beta_L$. Similar to the proof of Proposition 4 , if $L^2/C_L \to 0$, then we get $\tilde{\Psi}_{L,2}(t) \xrightarrow{p} \tilde{\Psi}_{\infty,2}(0)$ uniformly for $t \in [0,T]$.

On the other hand, Lemma 26 also implies

$$\begin{aligned}
&\sum_{j=1}^8 \tilde{\Phi}_{L,j}(x,t) \\
&\quad \le \sum_{j=1}^8 \tilde{\Phi}_{L,j}(x,0) + \int_0^t \frac{1296d^4\max_l\{n_l^4\}N^{1/2}(L-1)}{C_L}\left(\sum_{j=1}^8 \tilde{\Phi}_{L,j-1}(x,s)\right)\tilde{\Phi}_{L,8}(x,s)\tilde{\Psi}_{L,2}(x,s)\Phi_{L,8}(s)\Psi_{L,8}(s)\left\|\delta^L_s\right\|_{p^{in}} ds \\
&\quad \le \sum_{j=1}^8 \tilde{\Phi}_{L,j}(x,0) + \int_0^t \frac{1296d^42\max_l\{n_l^4\}N^{1/2}(L-1)}{C_L}\left(\sum_{j=1}^8 \tilde{\Phi}_{L,j}(x,s)\right)^2 \tilde{\Psi}_{L,2}(x,s)\Phi_{L,8}(s)\Psi_{L,8}(s)\left\|\delta^L_s\right\|_{p^{in}} ds
\end{aligned}$$

$$\tilde{\Psi}_{L,8}(x,t) \le \tilde{\Psi}_{L,8}(x,0) + \int_0^t \frac{72d^4(L-1)}{C_L}\left(\tilde{\Psi}_{L,8}(x,s)\right)^2\tilde{\Phi}_{L,4}(x,s)\tilde{\Psi}_{L,2}(x,s)\Psi_{L,8}(s)\left\|\delta^L_s\right\|_{p^{in}} ds.$$

We can similarly apply Grönwall's lemma to show that $\sum_{j=1}^{8}\tilde{\Phi}_{L,j}(x,t)$ and $\tilde{\Psi}_{L,8}(x,t)$ converge to constant values. $\square$

Now, we consider our last and core Lyapunov function which contains intermediate weight and pre-activation values from a single layer. Define

$$\tilde{\Psi}_L^k(x,t) = \left\|\mathfrak{W}_L^k(x,0)\right\| + \left\|\mathfrak{W}_L^k(x,t) - \mathfrak{W}_L^k(x,0)\right\|$$
$$\tilde{\Phi}_L^l(x,t) = \frac{1}{C_L}\left(\left\|f^l_{\theta^{(l)}(0)}(x)\right\| + \left\|f^l_{\theta^{(l)}(t)}(x) - f^l_{\theta^{(l)}(0)}(x)\right\|\right)$$

for $1 \le l \le L$ and $2 \le k \le L$ and we define

$$\left\|f^l_{\theta^{(l)}(0)}(x)\right\| + \left\|f^l_{\theta^{(l)}(t)}(x) - f^l_{\theta^{(l)}(0)}(x)\right\| = \left\|\mathrm{vec}\left(f^l_{\theta^{(l)}(0)}(x)\right)\right\| + \left\|\mathrm{vec}\left(f^l_{\theta^{(l)}(t)}(x)\right) - \mathrm{vec}\left(f^l_{\theta^{(l)}(0)}(x)\right)\right\|_{p^{in}}.$$

for all l in $\tilde{\Phi}_L^l(x,t)$.

At initialization, $\tilde{\Psi}_L^k(x,0)$ and $\tilde{\Phi}_L^l(x,0)$ are stochastically bounded. Then the following proposition implies the invariance of Lyapunov functions during training.

Proposition 6. *For $1 \le l \le L$, $2 \le k \le L$ and any $T > 0$,*

$$\tilde{\Psi}_L^k(x,t) \xrightarrow{p} \tilde{\Psi}_\infty^k(x,0) \quad \tilde{\Phi}_L^l(x,t) \to \tilde{\Phi}_\infty^l(x,0).$$

uniformly for $t \in [0,T]$ as $L \to \infty$.

Lemma 27. *For $t \ge 0$ and $x \in \mathbb{R}_+^{d\times d}$, if $f^l_{\theta^{(l)}(t)}(x)$ is element-wise nonzero for all $1 \le l \le L-1$, then*

$$\partial_t \tilde{\Psi}_L^k(x,t) \le \frac{9d^4(L-1)}{C_L}\Psi_{L,8}(t)\tilde{\Phi}_{L,4}(x,t)\tilde{\Psi}_{L,2}(x,t)\tilde{\Psi}_{L,8}(x,t)\left\|\delta_t^L\right\|_{p^{in}}$$
$$\partial_t \tilde{\Phi}_L^l(x,t) \le \frac{81d^4\max_l\{n_l^4\}N^{1/2}(L-1)}{C_L}\Phi_{L,8}(t)\Psi_{L,8}(t)\tilde{\Phi}_{L,8}(x,t)\tilde{\Psi}_{L,2}(x,t)\left\|\delta_t^L\right\|_{p^{in}}.$$

for $1 \le l \le L$ and $2 \le k \le L$.

The proof of Lemma 27 is given in Section N.

Proof of Proposition 6. Again, we can assume that $f^l_{\theta^{(l)}(t)}(x)$ is element-wise nonzero for large enough L for all $t \in [0,T]$, $1 \le l \le L-1$, and $x \in \mathbb{R}_+^{d\times d}$. Lemma 27 implies

$$\tilde{\Psi}_L^k(x,t) \le \tilde{\Psi}_L^k(x,0) + \int_0^t \frac{9d^4(L-1)}{C_L}\Psi_{L,8}(t)\tilde{\Phi}_{L,4}(x,t)\tilde{\Psi}_{L,2}(x,t)\tilde{\Psi}_{L,8}(x,t)\left\|\delta_t^L\right\|_{p^{in}}\, ds$$
$$\tilde{\Phi}_L^l(x,t) \le \tilde{\Phi}_L^l(x,0) + \int_0^t \frac{81d^4\max_l\{n_l^4\}N^{1/2}(L-1)}{C_L}\Phi_{L,8}(t)\Psi_{L,8}(t)\tilde{\Phi}_{L,8}(x,t)\tilde{\Psi}_{L,2}(x,t)\left\|\delta_t^L\right\|_{p^{in}}\, ds.$$

Unike previous proofs, we do not need Grönwall's Lemma. From Proposition 4 and Proposition 5 , all terms in RHS (such as $\Psi_{L,8}(t)$, $\tilde{\Phi}_{L,4}(t)$, $\tilde{\Psi}_{L,2}(t)$, etc.) converge to constants in probability. Since $L^2/C_L \to 0$, integral terms converge to zero as $L \to \infty$. On the other hand, it is clear that $\tilde{\Psi}_L^k(x,t) \ge \tilde{\Psi}_L^k(x,0)$ and $\tilde{\Phi}_L^l(x,t) \ge \tilde{\Phi}_L^l(x,0)$ by construction. Thus, $\tilde{\Psi}_L^k(x,t)$ and $\tilde{\Phi}_L^l(x,t)$ converges to $\tilde{\Psi}_L^k(x,0)$ and $\tilde{\Phi}_L^l(x,0)$ in probability, respectively. Since the integral terms are independent from the choice of k and l, we get uniform convergence.

$\square$

Proposition 6 implies the following desired result which implies that the variation of intermediate pre-activation values and weights must be negligible.

Corollary 28. *For any x, as $L \to \infty$, we have*

$$\sup_{1\le l\le L, t\in[0,T]} \frac{1}{C_L}\left\|\mathrm{vec}\left(f^l_{\theta^{(l)}(t)}(x)\right) - \mathrm{vec}\left(f^l_{\theta^{(l)}(0)}(x)\right)\right\| \xrightarrow{p} 0$$
$$\sup_{2\le k\le L, t\in[0,T]} \left\|\mathfrak{W}_L^k(x,t) - \mathfrak{W}_L^k(x,0)\right\| \xrightarrow{p} 0.$$

With this corollary, we are now ready to prove our main theorem, the invariance of scaled NTK.

Theorem 5 (Invariance of scaled NTK)**.** *Let $T > 0$. Suppose $\int_0^T \left\| \delta|_{f_\theta^L} \right\|_{p^{in}} dt$ is stochastically bounded as $L \to \infty$. Then, for any $x, x' \in \mathbb{R}^{d\times d}$,*

$$S\left(\tilde{\Theta}_t^L(x,x')\right) \xrightarrow{p} \tilde{\Theta}^\infty(x,x')$$

uniformly for $t \in [0,T]$ as $L \to \infty$.

Proof. By definition,

$$\tilde{\Theta}_t^L(x,x') = \frac{1}{LC_L^2}\sum_{l=0}^{L-1}\sum_{c'csu}\mathfrak{W}_L^{l+2}(x,t)P_{c'}\left(\sigma\left(\left(f_{\theta(t)}^{l-1}(x)\right)_c\right)_{\psi_{su}}\right)\left(\frac{1}{C_L}P_{c'}\left(\sigma\left(\left(f_{\theta(t)}^{l-1}(x')\right)_c\right)_{\psi_{su}}\right)\right)^\intercal\mathfrak{W}_L^{l+2}(x',t)^\intercal$$
$$+\frac{1}{LC_L^2}\sum_{l=0}^{L-1}\mathfrak{W}_L^{l+2}(x,t)B^l\mathfrak{W}_L^{l+2}(x',t)^\intercal$$

Informally, Proposition implies that $f_{\theta^{(l)}(t)}^l(x)$ and $\mathfrak{W}_L^{l+2}(x,t)$ are effectively invariant, and therefore the kernel $\tilde{\Theta}_t^L(x,x')$ is invariant. An additional effort is required to handle the summation of L terms since L also increases. More formal proof is given in the following.

Since ReLU is 1-Lipshitz, Corollary 28 implies

$$\sup_{1\le l\le L, 1\le c\le n_l, t\in[0,T]}\left\|\frac{1}{C_L}P_{c'}\left(\sigma\left(\left(f_{\theta(t)}^{l-1}(x)\right)_c\right)_{\psi_{su}}\right)\left(\frac{1}{C_L}P_{c'}\left(\sigma\left(\left(f_{\theta(t)}^{l-1}(x')\right)_c\right)_{\psi_{su}}\right)\right)^\intercal - \right.$$
$$\left.\frac{1}{C_L}P_{c'}\left(\sigma\left(\left(f_{\theta(0)}^{l-1}(x)\right)_c\right)_{\psi_{su}}\right)\left(\frac{1}{C_L}P_{c'}\left(\sigma\left(\left(f_{\theta(0)}^{l-1}(x')\right)_c\right)_{\psi_{su}}\right)\right)^\intercal\right\| \xrightarrow{p} 0$$

as $L \to \infty$ for all u,s,x and x'. On the other hand, the norm of the difference between kernels is bounded by

$$\left\|\tilde{\Theta}_t^L(x,x') - \tilde{\Theta}_0^L(x,x')\right\|$$
$$\le \frac{1}{L}\sum_{l=0}^{L}\sum_{c'csu}\left\|\mathfrak{W}_L^{l+2}(x,t)\left(\frac{1}{C_L}P_{c'}\left(\sigma\left(\left(f_{\theta(t)}^{l-1}(x)\right)_c\right)_{\psi_{su}}\right)\left(\frac{1}{C_L}P_{c'}\left(\sigma\left(\left(f_{\theta(t)}^{l-1}(x')\right)_c\right)_{\psi_{su}}\right)\right)^\intercal\right)\mathfrak{W}_L^{l+2}(x',t)^\intercal\right.$$
$$\left. - \mathfrak{W}_L^{l+2}(x,0)\left(\frac{1}{C_L}P_{c'}\left(\sigma\left(\left(f_{\theta(0)}^{l-1}(x)\right)_c\right)_{\psi_{su}}\right)\left(\frac{1}{C_L}P_{c'}\left(\sigma\left(\left(f_{\theta(0)}^{l-1}(x')\right)_c\right)_{\psi_{su}}\right)\right)^\intercal\right)\mathfrak{W}_L^{l+2}(x',0)^\intercal\right\|$$
$$+\frac{1}{LC_L^2}\sum_{l=0}^{L-1}\left\|\left[\mathfrak{W}_L^{l+2}(x,t)B^l\mathfrak{W}_L^{l+2}(x',t)^\intercal\right] - \left[\mathfrak{W}_L^{l+2}(x,0)B^l\mathfrak{W}_L^{l+2}(x',0)^\intercal\right]\right\|$$
$$\le 27\max_{l,c}\left\|\mathfrak{W}_L^{l+2}(x,t)\left(\frac{1}{C_L}P_{c'}\left(\sigma\left(\left(f_{\theta(t)}^{l-1}(x)\right)_c\right)_{\psi_{su}}\right)\left(\frac{1}{C_L}P_{c'}\left(\sigma\left(\left(f_{\theta(t)}^{l-1}(x')\right)_c\right)_{\psi_{su}}\right)\right)^\intercal\right)\mathfrak{W}_L^{l+2}(x',t)^\intercal\right.$$
$$\left. - \mathfrak{W}_L^{l+2}(x,0)\left(\frac{1}{C_L}P_{c'}\left(\sigma\left(\left(f_{\theta(0)}^{l-1}(x)\right)_c\right)_{\psi_{su}}\right)\left(\frac{1}{C_L}P_{c'}\left(\sigma\left(\left(f_{\theta(0)}^{l-1}(x')\right)_c\right)_{\psi_{su}}\right)\right)^\intercal\right)\mathfrak{W}_L^{l+2}(x',0)^\intercal\right\|$$
$$+\frac{1}{C_L^2}\max_l\left\|\left[\mathfrak{W}_L^{l+2}(x,t)B^l\mathfrak{W}_L^{l+2}(x',t)^\intercal\right] - \left[\mathfrak{W}_L^{l+2}(x,0)B^l\mathfrak{W}_L^{l+2}(x',0)^\intercal\right]\right\|.$$

Thus, as $L \to \infty$, we have

$$\sup_{t\in[0,T]}\left\|\tilde{\Theta}_t^L(x,x') - \tilde{\Theta}_0^L(x,x')\right\| \xrightarrow{p} 0.$$

Finally, Theorem 5 implies $\tilde{\Theta}_0^L$ converges to $\tilde{\Theta}_0^\infty$ in probability, and therefore $\tilde{\Theta}_t^L$ converges to $\tilde{\Theta}_0^\infty$ in probability. This also implies

$$\sup_{t\in[0,T]}\left\|S\left(\tilde{\Theta}_t^L(x,x')\right)-\left(\tilde{\Theta}_0^L(x,x')\right)\right\|\xrightarrow{p}0.$$

□

I.4. Proof of Theorem 6

When the loss function is quadratic loss, stochastically bounded assumption of Theorem 5 also can be proved by Lemma 16.

Theorem 6 (Equivalence between deep CNN and kernel regression)**.** *Let $\mathcal{L}(f)=\frac{1}{2}\|f-f^\star\|_{p^{in}}^2$. Let $\tilde{\Theta}^\infty$ be positive definite. If f_t follows the kernel gradient flow with respect to $\tilde{\Theta}^\infty$, then for any $x\in\mathbb{R}_+^{d\times d}$,*

$$\lim_{t\to\infty}f_t(x)=f_{\mathrm{ntk}}(x).$$

Proof of Theorem 6. This can be proved using the exact same argument in the proof of Theorem 3. The only difference is the definition of $\tilde{\Theta}^\infty$, which does not affect the proof. □

J. Zero-crossing does not happen during training CNN

Similar to Section C, we now show that all entries of $\{f^l_{\theta^{(l)}(t)}(x_1),\dots f^l_{\theta^{(l)}(t)}(x_N),f^l_{\theta^{(l)}(t)}(x'),f^l_{\theta^{(l)}(t)}(x'')\}$ are nonzero with high probability for any $x',x''\in\mathbb{R}_+^{d\times d}$ and $T>0$. For the sake of notational simplicity, we set $x_{N+1}=x'$ and $x_{N+2}=x''$ in the following lemma.

Lemma 29. *Suppose $\int_0^T\left\|\delta|_{f_\theta^L}\right\|_{p^{in}}dt$ is stochastically bounded as $L\to\infty$. For any $T>0$, let*

$$T_L=\bigwedge_{i=1}^{N+2}\inf_t\left\{\left(f^l_{\theta^{(l)}(t)}(x_i)\right)_{rpq}=0\,\textit{for some}\;1\le l\le L-1,\,r,\,p,\textit{ and }q\right\}\bigwedge T,$$

where $x_{N+1},x_{N+2}\in\mathbb{R}_+^{d\times d}$ are arbitrary inputs. Then, for any $\epsilon>0$, there exists $L_{\max}$ such that Pr $[T_L=T]\ge1-\epsilon$ for all $L>L_{\max}$.

Proof of Lemma 15. For $\epsilon'>0$, let $M=1+\epsilon'$. Since $\int_0^T\left\|\delta|_{f_\theta^L}\right\|_{p^{in}}dt$ is stochastically bounded as $L\to\infty$, there exists $K_0>0$ and large enough $L_0>0$ such that

$$\Pr\left[\int_0^T\left\|\delta|_{f_\theta^L}\right\|_{p^{in}}dt>K_0\right]<\frac{\epsilon}{3}$$

for all $L\ge L_0$. We define a complement of such event by

$$E_0=\left\{\int_0^T\left\|\delta|_{f_\theta^L}\right\|_{p^{in}}dt\le K_0\right\},$$

where $\Pr[E_0]\ge1-\frac{\epsilon}{3}$.

On the other hand, recall that $\tilde{\Psi}_{\infty,2}(x_i,0),\tilde{\Psi}_{\infty,8}(x,0),\tilde{\Phi}_{\infty,j}(x_i,0),\Gamma_\infty(0)$ are all (non-random) constants for all $1\le j\le8$

and $1 \le i \le N+2$. Since $L^2/C_L \to 0$ as $L \to \infty$, we have

$$\begin{aligned}
\frac{3888 \max_l\{n_l^4\} d^4 N^{3/2} M^3 (L-1)}{C_L} \Gamma_\infty(0)^3 K_0 \to& 0, \\
\frac{18 d^4 M^5 (L-1)}{C_L} \tilde{\Psi}_{\infty,2}(x_i,0) \Gamma_\infty^2(0) K_0 \to& 0, \\
\frac{M^7 1296 d^4 2 \max_l\{n_l^4\} N^{1/2} (L-1)}{C_L} \left(\sum_{j=1}^{8} \tilde{\Phi}_{\infty,j}(x_i,0) \right) \tilde{\Psi}_{\infty,2}(0) \Gamma_\infty^2(0) K_0 \to& 0, \\
\frac{M^7 72 d^4 (L-1)}{C_L} \tilde{\Psi}_{\infty,8}(x_i,0) \left(\sum_{j=1}^{8} \tilde{\Phi}_{\infty,j}(x_i,0) \right) \tilde{\Psi}_{\infty,2}(x_i,0) \Gamma_\infty(0) K_0 \to& 0,
\end{aligned}$$

as $L \to \infty$. Thus, there exists L_1 such that

$$\begin{aligned}
\frac{3888 \max_l\{n_l^4\} d^4 N^{3/2} M^3 (L-1)}{C_L} \Gamma_\infty(0)^3 K_0 <& 1 - M^{-3}, \\
\frac{18 d^4 M^5 (L-1)}{C_L} \tilde{\Psi}_{\infty,2}(x_i,0) \Gamma_\infty^2(0) K_0 <& 1 - M^{-1}, \\
\frac{M^7 1296 d^4 2 \max_l\{n_l^4\} N^{1/2} (L-1)}{C_L} \left(\sum_{j=1}^{8} \tilde{\Phi}_{\infty,j}(x_i,0) \right) \tilde{\Psi}_{\infty,2}(0) \Gamma_\infty^2(0) K_0 <& 1 - M^{-1}, \\
\frac{M^7 72 d^4 (L-1)}{C_L} \tilde{\Psi}_{\infty,8}(x_i,0) \left(\sum_{j=1}^{8} \tilde{\Phi}_{\infty,j}(x_i,0) \right) \tilde{\Psi}_{\infty,2}(x_i,0) \Gamma_\infty(0) K_0 <& 1 - M^{-1},
\end{aligned}$$

for all $L \ge L_1$.

Now, we define following events which indicate Lyapunov functions converge to constant values by the law of large numbers.

$$\begin{aligned}
E_1^{(L)} &= \{\Gamma_L(0) \le M \Gamma_\infty(0)\} \\
E_2^{(L)} &= \bigcup_{i=1}^{N+2} \left\{ \tilde{\Psi}_{L,8}(x_i,0) \le M \tilde{\Psi}_{\infty,8}(x_i,0) \right\} \\
E_3^{(L)} &= \bigcup_{i=1}^{N+2} \left\{ \tilde{\Psi}_{L,2}(x_i,0) \le M \tilde{\Psi}_{\infty,2}(x_i,0) \right\} \\
E_4^{(L)} &= \bigcup_{i=1}^{N+2} \left\{ \sum_{j=1}^{8} \tilde{\Phi}_{L,j}(x_i,0) \le M \sum_{j=1}^{8} \tilde{\Phi}_{\infty,j}(x_i,0) \right\}.
\end{aligned}$$

By the law of large numbers, all events $E_1^{(L)}, E_2^{(L)}, E_3^{(L)}, E_4^{(L)}$ have probabilities converge to 1 as $L \to \infty$. Thus, there exists L_2 such that

$$\Pr\left[E_1^{(L)} \cap E_2^{(L)} \cap E_3^{(L)} \cap E_4^{(L)} \right] \ge 1 - \frac{\epsilon}{3}$$

for all $L \ge L_2$.

Finally, we are interested in the event

$$E_5 = \bigcup_{i=1}^{N+2} \left\{ \left| \left(f^l_{\theta(0)}(x_i) \right)_{rpq} \right| \ge \frac{L}{C_L}(K_1 + 1) \text{ for all } 1 \le l \le L-1, r, p, \text{and } q \right\},$$

where we specify $K_1 > 0$ later. Consider a random element of intermediate pre-activation values

$$g^1_{pq}(x) = \left(f^1_{\theta(0)}(x)\right)_{2pq} = \frac{1}{C_L}\left\langle w^1_{2,1}, \phi_{pq}(x)\right\rangle + \frac{1}{C_L}v^1$$
$$g^l_{pq}(x) = \left(f^l_{\theta^{(l)}(0)}(x)\right)_{2pq} = \left\langle w^l_{2,1}, \phi_{pq}(x)\right\rangle + \frac{1}{C_L}v^l \text{ for } 2 \leq l \leq L-1.$$

Then, $g^1_{pq} \sim \mathcal{N}(0, \sigma^2_{1pq})$ and $g^l_{pq} \sim \mathcal{N}(0, \sigma^2_{gpq})$ for $2 \leq l \leq L-1$, p, and q are independent Gaussian random variables with finite variances. Since

$$\Pr\left[|g^l_{pq}| < \tfrac{L}{C_L}(K+1) \text{ for some } 1 \leq l \leq L-1\right] \leq 2d^2(K+1)\frac{L(L-2)}{C_L}\frac{1}{\sqrt{2\pi\sigma_g^2}} + 2d^2(K+1)\frac{L}{C_L}\frac{1}{\sqrt{2\pi\sigma_1^2}}$$

by union bound, the probability converges to 0 as $L \to \infty$. Thus, there exists L_3 such that $\Pr[E_5] \geq 1 - \frac{\epsilon}{3}$ holds for all $L \geq L_3$.

Then, we would like to show the following claim.

Claim 2. *If $L \geq L_1$, and all events $E_0, E^{(L)}_1, E^{(L)}_2, E^{(L)}_3, E^{(L)}_4, E_5$ are given, then all elements of intermediate pre-activation values never cross zeros while training for $t \in [0, T]$.*

If the claim holds, then for $L \geq L_{\max} = \max\{L_0, L_1, L_2, L_3\}$,

$$\Pr\left[E_0 \cap E^{(L)}_1 \cap E^{(L)}_2 \cap E^{(L)}_3 \cap E^{(L)}_4 \cap E_5\right] \geq 1 - \epsilon,$$

which concludes the proof of Lemma 15. In the rest of the proof, we will prove the above claim. Suppose $L \geq L_1$ and all events $E_0, E^{(L)}_1, E^{(L)}_2, E^{(L)}_3, E^{(L)}_4, E_5$ are given. Since all intermediate pre-activation values at initialization

$$f^1_1 = C_L x, \quad f^1_2 = w^1_{2,1} * x + \mathbb{1}_{d\times d}v^1, \quad f^1_3 = C_L\mathbb{1}_{d\times d}$$
$$f^l_1 = C_L x, \quad f^l_2 = w^l_{2,1} * C_L x + \mathbb{1}_{d\times d}v^l, \quad f^l_3 = C_L\mathbb{1}_{d\times d}$$

are nonzero for $2 \leq l \leq L-1$ and $1 \leq i \leq N+2$ with probability 1. Then, for all $t \in [0, T_L]$, Lemma 10 holds which is essential in the proof of Proposition 1. In the proof, using Grönwall's lemma, we showed

$$\Gamma_L(t) \leq \Gamma_L(0)\left\{1 - \frac{3888\max_l\{n_l^4\}d^4N^{3/2}(L-1)}{C_L}\Gamma_L(0)^3\Gamma_L(0)^3\int_0^{\beta_L}\left\|\delta^L_s\right\|_{p^{in}}\,ds\right\}^{-\frac{1}{3}}$$

for $t \in \left[0, \beta^{(1)}_L\right]$, where

$$\beta^{(1)}_L = \sup\left\{t : \frac{3888\max_l\{n_l^4\}d^4N^{3/2}(L-1)}{C_L}\Gamma_L(0)^3\int_0^t\left\|\delta^L_s\right\|_{p^{in}}\,ds < 1\right\}.$$

Since $E^{(L)}_1$ is given, $\Gamma_L(0) \leq M\Gamma_\infty(0)$. Then, since E_0 is given,

$$\frac{3888\max_l\{n_l^4\}d^4N^{3/2}(L-1)}{C_L}\Gamma_L(0)^3\int_0^t\left\|\delta^L_s\right\|_{p^{in}}\,ds \leq \frac{3888M^3\max_l\{n_l^4\}d^4N^{3/2}(L-1)}{C_L}\Gamma_\infty(0)^3K_0$$
$$< 1 - M^{-3}$$

by definition of L_1. This implies that $\beta^{(1)}_L = T_L$ and $\Gamma_L(t) \leq M^2\Gamma_\infty(0)$ if E_0 and $E^{(L)}_1$ are given and $L \geq L_1$. Recall that $\Gamma_L(t) = \sum_{j=1}^8\Phi_{L,j}(t) + \Psi_{L,2}(t) + \Psi_{L,8}(t)$, and therefore $\Gamma_L(t) \leq M^2\Gamma_\infty(0)$ also implies $\Psi_{L,8}(t) \leq M^2\Gamma_\infty(0)$, $\Psi_{L,2}(t) \leq M^2\Gamma_\infty(0)$, and $\Phi_{L,j}(t) \leq M^2\Gamma_\infty(0)$ for all $1 \leq j \leq 8$.

Similarly, in the proof of Proposition 2, we proved

$$\tilde{\Psi}_{L,2}(x_i, t) \leq \tilde{\Psi}_{L,2}(x_i, 0)\left\{1 - \frac{18d^4(L-1)}{C_L}\tilde{\Psi}_{L,2}(x_i, 0)\int_0^t\Phi_{L,4}(s)\Psi_{L,4}(s)\left\|\delta^L_s\right\|_{p^{in}}\,ds\right\}^{-1}$$

for all $t \in \left[0, \beta_L^{(2)}\right]$, where

$$\beta_L^{(2)} = \sup\left\{t : \frac{18d^4(L-1)}{C_L}\tilde{\Psi}_{L,2}(x_i,0)\int_0^t \Phi_{L,4}(s)\Psi_{L,8}(s)\left\|\delta_s^L\right\|_{p^{in}} ds < 1\right\}.$$

We have $\tilde{\Psi}_{L,8}(x_i,0) < M\tilde{\Psi}_{\infty,2}(x_i,0)$ if $E_3^{(L)}$ is given. In this case, with a definition of L_1,

$$\begin{aligned}\frac{18d^4(L-1)}{C_L}\tilde{\Psi}_{L,2}(x_i,0)\int_0^t \Phi_{L,4}(s)\Psi_{L,8}(s)\left\|\delta_s^L\right\|_{p^{in}} ds &\leq \frac{18M^5d^4(L-1)}{C_L}\tilde{\Psi}_{\infty,2}(x_i,0)\Gamma_\infty^2(0)K_0 \\ &< 1-M^{-1}\end{aligned}$$

if E_0 is also given. This implies that $\beta_L^{(2)} = T_L$ and $\tilde{\Psi}_{L,2}(x_i,t) \leq M^2\tilde{\Psi}_{\infty,2}(x_i,0)$, if $E_3^{(L)}$ and E_0 are given and $L \geq L_1$. In similar way, we obtain $\sum_{j=1}^8 \tilde{\Phi}_{\infty,j}(x_i,t) \leq M^2\sum_{j=1}^8 \tilde{\Phi}_{\infty,j}(x_i,0)$ for $t \in [0,T_L]$ if $E_4^{(L)}$ and E_0 are given and $L \geq L_1$. Also, we have $\tilde{\Psi}_{L,8}(x_i,t) \leq M^2\tilde{\Psi}_{\infty,8}(x_i,0)$ for $t \in [0,T_L]$ if $E_2^{(L)}$ and E_0 are given and $L \geq L_1$.

Finally, in the proof of Proposition 3, we proved

$$\tilde{\Phi}_L^l(x,t) \leq \tilde{\Phi}_L^l(x,0) + \int_0^t \frac{81d^4\max_l\{n_l^4\}N^{1/2}(L-1)}{C_L}\Phi_{L,8}(t)\Psi_{L,8}(t)\tilde{\Phi}_{L,8}(x,t)\tilde{\Psi}_{L,2}(x,t)\left\|\delta_t^L\right\|_{p^{in}} ds.$$

Then, if $E_0, E_1^{(L)}, E_2^{(L)}, E_3^{(L)}, E_4^{(L)}$ are given and $L \geq L_1$,

$$\begin{aligned}&\int_0^{T_L} \frac{81d^4\max_l\{n_l^4\}N^{1/2}(L-1)}{C_L}\Phi_{L,8}(t)\Psi_{L,8}(t)\tilde{\Phi}_{L,8}(x,t)\tilde{\Psi}_{L,2}(x,t)\left\|\delta_t^L\right\|_{p^{in}} ds. \\ &\leq \frac{81M^8d^4\max_l\{n_l^4\}N^{1/2}(L-1)}{C_L}\Gamma_\infty^2(0)\left(\sum_{j=1}^8 \tilde{\Phi}_{\infty,j}(x_i,0)\right)\tilde{\Psi}_{\infty,2}(x_i,0)K_0.\end{aligned}$$

If we let $K_1 = \sup_{1\leq i\leq N+2} 81M^8d^4\max_l\{n_l^4\}N^{1/2}\Gamma_\infty^2(0)\left(\sum_{j=1}^8 \tilde{\Phi}_{\infty,j}(x_i,0)\right)\tilde{\Psi}_{\infty,2}(x_i,0)K_0$, then this inequality implies

$$\sup_{1\leq l\leq L-1, t\in[0,T_L]} \frac{1}{C_L}\left\|\mathrm{vec}\left(f^l_{\theta^{(l)}(t)}(x_i)\right) - \mathrm{vec}\left(f^l_{\theta^{(l)}(0)}(x_i)\right)\right\| < \frac{L}{C_L}K_1$$

for all i. On the other hand, if E_5 is given,

$$\frac{1}{C_L}\left|\left(f^l_{\theta(0)}(x_i)\right)_{rpq}\right| > \frac{L}{C_L}(K_1+1)$$

for all i, l, r, p, and q, and therefore

$$\inf_{1\leq l\leq L-1, t\in[0,T_L]}\left|\left(f^l_{\theta^{(l)}(t)}(x_i)\right)_{rpq}\right| > L$$

for all i, r, p, and q. This implies $T_L = T$, which concludes the proof of Claim 2. □

Equipped with Lemma 29, we provide a brief outline of the rigorous proof of Theorem 5. For any $T > 0$, let

$$T_L = \bigwedge_{i=1}^{N+2}\inf_t\left\{\left(f^l_{\theta^{(l)}(t)}(x_i)\right)_r = 0 \text{ for some } 1 \leq l \leq L-1 \text{ and } r\right\}\bigwedge T,$$

where $x_{N+1}, x_{N+2} \in \mathbb{R}_+^{d_{in}}$ are arbitrary inputs. Also, for $\epsilon, \epsilon' > 0$, where $M = 1+\epsilon'$, we can define events $E_0, E_1^{(L)}, E_2^{(L)}, E_3^{(L)}, E_4^{(L)}, E_5$ and constants K_0, L_0, L_1, L_2, L_3 as defined in the proof of Lemma 29. We further let

$$K_2 = \max\left\{K_1, \sup_{1\leq i\leq N+2} M^8 9d^4\Gamma_\infty(0)\left(\sum_{j=1}^8 \tilde{\Phi}_{\infty,j}(x_i,0)\right)\tilde{\Psi}_{\infty,2}(x_i,0)\tilde{\Psi}_{\infty,8}(x_i,0)K_0\right\}.$$

and

$$E_6 = \Bigg\{ \frac{1}{L} \sum_l \Bigg[\left(1 + \frac{L}{C_L} K_2 + \left\| \frac{1}{C_L} P_{c'} \left(\sigma \left(\left(f_{\theta(0)}^{l-1}(x) \right)_c \right)_{\psi_{su}} \right) \right\| \right) \left(1 + \frac{L}{C_L} K_2 + \left\| \frac{1}{C_L} P_{c'} \left(\sigma \left(\left(f_{\theta(0)}^{l-1}(x') \right)_c \right)_{\psi_{su}} \right) \right\| \right)$$
$$\times \left(1 + \frac{L}{C_L} K_2 + \left\| \mathfrak{W}_L^{l+2}(x,0) \right\| \right) \left(1 + \frac{L}{C_L} K_2 + \left\| \mathfrak{W}_L^{l+2}(x',0) \right\| \right) \Bigg] \times \frac{L}{C_L} K_2$$
$$+ \frac{1}{LC_L^2} \sum_l \left(1 + \frac{L}{C_L} K_2 + \left\| \mathfrak{W}_L^{l+2}(x,0) \right\| \right) \left(1 + \frac{L}{C_L} K_2 + \left\| \mathfrak{W}_L^{l+2}(x',0) \right\| \right) \times \frac{L}{C_L} K_2 < \epsilon'' \Bigg\}$$

for $\epsilon'' > 0$. Since $L^2/C_L \to 0$ as $L \to \infty$, the probability of $E_6^{(L)}$ converges to 1 by the law of large numbers, and there exists L_4 such that

$$\Pr\left[E_6^{(L)}\right] > 1 - \epsilon$$

for all $L \geq L_4$. Suppose all events $E_0, E_1^{(L)}, E_2^{(L)}, E_3^{(L)}, E_4^{(L)}, E_5, E_6^{(L)}$ are given, and $L \geq L_{\max} = \max\{L_0, L_1, L_2, L_3, L_4\}$. Then, instead of Proposition 4, the following inequalities hold without having $L \to \infty$:

$$\Phi_{L,j}(t) \leq M^2 \Gamma_\infty(0)$$
$$\Psi_{L,2}(t) \leq M^2 \Gamma_\infty(0)$$
$$\Psi_{L,8}(t) \leq M^2 \Gamma_\infty(0)$$

for all $1 \leq j \leq 8$ and $t \in [0, T_L]$ since $\beta_L^{(1)} = T_L$. Similarly, we can rewrite Proposition 5 without having $L \to \infty$:

$$\sum_{j=1}^{8} \tilde{\Phi}_{\infty,j}(x_i, 0) \leq \sum_{j=1}^{8} \tilde{\Phi}_{L,j}(x_i, t) \leq M^2 \sum_{j=1}^{8} \tilde{\Phi}_{\infty,j}(x_i, 0)$$
$$\tilde{\Psi}_{\infty,2}(x_i, 0) \leq \tilde{\Psi}_{L,2}(x_i, t) \leq M^2 \tilde{\Psi}_{\infty,2}(x_i, 0)$$
$$\tilde{\Psi}_{\infty,8}(x_i, 0) \leq \tilde{\Psi}_{L,8}(x_i, t) \leq M^2 \tilde{\Psi}_{\infty,8}(x_i, 0)$$

for all $1 \leq i \leq N+2$ and $t \in [0, T_L]$. Finally, Corollary 28 can be replaced by

$$\sup_{1 \leq l \leq L-1, t \in [0, T_L]} \left\| P_{c'} \left(\sigma \left(\left(f_{\theta(t)}^{l-1}(x) \right)_c \right)_{\psi_{su}} \right) - P_{c'} \left(\sigma \left(\left(f_{\theta(0)}^{l-1}(x) \right)_c \right)_{\psi_{su}} \right) \right\| < \frac{L}{C_L} K_1$$

for all i, c, c', s, u, where K_1 is as defined in the proof of Lemma 29. This already implies $T_L = T$.

In addition, by Lemma 27,

$$\sup_{1 \leq l \leq L-1, t \in [0, T]} \left\| \mathfrak{W}_L^k(x_i, t) - \mathfrak{W}_L^k(x_i, 0) \right\| < \frac{L}{C_L} K_2.$$

Thus, for $x, x' \in \{x_1, \ldots, x_{N+2}\}$, we can bound the deviation of $\tilde{\Theta}_t^L(x, x')$ by

$$\begin{aligned}
&\left\|\tilde{\Theta}_t^L(x,x') - \tilde{\Theta}_0^L(x,x')\right\| \\
&\leq \frac{1}{L}\sum_{l=0}^{L}\sum_{c'csu}\Bigg\|\mathfrak{W}_L^{l+2}(x,t)\left(\frac{1}{C_L}P_{c'}\left(\sigma\left(\left(f_{\theta(t)}^{l-1}(x)\right)_c\right)_{\psi_{su}}\right)\left(\frac{1}{C_L}P_{c'}\left(\sigma\left(\left(f_{\theta(t)}^{l-1}(x')\right)_c\right)_{\psi_{su}}\right)\right)^\intercal\right)\mathfrak{W}_L^{l+2}(x',t)^\intercal \\
&\quad - \mathfrak{W}_L^{l+2}(x,0)\left(\frac{1}{C_L}P_{c'}\left(\sigma\left(\left(f_{\theta(0)}^{l-1}(x)\right)_c\right)_{\psi_{su}}\right)\left(\frac{1}{C_L}P_{c'}\left(\sigma\left(\left(f_{\theta(0)}^{l-1}(x')\right)_c\right)_{\psi_{su}}\right)\right)^\intercal\right)\mathfrak{W}_L^{l+2}(x',0)^\intercal\Bigg\| \\
&\quad + \frac{1}{LC_L^2}\sum_{l=0}^{L-1}\left\|\left[\mathfrak{W}_L^{l+2}(x,t)B^l\mathfrak{W}_L^{l+2}(x',t)^\intercal\right] - \left[\mathfrak{W}_L^{l+2}(x,0)B^l\mathfrak{W}_L^{l+2}(x',0)^\intercal\right]\right\| \\
&\leq \frac{1}{L}\sum_l\Bigg[\left(\frac{L}{C_L}K_2 + \left\|\frac{1}{C_L}P_{c'}\left(\sigma\left(\left(f_{\theta(t)}^{l-1}(x)\right)_c\right)_{\psi_{su}}\right)\right\|\right)\left(\frac{L}{C_L}K_2 + \left\|\frac{1}{C_L}P_{c'}\left(\sigma\left(\left(f_{\theta(t)}^{l-1}(x')\right)_c\right)_{\psi_{su}}\right)\right\|\right) \\
&\qquad \times\left(\frac{L}{C_L}K_2 + \left\|\mathfrak{W}_L^{l+2}(x,0)\right\|\right)\left(\frac{L}{C_L}K_2 + \left\|\mathfrak{W}_L^{l+2}(x',0)\right\|\right) \\
&\qquad - \left\|\frac{1}{C_L}P_{c'}\left(\sigma\left(\left(f_{\theta(0)}^{l-1}(x)\right)_c\right)_{\psi_{su}}\right)\right\|\left\|\frac{1}{C_L}P_{c'}\left(\sigma\left(\left(f_{\theta(0)}^{l-1}(x)\right)_c\right)_{\psi_{su}}\right)\right\|\left\|\mathfrak{W}_L^{l+2}(x,0)\right\|\left\|\mathfrak{W}_L^{l+2}(x',0)\right\|\Bigg] \\
&\quad + \frac{1}{LC_L^2}\sum_l\Bigg[\left(\frac{L}{C_L}K_2 + \left\|\mathfrak{W}_L^{l+2}(x,0)\right\|\right)\left(\frac{L}{C_L}K_2 + \left\|\mathfrak{W}_L^{l+2}(x',0)\right\|\right) \\
&\qquad - \left\|\mathfrak{W}_L^{l+2}(x,0)\right\|\left\|\mathfrak{W}_L^{l+2}(x',0)\right\|\Bigg] \\
&\leq \frac{1}{L}\sum_l\Bigg[\left(1 + \frac{L}{C_L}K_2 + \left\|\frac{1}{C_L}P_{c'}\left(\sigma\left(\left(f_{\theta(0)}^{l-1}(x)\right)_c\right)_{\psi_{su}}\right)\right\|\right)\left(1 + \frac{L}{C_L}K_2 + \left\|\frac{1}{C_L}P_{c'}\left(\sigma\left(\left(f_{\theta(0)}^{l-1}(x')\right)_c\right)_{\psi_{su}}\right)\right\|\right) \\
&\qquad \times\left(1 + \frac{L}{C_L}K_2 + \left\|\mathfrak{W}_L^{l+2}(x,0)\right\|\right)\left(1 + \frac{L}{C_L}K_2 + \left\|\mathfrak{W}_L^{l+2}(x',0)\right\|\right)\Bigg] \times \frac{L}{C_L}K_2 \\
&\quad + \frac{1}{LC_L^2}\sum_l\left(1 + \frac{L}{C_L}K_2 + \left\|\mathfrak{W}_L^{l+2}(x,0)\right\|\right)\left(1 + \frac{L}{C_L}K_2 + \left\|\mathfrak{W}_L^{l+2}(x',0)\right\|\right) \times \frac{L}{C_L}K_2 \\
&\leq \epsilon''
\end{aligned}$$

Finally,

$$\Pr\left[\left\|\tilde{\Theta}_t^L(x,x') - \tilde{\Theta}_0^L(x,x')\right\| < \epsilon''\right] \geq 1 - 2\epsilon$$

for all $L \geq L'_{\max}$. Since ϵ, ϵ', and ϵ'' are arbitrary, we can conclude $S\left(\tilde{\Theta}_t^L(x,x')\right) \xrightarrow{p} S\left(\tilde{\Theta}^\infty(x,x')\right)$.

K. Equalities and inequalities of sub layers for CNN

In this section, we introduce some useful equalities and inequalities in our problem setting. Like Section E, we use $f^l \equiv f^l_{\theta^{(l)}}$ if it is clear from the context.

First, following lemma provides bound of Jacobian.

Lemma 30. *For* $1 \leq l \leq L$*,*

$$\max\left\{\left\|\partial_{w^l}f^l\right\|_{p^{in}}, \left\|\partial_{w^l}f^l\right\|\right\} \leq 9n_l n_{l-1}\left\|\left(\sigma(f^{l-1}(x))\right)\right\|_{p^{in}}$$

Proof of Lemma 30.

$$\begin{aligned}\left\|\partial_{w^l} f^l\right\|_{p^{in}} &\leq \sum_{c'csu} \left\|\partial_{w_{c'csu}} f^l\right\|_{p^{in}} \\ &\leq \sum_{c'csu} \left\|P_{c'}\left(\sigma(f_c^{l-1}(x))_{\psi_{su}}\right)\right\|_{p^{in}} \\ &\leq 9n_l n_{l-1}\left\|\sigma(f^{l-1}(x))\right\|_{p^{in}}\end{aligned}$$

We can bound $\left\|\partial_{w^l} f^l\right\|$ similarly. □

Lemma 31. *For $1 \leq l \leq L$,*

$$\max\left\{\left\|\partial_{b^l} f^l\right\|_{p^{in}}, \left\|\partial_{b^l} f^l\right\|\right\} \leq d^2.$$

As previously mentioned, we set the scaling factor of gradient flow by $\frac{1}{(L-1)C_L^2}$ throughout the proof. The gradient flow is given by

$$\begin{aligned}\partial_t \theta^l &= -\frac{1}{(L-1)C_L^2}\mathbb{E}_{x\sim p^{in}}[(\partial_{\theta^l} f^l(x))^\intercal \delta_t^l] \\ \partial_t f^l(x') &= -\frac{1}{(L-1)C_L^2}\mathbb{E}_{x\sim p^{in}}[\partial_{\theta^l} f^l(x')(\partial_{\theta^l} f^l(x))^\intercal \delta_t^l],\end{aligned}$$

where $\delta_t^j = \left(\mathfrak{W}_L^{j+1}(x,t)\right)^\intercal \delta_t^L$ since

$$\partial_{\theta^l} f^L = \mathfrak{W}_L^{l+1}(x,t)\partial_{\theta^l} f^l \quad \text{for } L \geq l.$$

Then, by the chain rule, we obtain the following lemma.

Lemma 32. *For $t \leq 0$ and $1 \leq l \leq L$,*

$$\left\|\partial_t w^l\right\|_{p^{in}} \leq \frac{9n_l n_{l-1}}{(L-1)C_L^2}\left\|\delta_t^l\right\|_{p^{in}}\left\|\sigma(f^{l-1})\right\|_{p^{in}}, \quad \left\|\partial_t b^l\right\|_{p^{in}} \leq \frac{d^2}{(L-1)C_L^2}\left\|\delta_t^l\right\|_{p^{in}}$$

Proof of Lemma 32.

$$\begin{aligned}\left\|\partial_t w^l\right\|_{p^{in}} &= \left\|\frac{1}{(L-1)C_L^2}\mathbb{E}_{x\sim p^{in}}[(\partial_{w^l} f^l(x))^\intercal \delta_t^l]\right\| \\ &\leq \frac{1}{(L-1)C_L^2}\mathbb{E}_{x\sim p^{in}}\left[\|(\partial_{w^l} f^l(x))^\intercal \delta_t^l\|\right] \\ &\leq \frac{1}{(L-1)C_L^2}\mathbb{E}_{x\sim p^{in}}\left[\|\partial_{w^l} f^l(x)\|\|\delta_t^l\|\right] \\ &\leq \frac{1}{(L-1)C_L^2}\sqrt{\mathbb{E}_{x\sim p^{in}}\left[\|\partial_{w^l} f^l(x)\|^2\right]}\sqrt{\mathbb{E}_{x\sim p^{in}}\left[\|\delta_t^l\|^2\right]} \\ &= \frac{1}{(L-1)C_L^2}\left\|\partial_{w^l} f^l(x)\right\|_{p^{in}}\left\|\delta_t^l\right\|_{p^{in}} \\ &\leq \frac{9n_l n_{l-1}}{(L-1)C_L^2}\left\|\sigma(f^{l-1})\right\|_{p^{in}}\left\|\delta_t^l\right\|_{p^{in}},\end{aligned}$$

where the first inequality is from Jensen's inequality, third inequality comes from Cauchy's inequality and last inequality follows from Lemma 30. Finally, we can bound $\left\|\partial_t b^l\right\|_{p^{in}}$ similarly. □

Also we can induce following lemma.

Lemma 33. *Suppose all elements of $f^i_{\theta^{(i)}(t)}(x')$ are nonzero for all $1 \le l \le L-1$ and x' in dataset. Then, for $2 \le l \le L$,*

$$\max\left\{\left\|\partial_t \mathfrak{W}^l_l\right\|_{p^{in}}, \left\|\partial_t \mathfrak{W}^l_l\right\|\right\} \le \frac{9d^4}{(L-1)C_L^2}\left\|\sigma(f^{l-1}(x))\right\|_{p^{in}}\left\|\delta^l_t\right\|_{p^{in}}.$$

Proof of Lemma 33. By the chain rule, we have

$$\begin{aligned}
&\left\|\partial_t\left(\partial_{f^{l-1}_{c,i,j}} f^l_{c',s,u}(x)\right)\right\|_{p^{in}} \\
&= \left\|\partial_t\left(w^l_{c',c,i-s+2,j-u+2}\dot\sigma(f^{l-1}_{c,i,j})\mathbf{1}_{\{s-1\le i\le s+1\}}\mathbf{1}_{\{u-1\le j\le u+1\}}\right)\right\|_{p^{in}} \\
&= \frac{1}{(L-1)C_L^2}\left\|\mathbb{E}_{x\sim p^{in}}\left[(\partial_{w^l_{c',c,i-s+2,j-u+2}} f^l(x))^\intercal \delta^l_t\right]\right\|\left\|\dot\sigma(f^{l-1}_{c,i,j}(x))\right\|_{p^{in}}\mathbf{1}_{\{s-1\le i\le s+1\}}\mathbf{1}_{\{u-1\le j\le u+1\}} \\
&\le \frac{1}{(L-1)C_L^2}\left\|\sigma(f^{l-1}_c(x))_{\psi_{i-s+2,j-u+2}}\right\|_{p^{in}}\left\|\delta^l_t\right\|_{p^{in}}\mathbf{1}_{\{s-1\le i\le s+1\}}\mathbf{1}_{\{u-1\le j\le u+1\}}.
\end{aligned}$$

Hence

$$\left\|\partial_t \mathfrak{W}^l_l\right\|_{p^{in}} \le \frac{9d^4}{(L-1)C_L^2}\left\|\sigma(f^{l-1}(x))\right\|_{p^{in}}\left\|\delta^l_t\right\|_{p^{in}}$$

and we can bound $\left\|\partial_t \mathfrak{W}^l_l\right\|_{p^{in}}$ similarly. ∎

From gradient flow, we get following lemma provides a bound of the norm.

Lemma 34. *For $1 \le l \le L$ and $t \le 0$,*

$$\left\|\partial_t f^l(x')\right\|_{p^{in}} \le \frac{N}{(L-1)C_L^2}\mathbb{E}_{x,x'\sim p^{in}}\left[\left\|(\partial_{\theta^l} f^l(x'))(\partial f^l_{\theta^l}(x))^\intercal \delta^L_t(x)\right\|\right]$$

Following lemma is the key lemma to prove invariances of Lyapunov functions and it's only factor different compared to MLP setting.

Lemma 35. *Suppose all elements of $f^i_{\theta^{(i)}(t)}(x')$ are nonzero for all $1 \le i \le L-1$ and x' in dataset. Then, for $t \in [0,T]$, $1 \le k \le j \le L$, and $1 \le i \le L$,*

$$\begin{aligned}
\partial_t\left\|\mathfrak{W}^k_j(x,t)\right\|_{p^{in}} &\le \frac{9d^4N}{(L-1)C_L^2}\sum_{i=k}^{j}\left\|\mathfrak{W}^{i+1}_j(x,t)\right\|_{p^{in}}\left\|\mathfrak{W}^k_{i-1}(x,t)\right\|_{p^{in}}\left\|\delta^i_t\right\|_{p^{in}}\left\|\sigma\left(f^{i-1}_{\theta^{(i-1)}(t)}(x)\right)\right\|_{p^{in}} \\
\partial_t\left\|f^i_{\theta^{(i)}(t)}(x')\right\|_{p^{in}} &\le \frac{81n_i^2n_{i-1}^2N^{3/2}}{(L-1)C_L^2}\sum_{l=0}^{i-1}\left\|f^l_{\theta^{(l)}(t)}(x)\right\|_{p^{in}}\left\|f^l_{\theta^{(l)}(t)}(x')\right\|_{p^{in}}\left\|\mathfrak{W}^{l+2}_i\right\|_{p^{in}}\left\|(\mathfrak{W}^{l+2}_L)^\intercal\right\|_{p^{in}}\left\|\delta^L_t\right\|_{p^{in}} \\
&\quad + \frac{d^4N}{(L-1)C_L^2}\sum_{l=0}^{i-1}\left\|\mathfrak{W}^{l+2}_i\right\|_{p^{in}}\left\|(\mathfrak{W}^{l+2}_L)^\intercal\right\|_{p^{in}}\left\|\delta^L_t\right\|_{p^{in}}.
\end{aligned}$$

Proof of Lemma 22 . By the chain rule,

$$\begin{aligned}
\partial_t\left\|\mathfrak{W}^k_j(x,t)\right\|_{p^{in}} &\le \sum_{i=k}^{j}\left\|\mathfrak{W}^{i+1}_j \partial_t \mathfrak{W}^i_i \mathfrak{W}^k_{i-1}\right\|_{p^{in}} \\
&\le N\sum_{i=k}^{j}\left\|\mathfrak{W}^{i+1}_j\right\|_{p^{in}}\left\|\partial_t \mathfrak{W}^i_i\right\|_{p^{in}}\left\|\mathfrak{W}^k_{i-1}\right\|_{p^{in}},
\end{aligned}$$

where $\mathfrak{W}^{i+1}_j$ denotes $\mathfrak{W}^{i+1}_j(x,t)$ if it is clear from the context. Then, by previous Lemma 33,

$$\partial_t\left\|\mathfrak{W}^k_j(x,t)\right\|_{p^{in}} \le \frac{9d^4N}{(L-1)C_L^2}\sum_{i=k}^{j}\left\|\mathfrak{W}^{i+1}_j(x,t)\right\|_{p^{in}}\left\|\mathfrak{W}^k_{i-1}(x,t)\right\|_{p^{in}}\left\|\delta^i_t\right\|_{p^{in}}\left\|\sigma(f^{i-1}_{\theta^{(i-1)}(t)}(x)\right\|_{p^{in}}.$$

On the other hand, from Lemma 34 and 24,

$$\begin{aligned}
\left\|\partial_t f^l_{\theta^{(l)}(t)}(x)\right\|_{p^{in}} &\leq \frac{N}{(L-1)C_L^2}\mathbb{E}_{x,x'\sim p^{in}}\left[\left\|\partial_{\theta^l} f^l_{\theta^{(l)}(t)}(x')\left(\partial_{\theta^l} f^l_{\theta^{(l)}(t)}(x)\right)^\intercal \left(\mathfrak{W}_L^{l+1}(x)\right)^\intercal \delta_t^L(x)\right\|\right] \\
&\leq \frac{N}{(L-1)C_L^2}\mathbb{E}_{x,x'\sim p^{in}}\Bigg[\sum_{i=0}^{l-1}\Bigg\|\sum_{c'csu}\mathfrak{W}_L^{l+2}(x,t)P_{c'}\left(\sigma(f^l_c(x))_{\psi_{su}}\right)\left(P_{c'}\left(\sigma(f^l_c(x'))_{\psi_{su}}\right)\right)^\intercal \mathfrak{W}_L^{l+2}(x',t)^\intercal + \\
&\mathfrak{W}_L^{l+2}(x,t)B^l\mathfrak{W}_L^{l+2}(x',t)^\intercal\delta_t^L(x)\Bigg\|\Bigg] \\
&\leq \frac{81\max_l\{n_l^4\}N}{(L-1)C_L^2}\sum_{i=0}^{l-1}\mathbb{E}_{x,x'\sim p^{in}}\left[\|(\sigma(f^i(x))\|\|\sigma(f^i(x'))\|\|\mathfrak{W}_l^{i+2}(x')\|\|(\mathfrak{W}_L^{i+2}(x))^\intercal\|\|\delta_t^L(x)\|\right] \\
&\quad + \frac{d^4N}{(L-1)C_L^2}\sum_{i=0}^{l-1}\mathbb{E}_{x,x'\sim p^{in}}\left[\|\mathfrak{W}_l^{i+2}(x')\|\|(\mathfrak{W}_L^{i+2}(x))^\intercal\|\|\delta_t^L(x)\|\right] \\
&\leq \frac{81\max_l\{n_l^4\}N^{3/2}}{(L-1)C_L^2}\sum_{i=0}^{l-1}\left\|\sigma(f^i(x))\right\|_{p^{in}}\left\|\sigma(f^i(x'))\right\|_{p^{in}}\left\|\mathfrak{W}_l^{i+2}\right\|_{p^{in}}\left\|(\mathfrak{W}_L^{i+2})^\intercal\right\|_{p^{in}}\left\|\delta_t^L(x)\right\|_{p^{in}} \\
&\quad + \frac{d^4N}{(L-1)C_L^2}\sum_{i=0}^{l-1}\left\|\mathfrak{W}_l^{i+2}\right\|_{p^{in}}\left\|(\mathfrak{W}_L^{i+2})^\intercal\right\|_{p^{in}}\left\|\delta_t^L(x)\right\|_{p^{in}},
\end{aligned}$$

where third inequality comes from Lemma 30 and 31 and the last inequality holds due to the property of p^{in}-norm

$$\mathbb{E}_{x\sim p^{in}}\left[\|(\sigma(f^i(x))\|\|(\mathfrak{W}_L^{i+2})^\intercal\|\|\delta_t^L(x)\|\right] \leq \sqrt{N}\left\|(\sigma(f^i(x))\right\|_{p^{in}}\left\|(\mathfrak{W}_L^{i+2})^\intercal\right\|_{p^{in}}\left\|\delta_t^L(x)\right\|_{p^{in}}.$$

□

We have a similar bounds for derivatives of ℓ_2-norms.

Lemma 36. *For $x' \in \mathbb{R}_+^{d\times d}$, if all elements of $f^i_{\theta^{(i)}(t)}(x')$ are nonzero for all $1 \leq i \leq L-1$, then, for $t \in [0,T]$, $1 \leq l \leq j \leq L$, and $1 \leq i \leq L$*

$$\begin{aligned}
\partial_t\left\|\mathfrak{W}_j^k(x',t)\right\| &\leq \frac{9d^4}{(L-1)C_L^2}\sum_{l\geq i\geq k}\left\|\mathfrak{W}_l^{i+1}(x')\right\|\left\|\mathfrak{W}_{i-1}^k(x')\right\|\left\|\delta_t^i\right\|_{p^{in}}\left\|\sigma(f^{i-1})^\intercal\right\|_{p^{in}} \\
\partial_t\left\|f^i_{\theta^{(i)}(t)}(x')\right\| &\leq \frac{81n_i^2n_{i-1}^2N^{1/2}}{(L-1)C_L^2}\sum_{i=0}^{l-1}\left\|\sigma(f^i(x)\right\|_{p^{in}}\left\|\sigma(f^i(x')\right\|\left\|\mathfrak{W}_l^{i+2}(x')\right\|\left\|(\mathfrak{W}_L^{i+2})^\intercal\right\|_{p^{in}}\left\|\delta_t^L(x)\right\|_{p^{in}} \\
&\quad + \frac{d^4}{(L-1)C_L^2}\sum_{i=0}^{l-1}\left\|\mathfrak{W}_l^{i+2}(x')\right\|\left\|(\mathfrak{W}_L^{i+2})^\intercal\right\|_{p^{in}}\left\|\delta_t^L(x)\right\|_{p^{in}}.
\end{aligned}$$

Proof of Lemma 36. The proof is similar to that of Lemma 35.

$$\begin{aligned}
\partial_t\left\|\mathfrak{W}_l^k(x)\right\| &\leq \left\|\partial_t\mathfrak{W}_l^k(x)\right\| \\
&\leq \sum_{i=k}^{j}\left\|\mathfrak{W}_j^{i+1}(x)\right\|\left\|\partial_t\mathfrak{W}_i^i(x)\right\|_{p^{in}}\left\|\mathfrak{W}_{i-1}^k(x)\right\| \\
&\leq \frac{9d^4}{(L-1)C_L^2}\sum_{i=k}^{j}\left\|\mathfrak{W}_j^{i+1}(x,t)\right\|_{p^{in}}\left\|\mathfrak{W}_{i-1}^k(x,t)\right\|_{p^{in}}\left\|\delta_t^i\right\|_{p^{in}}\left\|\sigma(f^{i-1}_{\theta^{(i-1)}(t)}(x)\right\|_{p^{in}}
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\left\|\partial_t f^l(t)\right\| &\leq \frac{1}{(L-1)C_L^2}\left[\left\|\partial_{\theta^l} f^l(x')\mathbb{E}_{x\sim p^{in}}\left[\left(\partial f^l_{\theta^l}(x)\right)^\intercal\left(\mathfrak{W}_L^{l+1}(x)\right)^\intercal \delta_t^L(x)\right]\right\|\right]\\
&\leq \frac{1}{(L-1)C_L^2}\mathbb{E}_{x\sim p^{in}}\left[\sum_{i=0}^{l-1}\left\|\sum_{c'csu}\mathfrak{W}_L^{l+2}(x,t)P_{c'}\left(\sigma(f_c^l(x))_{\psi_{su}}\right)\left(P_{c'}\left(\sigma(f_c^l(x'))_{\psi_{su}}\right)\right)^\intercal\mathfrak{W}_L^{l+2}(x',t)^\intercal+\right.\right.\\
&\left.\left.\mathfrak{W}_L^{l+2}(x,t)B^l\mathfrak{W}_L^{l+2}(x',t)^\intercal\delta_t^L(x)\right\|\right]\\
&\leq \frac{81\max_l\{n_l^4\}}{(L-1)C_L^2}\sum_{i=0}^{l-1}\mathbb{E}_{x\sim p^{in}}\left[\|(\sigma(f^i(x))\|\|\sigma(f^i(x')\|\|\mathfrak{W}_l^{i+2}(x')\|\|(\mathfrak{W}_L^{i+2}(x))^\intercal\|\|\delta_t^L(x)\|\right]\\
&\quad+\frac{d^4}{(L-1)C_L^2}\sum_{i=0}^{l-1}\mathbb{E}_{x\sim p^{in}}\left[\|\mathfrak{W}_l^{i+2}(x')\|\|(\mathfrak{W}_L^{i+2}(x))^\intercal\|\|\delta_t^L(x)\|\right]\\
&\leq \frac{81\max_l\{n_l^4\}N^{1/2}}{(L-1)C_L^2}\sum_{i=0}^{l-1}\left\|\sigma(f^i(x)\right\|_{p^{in}}\left\|\sigma(f^i(x')\right\|_{p^{in}}\left\|\mathfrak{W}_l^{i+2}\right\|_{p^{in}}\left\|(\mathfrak{W}_L^{i+2})^\intercal\right\|_{p^{in}}\left\|\delta_t^L(x)\right\|_{p^{in}}\\
&\quad+\frac{d^4}{(L-1)C_L^2}\sum_{i=0}^{l-1}\left\|\mathfrak{W}_l^{i+2}\right\|_{p^{in}}\left\|(\mathfrak{W}_L^{i+2})^\intercal\right\|_{p^{in}}\left\|\delta_t^L(x)\right\|_{p^{in}}.
\end{aligned}$$

□

L. Proof of Lemma 25 for Proposition 4

Lemma 25. *For $t \geq 0$, if $f^l_{\theta^{(l)}(t)}(x)$ is element-wise nonzero for all $1 \leq l \leq L-1$ and $x \in \{x_1, \ldots, x_N\}$, then*

$$\begin{aligned}
\partial_t\Phi_{L,j}(t) &\leq \frac{162\max_l\{n_l^4\}d^4 jN^{3/2}(L-1)}{C_L}\Phi_{L,j-1}(t)\Phi_{L,8}(t)\Psi_{L,2}(t)\Psi_{L,8}(t)\left\|\delta_t^L\right\|_{p^{in}}\\
\partial_t\Psi_{L,2}(t) &\leq \frac{18d^4N(L-1)}{C_L}\Phi_{L,4}(t)\left(\Psi_{L,2}(t)\right)^2(t)\Psi_{L,8}(t)\left\|\delta_t^L\right\|_{p^{in}}\\
\partial_t\Psi_{L,8}(t) &\leq \frac{72d^4N(L-1)}{C_L}\Phi_{L,4}(t)\Psi_{L,2}(t)\left(\Psi_{L,8}(t)\right)^2\left\|\delta_t^L\right\|_{p^{in}}
\end{aligned}$$

$1 \leq j \leq 8$.

Proof of Lemma 25. Consider the derivative of $\Psi_{L,8}$. Like Section C, we show that the intermediate pre-activation values never reach zero while training, and therefore we can apply Lemma 35.

$$\begin{aligned}
&\partial_t\Psi_{L,8}(t)\\
&\leq \frac{72Nd^4}{(L-1)^2C_L^2}\sum_{l=2}^{L}\sum_{l=i}^{L}\left\|\mathfrak{W}_L^{l+1}(t)\right\|_{p^{in}}^2\left\|\mathfrak{W}_{l-1}^i(t)\right\|_{p^{in}}\left(\left\|\mathfrak{W}_L^i(0)\right\|_{p^{in}}+\left\|\mathfrak{W}_L^i(t)-\mathfrak{W}_L^i(0)\right\|_{p^{in}}\right)^7\\
&\quad\times\left\|f^{l-1}_{\theta^{(l-1)}(t)}(x)\right\|_{p^{in}}\left\|\delta_t^L\right\|_{p^{in}}.
\end{aligned}$$

First, the following rough bound holds.

$$\begin{aligned}
&\sum_{l=2}^{L}\sum_{l=i}^{L}\left\|\mathfrak{W}_L^{l+1}(t)\right\|_{p^{in}}^2\left\|\mathfrak{W}_{l-1}^{i}(t)\right\|_{p^{in}}\left(\left\|\mathfrak{W}_L^{i}(0)\right\|_{p^{in}}+\left\|\mathfrak{W}_L^{i}(t)-\mathfrak{W}_L^{i}(0)\right\|_{p^{in}}\right)^7\left\|f_{\theta^{(l-1)}(t)}^{l-1}(x)\right\|_{p^{in}}\\
&\leq\left(\sum_{i=2}^{L}\sum_{l=i}^{L}\left\|\mathfrak{W}_{l-1}^{i}(t)\right\|_{p^{in}}\left(\left\|f_{\theta^{(l-1)}(0)}^{l-1}\right\|_{p^{in}}+\left\|f_{\theta^{(l-1)}(t)}^{l-1}-f_{\theta^{(l-1)}(0)}^{l-1}\right\|_{p^{in}}\right)\right.\\
&\quad\left.\times\left(\left\|\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}+\left\|\mathfrak{W}_L^{l+1}(t)-\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}\right)^2\right)\sum_{i=2}^{L}\left(\left\|\mathfrak{W}_L^{i}(0)\right\|_{p^{in}}+\left\|\mathfrak{W}_L^{i}(t)-\mathfrak{W}_L^{i}(0)\right\|_{p^{in}}\right)^8\\
&\leq(L-1)\Psi_{L,8}(L-1)^2C_L\left(\Psi_{L,2}\right)^{(1/2)}\left(\Phi_{L,4}\right)^{(1/4)}\left(\Psi_{L,8}\right)^{(1/4)},
\end{aligned}$$

where the last inequality is from

$$\begin{aligned}
&\left(\sum_{i=2}^{L}\sum_{l=i}^{L}\left\|\mathfrak{W}_{l-1}^{i}\right\|_{p^{in}}\left(\left\|f_{\theta^{(l-1)}(0)}^{l-1}\right\|_{p^{in}}+\left\|f_{\theta^{(l-1)}(t)}^{l-1}-f_{\theta^{(l-1)}(0)}^{l-1}\right\|_{p^{in}}\right)\left(\left\|\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}+\left\|\mathfrak{W}_L^{l+1}(t)-\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}\right)^2\right)^2\\
&\leq\left(\sum_{i=2}^{L}\sum_{l=i}^{L}\frac{L-1}{l-1}\left\|\mathfrak{W}_l^{i}\right\|_{p^{in}}^2\right)\\
&\quad\times\left((L-1)\sum_{l=2}^{L}\left(\left\|f_{\theta^{(l-1)}(0)}^{l-1}\right\|_{p^{in}}+\left\|f_{\theta^{(l-1)}(t)}^{l-1}-f_{\theta^{(l-1)}(0)}^{l-1}\right\|_{p^{in}}\right)^2\left(\left\|\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}+\left\|\mathfrak{W}_L^{l+1}(t)-\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}\right)^4\right)\\
&\leq(L-1)^2\Psi_{L,2}(L-1)\\
&\quad\times\sqrt{\left(\sum_{l=1}^{L-1}\left(\left\|f_{\theta^{(l)}(0)}^{l}\right\|_{p^{in}}+\left\|f_{\theta^{(l)}(t)}^{l}-f_{\theta^{(l)}(0)}^{l}\right\|_{p^{in}}\right)^4\right)\left(\sum_{l=2}^{L}\left(\left\|\mathfrak{W}_L^{l}(0)\right\|_{p^{in}}+\left\|\mathfrak{W}_L^{l}(t)-\mathfrak{W}_L^{l}(0)\right\|_{p^{in}}\right)^8\right)}\\
&\leq(L-1)^2\Psi_{L,2}(L-1)\sqrt{(L-1)^2C_L^4\Phi_{L,4}\Psi_{L,8}}.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\partial_t\Psi_{L,8}(t)\\
&\leq\frac{72Nd^4}{(L-1)^2C_L^2}\sum_{l=2}^{L}\sum_{l=i}^{L}\left\|\mathfrak{W}_L^{l+1}(t)\right\|_{p^{in}}^2\left\|\mathfrak{W}_{l-1}^{i}(t)\right\|_{p^{in}}\left(\left\|\mathfrak{W}_L^{i}(0)\right\|_{p^{in}}+\left\|\mathfrak{W}_L^{i}(t)-\mathfrak{W}_L^{i}(0)\right\|_{p^{in}}\right)^7\\
&\quad\times\left\|f_{\theta^{(l-1)}(t)}^{l-1}(x)\right\|_{p^{in}}\left\|\delta_t^L\right\|_{p^{in}}\\
&\leq\frac{72Nd^4(L-1)}{C_L}\left(\Psi_{L,2}\right)^{1/2}\left(\Phi_{L,4}\right)^{1/4}\left(\Psi_{L,8}\right)^{5/4}\left\|\delta_t^L\right\|_{p^{in}}\\
&\leq\frac{72Nd^4(L-1)}{C_L}\Psi_{L,2}\Phi_{L,4}\left(\Psi_{L,8}\right)^2\left\|\delta_t^L\right\|_{p^{in}}.
\end{aligned}$$

Then, let consider the derivative of $\Psi_{L,2}$. From Lemma 35,

$$
\begin{aligned}
&\partial_t \Psi_{L,2}(t) \\
&\leq \frac{18d^4 N}{(L-1)^3 C_L^2} \sum_{i=2}^{L} \sum_{k=2}^{i} \sum_{l=k}^{i} \frac{L-1}{i-1} \left\| \mathfrak{W}_L^{l+1} \right\|_{p^{in}} \left\| \mathfrak{W}_i^{l+1} \right\|_{p^{in}} \left\| \mathfrak{W}_{l-1}^{k} \right\|_{p^{in}} \left(\left\| \mathfrak{W}_i^{k}(0) \right\|_{p^{in}} + \left\| \mathfrak{W}_i^{k}(t) - \mathfrak{W}_i^{k}(0) \right\|_{p^{in}} \right) \\
&\quad \times \left\| f^{l-1}_{\theta^{(l-1)}(t)} \right\|_{p^{in}} \left\| \delta_t^L \right\|_{p^{in}} \\
&\leq \frac{9d^4 N}{(L-1)^3 C_L^2} \sum_{i=2}^{L} \sum_{k=2}^{i} \sum_{l=k}^{i} \frac{L-1}{i-1} \left\| \mathfrak{W}_L^{l+1} \right\|_{p^{in}} \left\| \mathfrak{W}_{l-1}^{k} \right\|_{p^{in}} \left(\left\| \mathfrak{W}_i^{l+1} \right\|_{p^{in}}^2 + \left(\left\| \mathfrak{W}_i^{k}(0) \right\|_{p^{in}} + \left\| \mathfrak{W}_i^{k}(t) - \mathfrak{W}_i^{k}(0) \right\|_{p^{in}} \right)^2 \right) \\
&\quad \times \left\| f^{l-1}_{\theta^{(l-1)}(t)} \right\|_{p^{in}} \left\| \delta_t^L \right\|_{p^{in}} \\
&= \frac{9d^4 N}{(L-1)^3 C_L^2} \sum_{i=2}^{L} \sum_{k=2}^{i} \sum_{l=k}^{i} \frac{L-1}{i-1} \left\| \mathfrak{W}_L^{l+1} \right\|_{p^{in}} \left\| \mathfrak{W}_{l-1}^{k} \right\|_{p^{in}} \left\| \mathfrak{W}_i^{l+1} \right\|_{p^{in}}^2 \left\| f^{l-1}_{\theta^{(l-1)}(t)} \right\|_{p^{in}} \left\| \delta_t^L \right\|_{p^{in}} \\
&\quad + \frac{9d^4 N}{(L-1)^3 C_L^2} \sum_{i=2}^{L} \sum_{k=2}^{i} \sum_{l=k}^{i} \frac{L-1}{i-1} \left\| \mathfrak{W}_L^{l+1} \right\|_{p^{in}} \left\| \mathfrak{W}_{l-1}^{k} \right\|_{p^{in}} \left(\left\| \mathfrak{W}_i^{k}(0) \right\|_{p^{in}} + \left\| \mathfrak{W}_i^{k}(t) - \mathfrak{W}_i^{k}(0) \right\|_{p^{in}} \right)^2 \\
&\quad \times \left\| f^{l-1}_{\theta^{(l-1)}(t)} \right\|_{p^{in}} \left\| \delta_t^L \right\|_{p^{in}} .
\end{aligned}
$$

From Cauchy-Schwartz inequality,

$$
\begin{aligned}
&\left(\sum_{k=2}^{L} \sum_{l=k}^{L} \left\| \mathfrak{W}_{l-1}^{k} \right\|_{p^{in}} \left(\left\| f^{l-1}_{\theta^{(l-1)}(0)} \right\|_{p^{in}} + \left\| f^{l-1}_{\theta^{(l-1)}(t)} - f^{l-1}_{\theta^{(l-1)}(0)} \right\|_{p^{in}} \right) \left(\left\| \mathfrak{W}_L^{l+1}(0) \right\|_{p^{in}} + \left\| \mathfrak{W}_L^{l+1}(t) - \mathfrak{W}_L^{l+1}(0) \right\|_{p^{in}} \right) \right)^2 \\
&\leq \left(\sum_{k=2}^{L} \sum_{l=k}^{L} \frac{L-1}{l-1} \left\| \mathfrak{W}_l^{k} \right\|_{p^{in}}^2 \right) \\
&\quad \times \left((L-1) \sum_{l=2}^{L} \left(\left\| f^{l-1}_{\theta^{(l-1)}(0)} \right\|_{p^{in}} + \left\| f^{l-1}_{\theta^{(l-1)}(t)} - f^{l-1}_{\theta^{(l-1)}(0)} \right\|_{p^{in}} \right)^2 \left(\left\| \mathfrak{W}_L^{l+1}(0) \right\|_{p^{in}} + \left\| \mathfrak{W}_L^{l+1}(t) - \mathfrak{W}_L^{l+1}(0) \right\|_{p^{in}} \right)^2 \right) \\
&\leq (L-1)^2 \Psi_{L,2}(L-1) \\
&\quad \times \sqrt{ \left(\sum_{l=1}^{L-1} \left(\left\| f^{l}_{\theta^{(l)}(0)} \right\|_{p^{in}} + \left\| f^{l}_{\theta^{(l)}(t)} - f^{l}_{\theta^{(l)}(0)} \right\|_{p^{in}} \right)^4 \right) \left(\sum_{l=2}^{L} \left(\left\| \mathfrak{W}_L^{l}(0) \right\|_{p^{in}} + \left\| \mathfrak{W}_L^{l}(t) - \mathfrak{W}_L^{l}(0) \right\|_{p^{in}} \right)^4 \right) } \\
&\leq (L-1)^2 \Psi_{L,2}(L-1) \\
&\quad \times \sqrt{ \left(\sum_{l=1}^{L-1} \left(\left\| f^{l}_{\theta^{(l)}(0)} \right\|_{p^{in}} + \left\| f^{l}_{\theta^{(l)}(t)} - f^{l}_{\theta^{(l)}(0)} \right\|_{p^{in}} \right)^4 \right) \left(\sum_{l=2}^{L} \left(\left\| \mathfrak{W}_L^{l}(0) \right\|_{p^{in}} + \left\| \mathfrak{W}_L^{l}(t) - \mathfrak{W}_L^{l}(0) \right\|_{p^{in}} \right)^8 \right) } \\
&\leq (L-1)^2 \Psi_{L,2}(L-1) \sqrt{(L-1)^2 C_L^4 \Phi_{L,4} \Psi_{L,8}} \\
&= (L-1)^4 C_L^2 \Psi_{L,2} \left(\Phi_{L,4} \right)^{(1/2)} \left(\Psi_{L,8} \right)^{(1/2)} .
\end{aligned}
$$

Using the above inequality, we have the following upper bounds:

$$\begin{aligned}
&\sum_{i=2}^{L}\sum_{k=2}^{i}\sum_{l=k}^{i}\frac{L-1}{i-1}\left\|\mathfrak{W}_L^{l+1}\right\|_{p^{in}}\left\|\mathfrak{W}_{l-1}^{k}\right\|_{p^{in}}\left\|\mathfrak{W}_i^{l+1}\right\|_{p^{in}}^2\left\|f_{\theta^{(l-1)}(t)}^{l-1}\right\|_{p^{in}}\\
&\leq\left(\sum_{k=2}^{L}\sum_{l=k}^{L}\left\|\mathfrak{W}_{l-1}^{k}\right\|_{p^{in}}\left(\left\|f_{\theta^{(l-1)}(0)}^{l-1}\right\|_{p^{in}}+\left\|f_{\theta^{(l-1)}(t)}^{l-1}-f_{\theta^{(l-1)}(0)}^{l-1}\right\|_{p^{in}}\right)\right.\\
&\quad\left.\times\left(\left\|\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}+\left\|\mathfrak{W}_L^{l+1}(t)-\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}\right)\right)\left(\sum_{i=2}^{L}\sum_{l=2}^{i}\frac{L-1}{i-1}\left\|\mathfrak{W}_i^{l+1}\right\|_{p^{in}}^2\right)\\
&\leq(L-1)^2\Psi_{L,2}(L-1)^2C_L\left(\Psi_{L,2}\right)^{(1/2)}\left(\Phi_{L,4}\right)^{(1/4)}\left(\Psi_{L,8}\right)^{(1/4)}\\
&=(L-1)^4C_L\left(\Psi_{L,2}\right)^{(3/2)}\left(\Phi_{L,4}\right)^{(1/4)}\left(\Psi_{L,8}\right)^{(1/4)}.
\end{aligned}$$

and

$$\begin{aligned}
&\sum_{i=2}^{L}\sum_{k=2}^{i}\sum_{l=k}^{i}\frac{L-1}{i-1}\left\|\mathfrak{W}_L^{l+1}\right\|_{p^{in}}\left\|\mathfrak{W}_{l-1}^{k}\right\|_{p^{in}}\left(\left\|\mathfrak{W}_i^{k}(0)\right\|_{p^{in}}+\left\|\mathfrak{W}_i^{k}(t)-\mathfrak{W}_i^{k}(0)\right\|_{p^{in}}\right)^2\left\|f_{\theta^{(l-1)}(t)}^{l-1}\right\|_{p^{in}}\\
&\leq\left(\sum_{k=2}^{L}\sum_{l=k}^{L}\left\|\mathfrak{W}_{l-1}^{k}\right\|_{p^{in}}\left(\left\|f_{\theta^{(l-1)}(0)}^{l-1}\right\|_{p^{in}}+\left\|f_{\theta^{(l-1)}(t)}^{l-1}-f_{\theta^{(l-1)}(0)}^{l-1}\right\|_{p^{in}}\right)\right.\\
&\quad\left.\times\left(\left\|\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}+\left\|\mathfrak{W}_L^{l+1}(t)-\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}\right)\right)\left(\sum_{i=2}^{L}\sum_{k=2}^{i}\frac{L-1}{i-1}\left(\left\|\mathfrak{W}_i^{k}(0)\right\|_{p^{in}}+\left\|\mathfrak{W}_i^{k}(t)-\mathfrak{W}_i^{k}(0)\right\|_{p^{in}}\right)^2\right)\\
&\leq(L-1)^2\Psi_{L,2}(L-1)^2C_L\left(\Psi_{L,2}\right)^{(1/2)}\left(\Phi_{L,4}\right)^{(1/4)}\left(\Psi_{L,8}\right)^{(1/4)}\\
&=(L-1)^4C_L\left(\Psi_{L,2}\right)^{(3/2)}\left(\Phi_{L,4}\right)^{(1/4)}\left(\Psi_{L,8}\right)^{(1/4)}.
\end{aligned}$$

Thus, we have

$$\begin{aligned}
\partial_t\Psi_L^{(2)}(t)&\leq\frac{9d^4N}{(L-1)^3C_L^2}\sum_{i=2}^{L}\sum_{k=2}^{i}\sum_{l=k}^{i}\frac{L-1}{i-1}\left\|\mathfrak{W}_L^{l+1}\right\|_{p^{in}}\left\|\mathfrak{W}_{l-1}^{k}\right\|_{p^{in}}\\
&\quad\times\left(\left\|\mathfrak{W}_i^{l+1}\right\|_{p^{in}}^2+\left(\left\|\mathfrak{W}_i^{k}(0)\right\|_{p^{in}}+\left\|\mathfrak{W}_i^{k}(t)-\mathfrak{W}_i^{k}(0)\right\|_{p^{in}}\right)^2\right)\left\|f_{\theta^{(l-1)}(t)}^{l-1}\right\|_{p^{in}}\left\|\delta_t^L\right\|_{p^{in}}\\
&\leq\frac{18d^4N(L-1)}{C_L}\left(\Psi_{L,2}\right)^{(3/2)}\left(\Phi_{L,4}\right)^{(1/4)}\left(\Psi_{L,8}\right)^{(1/4)}\left\|\delta_t^L\right\|_{p^{in}}\\
&\leq\frac{18d^4N(L-1)}{C_L}\left(\Psi_{L,2}\right)^2\Phi_{L,4}\Psi_{L,8}\left\|\delta_t^L\right\|_{p^{in}}.
\end{aligned}$$

Finally, let consider the derivatives of $\Phi_{L,j}$ for $1\leq j\leq 8$. From Lemma 35,

$$\begin{aligned}
\partial_t\Phi_{L,j}(t)\leq&\frac{81\max_l\{n_l^4\}jN^{3/2}}{(L-1)^2C_L^{2+j}}\sum_{i=1}^{L-1}\sum_{l=0}^{i-1}\left(\left\|f_{\theta^{(i)}(0)}^{i}\right\|_{p^{in}}+\left\|f_{\theta^{(i)}(t)}^{i}-f_{\theta^{(i)}(0)}^{i}\right\|_{p^{in}}\right)^{j-1}\left\|f_{\theta^{(l)}(t)}^{l}(x)\right\|_{p^{in}}\left\|f_{\theta^{(l)}(t)}^{l}(x')\right\|_{p^{in}}\\
&\times\left\|\mathfrak{W}_i^{l+2}\right\|_{p^{in}}\left\|(\mathfrak{W}_L^{l+2})^\intercal\right\|_{p^{in}}\left\|\delta_t^L\right\|_{p^{in}}\\
&+\frac{d^4jN}{(L-1)^2C_L^{2+j}}\sum_{i=1}^{L-1}\sum_{l=0}^{i-1}\left(\left\|f_{\theta^{(i)}(0)}^{i}\right\|_{p^{in}}+\left\|f_{\theta^{(i)}(t)}^{i}-f_{\theta^{(i)}(0)}^{i}\right\|_{p^{in}}\right)^{j-1}\left\|\mathfrak{W}_i^{l+2}\right\|_{p^{in}}\left\|(\mathfrak{W}_L^{l+2})^\intercal\right\|_{p^{in}}\left\|\delta_t^L\right\|_{p^{in}}.
\end{aligned}$$

Similar to the previous inequalities, we have

$$
\begin{aligned}
&\left(\sum_{i=1}^{L-1}\sum_{l=0}^{i-1}\|\mathfrak{W}_i^{l+2}\|_{p^{in}}\left(\left\|f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}+\left\|f^l_{\theta^{(l)}(t)}-f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}\right)^2\left(\|\mathfrak{W}_L^{l+2}(0)\|_{p^{in}}+\|\mathfrak{W}_L^{l+2}(t)-\mathfrak{W}_L^{l+2}(0)\|_{p^{in}}\right)\right)^2\\
&\leq\left(\sum_{i=1}^{L-1}\sum_{l=0}^{i-1}\frac{L-1}{i-1}\|\mathfrak{W}_i^{l+2}\|^2_{p^{in}}\right)\\
&\quad\times\left((L-1)\sum_{l=0}^{L-2}\left(\left\|f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}+\left\|f^l_{\theta^{(l)}(t)}-f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}\right)^4\left(\|\mathfrak{W}_L^{l+2}(0)\|_{p^{in}}+\|\mathfrak{W}_L^{l+2}(t)-\mathfrak{W}_L^{l+2}(0)\|_{p^{in}}\right)^2\right)\\
&\leq(L-1)^2\Psi_{L,2}(t)(L-1)\\
&\quad\times\sqrt{\left(\sum_{l=0}^{L-2}\left(\left\|f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}+\left\|f^l_{\theta^{(l)}(t)}-f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}\right)^8\right)\left(\sum_{l=2}^{L}\left(\|\mathfrak{W}_L^{l}(0)\|_{p^{in}}+\|\mathfrak{W}_L^{l}(t)-\mathfrak{W}_L^{l}(0)\|_{p^{in}}\right)^4\right)}\\
&\leq(L-1)^2\Psi_{L,2}(t)(L-1)\\
&\quad\times\sqrt{\left(\sum_{l=0}^{L-2}\left(\left\|f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}+\left\|f^l_{\theta^{(l)}(t)}-f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}\right)^8\right)\left(\sum_{l=2}^{L}\left(\|\mathfrak{W}_L^{l}(0)\|_{p^{in}}+\|\mathfrak{W}_L^{l}(t)-\mathfrak{W}_L^{l}(0)\|_{p^{in}}\right)^8\right)}\\
&\leq(L-1)^2\Psi_{L,2}(t)(L-1)\times(L-1)C_L^4\left(\Phi_{L,8}(t)\right)^{(1/2)}\left(\Psi_{L,8}(t)\right)^{(1/2)}\\
&=(L-1)^4C_L^4\Psi_{L,2}(t)\left(\Phi_{L,8}(t)\right)^{(1/2)}\left(\Psi_{L,8}(t)\right)^{(1/2)}.
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
&\sum_{i=1}^{L-1}\sum_{l=0}^{i-1}\left(\left\|f^i_{\theta^{(i)}(0)}\right\|_{p^{in}}+\left\|f^i_{\theta^{(i)}(t)}-f^i_{\theta^{(i)}(0)}\right\|_{p^{in}}\right)^{j-1}\left\|f^l_{\theta^{(l)}(t)}(x)\right\|_{p^{in}}\left\|f^l_{\theta^{(l)}(t)}(x')\right\|_{p^{in}}\|\mathfrak{W}_i^{l+2}\|_{p^{in}}\|(\mathfrak{W}_L^{l+2})^\intercal\|_{p^{in}}\\
&\leq\sum_{i=1}^{L-1}\sum_{l=0}^{i-1}\|\mathfrak{W}_i^{l+2}\|_{p^{in}}\left(\left\|f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}+\left\|f^l_{\theta^{(l)}(t)}-f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}\right)^2\left(\|\mathfrak{W}_L^{l+2}(0)\|_{p^{in}}+\|\mathfrak{W}_L^{l+2}(t)-\mathfrak{W}_L^{l+2}(0)\|_{p^{in}}\right)\\
&\quad\times\sum_{i=1}^{L-1}\left(\left\|f^i_{\theta^{(i)}(0)}\right\|_{p^{in}}+\left\|f^i_{\theta^{(i)}(t)}-f^i_{\theta^{(i)}(0)}\right\|_{p^{in}}\right)^{j-1}\\
&\leq(L-1)C_L^{j-1}\Phi_{L,j-1}(t)(L-1)^2C_L^2\left(\Psi_{L,2}(t)\right)^{1/2}\left(\Phi_{L,8}(t)\right)^{(1/4)}\left(\Psi_{L,8}(t)\right)^{(1/4)}\\
&=(L-1)^3C_L^{j+1}\Phi_{L,j-1}(t)\left(\Psi_{L,2}(t)\right)^{1/2}\left(\Phi_{L,8}(t)\right)^{(1/4)}\left(\Psi_{L,8}(t)\right)^{(1/4)}
\end{aligned}
$$

and

$$
\begin{aligned}
&\frac{d^4jN}{(L-1)^2C_L^{2+j}}\sum_{i=1}^{L-1}\sum_{l=0}^{i-1}\left(\left\|f^i_{\theta^{(i)}(0)}\right\|_{p^{in}}+\left\|f^i_{\theta^{(i)}(t)}-f^i_{\theta^{(i)}(0)}\right\|_{p^{in}}\right)^{j-1}\|\mathfrak{W}_i^{l+2}\|_{p^{in}}\|(\mathfrak{W}_L^{l+2})^\intercal\|_{p^{in}}\\
&\leq\frac{81\max_l\{n_l^4\}d^4jN^{3/2}}{(L-1)^2C_L^{2+j}}\sum_{i=1}^{L-1}\\
&\quad\times\sum_{l=0}^{i-1}\left(\left\|f^i_{\theta^{(i)}(0)}\right\|_{p^{in}}+\left\|f^i_{\theta^{(i)}(t)}-f^i_{\theta^{(i)}(0)}\right\|_{p^{in}}\right)^{j-1}\left\|f^l_{\theta^{(l)}(t)}(x)\right\|_{p^{in}}\left\|f^l_{\theta^{(l)}(t)}(x')\right\|_{p^{in}}\|\mathfrak{W}_i^{l+2}\|_{p^{in}}\|(\mathfrak{W}_L^{l+2})^\intercal\|_{p^{in}}.
\end{aligned}
$$

By combining the above inequalities, we have

$$\begin{aligned}
\partial_t \Phi_{L,j}(t) \leq & \frac{81 \max_l \{n_l^4\} j N^{3/2}}{(L-1)^2 C_L^{2+j}} \sum_{i=1}^{L-1} \sum_{l=0}^{i-1} \left(\left\| f^i_{\theta^{(i)}(0)} \right\|_{p^{in}} + \left\| f^i_{\theta^{(i)}(t)} - f^i_{\theta^{(i)}(0)} \right\|_{p^{in}} \right)^{j-1} \left\| f^l_{\theta^{(l)}(t)}(x) \right\|_{p^{in}} \left\| f^l_{\theta^{(l)}(t)}(x') \right\|_{p^{in}} \\
& \times \left\| \mathfrak{W}_i^{l+2} \right\|_{p^{in}} \left\| (\mathfrak{W}_L^{l+2})^\intercal \right\|_{p^{in}} \left\| \delta_t^L \right\|_{p^{in}} \\
& + \frac{d^4 j N}{(L-1)^2 C_L^{2+j}} \sum_{i=1}^{L-1} \sum_{l=0}^{i-1} \left(\left\| f^i_{\theta^{(i)}(0)} \right\|_{p^{in}} + \left\| f^i_{\theta^{(i)}(t)} - f^i_{\theta^{(i)}(0)} \right\|_{p^{in}} \right)^{j-1} \left\| \mathfrak{W}_i^{l+2} \right\|_{p^{in}} \left\| (\mathfrak{W}_L^{l+2})^\intercal \right\|_{p^{in}} \left\| \delta_t^L \right\|_{p^{in}} \\
\leq & \frac{162 \max_l \{n_l^4\} d^4 j N^{3/2} (L-1)}{C_L} \Phi_{L,j-1}(t) (\Psi_{L,2}(t))^{1/2} (\Phi_{L,8}(t))^{(1/4)} (\Psi_{L,8}(t))^{(1/4)} \left\| \delta_t^L \right\|_{p^{in}} \\
\leq & \frac{162 \max_l \{n_l^4\} d^4 j N^{3/2} (L-1)}{C_L} \Phi_{L,j-1}(t) \Psi_{L,2}(t) \Phi_{L,8}(t) \Psi_{L,8}(t) \left\| \delta_t^L \right\|_{p^{in}}.
\end{aligned}$$

□

M. Proof of Lemma 26 for Proposition 5

Lemma 26. *For $t \geq 0$ and $x \in \mathbb{R}_+^{d \times d}$, if $f^l_{\theta^{(l)}(t)}(x)$ is element-wise nonzero for all $1 \leq l \leq L-1$, then*

$$\begin{aligned}
\partial_t \tilde{\Psi}_{L,2}(x,t) &\leq \frac{j 162 d^4 \max_l \{n_l^4\} N^{1/2} (L-1)}{C_L} \left(\tilde{\Psi}_{L,2}(x,t) \right)^2 \Phi_{L,4}(t) \Psi_{L,8}(t) \left\| \delta_t^L \right\|_{p^{in}} \\
\partial_t \tilde{\Phi}_{L,j}(x,t) &\leq \frac{18 d^4 (L-1)}{C_L} \tilde{\Phi}_{L,j-1}(t) \tilde{\Phi}_{L,8}(x,t) \tilde{\Psi}_{L,2}(x,t) \Phi_{L,8}(t) \Psi_{L,8}(t) \left\| \delta_t^L \right\|_{p^{in}} \\
\partial_t \tilde{\Psi}_{L,8}(x,t) &\leq \frac{72 d^4 (L-1)}{C_L} \tilde{\Phi}_{L,4}(x,t) \tilde{\Psi}_{L,2}(x,t) \left(\tilde{\Psi}_{L,8}(x,t) \right)^2 \Psi_{L,8}(t) \left\| \delta_t^L \right\|_{p^{in}}.
\end{aligned}$$

for $1 \leq j \leq 8$.

Proof of Lemma 26. Consider the derivative of $\tilde{\Psi}_{L,8}$ first. Like Section C, we show that the intermediate pre-activation values never reach zero while training, and therefore we can apply Lemma 36.

$$\begin{aligned}
& \partial_t \tilde{\Psi}_{L,8}(t) \\
& \leq \frac{72 d^4}{(L-1)^2 C_L^2} \sum_{l=2}^{L} \sum_{i=l}^{L} \left\| \mathfrak{W}_L^{l+1}(t) \right\|_{p^{in}} \left\| \mathfrak{W}_L^{l+1}(t) \right\| \left\| \mathfrak{W}_{l-1}^{i}(t) \right\| \left(\left\| \mathfrak{W}_L^i(0) \right\| + \left\| \mathfrak{W}_L^i(t) - \mathfrak{W}_L^i(0) \right\| \right)^7 \\
& \quad \times \left\| f^{l-1}_{\theta^{(l-1)}(t)}(x) \right\|_{p^{in}} \left\| \delta_t^L \right\|_{p^{in}}.
\end{aligned}$$

Then,

$$
\begin{aligned}
&\sum_{l=2}^{L}\sum_{i=l}^{L}\left\|\mathfrak{W}_L^{l+1}(t)\right\|_{p^{in}}\left\|\mathfrak{W}_L^{l+1}(t)\right\|\left\|\mathfrak{W}_{l-1}^{i}(t)\right\|\left(\left\|\mathfrak{W}_L^{i}(0)\right\|+\left\|\mathfrak{W}_L^{i}(t)-\mathfrak{W}_L^{i}(0)\right\|\right)^7\left\|f^{l-1}_{\theta^{(l-1)}(t)}(x)\right\|_{p^{in}}\\
&\leq\left(\sum_{i=2}^{L}\sum_{i=l}^{L}\left\|\mathfrak{W}_{l-1}^{i}\right\|\left(\left\|f^{l-1}_{\theta^{(l-1)}(0)}\right\|_{p^{in}}+\left\|f^{l-1}_{\theta^{(l-1)}(t)}-f^{l-1}_{\theta^{(l-1)}(0)}\right\|_{p^{in}}\right)\left(\left\|\mathfrak{W}_L^{l+1}(0)\right\|+\left\|\mathfrak{W}_L^{l+1}(t)-\mathfrak{W}_L^{l+1}(0)\right\|\right)\right.\\
&\quad\left.\times\left(\left\|\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}+\left\|\mathfrak{W}_L^{l+1}(t)-\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}\right)\right)\left(\sum_{i=2}^{L}\left(\left\|\mathfrak{W}_L^{i}(0)\right\|_{p^{in}}+\left\|\mathfrak{W}_L^{i}(t)-\mathfrak{W}_L^{i}(0)\right\|_{p^{in}}\right)^8\right)\\
&\leq\left(\sum_{l=2}^{L}\sum_{i=2}^{l}\frac{L-1}{l-1}\left\|\mathfrak{W}_l^{i}\right\|^2\right)^{1/2}\left((L-1)\sum_{l=2}^{L}\left(\left\|f^{l-1}_{\theta^{(l-1)}(0)}\right\|_{p^{in}}+\left\|f^{l-1}_{\theta^{(l-1)}(t)}-f^{l-1}_{\theta^{(l-1)}(0)}\right\|_{p^{in}}\right)^2\right.\\
&\quad\left.\times\left(\left\|\mathfrak{W}_L^{l+1}(0)\right\|+\left\|\mathfrak{W}_L^{l+1}(t)-\mathfrak{W}_L^{l+1}(0)\right\|\right)^2\left(\left\|\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}+\left\|\mathfrak{W}_L^{l+1}(t)-\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}\right)^2\right)^{1/2}(L-1)\tilde{\Psi}_{L,8}\\
&\leq\left((L-1)^2\tilde{\Psi}_{L,2}\right)^{1/2}\sqrt{L-1}\left(\sum_{l=2}^{L}\left(\left\|f^{l-1}_{\theta^{(l-1)}(0)}\right\|_{p^{in}}+\left\|f^{l-1}_{\theta^{(l-1)}(t)}-f^{l-1}_{\theta^{(l-1)}(0)}\right\|_{p^{in}}\right)^4\right)^{1/4}\\
&\quad\times\left(\sum_{l=2}^{L}\left(\left\|\mathfrak{W}_L^{l+1}(0)\right\|+\left\|\mathfrak{W}_L^{l+1}(t)-\mathfrak{W}_L^{l+1}(0)\right\|\right)^4\left(\left\|\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}+\left\|\mathfrak{W}_L^{l+1}(t)-\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}\right)^4\right)^{1/4}(L-1)\tilde{\Psi}_{L,8}\\
&\leq\left((L-1)^2\tilde{\Psi}_{L,2}\right)^{1/2}\sqrt{L-1}(LC_L^4)^{1/4}(\Phi_{L,4})^{1/4}(L-1)\tilde{\Psi}_{L,8}\\
&\quad\times\left(\sum_{l=2}^{L}\left(\left\|\mathfrak{W}_L^{l+1}(0)\right\|+\left\|\mathfrak{W}_L^{l+1}(t)-\mathfrak{W}_L^{l+1}(0)\right\|\right)^8\right)^{1/8}\left(\sum_{l=2}^{L}\left(\left\|\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}+\left\|\mathfrak{W}_L^{l+1}(t)-\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}\right)^8\right)^{1/8}\\
&=(L-1)\tilde{\Psi}_{L,8}(L-1)^2C_L\left(\tilde{\Psi}_{L,2}\right)^{1/2}\left(\tilde{\Phi}_{L,4}\right)^{(1/4)}\left(\Psi_{L,8}\right)^{(1/8)}\left(\tilde{\Psi}_{L,8}\right)^{(1/8)}
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
&\partial_t\tilde{\Psi}_{L,8}(t)\\
&\leq\frac{72d^4}{(L-1)^2C_L^2}\sum_{l=2}^{L}\sum_{i=l}^{L}\left\|\mathfrak{W}_L^{l+1}(t)\right\|_{p^{in}}\left\|\mathfrak{W}_L^{l+1}(t)\right\|\left\|\mathfrak{W}_{l-1}^{i}(t)\right\|\left(\left\|\mathfrak{W}_L^{i}(0)\right\|+\left\|\mathfrak{W}_L^{i}(t)-\mathfrak{W}_L^{i}(0)\right\|\right)^7\left\|f^{l-1}_{\theta^{(l-1)}(t)}(x)\right\|_{p^{in}}\left\|\delta_t^L\right\|_{p^{in}}\\
&\leq\frac{72d^4(L-1)}{C_L}\left(\tilde{\Psi}_{L,2}\right)^{1/2}\left(\tilde{\Phi}_{L,4}\right)^{1/4}\left(\Psi_{L,8}\right)^{1/8}\left(\tilde{\Psi}_{L,8}\right)^{9/8}\left\|\delta_t^L\right\|_{p^{in}}\\
&\leq\frac{72d^4L}{C_L}\tilde{\Psi}_{L,2}\tilde{\Phi}_{L,4}\Psi_{L,8}\left(\tilde{\Psi}_{L,8}\right)^2\left\|\delta_t^L\right\|_{p^{in}}.
\end{aligned}
$$

Now, let consider the derivative of $\tilde{\Psi}_{L,2}$. From Lemma 36,

$$\begin{aligned}
&\partial_t \tilde{\Psi}_{L,2}(t) \\
&\le \frac{18d^4}{(L-1)^3 C_L^2} \sum_{i=2}^{L}\sum_{k=2}^{i}\sum_{l=k}^{i} \frac{L-1}{i-1} \left\|\mathfrak{W}_L^{l+1}\right\|_{p^{in}} \left\|\mathfrak{W}_i^{l+1}\right\| \left\|\mathfrak{W}_{l-1}^{k}\right\| \left(\left\|\mathfrak{W}_i^k(0)\right\| + \left\|\mathfrak{W}_i^k(t) - \mathfrak{W}_i^k(0)\right\|\right) \\
&\quad \times \left\|f^{l-1}_{\theta^{(l-1)}(t)}\right\|_{p^{in}} \left\|\delta_t^L\right\|_{p^{in}} \\
&\le \frac{9d^4}{(L-1)^3 C_L^2} \sum_{i=2}^{L}\sum_{k=2}^{i}\sum_{l=k}^{i} \frac{L-1}{i-1} \left\|\mathfrak{W}_L^{l+1}\right\|_{p^{in}} \left\|\mathfrak{W}_{l-1}^{k}\right\| \left(\left\|\mathfrak{W}_i^{l+1}\right\|^2 + \left(\left\|\mathfrak{W}_i^k(0)\right\| + \left\|\mathfrak{W}_i^k(t) - \mathfrak{W}_i^k(0)\right\|\right)^2\right) \\
&\quad \times \left\|f^{l-1}_{\theta^{(l-1)}(t)}\right\|_{p^{in}} \left\|\delta_t^L\right\|_{p^{in}} \\
&= \frac{9d^4}{(L-1)^3 C_L^2} \sum_{i=2}^{L}\sum_{k=2}^{i}\sum_{l=k}^{i} \frac{L-1}{i-1} \left\|\mathfrak{W}_L^{l+1}\right\|_{p^{in}} \left\|\mathfrak{W}_{l-1}^{k}\right\| \left\|\mathfrak{W}_i^{l+1}\right\|^2 \left\|f^{l-1}_{\theta^{(l-1)}(t)}\right\|_{p^{in}} \left\|\delta_t^L\right\|_{p^{in}} \\
&\quad + \frac{9d^4}{(L-1)^3 C_L^2} \sum_{i=2}^{L}\sum_{k=2}^{i}\sum_{l=k}^{i} \frac{L-1}{i-1} \left\|\mathfrak{W}_L^{l+1}\right\|_{p^{in}} \left\|\mathfrak{W}_{l-1}^{k}\right\| \left(\left\|\mathfrak{W}_i^k(0)\right\| + \left\|\mathfrak{W}_i^k(t) - \mathfrak{W}_i^k(0)\right\|\right)^2 \left\|f^{l-1}_{\theta^{(l-1)}(t)}\right\|_{p^{in}} \left\|\delta_t^L\right\|_{p^{in}}.
\end{aligned}$$

By Cauchy-Schwartz inequality,

$$\begin{aligned}
&\left(\sum_{k=2}^{L}\sum_{l=k}^{L} \left\|\mathfrak{W}_{l-1}^{k}\right\| \left(\left\|f^{l-1}_{\theta^{(l-1)}(0)}\right\|_{p^{in}} + \left\|f^{l-1}_{\theta^{(l-1)}(t)} - f^{l-1}_{\theta^{(l-1)}(0)}\right\|_{p^{in}}\right) \left(\left\|\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}} + \left\|\mathfrak{W}_L^{l+1}(t) - \mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}\right)\right)^2 \\
&\le \left(\sum_{k=2}^{L}\sum_{l=k}^{L} \frac{L-1}{l-1} \left\|\mathfrak{W}_l^k\right\|^2\right) \\
&\quad \times \left((L-1)\sum_{l=2}^{L} \left(\left\|f^{l-1}_{\theta^{(l-1)}(0)}\right\|_{p^{in}} + \left\|f^{l-1}_{\theta^{(l-1)}(t)} - f^{l-1}_{\theta^{(l-1)}(0)}\right\|_{p^{in}}\right)^2 \left(\left\|\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}} + \left\|\mathfrak{W}_L^{l+1}(t) - \mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}\right)^2\right) \\
&\le (L-1)^2 \tilde{\Psi}_{L,2} (L-1) \sqrt{(L-1) C_L^4 \Phi_{L,4} L \Psi_{L,8}} \\
&\le (L-1)^4 C_L^2 \tilde{\Psi}_{L,2} \left(\Phi_{L,4}\right)^{(1/2)} \left(\Psi_{L,8}\right)^{(1/2)}.
\end{aligned}$$

Then, we have

$$\begin{aligned}
&\sum_{i=2}^{L}\sum_{k=2}^{i}\sum_{l=k}^{i} \frac{L-1}{i-1} \left\|\mathfrak{W}_L^{l+1}\right\|_{p^{in}} \left\|\mathfrak{W}_{l-1}^{k}\right\| \left\|\mathfrak{W}_i^{l+1}\right\|^2 \left\|f^{l-1}_{\theta^{(l-1)}(t)}\right\|_{p^{in}} \\
&\le \left(\sum_{k=2}^{L}\sum_{l=k}^{L} \left\|\mathfrak{W}_{l-1}^{k}\right\| \left(\left\|f^{l-1}_{\theta^{(l-1)}(0)}\right\|_{p^{in}} + \left\|f^{l-1}_{\theta^{(l-1)}(t)} - f^{l-1}_{\theta^{(l-1)}(0)}\right\|_{p^{in}}\right) \left(\left\|\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}} + \left\|\mathfrak{W}_L^{l+1}(t) - \mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}\right)\right) \\
&\quad \times \left(\sum_{i=2}^{L}\sum_{l=2}^{i} \frac{L-1}{i-1} \left\|\mathfrak{W}_i^{l+1}\right\|^2\right) \\
&\le (L-1)^2 \tilde{\Psi}_{L,2} (L-1)^2 C_L \left(\tilde{\Psi}_{L,2}\right)^{(1/2)} \left(\Phi_{L,2}\right)^{(1/4)} \left(\Psi_{L,8}\right)^{(1/4)} \\
&= (L-1)^4 C_L \left(\tilde{\Psi}_{L,2}\right)^{(3/2)} \left(\Phi_{L,4}\right)^{(1/2)} \left(\Psi_{L,8}\right)^{(1/4)}
\end{aligned}$$

and

$$\begin{aligned}
&\sum_{i=2}^{L}\sum_{k=2}^{i}\sum_{l=k}^{i}\frac{L-1}{i-1}\left\|\mathfrak{W}_L^{l+1}\right\|_{p^{in}}\left\|\mathfrak{W}_{l-1}^{k}\right\|\left(\left\|\mathfrak{W}_i^k(0)\right\|+\left\|\mathfrak{W}_i^k(t)-\mathfrak{W}_i^k(0)\right\|\right)^2\left\|f^{l-1}_{\theta^{(l-1)}(t)}\right\|_{p^{in}}\\
&\le\left(\sum_{k=2}^{L}\sum_{l=k}^{L}\left\|\mathfrak{W}_{l-1}^{k}\right\|\left(\left\|f^{l-1}_{\theta^{(l-1)}(0)}\right\|_{p^{in}}+\left\|f^{l-1}_{\theta^{(l-1)}(t)}-f^{l-1}_{\theta^{(l-1)}(0)}\right\|_{p^{in}}\right)\left(\left\|\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}+\left\|\mathfrak{W}_L^{l+1}(t)-\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}\right)\right)\\
&\quad\times\left(\sum_{i=2}^{L}\sum_{k=2}^{i}\frac{L-1}{i-1}\left(\left\|\mathfrak{W}_i^k(0)\right\|+\left\|\mathfrak{W}_i^k(t)-\mathfrak{W}_i^k(0)\right\|\right)^2\right)\\
&\le(L-1)^2\tilde{\Psi}_{L,2}(L-1)^2C_L\left(\tilde{\Psi}_{L,2}\right)^{(1/2)}\left(\Phi_{L,4}\right)^{(1/4)}\left(\Psi_{L,8}\right)^{(1/4)}\\
&=(L-1)^4C_L\left(\tilde{\Psi}_{L,2}\right)^{(3/2)}\left(\Phi_{L,4}\right)^{(1/4)}\left(\Psi_{L,8}\right)^{(1/4)}.
\end{aligned}$$

Thus, by combining the above two inequalities,

$$\begin{aligned}
&\partial_t\tilde{\Psi}_{L,2}(t)\\
&\le\frac{9d^4}{(L-1)^3C_L^2}\sum_{i=2}^{L}\sum_{k=2}^{i}\sum_{l=k}^{i}\frac{L-1}{i-1}\left\|\mathfrak{W}_L^{l+1}\right\|_{p^{in}}\left\|\mathfrak{W}_{l-1}^{k}\right\|\left(\left\|\mathfrak{W}_i^{l+1}\right\|^2+\left(\left\|\mathfrak{W}_i^k(0)\right\|+\left\|\mathfrak{W}_i^k(t)-\mathfrak{W}_i^k(0)\right\|\right)^2\right)\\
&\quad\times\left\|f^{l-1}_{\theta^{(l-1)}(t)}\right\|_{p^{in}}\left\|\delta_t^L\right\|_{p^{in}}\\
&\le\frac{18d^4(L-1)}{C_L}\left(\tilde{\Psi}_{L,2}\right)^{3/2}\left(\Phi_{L,4}\right)^{1/4}\left(\Psi_{L,8}\right)^{1/4}\left\|\delta_t^L\right\|_{p^{in}}\\
&\le\frac{18d^4(L-1)}{C_L}\left(\tilde{\Psi}_{L,2}\right)^{2}\Phi_{L,4}\Psi_{L,8}\left\|\delta_t^L\right\|_{p^{in}}
\end{aligned}$$

Finally, let consider the derivatives of $\tilde{\Phi}_{L,j}$ for $1\le j\le 8$. From Lemma 36,

$$\begin{aligned}
&\partial_t\tilde{\Phi}_{L,j}(t)\\
&\le\frac{j81\max_l\{n_l^4\}N^{1/2}}{(L-1)^2C_L^{2+j}}\sum_{i=1}^{L-1}\sum_{l=0}^{i-1}\left(\left\|f^i_{\theta^{(i)}(0)}\right\|+\left\|f^i_{\theta^{(i)}(t)}-f^i_{\theta^{(i)}(0)}\right\|\right)^{j-1}\left\|f^l_{\theta^{(l)}(t)}(x)\right\|_{p^{in}}\left\|f^l_{\theta^{(l)}(t)}(x')\right\|\\
&\quad\times\left\|\mathfrak{W}_i^{l+2}\right\|\left\|(\mathfrak{W}_L^{l+2})^{\intercal}\right\|_{p^{in}}\left\|\delta_t^L\right\|_{p^{in}}\\
&\quad+\frac{jd^4}{(L-1)^2C_L^{2+j}}\sum_{i=1}^{L-1}\sum_{l=0}^{i-1}\left(\left\|f^i_{\theta^{(i)}(0)}\right\|+\left\|f^i_{\theta^{(i)}(t)}-f^i_{\theta^{(i)}(0)}\right\|\right)^{j-1}\left\|\mathfrak{W}_i^{l+2}\right\|\left\|(\mathfrak{W}_L^{l+2})^{\intercal}\right\|_{p^{in}}\left\|\delta_t^L\right\|_{p^{in}}\\
&\le\frac{j162d^4\max_l\{n_l^4\}N^{1/2}}{(L-1)^2C_L^{2+j}}\sum_{i=1}^{L-1}\sum_{l=0}^{i-1}\left(\left\|f^i_{\theta^{(i)}(0)}\right\|+\left\|f^i_{\theta^{(i)}(t)}-f^i_{\theta^{(i)}(0)}\right\|\right)^{j-1}\left\|f^l_{\theta^{(l)}(t)}(x)\right\|_{p^{in}}\left\|f^l_{\theta^{(l)}(t)}(x')\right\|\\
&\quad\times\left\|\mathfrak{W}_i^{l+2}\right\|\left\|(\mathfrak{W}_L^{l+2})^{\intercal}\right\|_{p^{in}}\left\|\delta_t^L\right\|_{p^{in}}
\end{aligned}$$

since $\left\|f^l_{\theta^{(l)}(t)}(x)\right\|\ge 1$.

By Cauchy-Schwartz inequality,

$$
\begin{aligned}
&\left(\sum_{l=0}^{L-2}\left(\left\|f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}+\left\|f^l_{\theta^{(l)}(t)}-f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}\right)^2\left(\left\|f^l_{\theta^{(l)}(0)}\right\|+\left\|f^l_{\theta^{(l)}(t)}-f^l_{\theta^{(l)}(0)}\right\|\right)^2\right.\\
&\quad\left.\times\left(\left\|\mathfrak{W}^{l+2}_L(0)\right\|_{p^{in}}+\left\|\mathfrak{W}^{l+2}_L(t)-\mathfrak{W}^{l+2}_L(0)\right\|_{p^{in}}\right)^2\right)^2\\
&\le\left(\sum_{l=0}^{L-2}\left(\left\|f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}+\left\|f^l_{\theta^{(l)}(t)}-f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}\right)^4\left(\left\|f^l_{\theta^{(l)}(0)}\right\|+\left\|f^l_{\theta^{(l)}(t)}-f^l_{\theta^{(l)}(0)}\right\|\right)^4\right)\\
&\le\left(\sum_{l=0}^{L-2}\left(\left\|f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}+\left\|f^l_{\theta^{(l)}(t)}-f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}\right)^4\left(\left\|f^l_{\theta^{(l)}(0)}\right\|+\left\|f^l_{\theta^{(l)}(t)}-f^l_{\theta^{(l)}(0)}\right\|\right)^4\right)\\
&\quad\times\left(\sum_{l=2}^{L}\left(\left\|\mathfrak{W}^{l}_L(0)\right\|_{p^{in}}+\left\|\mathfrak{W}^{l}_L(t)-\mathfrak{W}^{l}_L(0)\right\|_{p^{in}}\right)^4\right)\\
&\le\left(\sum_{l=0}^{L-2}\left(\left\|f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}+\left\|f^l_{\theta^{(l)}(t)}-f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}\right)^8\right)^{1/2}\left(\left(\left\|f^l_{\theta^{(l)}(0)}\right\|+\left\|f^l_{\theta^{(l)}(t)}-f^l_{\theta^{(l)}(0)}\right\|\right)^8\right)^{1/2}\\
&\quad\times\left(\sum_{l=2}^{L}\left(\left\|\mathfrak{W}^{l}_L(0)\right\|_{p^{in}}+\left\|\mathfrak{W}^{l}_L(t)-\mathfrak{W}^{l}_L(0)\right\|_{p^{in}}\right)^8\right)\\
&\le(L-1)\Psi_{L,8}(L-1)C_L^8\left(\Phi_{L,8}\right)^{1/2}\left(\tilde{\Phi}_{L,8}\right)^{1/2}.
\end{aligned}
$$

Then, we have

$$
\begin{aligned}
&\sum_{i=1}^{L-1}\sum_{l=0}^{i-1}\left(\left\|f^i_{\theta^{(i)}(0)}\right\|+\left\|f^i_{\theta^{(i)}(t)}-f^i_{\theta^{(i)}(0)}\right\|\right)^{j-1}\left\|f^l_{\theta^{(l)}(t)}(x)\right\|_{p^{in}}\left\|f^l_{\theta^{(l)}(t)}(x')\right\|\left\|\mathfrak{W}^{l+2}_i\right\|\left\|(\mathfrak{W}^{l+2}_L)^\intercal\right\|_{p^{in}}\\
&\le\left(\sum_{i=1}^{L-1}\left(\left\|f^i_{\theta^{(i)}(0)}\right\|+\left\|f^i_{\theta^{(i)}(t)}-f^i_{\theta^{(i)}(0)}\right\|\right)^{j-1}\right)\left(\sum_{i=1}^{L-1}\sum_{l=0}^{i-1}\frac{L-1}{i-1}\left\|\mathfrak{W}^{l+2}_i\right\|\left(\left\|f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}+\left\|f^l_{\theta^{(l)}(t)}-f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}\right)\right.\\
&\quad\left.\times\left(\left\|f^l_{\theta^{(l)}(0)}\right\|+\left\|f^l_{\theta^{(l)}(t)}-f^l_{\theta^{(l)}(0)}\right\|\right)\left(\left\|\mathfrak{W}^{l+2}_L(0)\right\|_{p^{in}}+\left\|\mathfrak{W}^{l+2}_L(t)-\mathfrak{W}^{l+2}_L(0)\right\|_{p^{in}}\right)\right)\\
&\le(L-1)C_L^{j-1}\tilde{\Phi}_{L,j-1}\left(\sum_{i=1}^{L-1}\sum_{l=0}^{i-1}\frac{L-1}{i-1}\left\|\mathfrak{W}^{l+2}_i\right\|\left(\left\|f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}+\left\|f^l_{\theta^{(l)}(t)}-f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}\right)\right.\\
&\quad\left.\times\left(\left\|f^l_{\theta^{(l)}(0)}\right\|+\left\|f^l_{\theta^{(l)}(t)}-f^l_{\theta^{(l)}(0)}\right\|\right)\left(\left\|\mathfrak{W}^{l+2}_L(0)\right\|_{p^{in}}+\left\|\mathfrak{W}^{l+2}_L(t)-\mathfrak{W}^{l+2}_L(0)\right\|_{p^{in}}\right)\right)\\
&\le(L-1)C_L^{j-1}\tilde{\Phi}_{L,j-1}\left(\sum_{i=1}^{L-1}\sum_{l=0}^{i-1}\frac{L-1}{i-1}\left\|\mathfrak{W}^{l+2}_i\right\|^2\right)^{1/2}\left((L-1)\sum_{l=0}^{L-2}\left(\left\|f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}+\left\|f^l_{\theta^{(l)}(t)}-f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}\right)^2\right.\\
&\quad\left.\left(\left\|f^l_{\theta^{(l)}(0)}\right\|+\left\|f^l_{\theta^{(l)}(t)}-f^l_{\theta^{(l)}(0)}\right\|\right)^2\times\left(\left\|\mathfrak{W}^{l+2}_L(0)\right\|_{p^{in}}+\left\|\mathfrak{W}^{l+2}_L(t)-\mathfrak{W}^{l+2}_L(0)\right\|_{p^{in}}\right)^2\right)^{1/2}\\
&\le(L-1)C_L^{j-1}\tilde{\Phi}_{L,j-1}\left((L-1)^2\tilde{\Psi}_{L,2}\right)^{1/2}\left((L-1)^2C_L^4\left(\Phi_{L,8}\right)^{1/4}\left(\Psi_{L,8}\right)^{1/2}\left(\tilde{\Phi}_{L,8}\right)^{1/4}\right)^{1/2}\\
&=(L-1)^3C_L^{j+1}\tilde{\Phi}_{L,j-1}\left(\tilde{\Psi}_{L,2}\right)^{1/2}\left(\Phi_{L,8}\right)^{1/8}\left(\Psi_{L,8}\right)^{1/4}\left(\tilde{\Phi}_{L,8}\right)^{1/8}
\end{aligned}
$$

Thus, by combining the above inqualities,

$$
\begin{aligned}
&\partial_t \tilde{\Phi}_{L,j}(t) \\
&\leq \frac{j162d^2 \max_l\{n_l^4\}N^{1/2}}{(L-1)^2 C_L^{2+j}} \sum_{i=1}^{L-1}\sum_{l=0}^{i-1} \left(\left\|f^i_{\theta^{(i)}(0)}\right\| + \left\|f^i_{\theta^{(i)}(t)} - f^i_{\theta^{(i)}(0)}\right\|\right)^{j-1} \left\|f^l_{\theta^{(l)}(t)}(x)\right\|_{p^{in}} \left\|f^l_{\theta^{(l)}(t)}(x')\right\| \\
&\quad \times \left\|\mathfrak{W}_i^{l+2}\right\| \left\|(\mathfrak{W}_L^{l+2})^\intercal\right\|_{p^{in}} \left\|\delta_t^L\right\|_{p^{in}} \\
&\leq \frac{j162d^2 \max_l\{n_l^4\}N^{1/2}(L-1)}{C_L} \tilde{\Phi}_{L,j-1}\left(\tilde{\Psi}_{L,2}\right)^{1/2} (\Phi_{L,8})^{1/8} (\Psi_{L,8})^{1/4} \left(\tilde{\Phi}_{L,8}\right)^{1/8} \left\|\delta_t^L\right\|_{p^{in}} \\
&\leq \frac{j162d^2 \max_l\{n_l^4\}N^{1/2}(L-1)}{C_L} \tilde{\Phi}_{L,j-1}\tilde{\Psi}_{L,2}\Phi_{L,8}\Psi_{L,8}\tilde{\Phi}_{L,8} \left\|\delta_t^L\right\|_{p^{in}}.
\end{aligned}
$$

□

N. Proof of Lemma 27 for Proposition 6

Lemma 27. *For $t \geq 0$ and $x \in \mathbb{R}_+^{d\times d}$, if $f^l_{\theta^{(l)}(t)}(x)$ is element-wise nonzero for all $1 \leq l \leq L-1$, then*

$$
\begin{aligned}
\partial_t \tilde{\Psi}_L^k(x,t) &\leq \frac{9d^4(L-1)}{C_L}\Psi_{L,8}(t)\tilde{\Phi}_{L,4}(x,t)\tilde{\Psi}_{L,2}(x,t)\tilde{\Psi}_{L,8}(x,t)\left\|\delta_t^L\right\|_{p^{in}} \\
\partial_t \tilde{\Phi}_L^l(x,t) &\leq \frac{81d^4 \max_l\{n_l^4\}N^{1/2}(L-1)}{C_L}\Phi_{L,8}(t)\Psi_{L,8}(t)\tilde{\Phi}_{L,8}(x,t)\tilde{\Psi}_{L,2}(x,t)\left\|\delta_t^L\right\|_{p^{in}}.
\end{aligned}
$$

for $1 \leq l \leq L$ and $2 \leq k \leq L$.

Proof of Lemma 27. Consider the derivative of $\tilde{\Psi}_L^k$ first. In the proof of Lemma 26, we showed

$$
\begin{aligned}
&\left(\sum_{i=2}^{L}\sum_{l=i}^{L}\left\|\mathfrak{W}_{l-1}^i\right\|\left(\left\|f^{l-1}_{\theta^{(l-1)}(0)}\right\|_{p^{in}} + \left\|f^{l-1}_{\theta^{(l-1)}(t)} - f^{l-1}_{\theta^{(l-1)}(0)}\right\|_{p^{in}}\right)\left(\left\|\mathfrak{W}_L^{l+1}(0)\right\| + \left\|\mathfrak{W}_L^{l+1}(t) - \mathfrak{W}_L^{l+1}(0)\right\|\right)\right. \\
&\quad \left. \times \left(\left\|\mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}} + \left\|\mathfrak{W}_L^{l+1}(t) - \mathfrak{W}_L^{l+1}(0)\right\|_{p^{in}}\right)\right)^2 \\
&\leq (L-1)^2\tilde{\Psi}_{L,2}(L-1)\times(L-1)C_L^2\left(\Phi_{L,4}\right)^{1/2}\left(\Psi_{L,8}(t)\right)^{1/4}\left(\tilde{\Psi}_{L,8}(t)\right)^{1/4}.
\end{aligned}
$$

Thus, from Lemma 36,

$$
\begin{aligned}
\partial_t \tilde{\Psi}_L^k(t) &\leq \frac{9d^4}{(L-1)C_L^2}\sum_{l=k}^{L}\left\|\mathfrak{W}_L^{l+1}(t)\right\|_{p^{in}}\left\|\mathfrak{W}_L^{l+1}(t)\right\|\left\|\mathfrak{W}_{l-1}^k(t)\right\|\left\|f^{l-1}_{\theta^{(l-1)}(t)}(x)\right\|_{p^{in}}\left\|\delta_t^L\right\|_{p^{in}} \\
&\leq \frac{9d^4(L-1)}{C_L}\left(\tilde{\Psi}_{L,2}\right)^{1/2}\left(\tilde{\Phi}_{L,4}\right)^{(1/4)}\left(\Psi_{L,8}\right)^{(1/8)}\left(\tilde{\Psi}_{L,8}\right)^{(1/8)}\left\|\delta_t^L\right\|_{p^{in}} \\
&\leq \frac{9d^4(L-1)}{C_L}\tilde{\Psi}_{L,2}\tilde{\Phi}_{L,4}\Psi_{L,8}\tilde{\Psi}_{L,8}\left\|\delta_t^L\right\|_{p^{in}}.
\end{aligned}
$$

On the other hand, let consider the derivative of $\tilde{\Phi}_L^l(t)$. We also showed the following inequality from the proof of Lemma 26.

$$
\begin{aligned}
&\sum_{i=1}^{L-1}\sum_{l=0}^{i-1}\frac{L-1}{i-1}\left\|\mathfrak{W}_i^{l+2}\right\|\left(\left\|f^l_{\theta^{(l)}(0)}\right\|_{p^{in}} + \left\|f^l_{\theta^{(l)}(t)} - f^l_{\theta^{(l)}(0)}\right\|_{p^{in}}\right)\left(\left\|f^l_{\theta^{(l)}(0)}\right\| + \left\|f^l_{\theta^{(l)}(t)} - f^l_{\theta^{(l)}(0)}\right\|\right) \\
&\quad \times\left(\left\|\mathfrak{W}_L^{l+2}(0)\right\|_{p^{in}} + \left\|\mathfrak{W}_L^{l+2}(t) - \mathfrak{W}_L^{l+2}(0)\right\|_{p^{in}}\right) \\
&\leq (L-1)^2C_L^2\left(\tilde{\Psi}_{L,2}\right)^{1/2}\left(\Phi_{L,8}\right)^{1/8}\left(\Psi_{L,8}\right)^{1/4}\left(\tilde{\Phi}_{L,8}\right)^{1/8}.
\end{aligned}
$$

Thus, from Lemma 36,

$$\begin{aligned}
\partial_t \tilde{\Phi}_L^l(t) &\leq \frac{81 n_i^2 n_{i-1}^2 N^{1/2}}{(L-1)C_L^3} \sum_{i=0}^{l-1} \left\| f^i_{\theta^{(i)}(t)}(x) \right\|_{p^{in}} \left\| f^i_{\theta^{(i)}(t)}(x') \right\| \left\| \mathfrak{W}_l^{i+2} \right\| \left\| (\mathfrak{W}_L^{i+2})^\intercal \right\|_{p^{in}} \left\| \delta_t^L \right\|_{p^{in}} \\
&\quad + \frac{d^4}{(L-1)C_L^3} \sum_{i=0}^{l-1} \left\| \mathfrak{W}_l^{i+2} \right\| \left\| (\mathfrak{W}_L^{i+2})^\intercal \right\|_{p^{in}} \left\| \delta_t^L \right\|_{p^{in}} \\
&\leq \frac{81 d^4 \max_l \{n_l^4\} N^{1/2}(L-1)}{C_L} \left(\tilde{\Psi}_{L,2}\right)^{1/2} \left(\Phi_{L,8}\right)^{1/8} \left(\Psi_{L,8}\right)^{1/4} \left(\tilde{\Phi}_{L,8}\right)^{1/8} \left\| \delta_t^L \right\|_{p^{in}} \\
&\leq \frac{81 d^4 \max_l \{n_l^4\} N^{1/2}(L-1)}{C_L} \tilde{\Psi}_{L,2} \Phi_{L,8} \Psi_{L,8} \tilde{\Phi}_{L,8} \left\| \delta_t^L \right\|_{p^{in}}.
\end{aligned}$$

□