Modeling from Features: a Mean-field Framework for Over-parameterized Deep Neural Networks

Cong Fang FANGCONG@PKU.EDU.CN
Department of Machine Intelligence, Peking University

Jason D. Lee JASONLEE@PRINCETON.EDU
Electrical and Computer Engineering, Princeton University

Pengkun Yang YANGPENGKUN@TSINGHUA.EDU.CN
Center for Statistical Science, Tsinghua University

Tong Zhang TONGZHANG@TONGZHANG-ML.ORG
Department of Mathematics and Computer Science, HKUST

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

This paper proposes a new mean-field framework for over-parameterized deep neural networks (DNNs), which can be used to analyze neural network training. In this framework, a DNN is represented by probability measures and functions over its features (that is, the function values of the hidden units over the training data) in the continuous limit, instead of the neural network parameters as most existing studies have done. This new representation overcomes the degenerate situation where all the hidden units essentially have only one meaningful hidden unit in each middle layer, leading to a simpler representation of DNNs. Moreover, we construct a non-linear dynamics called *neural feature flow*, which captures the evolution of an over-parameterized DNN trained by Gradient Descent. We illustrate the framework via the Residual Network (Res-Net) architecture. It is shown that when the neural feature flow process converges, it reaches a *global* minimal solution under suitable conditions.

Keywords: deep residual network, mean-field theory, non-linear dynamics, global minimum.

1. Introduction

In recent years, deep neural networks (DNNs) have achieved great success empirically. However, the theoretical understanding of the practical success is still limited. One main conceptual difficulty is the non-convexity of DNN models. More recently, there has been remarkable progress in understanding the over-parameterized neural networks (NNs), which are NNs with massive hidden units. The over-parameterization is capable of circumventing the hurdles in analyzing non-convex functions under specific settings:

(i) Under a specific scaling and initialization, it is sufficient to study the NN weights in a small region around the initial values given sufficiently many hidden units - the aptly named "lazy training" regime (Jacot et al., 2018; Li and Liang, 2018; Du et al., 2019a; Arora et al., 2019; Du et al., 2019b; Allen-Zhu et al., 2018; Allen-Zhu and Li, 2019; Zou et al., 2018; Chizat et al., 2019). The NN in this regime is nearly a linear model fitted with a random kernel in the tangent space, and provably achieves minimum training error. However, this regime does not explain why NNs can effectively learn representative features, and the expressive power of random kernels is limited (Yehudai and Shamir, 2019).

(ii) Another line of research applies the mean-field analysis for NNs (Mei et al., 2018; Chizat and Bach, 2018; Sirignano and Spiliopoulos, 2019a; Rotskoff and Vanden-Eijnden, 2018; Mei et al., 2019; Dou and Liang, 2019; Wei et al., 2018; Sirignano and Spiliopoulos, 2019b; Fang et al., 2019; Araújo et al., 2019; Nguyen and Pham, 2020; Chen et al., 2020). Learning a two-layer over-parameterized NN can be approximately described as optimizing a functional over probability distributions of the NN weights. The evolution of NN weights trained by the (noisy) Gradient Descent algorithm corresponds to a Wasserstein gradient flow called "distributional dynamics", solution to a non-linear partial differential equation (PDE) of McKean-Vlasov type (Sznitman, 1991). In the mean-field limit, the Wasserstein gradient flow converges to the globally optimal solution for two-layer NNs (Mei et al., 2018; Chizat and Bach, 2018; Fang et al., 2019). Compared with lazy training, the mean-field view can characterize the entire training process of NNs.

However, the mean-field analysis on DNNs is a challenging task. First of all, it is not easy to formulate the mean-field limit of DNNs. As we will discuss in Section 2.1, extending existing formulations to DNNs, hidden units in a middle layer essentially behave as a single unit along the training. This degenerate situation arguably cannot fully characterize the training process of actual DNNs. Moreover, understandings for the global convergence of Gradient Descent on DNNs are still required in the mean-field regime.

In this paper, we propose a new mean-field framework for over-parameterized DNNs to analyze NN training. In contrast to existing studies focusing on the NN weights, this framework represents a DNN in the continuous limit by probability measures and functions over its features, that is, the outputs of the hidden units over the training data. This new representation overcomes the degenerate situation in previous studies (Araújo et al., 2019; Nguyen and Pham, 2020).

We further describe a non-linear dynamic called *neural feature flow* that captures the evolution of a DNN trained by Gradient Descent. We illustrate the framework by Res-Nets (He et al., 2016). Neural feature flow involves the evolution of the features and does not require the boundedness of the weights. Under the standard initialization method of discrete Res-Nets (Glorot and Bengio, 2010; He et al., 2015), the NN weights scale to infinity with the growth of the number of hidden units. There are empirical studies, e.g. Zhang et al. (2019), which show that properly rescaling the standard initialization stabilizes training. We introduce a simple ℓ_2-regression at initialization (see Algorithm 2). We prove that Gradient Descent from the regularized initialization with a suitable time scale on Res-Nets can be well-approximated by its limit, i.e., neural feature flow, when the number of hidden units is sufficiently large.

Finally, we consider the global convergence of neural feature flow for Res-Nets. Surprisingly, we show that when the neural feature flow process converges, it reaches a *globally* optimal solution under suitable conditions. We summarize the contributions of the paper below:

(A) We propose a new mean-field framework of DNNs which characterizes DNNs via probability measures and functions over the features and introduce neural feature flow to capture the evolution of DNNs trained by the Gradient Descent algorithm.

(B) We illustrate our framework by Res-Net model. We show that neural feature flow can find a global minimal solution of the learning task under certain conditions.

Our mean-field description can also be used to study the standard DNNs, which is discussed in Appendix E. However, it still remains open to achieve the global convergence of neural feature flow for standard DNNs.

1.1. Notations

Let $[m_1 : m_2] := \{m_1, m_1 + 1, \ldots, m_2\}$ for $m_1, m_2 \in \mathbb{N}$ with $m_1 \leq m_2$ and $[m_2] := [1 : m_2]$ for $m_2 \geq 1$. Let $\mathcal{P}^n$ be the set of probability distributions over $\mathbb{R}^n$. For a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times m}$, let $\|\boldsymbol{A}\|_2$, $\|\boldsymbol{A}\|_F$, and $\|\boldsymbol{A}\|_\infty$ denote its operator, Frobenius, max norms, respectively. If $\boldsymbol{A}$ is symmetric, let $\lambda_{\min}(\boldsymbol{A})$ be its smallest eigenvalue. Vectors are treated as columns. For a vector $\boldsymbol{a} \in \mathbb{R}^n$, let $\|\boldsymbol{a}\|_2$ and $\|\boldsymbol{a}\|_\infty$ denote its ℓ_2 and ℓ_∞ norms, respectively. The i-th coordinate is denoted by $\boldsymbol{a}(i)$. For $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$, denote the entrywise product by $\boldsymbol{a} \circ \boldsymbol{b}$ that $[\boldsymbol{a} \circ \boldsymbol{b}](i) := \boldsymbol{a}(i) \circ \boldsymbol{b}(i)$ for $i \in [n]$. For $c > 0$ and $p \in [1, \infty]$, let $\mathcal{B}_p(\boldsymbol{a}, c)$ denote the ℓ_p-ball centered at $\boldsymbol{a}$ of radius c. For an unary function $f : \mathbb{R} \to \mathbb{R}$, define $\dot{f} : \mathbb{R}^n \to \mathbb{R}^n$ as the entrywise operation that $\dot{f}(\boldsymbol{a})(i) = f(\boldsymbol{a}(i))$ for $i \in [n]$ and $\boldsymbol{a} \in \mathbb{R}^n$. Denote n-dimensional identity matrix by $\boldsymbol{I}^n$. Denote m-by-n zero matrix and n-dimensional zero vector by $\boldsymbol{0}^{n\times m}$ and $\boldsymbol{0}^n$, respectively. We say a univariate distribution p is σ-sub-gaussian if $\mathbb{E}_{x\sim p} \exp(x^2/\sigma^2) \leq e$[1]; we say a d-dimensional distribution p is σ-sub-gaussian if the law of $u^\top \boldsymbol{x}$ is σ-sub-gaussian for $\boldsymbol{x} \sim p$ and any $u \in \mathbb{S}^{d-1}$. For two positive sequences $\{p_n\}$ and $\{q_n\}$, $p_n = \mathcal{O}(q_n)$ if $p_n \leq Cq_n$ for some positive constant C, and $p_n = \Omega(q_n)$ if $q_n = \mathcal{O}(p_n)$.

2. Related Deep Learning Theory

In recent years, there have been a number of significant developments to obtain better theoretical understandings of NNs. One remarkable direction is to restrict the NN training in a small region. In this lazy training regime, the analysis cannot explain how NNs learn discriminative features. This is observed in real applications and argued to be one of the contributors to the success of deep learning. Beyond lazy training, one promising direction is to conduct mean-field analysis. However, in section 2.1, we show the challenges of analysis on DNNs. Specifically, if we still model from the weights, the standard initialization,e.g., (Glorot and Bengio, 2010; He et al., 2015) scales the weights to $\sqrt{m}$, which diverges in the mean-field limit, where m is the number of hidden units. On the other side, if we initialize the weights from a fixed distribution that is independent of m as existing mean-field works (Araújo et al., 2019; Nguyen and Pham, 2020) considered, DNNs would be stuck in a degenerate situation where the middle layers essentially only have one single feature. Both the issues motivate us to study DNN directly from the features. As a result, we propose a mean-field framework from tracking the distributions of features to analyze the DNN training.

2.1. Challenges on Mean-field Theory for DNNs

We discuss related mean-field studies and point out the challenges in modeling DNNs. For two-layer NNs, most of the existing works (Mei et al., 2018; Chizat and Bach, 2018; Sirignano and Spiliopoulos, 2019a; Rotskoff and Vanden-Eijnden, 2018) formulate the continuous limit as

$$f(\boldsymbol{x}; p) = \int w_2 \, h\left(\boldsymbol{w}_1^\top \boldsymbol{x}\right) dp\left(w_2, \boldsymbol{w}_1\right),$$

1. Here the value e can be replaced by any number greater than one. See Vershynin (2010, Remark 5.6).

where p is the probability distribution over the pair of weights $(w_2, \boldsymbol{w}_1)$. The weights of the second layer w_2 can be viewed as functions of $\boldsymbol{w}_1$, which is a d-dimensional vector. However, following the approach, the higher-layer weights, say w_3, are functions over features of the hidden layer, with a diverging dimensionality in the mean-field limit. For 3-layer NNs, w_3 as the last hidden layer is indexed by the connection to the output units in Nguyen and Pham (2020), which is not generalizable when middle layers present. An alternative approach is to model DNNs with nested measures (also known as multi-layer measures; see Dawson et al. (1982); Dawson (2018) and references therein), which however suffers the closure problem to establish a well-defined limit (see discussions in Sirignano and Spiliopoulos (2019b, Section 4.3)).

The continuous limit of DNNs is investigated by Araújo et al. (2019); Nguyen and Pham (2020) under the initialization that all weights are i.i.d. realizations of a *fixed* distribution independent of the number of hidden units. However, under that setting, all neurons in a middle layer essentially behave as a single neuron. Consider the output $\hat{\boldsymbol{\beta}}$ of a middle-layer neuron connecting to m hidden neurons in the previous layer:

$$\hat{\boldsymbol{\beta}} = \frac{1}{m} \sum_{i=1}^{m} h(\hat{\boldsymbol{\beta}}_i') \, w_i, \tag{1}$$

where $\hat{\boldsymbol{\beta}}_i'$ is the output of i-th hidden neuron in the previous layer with bounded variance, w_i is the connecting weight, and h is the activation function. If w_i is initialized independently from $\mathcal{N}(0,1)$, it is clear that $\mathsf{var}[\hat{\boldsymbol{\beta}}] \to 0$ as $m \to \infty$, and thus the hidden neurons in middle layers are indistinguishable at the initialization. Moreover, the phenomenon sustains along the entire training process, as shown in Proposition 1. This phenomenon serves as the basis of Araújo et al. (2019); Nguyen and Pham (2020) to characterize the mean-field limit using finite-dimensional probability distributions. This degenerate situation arguably does not fully characterize the actual DNN training. In fact, similar calculations to (1) are carried out by Glorot and Bengio (2010); He et al. (2015) and motivate the popular initialization strategy with $\mathcal{N}(0, \mathcal{O}(m))$ such that the variance of $\hat{\boldsymbol{\beta}}$ is non-vanishing.

Proposition 1 *Consider fully-connected L-layer DNNs with m units in each hidden layer trained by Gradient Descent. Suppose the activation and loss functions satisfy Assumption 1. Let the weights be initialized from a distribution with $O(1)$ variance. Let $\hat{\boldsymbol{\beta}}_{\ell,i}^k$ denote the output of i-th hidden neuron at ℓ-th layer and k-th iteration, and define $\Delta_{\ell,m} := \max_{i \neq i', k \in [K]} \|\hat{\boldsymbol{\beta}}_{\ell,i}^k - \hat{\boldsymbol{\beta}}_{\ell,i'}^k\|_\infty$. Then, for every $\ell \in [2 : L-1]$, almost surely,*

$$\lim_{m \to \infty} \Delta_{\ell,m} = 0.$$

To overcome this degenerate situation, we consider the popular initialization strategy with a simple ℓ_2 regression (See Algorithm 2). The regression preserves all initial features, thus the variance of the features is now non-vanishing. Moreover, to accurately characterize DNNs in the mean-field regime, we introduce a probability measure over the features instead of the weights, which leads to a new dynamic system called neural feature flow to describe the neural nets trained by Gradient Descent. We achieve to describe a more realistic learning process.

2.2. Comparisons of Dynamics with Other Mean-field Works

It is known that the evolution of a two-level NN trained by Gradient Descent can be described as a Mckean-Vlason process (Mei et al., 2018; Chizat and Bach, 2018). There are lots of works that

studies the evolution of DNN trained by Gradient Descent. One important work from Araújo et al. (2019) shows that the evolution of DNNs can also be characterized by PDEs of Mckean-Vlason type when weights of DNNs in the first and the last layers are not updated. More recently, Nguyen and Pham (2020); Pham and Nguyen (2020) proposed another very interesting attempt by directly tracking the trajectories of the weights. In their description called neuronal embedding, the evolution of Gradient Descent is characterized by systems of ODEs, avoiding the presence of the conditional probabilities which is the main issue proposed by Araújo et al. (2019). Our description follows from the idea of Nguyen and Pham (2020). However, our dynamic further involves the evolution of the features and does not require the boundedness of the weights. Moreover, we introduce the conception of skip-connected paths to deal with the Res-Net architecture. Neural feature flow is more meaningful training dynamics as it is no longer restricted in the degenerate situation as Section 2.1 described.

For the analysis to achieve a global minimum for the dynamics, the work from Lu et al. (2020) considers the DNN as a relatively simple composition of multiple two-layer NNs. Their global convergence result requires a very restricted assumption that the limiting distribution has full support. The work form Nguyen and Pham (2020); Pham and Nguyen (2020) proved the global convergence for DNNs by a very novel topology argument under the degenerate initialization. Pham and Nguyen (2020) also mentioned the possibility of non-degenerate initialization leading to global convergence guarantees. Our proof idea is similar to Nguyen and Pham (2020), whereas, we take our concentration on the features.

2.3. Beyond Lazy Training

In the "lazy training" regime, e.g. Jacot et al. (2018); Du et al. (2019a); Allen-Zhu and Li (2019); Zou et al. (2018), the weights are restricted in an infinitely small region. The DNN in this regime essentially corresponds to a linear model on random features associated with a kernel termed neural tangent kernel. In the limit, the features are fixed. Encouragingly, we consider NNs beyond the lazy training regime and further allow the feature to move in a constant region. To arrive the goal, we study a special Res-Net architecture that bounds residuals by a bounded mapping h_2 (see Section 3.1). Note that in our analysis, the bound is not needed to be small enough but can be arbitrarily large and less than infinity. Therefore, our setting allows the DNN to learn the targeted features. From the technical aspect, we require a different treatment to show the full rank of the feature matrices; this is achieved by Brouwer's fixed-point theorem (see the Proof of Theorem 8 in Appendix B.2).

3. Formulation of Continuous Res-Nets

We consider the empirical minimization problem over N training samples $\{\boldsymbol{x}^i, y^i\}_{i=1}^N$, where $\boldsymbol{x}^i \in \mathbb{R}^d$ and $y^i \in \mathcal{Y}$. For regression problems, $\mathcal{Y}$ is typically $\mathbb{R}$; for classification problems, $\mathcal{Y}$ is often $[K]$ for an integer K. We first present the formulation of L-layer Res-Nets.

3.1. Discrete Res-Nets

For discrete Res-Nets, let m_ℓ denote the number of units at layer ℓ for $\ell = [0 : L+1]$. Suppose each hidden layer has m hidden units that $m_\ell = m$ for $\ell \in [L]$. Let $m_0 = d$ and node i outputs the value of i-th coordinate of the training data for $i \in [d]$. Let $m_{L+1} = 1$ that is the unit of the final network output. For $\ell \in [L+1]$, the output of node i for the N training samples in layer ℓ

is denoted by $\hat{\boldsymbol{\beta}}_{\ell,i} \in \mathbb{R}^N$; the weight that connects the node i at layer $\ell-1$ to node j at layer ℓ is denoted by $\hat{v}_{\ell,i,j} \in \mathbb{R}$.

(1) At the input layer, for $i \in [d]$, let $\hat{\boldsymbol{\beta}}_{0,i} := \left[\boldsymbol{x}^1(i), \boldsymbol{x}^2(i), \ldots, \boldsymbol{x}^N(i)\right]^\top$.

(2) At the first layer, for $j \in [m]$, let

$$\hat{\boldsymbol{\beta}}_{1,j} = \frac{1}{m_0} \sum_{i=1}^{m_0} \hat{v}_{1,i,j}\, \hat{\boldsymbol{\beta}}_{0,i}. \tag{2}$$

(3) We recursively define the upper layers for $\ell \in [2:L]$. Let $\hat{\boldsymbol{\alpha}}_{\ell,j} \in \mathbb{R}^N$ be the residual term at node j at layer ℓ:

$$\hat{\boldsymbol{\alpha}}_{\ell,j} = \frac{1}{m} \sum_{i=1}^{m} \hat{v}_{\ell,i,j}\, \dot{h}_1\left(\hat{\boldsymbol{\beta}}_{\ell-1,i}\right), \quad j \in [m], \tag{3}$$

where $h_1 : \mathbb{R} \to \mathbb{R}$ is the activation function and $\dot{h}_1 : \mathbb{R}^N \to \mathbb{R}^N$ is the entrywise operation for h_1, which satisfies $\dot{h}_1(\boldsymbol{a})(i) = h_1(\boldsymbol{a}(i))$ for $i \in [N]$ and $\boldsymbol{a} \in \mathbb{R}^N$. Furthermore, we consider the following coupling between the residual and the previous feature:

$$\hat{\boldsymbol{\beta}}_{\ell,j} = \dot{h}_2\left(\hat{\boldsymbol{\alpha}}_{\ell,j}\right) + \hat{\boldsymbol{\beta}}_{\ell-1,j}, \quad j \in [m]. \tag{4}$$

where $h_2 : \mathbb{R} \to \mathbb{R}$.

(4) For the output,

$$\hat{\boldsymbol{\beta}}_{L+1,1} = \frac{1}{m} \sum_{i=1}^{m} \hat{v}_{L+1,i,1} \dot{h}_1\left(\hat{\boldsymbol{\beta}}_{L,i}\right). \tag{5}$$

We collect the weights from all layers into a single vector $\hat{\boldsymbol{v}} \in \mathbb{R}^{D_1}$ with $D_1 := m^2(L-1) + (d+1)m$. We also collect the residuals, and features from layers 2 to L into single vectors $\hat{\boldsymbol{\alpha}} \in \mathbb{R}^{D_2}$, and $\hat{\boldsymbol{\beta}} \in \mathbb{R}^{D_2}$, respectively, where $D_2 := Nm(L-1)$. The learning problem for Res-Nets is given by

$$\min_{\hat{\boldsymbol{v}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}} \hat{\mathcal{L}}(\hat{\boldsymbol{v}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) = \frac{1}{N} \sum_{n=1}^{N} \phi\left(\hat{\boldsymbol{\beta}}_{L+1,1}(n),\ y^n\right), \tag{6}$$

where $(\hat{\boldsymbol{v}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ satisfies (2) – (5), and $\phi : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ denotes the loss function. One noteworthy feature in the architecture is (4), where we introduce a mapping h_2 on the residual $\hat{\boldsymbol{\alpha}}_{\ell,j}$ before fusing it with $\hat{\boldsymbol{\beta}}_{\ell-1,j}$. As have been mentioned, we assume that h_2 is bounded by a constant L_1, and hence $\|\hat{\boldsymbol{\beta}}_{\ell,j} - \hat{\boldsymbol{\beta}}_{\ell-1,j}\|_\infty \le L_1$. Therefore, the high-layer features can be regarded as perturbations of the low-layer ones. Similar ideas have also appeared in Du et al. (2019a); Hardt and Ma (2016), but realized in a different way. For example, in the lazy training regime, Du et al. (2019a) achieved it by scaling $\hat{\boldsymbol{\alpha}}_{\ell,j}$ with a vanishing $\mathcal{O}(\frac{1}{\sqrt{m}})$ factor.

3.2. Continuous Res-Nets Formulation

We propose our formulation for the continuous Res-Nets. We consider the Res-Net with the architecture described in Section 3.1 that is initialized by a combination of a standard initialization and an additional regression procedure (See Algorithm 2). One can find that the regression procedure preserves all initial features but reduces the redundancy of the weights, making us introduce real-value functions to characterize the weights in the mean-field limit.

In fact, displayed in Appendix E.2, the continuous limit of a standard DNN can be described by introducing probability measures over features for individual layers. This idea is much clear. Here, we study Res-Net with the skip connections as in (4). In the continuous case, the discrete index j no longer makes sense and the skip connections need to be properly parametrized by an infinite set. To overcome the hurdle of infinite skip connections, we introduce $\boldsymbol{\Theta} = (\boldsymbol{v}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_L) \in \mathbb{R}^D$ for $D = d + (N-1)L$ to parametrize the skip connections that are described in (8) and (9) below. Each $\boldsymbol{\Theta}$ consists of $\boldsymbol{v}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_L$ that can be regarded as an input-output path $\boldsymbol{v}_1 \to \boldsymbol{\alpha}_2 \to \cdots \to \boldsymbol{\alpha}_L$ and is called a *skip-connected path*. Our main technique is to characterize the overall state of the continuous Res-Nets by the density p over skip-connected paths. Thus the joint distribution p can be regarded as a description of the overall topological structure about the skip connections. We represent the features $\boldsymbol{\beta}_\ell$ in the hidden layer $\ell \in [L]$ as functions of $\boldsymbol{\Theta}$ that we introduce next:

(1) At the input layer, let $\boldsymbol{X} = \left[\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^N\right]^\top \in \mathbb{R}^{N \times d}$.

(2) At the first layer, let the features be

$$\boldsymbol{\beta}_1(\boldsymbol{\Theta}) = \frac{1}{d}(\boldsymbol{X}\boldsymbol{v}_1). \tag{7}$$

(3) At layer $\ell \in [2:L]$, let $v_\ell : \mathrm{supp}(p) \times \mathrm{supp}(p) \to \mathbb{R}$ denote the weights on the connections from layer $\ell - 1$ to ℓ, then for all $\boldsymbol{\Theta} = (\boldsymbol{v}_1, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_L) \in \mathbb{R}^D$, we have the forward-propagation constraint for v_ℓ and p:

$$\boldsymbol{\alpha}_\ell = \int v_\ell\left(\boldsymbol{\Theta}, \bar{\boldsymbol{\Theta}}\right) \dot{h}_1\left(\boldsymbol{\beta}_{\ell-1}(\bar{\boldsymbol{\Theta}})\right) dp\left(\bar{\boldsymbol{\Theta}}\right), \tag{8}$$

$$\boldsymbol{\beta}_\ell(\boldsymbol{\Theta}) = \dot{h}_2(\boldsymbol{\alpha}_\ell) + \boldsymbol{\beta}_{\ell-1}(\boldsymbol{\Theta}). \tag{9}$$

Here, $\boldsymbol{\Theta}$ takes on values in $\mathbb{R}^D$ and for each $\boldsymbol{\Theta}$, $\boldsymbol{\alpha}_\ell$ is one part of $\boldsymbol{\Theta}$ and $\boldsymbol{\beta}_\ell$ is a function of $\boldsymbol{\Theta}$. $\boldsymbol{\alpha}_\ell$ represents the residual at layer ℓ on the skip connected path described by $\boldsymbol{\Theta}$. And $\boldsymbol{\beta}_\ell(\boldsymbol{\Theta})$ represents the corresponding feature.

(4) At the output layer, let $v_{L+1} : \mathrm{supp}(p) \to \mathbb{R}$ be the weights in the layer $L+1$, and we have

$$\boldsymbol{\beta}_{L+1} = \int v_{L+1}(\boldsymbol{\Theta}) \dot{h}_1(\boldsymbol{\beta}_L(\boldsymbol{\Theta})) dp(\boldsymbol{\Theta}). \tag{10}$$

In our continuous formulation, a static Res-Net is characterized by p and $v_2 \ldots v_{L+1}$. We will show in the next section that the continuous Res-Net $(\{v_\ell\}_{\ell=2}^{L+1}, p)$ that satisfies the (7) – (10) will serve as a feasible initialization for neural feature flow.

Algorithm 1 Scaled Gradient Descent for Training a Res-Net.

1: Input the data $\{\boldsymbol{x}^i, y^i\}_{i=1}^N$, step size η, and initial weights $\hat{\boldsymbol{v}}^0$.
2: **for** $k = 0, 1, \ldots, K-1$ **do**
3: Perform forward-propagation (2) – (5) to compute $\hat{\boldsymbol{\beta}}_{L+1,1}^k$.
4: Perform backward-propagation to compute the gradient $\hat{\mathcal{G}}_{\ell,i,j}^k = \frac{\partial \hat{\mathcal{L}}}{\partial \hat{v}_{\ell,i,j}^k}$.
5: Perform Scaled Gradient Descent:

$$\hat{v}_{\ell,i,j}^{k+1} \quad = \quad \hat{v}_{\ell,i,j}^k - [\eta m_{\ell-1} m_\ell]\, \hat{\mathcal{G}}_{\ell,i,j}^k, \quad \ell \in [L+1],\ i \in [m_{\ell-1}],\ j \in [m_\ell].$$

6: **end for**
7: Output the weights $\hat{\boldsymbol{v}}^K$.

4. Scaled Gradient Descent and Neural Feature Flow for Res-Nets

We focus on the dynamic of the Res-Net trained by Gradient Descent. We consider the scaled Gradient Descent algorithm[2]. Given initial weights $\hat{\boldsymbol{v}}^0$, the meta-algorithm of scaled Gradient Descent is presented in Algorithm 1. Note that Algorithm 1 differs from the standard Gradient Descent only on the step sizes (time scales). Such scaling is also adopted in existing works (Araújo et al., 2019; Nguyen and Pham, 2020).

Now we describe the continuous limit of the Res-Net trained by Algorithm 1 by the continuous trajectories of the Res-Nets. This idea follows from Nguyen and Pham (2020) for analyzing there-layer DNNs. A trajectory is denoted by Φ that maps the initial Res-Nets at $t = 0$ to a Res-Net process over $[0, T]$. Specifically, it consists of the following parts:

- $\Phi_\ell^{\boldsymbol{\beta}} : \mathrm{supp}(p) \to C([0,T], \mathbb{R}^N)$ is the trajectory of $\boldsymbol{\beta}_\ell$ for $\ell \in [L]$;
- $\Phi_\ell^{\boldsymbol{\alpha}} : \mathrm{supp}(p) \to C([0,T], \mathbb{R}^N)$ is the trajectory of $\boldsymbol{\alpha}_\ell$ for $\ell \in [2:L]$;
- $\Phi_1^{\boldsymbol{v}} : \mathrm{supp}(p) \to C([0,T], \mathbb{R}^d)$ and $\Phi_{L+1}^{\boldsymbol{v}} : \mathrm{supp}(p) \to C([0,T], \mathbb{R})$ are the trajectories of $\boldsymbol{v}_1$ and v_{L+1}, respectively;
- $\Phi_\ell^{\boldsymbol{v}} : \mathrm{supp}(p) \times \mathrm{supp}(p) \to C([0,T], \mathbb{R})$ is the trajectory of v_ℓ for $\ell \in [2:L]$.

The continuous gradient for the weight can be obtained from the backward-propagation algorithm. For a given trajectory Φ, the gradients of weights at time $t \in [0,T]$ can be obtained from the backward-propagation algorithm. Similar to the usual backward-propagation, we first define gradients with respect to the features and residuals. Specifically, for all $\boldsymbol{\Theta} = (\boldsymbol{v}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_L) \in$

2. In practice, one often use stochastic gradient instead of the full counterpart for training. Under mild conditions, the dynamic of scaled Stochastic Gradient Descent will also converge to the neural feature flow in the continuous limit.

$\mathrm{supp}(p)$, $t \in [0,T]$, and $\ell \in [2:L]$, let

$$\boldsymbol{\beta}_{L+1}(\Phi,t) := \int \Phi^{\boldsymbol{v}}_{L+1}(\boldsymbol{\Theta})(t)\, \dot{h}_1\left(\Phi^{\boldsymbol{\beta}}_L(\boldsymbol{\Theta})(t)\right) dp(\boldsymbol{\Theta}), \tag{11a}$$

$$\underline{\mathcal{D}}_{L+1}(\Phi,t) := \left[\phi_1'\left(\boldsymbol{\beta}_{L+1}(\Phi,t)(1), y^1\right), \ldots, \phi_1'\left(\boldsymbol{\beta}_{L+1}(\Phi,t)(N), y^N\right)\right]^\top,$$

$$\underline{\mathcal{D}}^{\boldsymbol{\beta}}_L(\boldsymbol{\Theta};\Phi,t) := \left[\Phi^{\boldsymbol{v}}_{L+1}(\boldsymbol{\Theta})(t)\, \underline{\mathcal{D}}_{L+1}(\Phi,t)\right] \circ \dot{h}_1'\left(\Phi^{\boldsymbol{\beta}}_L(\boldsymbol{\Theta})(t)\right), \tag{11b}$$

$$\underline{\mathcal{D}}^{\boldsymbol{\alpha}}_\ell(\boldsymbol{\Theta};\Phi,t) := \underline{\mathcal{D}}^{\boldsymbol{\beta}}_\ell(\boldsymbol{\Theta};\Phi,t) \circ \dot{h}_2'\left(\Phi^{\boldsymbol{\alpha}}_\ell(\boldsymbol{\Theta})(t)\right), \tag{11c}$$

$$\underline{\mathcal{D}}^{\boldsymbol{\beta}}_{\ell-1}(\boldsymbol{\Theta};\Phi,t) := \underline{\mathcal{D}}^{\boldsymbol{\beta}}_\ell(\boldsymbol{\Theta};\Phi,t) + \left[\int \Phi^{\boldsymbol{v}}_\ell\left(\boldsymbol{\Theta},\bar{\boldsymbol{\Theta}}\right)(t)\, \underline{\mathcal{D}}^{\boldsymbol{\alpha}}_\ell(\bar{\boldsymbol{\Theta}};\Phi,t) dp\left(\bar{\boldsymbol{\Theta}}\right)\right] \circ \dot{h}_1'\left(\Phi^{\boldsymbol{\beta}}_{\ell-1}(\boldsymbol{\Theta})(t)\right).$$

For all $\boldsymbol{\Theta}, \bar{\boldsymbol{\Theta}} \in \mathrm{supp}(p)$, the drift term for the weights is given by

$$\underline{\mathcal{G}}^{\boldsymbol{v}}_{L+1}(\boldsymbol{\Theta};\Phi,t) := \frac{1}{N}\left[\underline{\mathcal{D}}_{L+1}(\Phi,t)\right]^\top \dot{h}_1\left(\Phi^{\boldsymbol{\beta}}_L(\boldsymbol{\Theta})(t)\right), \tag{12a}$$

$$\underline{\mathcal{G}}^{\boldsymbol{v}}_\ell\left(\boldsymbol{\Theta},\bar{\boldsymbol{\Theta}};\Phi,t\right) := \frac{1}{N}\left[\underline{\mathcal{D}}^{\boldsymbol{\alpha}}_\ell(\bar{\boldsymbol{\Theta}};\Phi,t)\right]^\top \dot{h}_1\left(\Phi^{\boldsymbol{\beta}}_{\ell-1}(\boldsymbol{\Theta})(t)\right), \quad \ell \in [2:L], \tag{12b}$$

$$\underline{\mathcal{G}}^{\boldsymbol{v}}_1(\boldsymbol{\Theta};\Phi,t) := \frac{1}{N}\boldsymbol{X}\, \underline{\mathcal{D}}^{\boldsymbol{\beta}}_1(\boldsymbol{\Theta};\Phi,t). \tag{12c}$$

Moreover, the changes of the weights will induce a change of the residuals and features. By the chain rule, we can obtain the drift term for the residuals and features: for $\ell \in [L-1]$ and $\boldsymbol{\Theta} \in \mathrm{supp}(p)$,

$$\underline{\mathcal{G}}^{\boldsymbol{\beta}}_1(\boldsymbol{\Theta};\Phi,t) := \frac{1}{d}\left[X \underline{\mathcal{G}}^{\boldsymbol{v}}_1(\boldsymbol{\Theta};\Phi,t)\right], \tag{13a}$$

$$\underline{\mathcal{G}}^{\boldsymbol{\alpha}}_{\ell+1}(\boldsymbol{\Theta};\Phi,t) := \int \Phi^{\boldsymbol{v}}_{\ell+1}\left(\bar{\boldsymbol{\Theta}},\boldsymbol{\Theta}\right)(t)\ \left[\dot{h}_1'\left(\Phi^{\boldsymbol{\beta}}_\ell(\bar{\boldsymbol{\Theta}})(t)\right) \circ \underline{\mathcal{G}}^{\boldsymbol{\beta}}_\ell\left(\bar{\boldsymbol{\Theta}};\Phi,t\right)\right] dp\left(\bar{\boldsymbol{\Theta}}\right) + \\ + \int \dot{h}_1\left(\Phi^{\boldsymbol{\beta}}_\ell\left(\bar{\boldsymbol{\Theta}}\right)(t)\right) \circ \underline{\mathcal{G}}^{\boldsymbol{v}}_{\ell+1}\left(\bar{\boldsymbol{\Theta}},\boldsymbol{\Theta};\Phi,t\right) dp\left(\bar{\boldsymbol{\Theta}}\right), \tag{13b}$$

$$\underline{\mathcal{G}}^{\boldsymbol{\beta}}_{\ell+1}(\boldsymbol{\Theta};\Phi,t) := \underline{\mathcal{G}}^{\boldsymbol{\beta}}_\ell(\boldsymbol{\Theta};\Phi,t) + \underline{\mathcal{G}}^{\boldsymbol{\alpha}}_{\ell+1}(\boldsymbol{\Theta};\Phi,t) \circ \dot{h}_2'\left(\Phi^{\boldsymbol{\alpha}}_{\ell+1}(\boldsymbol{\Theta})(t)\right). \tag{13c}$$

In all, the process of a continuous Res-Net trained by scaled Gradient Descent can be defined below.

Definition 2 (Neural Feature Flow for Res-Net) *Given an initial continuous Res-Net $(\{v_\ell\}_{\ell=2}^{L+1}, p)$ that satisfies the (7) – (10) and $T < \infty$, we say a trajectory Φ_* is a neural feature flow if for all $\boldsymbol{\Theta} = (\boldsymbol{v}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_L) \in \mathrm{supp}(p)$, $\bar{\boldsymbol{\Theta}} \in \mathrm{supp}(p)$, and $t \in [0,T]$,*

$$\Phi^{\boldsymbol{\beta}}_{*,\ell}(\boldsymbol{\Theta})(t) = \left[\frac{1}{d}\boldsymbol{X}\boldsymbol{v}_1 + \sum_{i=2}^{\ell} \dot{h}_2(\boldsymbol{\alpha}_i)\right] - \int_0^t \underline{\mathcal{G}}^{\boldsymbol{\beta}}_\ell(\boldsymbol{\Theta};\Phi_*,s), \quad \ell \in [L],$$

$$\Phi^{\boldsymbol{\alpha}}_{*,\ell}(\boldsymbol{\Theta})(t) = \boldsymbol{\alpha}_\ell - \int_0^t \underline{\mathcal{G}}^{\boldsymbol{\alpha}}_\ell(\boldsymbol{\Theta};\Phi_*,s)\, ds, \quad \ell \in [2:L],$$

$$\Phi^{\boldsymbol{v}}_{*,1}(\boldsymbol{\Theta})(t) = \boldsymbol{v}_1 - \int_0^t \underline{\mathcal{G}}^{\boldsymbol{v}}_1(\boldsymbol{\Theta};\Phi_*,s)\, ds,$$

$$\Phi^{\boldsymbol{v}}_{*,\ell}\left(\boldsymbol{\Theta},\bar{\boldsymbol{\Theta}}\right)(t) = v_\ell(\boldsymbol{\Theta},\bar{\boldsymbol{\Theta}}) - \int_0^t \underline{\mathcal{G}}^{\boldsymbol{v}}_\ell\left(\boldsymbol{\Theta},\bar{\boldsymbol{\Theta}};\Phi_*,s\right) ds, \quad \ell \in [2:L],$$

$$\Phi^{\boldsymbol{v}}_{*,L+1}(\boldsymbol{\Theta})(t) = v_{L+1}(\boldsymbol{\Theta}) - \int_0^t \underline{\mathcal{G}}^{\boldsymbol{v}}_{L+1}(\boldsymbol{\Theta};\Phi_*,s)\, ds.$$

We call the process as neural feature flow because it characterizes the evolution of both weights and features.

5. Main Results

5.1. Assumptions

We first present the assumptions that are needed in our analysis.

Assumption 1 (Activation Functions and Loss Function) *For the activation functions, we assume that there exist constants $L_1, L_2, L_3 > 0$ such that, for all $x \in \mathbb{R}$,*

$$|h_1(x)| \leq L_1, \quad |h_2(x)| \leq L_1, \quad \left|h_1'(x)\right| \leq L_2, \quad \left|h_2'(x)\right| \leq L_2.$$

Moreover, for all $x, y \in \mathbb{R}$,

$$\left|h_1'(x) - h_1'(y)\right| \leq L_3|x-y|, \quad \left|h_2'(x) - h_2'(y)\right| \leq L_3|x-y|.$$

For the loss function, we assume that there exist constants $L_4, L_5 > 0$ such that, for all $y \in \mathcal{Y}$, $x_1 \in \mathbb{R}$, and $x_2 \in \mathbb{R}$,

$$\left|\phi_1'(x_1, y)\right| \leq L_4, \qquad \left|\phi_1'(x_1, y) - \phi_1'(x_2, y)\right| \leq L_5|x_1 - x_2|.$$

Assumption 1 is easy to be satisfied. It only requires some boundedness, continuity, and smoothness for the activation and loss functions. It is adopted in most mean-field analysis, such as Mei et al. (2018); Araújo et al. (2019).

Assumption 2 (Strong Universal Approximation Property) *Assume that for any function $f_2 : \mathbb{R}^d \to \mathbb{R}^N$ that is bounded by C_B, i.e., for all $\boldsymbol{v}_1 \in \mathbb{R}^d$, $\|f_2(\boldsymbol{v}_1)\|_\infty \leq C_B$, we have*

$$\lambda_{\min}\left[\int \left[\dot{h}_1\left(\frac{1}{d}\boldsymbol{X}\boldsymbol{v}_1 + f_2\left(\boldsymbol{v}_1\right)\right)\right]\left[\dot{h}_1\left(\frac{1}{d}\boldsymbol{X}\boldsymbol{v}_1 + f_2\left(\boldsymbol{v}_1\right)\right)\right]^\top dp_1\left(\boldsymbol{v}_1\right)\right] \geq \bar{\lambda} > 0. \tag{14}$$

where $\bar{\lambda}$ only depends on $\boldsymbol{X}$, C_B, and h_1, and $p_1 = \mathcal{N}\left(\boldsymbol{0}^d, \boldsymbol{I}^d\right)$.

Assumption 2 is a technical assumption that we conjecture to hold under fairly general conditions. Notably when $C_B = 0$, it is shown in Du et al. (2019a, Lemma F.1) that the assumption holds for all analytic non-polynomial h_1. Lemma 3 affords many examples that satisfy the assumption for constant C_B.

Lemma 3 *Suppose that the data is non-parallel, i.e., $\boldsymbol{x}_i \notin \mathrm{Span}(\boldsymbol{x}_j)$ for all $i \neq j$.*

(i) If $g : \mathbb{R} \to \mathbb{R}$ is a non-polynomial function that is bounded and has Lipschitz continuous gradient, then $h_1(x) := g(cx)$ satisfies Assumption 2 when $c > 0$ is sufficiently small.

(ii) The Relu-type function $h_1(x) = (x)_+^\alpha$ for $\alpha > 0$ satisfies Assumption 2.

(iii) If $h_1(x) = c|x|^{-\alpha}$ or $h_1(x) = c(x)_+^{-\alpha}$ for $|x| > c'$, where $c, c', \alpha > 0$, then h_1 satisfies Assumption 2.

The condition in Lemma 3 is standard and widely used in the analysis of lazing training (see Du et al. (2019a, Lemma F.1)). It only requires the data are not parallel: for every $j \neq i$, $\boldsymbol{x}_j \neq c\boldsymbol{x}_i$ for any scalar c. In the following, we propose the conditions for the initial continuous Res-Net $(\{v_\ell\}_{\ell=2}^{L+1}, p)$. In the next section, we will show concrete examples that realize these assumptions.

Assumption 3 (Initialization for Continuous Res-Net) *We first assume that the initial continuous Res-Net $(\{v_\ell\}_{\ell=2}^{L+1}, p)$ is a feasible continuous Res-Net that satisfies the forward propagation constraints, i.e., (7) – (10). Moreover, p is σ-sub-gaussian distribution and has a full support[3]. For all $\ell \in [2:L]$, $v_\ell(\cdot,\cdot)$ has sublinear growth on the second argument, that is, there is a constant C_1 such that*

$$\left|v_\ell\left(\boldsymbol{\Theta}, \bar{\boldsymbol{\Theta}}\right)\right| \;\leq\; C_1\left(1 + \left\|\bar{\boldsymbol{\Theta}}\right\|_\infty\right), \quad \text{for all } \boldsymbol{\Theta}, \bar{\boldsymbol{\Theta}} \in \operatorname{supp}(p),\ \ell \in [2:L]. \tag{15}$$

Besides, $v_\ell(\cdot,\cdot)$ are locally Lipschitz continuous where the Lipschitz constant has sub-linear growth on the second argument. In detail, there is a constant C_2, such that for $\boldsymbol{\Theta}_1 \in \operatorname{supp}(p)$, $\tilde{\boldsymbol{\Theta}}_1 \in \operatorname{supp}(p) \cap \mathcal{B}_\infty(\boldsymbol{\Theta}_1, 1)$, $\boldsymbol{\Theta}_2 \in \operatorname{supp}(p)$, and $\tilde{\boldsymbol{\Theta}}_2 \in \operatorname{supp}(p) \cap \mathcal{B}_\infty(\boldsymbol{\Theta}_2, 1)$, we have

$$\left|v_\ell(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2) - v_\ell(\tilde{\boldsymbol{\Theta}}_1, \tilde{\boldsymbol{\Theta}}_2)\right| \;\leq\; C_2\left(1 + \|\boldsymbol{\Theta}_2\|_\infty\right)\left(\left\|\boldsymbol{\Theta}_1 - \tilde{\boldsymbol{\Theta}}_1\right\|_\infty + \left\|\boldsymbol{\Theta}_2 - \tilde{\boldsymbol{\Theta}}_2\right\|_\infty\right).$$

Finally, for the last layer, there exist constants C_3 and C_4, such that for all $\boldsymbol{\Theta}, \bar{\boldsymbol{\Theta}} \in \operatorname{supp}(p)$, we have

$$\left|v_{L+1}(\boldsymbol{\Theta})\right| \leq C_3 \quad \text{and} \quad \left|v_{L+1}(\boldsymbol{\Theta}) - v_{L+1}(\bar{\boldsymbol{\Theta}})\right| \;\leq\; C_4\left\|\boldsymbol{\Theta} - \bar{\boldsymbol{\Theta}}\right\|_\infty.$$

5.2. Property and Approximation of Neural Feature Flow

We analyze the neural feature flow for the continuous Res-Net under the initial conditions in Assumption 3. The following theorem guarantees the existence and uniqueness.

Theorem 4 (Existence and Uniqueness of Neural Feature Flow on Res-Net) *Under Assumptions 1 and 3, for any $T < \infty$, there exists an unique neural feature flow Φ_*.*

In fact, we also have that Φ_* is a continuous mapping on $\boldsymbol{\Theta}$ given time t (see Theorem 13 in Appendix C.1). A similar continuity argument has also been observed by Nguyen and Pham (2020). The proofs of Theorems 4 and 13 can be obtained by the technique of Picard iterations (see, e.g., Hartman (1964)) with a special consideration on the search space to deal with the unboundedness of parameters. The latter differs from the former by introducing a more restrictive space in which all the candidates satisfy the desired property. We defer the proofs of this paper to Appendix.

Now we consider the approximation between a discrete DNN trained by scaled Gradient Descent and a continuous one evolving as neural feature flow. Based on the initial condition for the continuous Res-Net, we introduce the initial condition for discrete Res-Net.

Definition 5 (ε_1-independent Initial Discrete Res-Net) *We say an initial discrete Res-Net $(\hat{\boldsymbol{v}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ is ε_1-independent if there exists a continuous initial Res-Net $(\{v_\ell\}_{\ell=2}^{L+1}, p)$ satisfying Assumption 3 and $(\bar{\boldsymbol{v}}, \bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\beta}})$ such that*

3. The assumption that p has a full support will only be used in Theorem 8. It can be replaced by a slightly weaker assumption that there exists a continuous function $f_1 : \mathbb{R}^d \to \mathbb{R}^{D-d}$ such that $\operatorname{supp}(p) \supseteq \{(\boldsymbol{v}_1, f_1(\boldsymbol{v}_1)) : \boldsymbol{v}_1 \in \mathbb{R}^d\}$. When p_1 has a full support, f_1 can be simply chosen as $f_1(\boldsymbol{v}_1) \equiv \mathbf{0}^{D-d}$.

(1) $\bar{\boldsymbol{\Theta}}_i = (\bar{\boldsymbol{v}}_{1,i}, \bar{\boldsymbol{\alpha}}_{2,i}, \ldots, \bar{\boldsymbol{\alpha}}_{L,i}) \overset{i.i.d.}{\sim} p$;

(2) For $\bar{\boldsymbol{\beta}}$ *and* $\bar{\boldsymbol{v}}$,

- $\bar{\boldsymbol{\beta}}_{\ell,i} = \frac{1}{d}(\boldsymbol{X}\bar{\boldsymbol{v}}_{1,i}) + \sum_{\ell_1=2}^{\ell} \dot{h}_2(\bar{\boldsymbol{\alpha}}_{\ell_1,i})$ *for* $\ell \in [L]$ *and* $i \in [m]$;
- $\bar{v}_{\ell,i,j} = v_\ell\left(\bar{\boldsymbol{\Theta}}_i, \bar{\boldsymbol{\Theta}}_j\right)$ *for* $\ell \in [2:L]$, $i, j \in [m]$;
- $\bar{v}_{L+1,i,1} = v_{L+1}\left(\bar{\boldsymbol{\Theta}}_i\right)$ *for* $i \in [m]$;

(3) ε_1*-closeness:*

- $\|\bar{\boldsymbol{v}}_{1,i} - \hat{\boldsymbol{v}}_{1,i}\|_\infty \le \left(1 + \left\|\bar{\boldsymbol{\Theta}}_i\right\|_\infty\right)\varepsilon_1$ *for* $i \in [m]$;
- $|\bar{v}_{\ell+1,i,j} - \hat{v}_{\ell+1,i,j}| \le \left(1 + \left\|\bar{\boldsymbol{\Theta}}_i\right\|_\infty + \left\|\bar{\boldsymbol{\Theta}}_j\right\|_\infty\right)\varepsilon_1$ *for* $\ell \in [L-1]$, $i, j \in [m]$;
- $|\bar{v}_{L+1,i,1} - \hat{v}_{L+1,i,1}| \le \left(1 + \left\|\bar{\boldsymbol{\Theta}}_i\right\|_\infty\right)\varepsilon_1$ *for* $i \in [m]$.

As an ε_1-independent initialization relates to a continuous Res-Net satisfying Assumption 3, which yields an unique neural feature flow Φ_* by Theorem 4, we show that scaled Gradient Descent from an ε_1-independent initialization is well-approximated by the corresponding neural feature flow when the number of hidden units is $\tilde{\Omega}(\varepsilon_1^{-2})$, where $\tilde{\Omega}$ hides poly-logarithmic factors. This resembles a "propagation of chaos" argument from a Kac's chaotic initial system (Sznitman, 1991). We compare the scaled Gradient Descent with an ideal discrete process determined by Φ_* as specified below:

- Actual process $(\hat{\boldsymbol{v}}^{[0:K]}, \hat{\boldsymbol{\alpha}}^{[0:K]}, \hat{\boldsymbol{\beta}}^{[0:K]})$ by executing Algorithm 1 in $K = \frac{T}{\eta}$ steps on the discrete Res-Net from $(\hat{\boldsymbol{v}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$;
- Ideal process $\left(\bar{\boldsymbol{v}}^{[0,T]}, \bar{\boldsymbol{\alpha}}^{[0,T]}, \bar{\boldsymbol{\beta}}^{[0,T]}\right)$that evolves as neural feature flow:

$$\begin{aligned}
\bar{\boldsymbol{\beta}}_{\ell,i}^t &= \Phi_{*,\ell}^{\boldsymbol{\beta}}\left(\bar{\boldsymbol{\Theta}}_i\right)(t), \quad \ell \in [L],\ i \in [m],\ t \in [0,T],\\
\bar{\boldsymbol{\alpha}}_{\ell,i}^t &= \Phi_{*,\ell}^{\boldsymbol{\alpha}}\left(\bar{\boldsymbol{\Theta}}_i\right)(t), \quad \ell \in [2:L],\ i \in [m],\ t \in [0,T],\\
\bar{\boldsymbol{v}}_{1,i}^t &= \Phi_{*,1}^{\boldsymbol{v}}\left(\bar{\boldsymbol{\Theta}}_i\right)(t), \quad i \in [m],\ t \in [0,T],\\
\bar{v}_{\ell,i,j}^t &= \Phi_{*,\ell}^{\boldsymbol{v}}\left(\bar{\boldsymbol{\Theta}}_i, \bar{\boldsymbol{\Theta}}_j\right)(t), \quad \ell \in [2:L],\ i \in [m], i \in [m],\ t \in [0,T],\\
\bar{v}_{L+1,i,1}^t &= \Phi_{*,L+1}^{\boldsymbol{v}}\left(\bar{\boldsymbol{\Theta}}_i\right)(t), \quad i \in [m],\ t \in [0,T].
\end{aligned}$$

We also compare the discrete and the continuous losses denoted by $\hat{\mathcal{L}}^k := \frac{1}{n}\sum_{n=1}^N \phi(\hat{\boldsymbol{\beta}}_{L+1,1}^k(n), y^n)$ and $\mathcal{L}^t := \frac{1}{N}\sum_{n=1}^N \phi\left(\boldsymbol{\beta}_{L+1}(\Phi_*, t)(n), y^n\right)$, respectively. We have the theorem below.

Theorem 6 *Under Assumption 1, suppose* $\varepsilon_1 \le \mathcal{O}(1)$ *and* $m \ge \tilde{\Omega}(\varepsilon_1^{-2})$*, and treat the parameters in assumptions and* T *as constants. Consider the actual process from an* ε_1*-independent initialization in Definition 5 with step size* $\eta \le \tilde{\mathcal{O}}(\varepsilon_1)$*. Then, the following holds with probability* $1-\delta$*:*

(1) The two processes are close to each other:

$$\sup_{k\in[0:K]}\left\{\sup_{i\in[m]}\left\|\hat{\boldsymbol{v}}_{1,i}^k - \bar{\boldsymbol{v}}_{1,i}^{k\eta}\right\|_\infty, \sup_{\ell\in[2:L],\ i,j\in[m]}\left|\hat{v}_{\ell,i,j}^k - \bar{v}_{\ell,i,j}^{k\eta}\right|\right\} \le \tilde{\mathcal{O}}(\varepsilon_1),$$

$$\sup_{k\in[0:K],\ i\in[m]}\left\{\left|\hat{v}_{L+1,i,1}^k - \bar{v}_{L+1,i,1}^{k\eta}\right|, \sup_{\ell\in[2:L]}\left\|\hat{\boldsymbol{\alpha}}_{\ell,i}^k - \bar{\boldsymbol{\alpha}}_{\ell,i}^{k\eta}\right\|_\infty, \sup_{\ell\in[L]}\left\|\hat{\boldsymbol{\beta}}_{\ell,i}^k - \bar{\boldsymbol{\beta}}_{\ell,i}^{k\eta}\right\|_\infty\right\} \le \tilde{\mathcal{O}}(\varepsilon_1).$$

Algorithm 2 Initialize a Discrete Res-Net.

Input the data $\boldsymbol{X}$, variance σ_1, and a constant C_3.
Independently draw $\hat{v}_{1,i,j} \sim p_0 = \mathcal{N}\left(0, d\sigma_1^2\right)$ for $i \in [d]$ and $j \in [m]$.
Set $\hat{\boldsymbol{\beta}}_{1,j} = \frac{1}{d}\sum_{i=1}^{d} \hat{v}_{1,i,j}\, \hat{\boldsymbol{\beta}}_{0,i}$ where $j \in [m]$. ◇ Standard Initialization for layer 1
for $\ell = 2, \ldots, L$ **do**
  Independently draw $\tilde{v}_{\ell,i,j} \sim \mathcal{N}\left(0, m\sigma_1^2\right)$ for $i, j \in [m]$.
  Set $\hat{\boldsymbol{\alpha}}_{\ell,j} = \frac{1}{m}\sum_{i=1}^{m} \tilde{v}_{\ell,i,j}\, \dot{h}_1(\hat{\boldsymbol{\beta}}_{\ell-1,i})$ where $j \in [m]$.
  Set $\hat{\boldsymbol{\beta}}_{\ell,j} = \hat{\boldsymbol{\beta}}_{\ell-1,j} + \dot{h}_2\left(\hat{\boldsymbol{\alpha}}_{\ell,j}\right)$ for $j \in [m]$. ◇ Standard Initialization for layer ℓ
end for
Set $\hat{v}_{L+1,i,1} = C_3$ where $i \in [m]$. ◇ Simply initialize $\{\hat{v}_{L+1,i,1}\}_{i=1}^{m}$ by a constant
for $\ell = 2, \ldots, L$ **do**
  for $j = 1, \ldots, m$ **do**
    Solve convex optimization problem: ◇ Perform ℓ_2-regression to reduce redundancy

$$\min_{\{\hat{v}_{\ell,i,j}\}_{i=1}^{m}} \frac{1}{m}\sum_{i=1}^{m} \left(\hat{v}_{\ell,i,j}\right)^2, \quad \text{s.t. } \hat{\boldsymbol{\alpha}}_{\ell,j} = \frac{1}{m}\sum_{i=1}^{m} \hat{v}_{\ell,i,j}\, \dot{h}_1(\hat{\boldsymbol{\beta}}_{\ell-1,i}). \tag{16}$$

  end for
end for
Output the discrete Res-Net parameters $(\hat{\boldsymbol{v}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$.

(2) The training losses are also close to each other:

$$\sup_{k\in[0:K]} \left| \hat{\mathcal{L}}^k - \mathcal{L}^{k\eta} \right| \le \tilde{\mathcal{O}}(\varepsilon_1).$$

Note in the discrete Res-Net, even though the connecting weights are independently initialized, $\hat{\boldsymbol{\alpha}}_{\ell,j}$ are not mutually independent since they all depend on a *common* set of random outputs from the previous layer. Therefore, Definition 5 restricts the skip-connected paths of the discrete Res-Net $\{\hat{\boldsymbol{v}}_{1,i}, \hat{\boldsymbol{\alpha}}_{2,i}, \ldots, \hat{\boldsymbol{\alpha}}_{L,i}\}_{i=1}^{m}$ are nearly independent, which makes it possible to construct an ideal initialization with independent skip-connected paths to approximate the discrete one. Then, using a "propagation of chaos" argument (Sznitman, 1991), we obtain Theorem 6.

Now we demonstrate real examples that can achieve our assumptions. We consider the Res-Net initialized by Algorithm 2, which is composed of a standard initialization (Glorot and Bengio, 2010; He et al., 2015) and an additional regression procedure while preserving all initial features.[4] The standard initialization strategy scales the weights to $\sqrt{m}$, which diverges in the mean-field limit. We perform the simple ℓ_2-regression to reduce the redundancy of the weights. The result in Theorem 7 shows that Algorithm 2 can produce an ε_1-independent initialization when m is sufficiently large.

Theorem 7 *Under Assumptions 1 and 2, treat the parameters in assumptions as constants. With probability at least $1-\delta$, Algorithm 2 produces an ε_1-independent initial discrete Res-Net with $\varepsilon_1 \le \tilde{\mathcal{O}}(\frac{1}{\sqrt{m}})$.*

4. In Algorithm 2, the weights in the last layer $\{v_{L+1,i,1}\}_{i=1}^{m}$ can also be initialized by the standard initialization followed by an ℓ_2-regression. The ℓ_2-regression (16) can be replaced by a soft version

$$\min_{\{\hat{v}_{\ell,i,j}\}_{i=1}^{m}} \frac{\lambda_m}{m}\sum_{i=1}^{m} \left(\hat{v}_{\ell,i,j}\right)^2 + \left\| \hat{\boldsymbol{\alpha}}_{\ell,j} - \frac{1}{m}\sum_{i=1}^{m} \hat{v}_{\ell,i,j}\, \dot{h}_1(\hat{\boldsymbol{\beta}}_{\ell-1,i}) \right\|^2.$$

5.3. Finding Global Minimum

We study the convergence of neural feature flow. We show in Theorem 8 that the neural feature flow always finds a globally optimal solution when it converges.

Theorem 8 *Under Assumptions 1 and 2, assume that the loss function ϕ is convex in the first argument. Let $\left(\{v_\ell\}_{\ell=2}^{L+1}, p\right)$ be the initial continuous Res-Net that satisfies Assumption 3 and Φ_* and $\mathcal{L}^t$ be the solution and loss of the neural feature flow, respectively. If $\Phi^{\boldsymbol{\beta}}_{*,L}(\boldsymbol{\Theta})(t)$ converges in $\ell_\infty(p)$ and $\Phi^{\boldsymbol{v}}_{*,L+1}(\boldsymbol{\Theta})(t)$ converges in $\ell_1(p)$ as $t \to \infty$, where $\boldsymbol{\Theta} \sim p$, then we have*

$$\lim_{t\to\infty} \mathcal{L}^t = \sum_{n=1}^{N} \left[\min_{y'} \phi\left(y', y^n\right)\right].$$

Theorem 8 is an important application of our mean-field framework, which shows that neural feature flow can find a global minimizer after it converges[5]. We prove that the distribution of the weights in the first layer always have a full support in any finite time by Brouwer's fixed-point theorem. Then, using a similar argument to Chizat and Bach (2018); Nguyen and Pham (2020), we show that all bad local minima are unstable. Our global convergence holds for Res-Nets with arbitrary (finite) depth. Before us, the global optimality was proved for three-layer ones (Nguyen and Pham, 2020) under a similar convergence assumption on the weights in the second layer.

6. Discussion

This paper proposed a new mean-field framework for DNNs where features in hidden layers have non-vanishing variance. We constructed a continuous dynamic called neural feature flow that captures the evolution of sufficiently over-parametrized Res-Nets trained by Gradient Descent. Furthermore, the neural feature flow reaches a globally optimal solution after it converges. We hope our new analytical tool pioneers better understandings for DNN training.

There are many interesting questions under our framework to be further investigated. First, the current work only focuses on the training part of NNs, and it still remains to study generalization. The generalization error in the mean-field regime has been studied for two-level NNs, e.g., Wei et al. (2018). Using our modeling, there are two potential directions: we may incorporate suitable regularizers on the DNNs to control the model complexity; implicit regularization is often observed in practice, which is hopefully preserved in our neural feature flow. However, a full treatment is left as a future study. Second, although our mean-field framework is applicable to standard DNNs (see Appendix E), it is still not answered how to prove that Gradient Descent can achieve the global minimum. Third, the implications of our theory in practice can be studied empirically as separate works. For example, this paper proposes a new interesting initialization strategy for DNNs and uses the scaled Gradient Descent to optimize DNNs. In Appendix F, we perform a toy simulation to validate the feasibility of this new training strategy. It is interesting to design new practical algorithms based on our learning strategy for large-scale data.

5. Remark: for ℓ_2 loss, i.e. $\phi(y', y) = \|y' - y\|^2$, as an example, Theorem 8 indicates converging to 0 loss.

References

Zeyuan Allen-Zhu and Yuanzhi Li. Can sgd learn recurrent neural networks with provable generalization? *arXiv:1902.01028*, 2019.

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *arXiv:1811.04918*, 2018.

Dyego Araújo, Roberto I Oliveira, and Daniel Yukimura. A mean-field limit for certain deep neural networks. *arXiv:1906.00193*, 2019.

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, 2019.

Rajendra Bhatia. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.

Zixiang Chen, Yuan Cao, Quanquan Gu, and Tong Zhang. Mean-field analysis of two-layer neural networks: Non-asymptotic rates and generalization bounds. *arXiv:2002.04026*, 2020.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in neural information processing systems*, pages 3036–3046, 2018.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, pages 2933–2943, 2019.

Donald A Dawson. Multilevel mutation-selection systems and set-valued duals. *Journal of mathematical biology*, 76(1-2):295–378, 2018.

Donald A Dawson, Kenneth J Hochberg, et al. Wandering random measures in the fleming-viot model. *The Annals of Probability*, 10(3):554–580, 1982.

Xialiang Dou and Tengyuan Liang. Training neural networks as learning data-adaptive kernels: Provable representation and approximation benefits. *arXiv:1901.07114*, 2019.

Simon S Du, Jason D Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, 2019a.

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

Cong Fang, Hanze Dong, and Tong Zhang. Over parameterized two-level neural networks can learn nearoptimal feature representations. *arXiv:1910.11508*, 2019.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.

Andrzej Granas and James Dugundji. *Fixed point theory*. Springer Science & Business Media, 2013.

Moritz Hardt and Tengyu Ma. Identity matters in deep learning. In *International Conference on Learning Representation*, 2016.

Philip Hartman. *Ordinary differential equations*. Wiley, 1964.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, 2018.

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, 2018.

Yiping Lu, Chao Ma, Yulong Lu, Jianfeng Lu, and Lexing Ying. A mean-field analysis of deep resnet and beyond: Towards provable optimization via overparameterization from depth. In *International Conference on Machine Learning*, 2020.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018. ISSN 0027-8424.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Annual Conference on Learning Theory*, 2019.

Phan-Minh Nguyen and Huy Tuan Pham. A rigorous framework for the mean field limit of multilayer neural networks. *arXiv:2001.11443*, 2020.

Huy Tuan Pham and Phan-Minh Nguyen. A note on the global convergence of multilayer neural networks in the mean field regime. *arXiv preprint arXiv:2006.09355*, 2020.

Grant M Rotskoff and Eric Vanden-Eijnden. Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. *arXiv:1805.00915*, 2018.

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 2019a.

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of deep neural networks. *arXiv:1903.04440*, 2019b.

Alain-Sol Sznitman. Topics in propagation of chaos. In *Ecole d'été de probabilités de Saint-Flour XIX—1989*, pages 165–251. Springer, 1991.

Joel A Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

Mihai Turinici. Abstract comparison principles and multivariable gronwall-bellman inequalities. *Journal of Mathematical Analysis and Applications*, 117(1):100–127, 1986.

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

Colin Wei, Jason D Lee, Qiang Liu, and Tengyu Ma. Regularization matters: Generalization and optimization of neural nets v.s. their induced kernel. *arXiv:1810.05369*, 2018.

Gilad Yehudai and Ohad Shamir. On the power and limitations of random features for understanding neural networks. In *Advances in Neural Information Processing Systems*, pages 6594–6604, 2019.

Hongyi Zhang, Yann N Dauphin, and Tengyu Ma. Fixup initialization: Residual learning without normalization. *arXiv:1901.09321*, 2019.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Stochastic gradient descent optimizes over-parameterized deep relu networks. In *Advances in neural information processing systems*, 2018.

Appendix A. Overview

The appendix is sketched as follows. Appendix B provides the key proofs of this paper. Especially, we will show how the initial condition (Theorem 7) and Assumption 2 (Lemma 3) can be realized. Besides, we will prove Theorem 8. Appendix C presents the rest proofs. Especially, we will follow the technique of Picard iterations to prove Theorem 4 and a continuity argument for neural feature flow (Theorem 13). Note that the latter will be used in Theorem 8. Moreover, we will follow a "propagation of chaos" argument (Sznitman, 1991) to prove Theorem 6. Appendix D presents some basic properties for sub-gaussian distributions. Finally, Appendix E introduces the extension of our mean-field framework to fully-connected DNNs.

In our proofs, we fix a set of training data and treat the parameters in the assumptions as constants. We use C to denote a generic constant; the value of C may change from line to line.

Appendix B. Key Proofs

B.1. Proof of Theorem 7

In this subsection, we prove Theorem 7 which states that Algorithm 2 produces an ε_1-independent initial discrete Res-Net with $\varepsilon_1 \leq \tilde{\mathcal{O}}(1/\sqrt{m})$. By Definition 5, this entails the construction of an initial distribution p, weight functions $\{v_\ell\}_{\ell=2}^{L+1}$, and an ideal discrete Res-Net satisfying the properties in Definition 5. We specify p, $\{v_\ell\}_{\ell=2}^{L+1}$, and the ideal discrete Res-Net below, and then we verify the properties in Theorem 7.

Initial distribution. We first define the distribution p:

(1) At the first layer, $\boldsymbol{\beta}_1 \sim p_1^{\boldsymbol{\beta}} = \mathcal{N}\left(\mathbf{0}^N, \sigma_1^2 \boldsymbol{K}_0\right)$, where $\boldsymbol{K}_0 := \frac{1}{d}\boldsymbol{X}\boldsymbol{X}^\top$.

(2) At the layer $\ell \in [L-1]$, let $\boldsymbol{K}_\ell^{\boldsymbol{\beta}} := \int \dot{h}_1\left(\boldsymbol{\beta}_\ell\right) \dot{h}_1\left(\boldsymbol{\beta}_\ell\right)^\top dp_\ell^{\boldsymbol{\beta}}\left(\boldsymbol{\beta}_\ell\right)$. We define the distribution of the residuals at layer $\ell+1$ as

$$p_{\ell+1}^{\boldsymbol{\alpha}} = \mathcal{N}\left(\mathbf{0}^N, \sigma_1^2 \boldsymbol{K}_\ell^{\boldsymbol{\beta}}\right). \tag{17}$$

Defining the mapping $\tilde{f}_{\ell+1}\left(\boldsymbol{\beta}_\ell, \boldsymbol{\alpha}_{\ell+1}\right) := \boldsymbol{\beta}_\ell + \dot{h}_2(\boldsymbol{\alpha}_{\ell+1})$, the features at layer $\ell+1$ is defined as the pushforward measure by $\tilde{f}_{\ell+1}$:

$$p_{\ell+1}^{\boldsymbol{\beta}} = \tilde{f}_{\ell+1} \# \left(p_\ell^{\boldsymbol{\beta}} \times p_{\ell+1}^{\boldsymbol{\alpha}}\right).$$

Finally, let p be a multivariate Gaussian distribution of the form

$$p\left(\boldsymbol{v}_1, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_L\right) := p_1^{\boldsymbol{v}}(\boldsymbol{v}_1) \times p_2^{\boldsymbol{\alpha}}(\boldsymbol{\alpha}_2) \times p_3^{\boldsymbol{\alpha}}(\boldsymbol{\alpha}_3) \times \cdots \times p_L^{\boldsymbol{\alpha}}(\boldsymbol{\alpha}_L). \tag{18}$$

Weight functions. Now we define the weight functions $\{v_\ell\}_{\ell=2}^{L+1}$. Note that those gram matrices $\boldsymbol{K}_\ell^{\boldsymbol{\beta}}$ are all positive definite under Assumption 2 (see Lemma 9) and thus are invertible. For $\boldsymbol{\Theta} = (\boldsymbol{v}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_L)$, $\boldsymbol{\Theta}' = (\boldsymbol{v}_1', \boldsymbol{\alpha}_2', \ldots, \boldsymbol{\alpha}_L')$, we define the connecting weights between consecutive layers by

$$v_\ell(\boldsymbol{\Theta}, \boldsymbol{\Theta}') = \dot{h}_1\left(\boldsymbol{\beta}_{\ell-1}(\boldsymbol{\Theta})\right)^\top \left[\boldsymbol{K}_{\ell-1}^{\boldsymbol{\beta}}\right]^{-1} \boldsymbol{\alpha}_\ell', \quad \ell \in [2:L], \tag{19}$$

where $\boldsymbol{\beta}_\ell(\boldsymbol{\Theta}) = \frac{1}{d}\boldsymbol{X}\boldsymbol{v}_1 + \sum_{i=2}^{\ell} \dot{h}_2(\boldsymbol{\alpha}_i)$ will satisfy (9). The weights at the output layer are initialized as a constant c. The forward propagation constraint (8) will also be satisfied by (19) and the definitions of $\boldsymbol{K}_\ell^{\boldsymbol{\beta}}$. Therefore $(\{v_\ell\}_{\ell=2}^{L+1}, p)$ constitutes a feasible continuous Res-Net.

Ideal discrete Res-Net. Finally we construct the initialization $(\bar{\boldsymbol{v}}, \bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\beta}})$ of the ideal discrete Res-Net. Recall Algorithm 2 that the corresponding variables are initialized as $(\hat{\boldsymbol{v}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$. Let $\bar{\boldsymbol{v}}_{1,i} := \hat{\boldsymbol{v}}_{i,1}$ for $i \in [m]$. For $\ell \in [L-1]$, define the empirical Gram matrix as

$$\hat{\boldsymbol{K}}_\ell^{\boldsymbol{\beta}} = \frac{1}{m}\sum_{i=1}^{m} \dot{h}_1\left(\hat{\boldsymbol{\beta}}_{\ell,i}\right) \dot{h}_1^\top\left(\hat{\boldsymbol{\beta}}_{\ell,i}\right).$$

Let $\bar{\boldsymbol{\alpha}}_{\ell+1,j} := \left(\boldsymbol{K}_\ell^{\boldsymbol{\beta}}\right)^{1/2} \left(\hat{\boldsymbol{K}}_\ell^{\boldsymbol{\beta}}\right)^{-1/2} \hat{\boldsymbol{\alpha}}_{\ell+1,j}$ for all $j \in [m]$ when $\hat{\boldsymbol{K}}_\ell^{\boldsymbol{\beta}}$ is invertible, and otherwise let $\bar{\boldsymbol{\alpha}}_{\ell+1,j} \overset{\text{i.i.d.}}{\sim} p_{\ell+1}^{\boldsymbol{\alpha}}$. We use Definition 5 (2) for the values of $\bar{\boldsymbol{\beta}}_\ell$ for $\ell \in [L]$ and $\bar{v}_\ell$ for all in $[2:L+1]$.

Proof [Proof of Theorem 7] We first show that the continuous Res-Net satisfies Assumptions 3. By definition p is a multivariate Gaussian distribution. By Lemma 9, we have $\left\|\left(\boldsymbol{K}_\ell^{\boldsymbol{\beta}}\right)^{-1}\right\|_2 \le C$ for a constant C. Since the activation function h_1 is bounded and Lipschitz continuous, the continuity conditions in Assumption 3 are all satisfied.

Now we consider the ideal discrete Res-Net. We first verify the independence. By definition, $\bar{\boldsymbol{\alpha}}_{\ell+1,j}$ are determined by the outputs of previous layer $\hat{\boldsymbol{\beta}}_{\ell,i}$ and the connecting weights $\hat{v}_{\ell+1,i,j}$.

Thus they are conditionally independent of $\bar{\boldsymbol{v}}_{1,i}$ and $\bar{\boldsymbol{\alpha}}_{2,i}, \ldots, \bar{\boldsymbol{\alpha}}_{\ell,i}$ for $i \in [m]$ given $\{\hat{\boldsymbol{\beta}}_{\ell,i}\}_{i\in[m]}$. Since $\hat{v}_{\ell+1,i,j}$ are independent Gaussian, $\hat{\boldsymbol{\alpha}}_{\ell+1,j}$ and thus $\bar{\boldsymbol{\alpha}}_{\ell+1,j}$ are conditionally independent Gaussian given $\{\hat{\boldsymbol{\beta}}_{\ell,i}\}_{i\in[m]}$. Furthermore, the conditional distribution of $\bar{\boldsymbol{\alpha}}_{\ell+1,j}$ given $\{\hat{\boldsymbol{\beta}}_{\ell,i}\}_{i\in[m]}$ is $\mathcal{N}\left(\mathbf{0}^N, \sigma_1^2 \boldsymbol{K}_\ell^{\boldsymbol{\beta}}\right) = p_{\ell+1}^{\boldsymbol{\alpha}}$. Therefore, marginally $\bar{\boldsymbol{\alpha}}_{\ell+1,j} \overset{\text{i.i.d.}}{\sim} p_{\ell+1}^{\boldsymbol{\alpha}}$ and they are independent of $\bar{\boldsymbol{v}}_{1,i}$ and $\bar{\boldsymbol{\alpha}}_{2,i}, \ldots, \bar{\boldsymbol{\alpha}}_{\ell,i}$ for $i \in [m]$. So $\{\bar{\boldsymbol{\Theta}}_i\}_{i\in[m]} \overset{\text{i.i.d.}}{\sim} p$. Since p is a product distribution, all $\bar{\boldsymbol{v}}_{1,i}, \bar{\boldsymbol{\alpha}}_{2,i}, \ldots, \bar{\boldsymbol{\alpha}}_{L,i}$ are all mutually independent.

Lastly we show the $\tilde{\mathcal{O}}(1/\sqrt{m})$-closeness specified in Definition 5 (3). By Lemma 10 we have the following events with probability $1-\delta$:

$$\|\bar{\boldsymbol{\alpha}}_{\ell+1,i}\|_2 \le \tilde{\mathcal{O}}(1), \tag{20}$$

$$\left\|\hat{\boldsymbol{K}}_\ell^{\boldsymbol{\beta}} - \boldsymbol{K}_\ell^{\boldsymbol{\beta}}\right\|_2 \le \varepsilon_2, \tag{21}$$

$$\|\bar{\boldsymbol{\alpha}}_{\ell+1,i} - \hat{\boldsymbol{\alpha}}_{\ell+1,i}\|_2 \le \varepsilon_2 \left\|\bar{\boldsymbol{\Theta}}_i\right\|_2, \tag{22}$$

$$\left\|\bar{\boldsymbol{\beta}}_{\ell+1,i} - \hat{\boldsymbol{\beta}}_{\ell+1,i}\right\|_2 \le \varepsilon_2 \left\|\bar{\boldsymbol{\Theta}}_i\right\|_2, \tag{23}$$

where $\varepsilon_2 \le \tilde{\mathcal{O}}(1/\sqrt{m})$. Under (21) the matrix $\hat{\boldsymbol{K}}_\ell^{\boldsymbol{\beta}}$ is invertible, and it follows from Lemma 11 that

$$\hat{v}_{\ell+1,i,j} = \hat{\boldsymbol{\alpha}}_{\ell+1,j}^\top \left[\hat{\boldsymbol{K}}_\ell^{\boldsymbol{\beta}}\right]^{-1} \dot{h}_1(\hat{\boldsymbol{\beta}}_{\ell,i}), \quad \ell \in [L-1],\ i,j \in [m].$$

By the triangle inequality,

$$\begin{aligned}
&|\bar{v}_{\ell+1,i,j} - \hat{v}_{\ell+1,i,j}| \\
&= \left\|\bar{\boldsymbol{\alpha}}_{\ell+1,j}^\top \left[\boldsymbol{K}_\ell^{\boldsymbol{\beta}}\right]^{-1} \dot{h}_1(\bar{\boldsymbol{\beta}}_{\ell,i}) - \hat{\boldsymbol{\alpha}}_{\ell+1,j}^\top \left[\hat{\boldsymbol{K}}_\ell^{\boldsymbol{\beta}}\right]^{-1} \dot{h}_1(\hat{\boldsymbol{\beta}}_{\ell,i})\right\|_2 \\
&\le \|\bar{\boldsymbol{\alpha}}_{\ell+1,j}\|_2 \left\|\left[\boldsymbol{K}_\ell^{\boldsymbol{\beta}}\right]^{-1}\right\|_2 \left\|\dot{h}_1(\bar{\boldsymbol{\beta}}_{\ell,i}) - \dot{h}_1(\hat{\boldsymbol{\beta}}_{\ell,i})\right\|_2 + \|\bar{\boldsymbol{\alpha}}_{\ell+1,j}\|_2 \left\|\left[\boldsymbol{K}_\ell^{\boldsymbol{\beta}}\right]^{-1} - \left[\hat{\boldsymbol{K}}_\ell^{\boldsymbol{\beta}}\right]^{-1}\right\|_2 \left\|\dot{h}_1(\hat{\boldsymbol{\beta}}_{\ell,i})\right\|_2 \\
&\quad + \|\bar{\boldsymbol{\alpha}}_{\ell+1,j} - \hat{\boldsymbol{\alpha}}_{\ell+1,j}\|_2 \left\|\left[\hat{\boldsymbol{K}}_\ell^{\boldsymbol{\beta}}\right]^{-1}\right\|_2 \left\|\dot{h}_1(\hat{\boldsymbol{\beta}}_{\ell,i})\right\|_2.
\end{aligned}$$

We upper bound three terms separately. By the Lipschitz continuity of h_1 and (23), the first term is at most $\tilde{\mathcal{O}}(1/\sqrt{m})\|\bar{\boldsymbol{\Theta}}_j\|_2$; the second term is also at most $\tilde{\mathcal{O}}(1/\sqrt{m})\|\bar{\boldsymbol{\Theta}}_j\|_2$ since $\|\bar{\boldsymbol{\alpha}}_{\ell+1,j}\|_2 \le \|\bar{\boldsymbol{\Theta}}_j\|_2$,

$$\left\|\left[\boldsymbol{K}_\ell^{\boldsymbol{\beta}}\right]^{-1} - \left[\hat{\boldsymbol{K}}_\ell^{\boldsymbol{\beta}}\right]^{-1}\right\|_2 \le \left\|\left[\boldsymbol{K}_\ell^{\boldsymbol{\beta}}\right]^{-1}\right\|_2 \left\|\boldsymbol{K}_\ell^{\boldsymbol{\beta}} - \hat{\boldsymbol{K}}_\ell^{\boldsymbol{\beta}}\right\|_2 \left\|\left[\hat{\boldsymbol{K}}_\ell^{\boldsymbol{\beta}}\right]^{-1}\right\|_2 \le C\varepsilon_2,$$

and h_1 is bounded; the third term is at most $\tilde{\mathcal{O}}(1/\sqrt{m})\|\bar{\boldsymbol{\Theta}}_j\|_2$ by (22). ■

B.1.1. PROOF OF ADDITIONAL LEMMAS

Lemma 9 $\min_{\ell\in[L-1]} \lambda_{\min}(\boldsymbol{K}_\ell^{\boldsymbol{\beta}}) \ge C > 0$.

Proof Fix $\ell \in [L-1]$. For $(\boldsymbol{v}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_L) \in \mathrm{supp}(p)$, given $\boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_L$, we have $\boldsymbol{\Theta} = \boldsymbol{\Theta}(\boldsymbol{v}_1)$ and

$$\left\| \boldsymbol{\beta}_\ell(\boldsymbol{\Theta}) - \frac{1}{d}\boldsymbol{X}\boldsymbol{v}_1 \right\|_\infty = \left\| \sum_{\ell_1=2}^{\ell} \dot{h}_2(\boldsymbol{\alpha}_{\ell_1}) \right\|_\infty \le LL_1. \tag{24}$$

Note that $\boldsymbol{v}_1$ is independent of $\boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_L$, and $\boldsymbol{v}_1 \sim \mathcal{N}\left(0, d\sigma_1^2 \boldsymbol{I}^d\right)$ which is equivalent to the standard Gaussian distribution. By Assumption 2 with $f_2(\boldsymbol{v}_1) \equiv \sum_{\ell_1=2}^{\ell} \dot{h}_2(\boldsymbol{\alpha}_{\ell_1})$ and the constant $C_B = LL_1$, we have

$$\mathbb{E}\left[\dot{h}_1\left(\boldsymbol{\beta}_\ell(\boldsymbol{\Theta})\right) \dot{h}_1^\top\left(\boldsymbol{\beta}_\ell(\boldsymbol{\Theta})\right) \mid \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_L \right] \succeq C\boldsymbol{I}^N.$$

Taking full expectation, we obtain Lemma 9. ■

In the sequel, we set $\bar{\lambda}_1 := \min_{\ell \in [L-1]} \lambda_{\min}(\boldsymbol{K}_\ell^{\boldsymbol{\beta}})$ that is strictly bounded away from zero.

Lemma 10 *Let $\varepsilon_2 \le \tilde{\mathcal{O}}(1/\sqrt{m})$. With probability $1-\delta$, for all $\ell \in [L-1]$ and $i \in [m]$,*

$$\left\| \hat{\boldsymbol{K}}_\ell^{\boldsymbol{\beta}} - \boldsymbol{K}_\ell^{\boldsymbol{\beta}} \right\|_2 \le \varepsilon_2, \qquad \left\| \bar{\boldsymbol{\alpha}}_{\ell+1,i} - \hat{\boldsymbol{\alpha}}_{\ell+1,i} \right\|_2 \le \varepsilon_2 \left\| \bar{\boldsymbol{\Theta}}_i \right\|_2,$$
$$\left\| \bar{\boldsymbol{\alpha}}_{\ell+1,i} \right\|_2 \le \tilde{\mathcal{O}}(1), \qquad \left\| \bar{\boldsymbol{\beta}}_{\ell+1,i} - \hat{\boldsymbol{\beta}}_{\ell+1,i} \right\|_2 \le \varepsilon_2 \left\| \bar{\boldsymbol{\Theta}}_i \right\|_2.$$

Proof In the proof of Lemma 7 we verified that $\bar{\boldsymbol{v}}_{1,i}, \bar{\boldsymbol{\alpha}}_{2,i}, \ldots, \bar{\boldsymbol{\alpha}}_{L,i}$ for all $i \in [m]$ are independent. Therefore, $\bar{\boldsymbol{\beta}}_{\ell,i} \overset{\text{i.i.d.}}{\sim} p_\ell^{\boldsymbol{\beta}}$ by the definitions of $\bar{\boldsymbol{\beta}}_{\ell,i}$ and $p_\ell^{\boldsymbol{\beta}}$. Consider auxiliary random matrices

$$\bar{\boldsymbol{K}}_\ell^{\boldsymbol{\beta}} := \frac{1}{m}\sum_{i=1}^{m} \dot{h}_1(\bar{\boldsymbol{\beta}}_{\ell,i}) \dot{h}_1^\top(\bar{\boldsymbol{\beta}}_{\ell,i}),$$

Since h_1 is bounded, by the matrix Bernstein inequality (Tropp, 2015), with probability $1 - \frac{\delta}{3}$,

$$\max_{\ell \in [L-1]} \left\| \bar{\boldsymbol{K}}_\ell^{\boldsymbol{\beta}} - \boldsymbol{K}_\ell^{\boldsymbol{\beta}} \right\|_2 \le \varepsilon_3 = \tilde{\mathcal{O}}(1/\sqrt{m}). \tag{25}$$

Due to the sub-gaussianness of p, we have $\|\bar{\boldsymbol{\alpha}}_{\ell+1,i}\|_2 \le C\sqrt{\log(m/\delta)} = \tilde{\mathcal{O}}(1)$ with probability $1-\delta/3$ (see Lemma 27 (1)) We will also use the following upper bound that happens with probability $1-\delta/3$ by the sub-gaussianness of p:

$$\frac{1}{m}\sum_{i=1}^{m} \|\bar{\boldsymbol{\Theta}}_i\|_2 \le \beta_1 = \tilde{\mathcal{O}}(1), \tag{26}$$

which can be obtained by the concentration inequality in Lemma 28).

Next we inductively prove that, for $\ell \in [L-1]$,

$$\left\| \boldsymbol{K}_\ell^{\boldsymbol{\beta}} - \hat{\boldsymbol{K}}_\ell^{\boldsymbol{\beta}} \right\|_2 \le (C\beta_1)^{\ell-1} C\varepsilon_3, \tag{27}$$
$$\left\| \hat{\boldsymbol{\alpha}}_{\ell+1,i} - \bar{\boldsymbol{\alpha}}_{\ell+1,i} \right\|_2 \le (C\beta_1)^{\ell-1} C\varepsilon_3 \left\| \bar{\boldsymbol{\Theta}}_i \right\|_2, \tag{28}$$
$$\left\| \hat{\boldsymbol{\beta}}_{\ell+1,i} - \bar{\boldsymbol{\beta}}_{\ell+1,i} \right\|_2 \le (C\beta_1)^{\ell-1} C\varepsilon_3 \left\| \bar{\boldsymbol{\Theta}}_i \right\|_2. \tag{29}$$

For $\ell = 1$, by definition $\bar{\boldsymbol{K}}_1^{\boldsymbol{\beta}} = \hat{\boldsymbol{K}}_1^{\boldsymbol{\beta}}$. The upper bound of $\left\| \left[\hat{\boldsymbol{K}}_1^{\boldsymbol{\beta}}\right]^{1/2} - \left[\boldsymbol{K}_1^{\boldsymbol{\beta}}\right]^{1/2} \right\|_2$ is achieved by matrix calculus (Bhatia, 2013, Section V.3). Since $\left\| \hat{\boldsymbol{K}}_1^{\boldsymbol{\beta}} - \boldsymbol{K}_1^{\boldsymbol{\beta}} \right\|_2 \le \frac{\bar{\lambda}_1}{2}$, then the eigenvalues of $\hat{\boldsymbol{K}}_1^{\boldsymbol{\beta}}$ are at least $\frac{\bar{\lambda}_1}{2}$. Let $f(x) := \sqrt{x}$. Then $|f'(x)| \ge \frac{1}{\sqrt{2\bar{\lambda}_1}}$ when x is the eigenvalue of $\hat{\boldsymbol{K}}_1^{\boldsymbol{\beta}}$. Applying (Bhatia, 2013, (V.20)) yields that

$$\left\| \left[\hat{\boldsymbol{K}}_1^{\boldsymbol{\beta}}\right]^{1/2} - \left[\boldsymbol{K}_1^{\boldsymbol{\beta}}\right]^{1/2} \right\|_2 \le \frac{N}{\sqrt{2\bar{\lambda}_1}} \left\| \hat{\boldsymbol{K}}_1^{\boldsymbol{\beta}} - \boldsymbol{K}_1^{\boldsymbol{\beta}} \right\|_2 \le C\varepsilon_3, \tag{30}$$

and

$$\|\hat{\boldsymbol{\alpha}}_{2,i} - \bar{\boldsymbol{\alpha}}_{2,i}\|_2 = \left\| \left(\left[\hat{\boldsymbol{K}}_1^{\boldsymbol{\beta}}\right]^{1/2} \left[\boldsymbol{K}_1^{\boldsymbol{\beta}}\right]^{-1/2} - \boldsymbol{I}^N \right) \bar{\boldsymbol{\alpha}}_{2,i} \right\|_2 \le C\varepsilon_3 \|\bar{\boldsymbol{\alpha}}_{2,i}\|_2 \le C\varepsilon_3 \|\bar{\boldsymbol{\Theta}}_i\|_2. \tag{31}$$

Then by the Lipschitz continuity of h_2, we have

$$\left\| \hat{\boldsymbol{\beta}}_{2,i} - \bar{\boldsymbol{\beta}}_{2,i} \right\|_2 = \left\| \left[\dot{h}_2(\hat{\boldsymbol{\alpha}}_{2,i}) - \dot{h}_2(\bar{\boldsymbol{\alpha}}_{2,i}) \right] \right\|_2 \le C\varepsilon_3 \|\bar{\boldsymbol{\Theta}}_i\|_2. \tag{32}$$

For $\ell \in [2 : L-1]$, suppose that

$$\left\| \hat{\boldsymbol{\beta}}_{\ell,i} - \bar{\boldsymbol{\beta}}_{\ell,i} \right\|_2 \le (C\beta_1)^{\ell-2} C\varepsilon_3 \left\| \bar{\boldsymbol{\Theta}}_i \right\|_2. \tag{33}$$

Then, by the boundedness of h_1 and the triangle inequality, we have

$$\left\| \hat{\boldsymbol{K}}_\ell^{\boldsymbol{\beta}} - \bar{\boldsymbol{K}}_\ell^{\boldsymbol{\beta}} \right\|_2 \le \frac{C}{m} \sum_{i=1}^m \left\| \dot{h}_1(\bar{\boldsymbol{\beta}}_{\ell,i}) - \dot{h}_1(\hat{\boldsymbol{\beta}}_{\ell,i}) \right\|_2.$$

Applying the Lipschitz continuity of h_1 and (33) yields that

$$\left\| \hat{\boldsymbol{K}}_\ell^{\boldsymbol{\beta}} - \bar{\boldsymbol{K}}_\ell^{\boldsymbol{\beta}} \right\|_2 \le \frac{(C\beta_1)^{\ell-2} C\varepsilon_3}{m} \sum_{i=1}^m \|\bar{\boldsymbol{\Theta}}_i\|_2 \le (C\beta_1)^{\ell-1} C\varepsilon_3. \tag{34}$$

where in the last inequality we used (26). Then we obtain (27) by triangle inequality from (25) and (34). The upper bound in (28) for $\ell + 1$ follows from a similar argument of (30) and (31). Finally (29) for $\ell + 1$ follows from (28) and

$$\left\| \hat{\boldsymbol{\beta}}_{\ell+1,i} - \bar{\boldsymbol{\beta}}_{\ell+1,i} \right\|_2 = \left\| \sum_{j=2}^{\ell+1} \left[\dot{h}_2(\hat{\boldsymbol{\alpha}}_{j,i}) - \dot{h}_2(\bar{\boldsymbol{\alpha}}_{j,i}) \right] \right\|_2 \le C(C\beta_1)^{\ell-1} \varepsilon_3 \left\| \bar{\boldsymbol{\Theta}}_i \right\|_2.$$

We finish the induction. Since $\beta = \tilde{\mathcal{O}}(1)$ and $\varepsilon_3 = \tilde{\mathcal{O}}(1/\sqrt{m})$, we complete the proof. ■

Lemma 11 *If $\hat{\boldsymbol{K}}_\ell^{\boldsymbol{\beta}}$ is invertible, then*

$$\hat{v}_{\ell+1,i,j} = \hat{\boldsymbol{\alpha}}_{\ell+1,j}^\top \left[\hat{\boldsymbol{K}}_\ell^{\boldsymbol{\beta}}\right]^{-1} \dot{h}_1(\hat{\boldsymbol{\beta}}_{\ell,i}), \quad \ell \in [L-1],\ i,j \in [m].$$

Proof For a given layer ℓ and j, the ℓ_2-regression problem in Algorithm 2 can be equivalently written as

$$\begin{aligned} \min_{\hat{\boldsymbol{v}}} \quad & \frac{1}{2}\|\hat{\boldsymbol{v}}\|^2 \\ \text{s.t.} \quad & \frac{1}{m}\hat{\boldsymbol{H}}\hat{\boldsymbol{v}} = \hat{\boldsymbol{\alpha}}_{\ell+1,j}, \end{aligned} \tag{35}$$

where $\hat{\boldsymbol{v}} = (\hat{v}_{\ell+1,1,j}, \ldots, \hat{v}_{\ell+1,m,j})^\top$ and $\hat{\boldsymbol{H}} = \left[\dot{h}_1(\hat{\boldsymbol{\beta}}_{\ell,1}), \ldots, \dot{h}_1(\hat{\boldsymbol{\beta}}_{\ell,m})\right]$. Decompose $\hat{\boldsymbol{v}}$ as

$$\hat{\boldsymbol{v}} = \hat{\boldsymbol{H}}^\top \boldsymbol{z} + \hat{\boldsymbol{v}}',$$

where $\boldsymbol{z} \in \mathbb{R}^m$ and $\hat{\boldsymbol{H}}\hat{\boldsymbol{v}}' = 0$. Then (35) is equivalent to

$$\begin{aligned} \min_{\boldsymbol{z},\, \hat{\boldsymbol{v}}'} \quad & \frac{1}{2}\left\|\hat{\boldsymbol{H}}^\top \boldsymbol{z}\right\|_2^2 + \frac{1}{2}\left\|\hat{\boldsymbol{v}}'\right\|_2^2 \\ \text{s.t.} \quad & \frac{1}{m}\hat{\boldsymbol{H}}\hat{\boldsymbol{H}}^\top \boldsymbol{z} = \hat{\boldsymbol{\alpha}}_{\ell+1,j}. \end{aligned}$$

Since $\frac{1}{m}\hat{\boldsymbol{H}}\hat{\boldsymbol{H}}^\top = \hat{\boldsymbol{K}}_\ell^{\boldsymbol{\beta}}$ is invertible, the optimal solution is $\boldsymbol{z} = \left[\hat{\boldsymbol{K}}_\ell^{\boldsymbol{\beta}}\right]^{-1}\hat{\boldsymbol{\alpha}}_{\ell+1,j}$ and $\hat{\boldsymbol{v}}' = \boldsymbol{0}^N$. ■

B.2. Proof of Theorem 8

Proof [Proof of Theorem 8] In the proof we use the following abbreviated notations: for $t \in [0, \infty)$ and $\boldsymbol{\Theta} \in \mathrm{supp}(p)$, let

$$\begin{aligned} \boldsymbol{\beta}_\ell^t(\boldsymbol{\Theta}) &= \Phi_{*,\ell}^{\boldsymbol{\beta}}(\boldsymbol{\Theta})(t), \quad \ell \in [L], \\ \boldsymbol{\alpha}_\ell^t(\boldsymbol{\Theta}) &= \Phi_{*,\ell}^{\boldsymbol{\alpha}}(\boldsymbol{\Theta})(t), \quad \ell \in [2:L], \\ \boldsymbol{v}_1^t(\boldsymbol{\Theta}) &= \Phi_{*,1}^{\boldsymbol{v}}(\boldsymbol{\Theta})(t), \\ v_{L+1}^t(\boldsymbol{\Theta}) &= \Phi_{*,L+1}^{\boldsymbol{v}}(\boldsymbol{\Theta})(t). \end{aligned}$$

From the convergence assumptions, it is clear that $\boldsymbol{\beta}_{L+1}^t$ converges as $t \to \infty$. Indeed, the convergence assumptions imply that, for any $\varepsilon_2 > 0$, there exists T, for any $t \geq T$,

$$\left\|\boldsymbol{\beta}_L^t(\boldsymbol{\Theta}) - \boldsymbol{\beta}_L^\infty(\boldsymbol{\Theta})\right\|_\infty \leq \varepsilon_2 \tag{36}$$

holds p-almost surely and

$$\int \left|v_{L+1}^t(\boldsymbol{\Theta}) - v_{L+1}^\infty(\boldsymbol{\Theta})\right| p(\boldsymbol{\Theta}) \leq \varepsilon_2. \tag{37}$$

Then, since h_1 is bounded and Lipschitz continuous, we have

$$\begin{aligned} & \left\|\boldsymbol{\beta}_{L+1}^t - \boldsymbol{\beta}_{L+1}^\infty\right\|_\infty \\ = & \left\|\int v_{L+1}^t(\boldsymbol{\Theta})\, \dot{h}_1\left(\boldsymbol{\beta}_L^t(\boldsymbol{\Theta})\right) - v_{L+1}^\infty(\boldsymbol{\Theta})\, \dot{h}_1\left(\boldsymbol{\beta}_L^\infty(\boldsymbol{\Theta})\right) dp(\boldsymbol{\Theta})\right\|_\infty \\ \leq & \int \left|v_{L+1}^\infty(\boldsymbol{\Theta})\right| \left\|\dot{h}_1\left(\boldsymbol{\beta}_L^t(\boldsymbol{\Theta})\right) - \dot{h}_1\left(\boldsymbol{\beta}_L^\infty(\boldsymbol{\Theta})\right)\right\|_\infty dp(\boldsymbol{\Theta}) \\ & + \int \left|v_{L+1}^t(\boldsymbol{\Theta}) - v_{L+1}^\infty(\boldsymbol{\Theta})\right| \left\|\dot{h}_1(\boldsymbol{\beta}_L^t(\boldsymbol{\Theta}))\right\|_\infty dp(\boldsymbol{\Theta}) \\ \leq & C\varepsilon_2. \end{aligned} \tag{38}$$

The goal of the proof is to show that

$$\left\|\dot{\phi}_1'\left(\boldsymbol{\beta}_{L+1}^{\infty}\right)\right\|_2 = 0. \tag{39}$$

To this end, for any $\varepsilon > 0$, we will construct a function

$$f_\varepsilon(\boldsymbol{v}_1) := \dot{\phi}_1'\left(\boldsymbol{\beta}_{L+1}^{\infty}\right)^\top \dot{h}_1\left(\frac{1}{d}\boldsymbol{X}\boldsymbol{v}_1 + g_\varepsilon(\boldsymbol{v}_1)\right), \tag{40}$$

where the functions g_ε is uniformly bounded, such that $|f_\varepsilon(\boldsymbol{v}_1)| < \varepsilon$. Then it follows from (40) that

$$\dot{\phi}_1'\left(\boldsymbol{\beta}_{L+1}^{\infty}\right) = \boldsymbol{K}^{-1}\int f_\varepsilon(\boldsymbol{v}_1)\,\dot{h}_1\left(\frac{1}{d}\boldsymbol{X}\boldsymbol{v}_1 + g_\varepsilon(\boldsymbol{v}_1)\right) d\tilde{p}_1(\boldsymbol{v}_1),$$

where $\tilde{p}_1 = \mathcal{N}(\mathbf{0}^d, \boldsymbol{I}^d)$ and $\boldsymbol{K} := \int \dot{h}_1\left(\frac{1}{d}\boldsymbol{X}\boldsymbol{v}_1 + g_{\varepsilon,\eta}(\boldsymbol{v}_1)\right)\dot{h}_1^\top\left(\frac{1}{d}\boldsymbol{X}\boldsymbol{v}_1 + g_{\varepsilon,\eta}(\boldsymbol{v}_1)\right) d\tilde{p}_1(\boldsymbol{v}_1)$ whose minimum eigenvalue is at least $\bar{\lambda}_1 > 0$ by Assumption 2. The boundedness of h_1 yields that

$$\left\|\dot{\phi}_1'\left(\boldsymbol{\beta}_{L+1}^{\infty}\right)\right\|_2 \le C\bar{\lambda}_1^{-1}\varepsilon.$$

Since $\bar{\lambda}_1$ is independent of ε, by letting $\varepsilon \to 0$, we obtain (39).

Next we construct g_ε and f_ε in (40). Let T be the time such that (36) and (37) hold with $\varepsilon_2 \le c\varepsilon$ for a constant c to be specified. Note that $\boldsymbol{v}_1^T$ is surjective by Lemma 12. Let $\tilde{g} : \mathbb{R}^d \to \mathrm{supp}(p)$ be the inverse function such that $\boldsymbol{v}_1^T(\tilde{g}(\boldsymbol{v}_1)) = \boldsymbol{v}_1$. Define

$$g_\varepsilon(\boldsymbol{v}_1) = \sum_{\ell=2}^{L}\dot{h}_2\left(\boldsymbol{\alpha}_\ell^T\left(\tilde{g}(\boldsymbol{v}_1)\right)\right), \qquad f_\varepsilon(\boldsymbol{v}_1) = \dot{\phi}_1'\left(\boldsymbol{\beta}_{L+1}^{\infty}\right)^\top \dot{h}_1\left(\boldsymbol{\beta}_L^T(\tilde{g}(\boldsymbol{v}_1))\right),$$

where g_ε is uniformly bounded by the boundedness of h_2. Suppose on the contrary that there exists $\boldsymbol{v}_1'$ such that $|f_\varepsilon(\boldsymbol{v}_1')| > \varepsilon$. Let $\boldsymbol{\Theta}' = \tilde{g}(\boldsymbol{v}_1')$. Since $\boldsymbol{\Theta} \mapsto \dot{\phi}_1'\left(\boldsymbol{\beta}_{L+1}^{\infty}\right)^\top \dot{h}_1\left(\boldsymbol{\beta}_L^T(\boldsymbol{\Theta})\right)$ is continuous by Theorem 13 (see Appendix C.1), there exists a ball around $\boldsymbol{\Theta}'$ denoted by S such that $p(S) > 0$ and $|\dot{\phi}_1'\left(\boldsymbol{\beta}_{L+1}^{\infty}\right)^\top \dot{h}_1\left(\boldsymbol{\beta}_L^T(\boldsymbol{\Theta})\right)| > \varepsilon/2$ with the same sign for all $\boldsymbol{\Theta} \in S$. However, for $t > T$,

$$\begin{aligned}
&\int \left|v_{L+1}^t(\boldsymbol{\Theta}) - v_{L+1}^T(\boldsymbol{\Theta})\right| dp(\boldsymbol{\Theta}) \\
\ge\ & \frac{1}{N}\int \mathcal{I}_S\left|\int_T^t \dot{\phi}_1'\left(\boldsymbol{\beta}_{L+1}^t\right)^\top \dot{h}_1\left(\boldsymbol{\beta}_L^t(\boldsymbol{\Theta})\right) dt\right| dp(\boldsymbol{\Theta}) \\
\ge\ & \frac{1}{N}\int \mathcal{I}_S\left(\left|\int_T^t \dot{\phi}_1'\left(\boldsymbol{\beta}_{L+1}^{\infty}\right)^\top \dot{h}_1\left(\boldsymbol{\beta}_L^T(\boldsymbol{\Theta})\right) dt\right| - \int_T^t C\varepsilon_2 dt\right) dp(\boldsymbol{\Theta}),
\end{aligned} \tag{41}$$

where in the last step we used $\|\boldsymbol{\beta}_L^t(\boldsymbol{\Theta}) - \boldsymbol{\beta}_L^T(\boldsymbol{\Theta})\|_\infty \le 2\varepsilon_2$ from (36), (38), and the boundedness and Lipschitz continuity of ϕ_1' and h_1. Let $c = \frac{1}{4C}$. The lower bound in (41) diverges with t, which contradicts (37).

Finally from (39) we show the convergence statement. Since ϕ is convex on the first argument, we obtain

$$\sum_{n=1}^{N}\phi\left(\boldsymbol{\beta}_{L+1}^{\infty}(n), y^n\right) = \sum_{n=1}^{N}\left[\min_{y'}\phi\left(y', y^n\right)\right].$$

Since $\boldsymbol{\beta}_{L+1}^t \to \boldsymbol{\beta}_{L+1}^\infty$ and ϕ is continuous, we obtain that

$$\lim_{t\to\infty} L^t = \sum_{n=1}^N \phi(\boldsymbol{\beta}_{L+1}^\infty(n), y^n) = \sum_{n=1}^N \left[\min_{y'} \phi\left(y', y^n\right)\right],$$

which completes the proof. ■

Lemma 12 *The function $t < \infty$, $\boldsymbol{v}_1^t : \mathrm{supp}(p) \to \mathbb{R}^d$ is a surjection.*

Proof Recall that at the initialization let $\boldsymbol{\Theta}(\boldsymbol{v}) = (\boldsymbol{v}, \mathbf{0}^{D-d}) \in \mathbb{R}^D = \mathrm{supp}(p)$, where the equality follows from Assumption 3 that p has a full support. Given $t < \infty$, consider $f_t : \mathbb{R}^d \to \mathbb{R}^d$ as

$$f_t(\boldsymbol{v}) = \boldsymbol{v}_1^t\left(\boldsymbol{\Theta}(\boldsymbol{v})\right).$$

It suffices to show that f_t is surjective. Note that f_t is continuous since $\boldsymbol{\Theta} \mapsto \boldsymbol{v}_1^t(\boldsymbol{\Theta})$ is continuous by Theorem 13. Furthermore, for any $\boldsymbol{v} \in \mathbb{R}^d$, by Lemma 21 which states that the gradient of the weights $\underline{\mathcal{G}}_1^{\boldsymbol{v}}$ is bounded (see Appendix C.1), we have

$$\left\|f_t(\boldsymbol{v}) - \boldsymbol{v}\right\|_\infty = \int_0^t \left\|\underline{\mathcal{G}}_1^{\boldsymbol{v}}\left(\left[\boldsymbol{v}; \mathbf{0}^{D-d}\right], \Phi_*, s\right)\right\|_\infty ds \le Ct.$$

For any $\boldsymbol{x} \in \mathbb{R}^d$, consider $g(\boldsymbol{v}) := \boldsymbol{x} - \left(f_t(\boldsymbol{v}) - \boldsymbol{v}\right)$ which continuously maps $\mathcal{B}_\infty(\boldsymbol{x}, Ct)$ to itself. By the Brouwer's fixed-point theorem (see, e.g. Granas and Dugundji (2013)), there exists $\boldsymbol{v}_* \in \mathcal{B}_\infty(\boldsymbol{x}, Ct)$ such that $g(\boldsymbol{v}_*) = \boldsymbol{v}_*$; equivalently, we have $f_t(\boldsymbol{v}_*) = \boldsymbol{x}$. ■

B.3. Proof of Lemma 3

Proof [Proof of Lemma 3] We first note the following results in Du et al. (2019a, Lemma F.1): suppose that $C_B = 0$ and the support of a random vector $\boldsymbol{V} \in \mathbb{R}^d$ denoted by R has positive Lebesgue measure. Moreover, h is an analytic non-polynomial function on R. Then

$$\min_{\|\boldsymbol{a}\|_2=1} \mathbb{E}\left\|\sum_{i=1}^N a_i h(\boldsymbol{x}_i \circ \boldsymbol{V})\right\|_2^2 = \lambda > 0,$$

where $\boldsymbol{a} = (a_1, \ldots, a_N)$. Lemma 3 shows that, for $\boldsymbol{V}' \sim p = \mathcal{N}\left(\mathbf{0}^d, \boldsymbol{I}^d\right)$, the same result holds with a constant perturbation of the functions h_1; namely, by letting $g_i(\boldsymbol{v}) = h_1(\boldsymbol{x}_i \circ \boldsymbol{v} + C_i(\boldsymbol{v}))$ where $\|C_i\|_\infty \le C_B$,

$$\min_{\|\boldsymbol{a}\|_2=1} \mathbb{E}\left\|\sum_{i=1}^N a_i g_i(\boldsymbol{V}')\right\|_2^2 = \lambda' > 0, \tag{42}$$

where λ' is uniform over all perturbations $\|C_i\|_\infty \le C_B$. It suffices to prove (42) for $\boldsymbol{V}' \sim q = \mathrm{Uniform}(R')$ where R' is determined by h and C_B, as the Radon–Nikodym derivative $\frac{dq}{dp}$ is bounded.

We first prove (i). Consider a compact region R such that, for $\boldsymbol{V} \sim \mathrm{Uniform}(R)$ and any unit vector $\boldsymbol{a}$,

$$\mathbb{E}\left\|\sum_{i=1}^{N} a_i g(\boldsymbol{x}_i \circ \boldsymbol{V})\right\|_2^2 \geq \lambda_R > 0.$$

Then for $\beta > 0$, since g is bounded and Lipschitz continuous, we have

$$\mathbb{E}\left\|\sum_{i=1}^{N} a_i g\left(\boldsymbol{x}_i \circ \boldsymbol{V} + \beta C_i(\boldsymbol{V}/\beta)\right)\right\|_2^2 \geq \lambda_R - CC_B\beta \geq \frac{\lambda_R}{2},$$

when $\beta \leq \frac{\lambda_R}{2C_BC}$. Let $h_1(x) = g(\beta x)$. Then $\mathbb{E}\|\sum_{i=1}^{N} a_i h_1\left(\boldsymbol{x}_i \circ \boldsymbol{V}/\beta + C_i(\boldsymbol{V}/\beta)\right)\|_2^2 \geq \frac{\lambda_R}{2}$. We achieve (42) by letting $\boldsymbol{V}' = \boldsymbol{V}/\beta$.

For (ii), consider $R = \{\boldsymbol{v} : 1/2 \leq \|\boldsymbol{v}\|_2 \leq 1\}$. Then, for $\boldsymbol{V} \sim \mathrm{Uniform}(R)$ and any unit vector $\boldsymbol{a}$,

$$\mathbb{E}\left\|\sum_{i=1}^{N} a_i h_1(\boldsymbol{x}_i \circ \boldsymbol{V})\right\|_2^2 \geq \lambda_R > 0.$$

Since $h_1(\beta x) = \beta^\alpha x$ for any $\beta > 0$, then we have $\mathbb{E}\|\sum_{i=1}^{N} a_i h_1(\boldsymbol{x}_i \circ \beta\boldsymbol{V})\|_2^2 \geq \beta^{2\alpha}\lambda_R$. Note that $|\boldsymbol{x}_i \circ \beta\boldsymbol{V}| = \Theta(\beta)$. For $x = \Theta(\beta)$, we have $|h_1(x)| \leq C\beta^\alpha$ and h_1 is $C\beta^{\alpha-1}$-Lipschitz continuous for a constant C. Therefore,

$$\mathbb{E}\left\|\sum_{i=1}^{N} a_i h_1(\boldsymbol{x}_i \circ \beta\boldsymbol{V} + C_i(\beta\boldsymbol{V}))\right\|_2^2 \geq \beta^{2\alpha}\lambda_R - C'\beta^{2\alpha-1}C_B \geq (C'C_B)^{2\alpha}\left(\frac{2}{\lambda_R}\right)^{2\alpha-1}.$$

for a constant C' when $\beta = \frac{2C'C_B}{\lambda_R}$. We achieve (42) by letting $\boldsymbol{V}' = \beta\boldsymbol{V}$.

For (iii), we only consider $h_1(x) = c|x|^{-\alpha}$. The case for $h_1(x) = c(x)_+^{-\alpha}$ can be obtained by a similar argument. We first show that there exists a compact set R such that, for all $\boldsymbol{v} \in R$ and $\boldsymbol{x}_i$,

$$|\boldsymbol{x}_i \circ \boldsymbol{v}| \geq c'. \tag{43}$$

This can be done by a simple probabilistic argument. Let $\boldsymbol{v}$ be drawn from the uniform distribution on S^{d-1}, for any fixed $\boldsymbol{x} \in \mathbb{R}^d$, we have

$$P\{|\boldsymbol{v}^\top \boldsymbol{x}| < t\|\boldsymbol{x}\|_2\} = \frac{2\pi^{\frac{d-1}{2}}/\Gamma(\frac{d-1}{2})}{2\pi^{\frac{d}{2}}/\Gamma(\frac{d}{2})}\int_{-t}^{t}(1-u^2)^{\frac{d-3}{2}}\,du < t\sqrt{d}.$$

By a union bound, we have $|\boldsymbol{x}_i \circ \boldsymbol{v}| \geq \frac{\|\boldsymbol{x}_i\|_2}{2N\sqrt{d}}$ with probability 0.5. Denote this set of $\boldsymbol{v} \in S^{d-1}$ by S'. Since $\min_i \|\boldsymbol{x}_i\|_2 := C_{mx} > 0$, we obtain (43) with $R = \{t\boldsymbol{v} : \boldsymbol{v} \in S', \frac{2c'N\sqrt{d}}{C_{mx}} \leq t \leq \frac{4c'N\sqrt{d}}{C_{mx}}\}$. Then, for $\boldsymbol{V} \sim \mathrm{Uniform}(R)$ and any unit vector $\boldsymbol{a}$,

$$\mathbb{E}\left\|\sum_{i=1}^{N} a_i h_1(\boldsymbol{x}_i \circ \boldsymbol{V})\right\|_2^2 \geq \lambda_R > 0.$$

Then, for any $\beta > 0$, we have $\mathbb{E}\|\sum_{i=1}^N a_i h_1(\boldsymbol{x}_i \circ \beta \boldsymbol{V})\|_2^2 \geq \beta^{-2\alpha}\lambda_R$. For $x = \Theta(\beta)$ we have $|h_1(x)| \leq C\beta^{-\alpha}$ and h_1 is $C\beta^{-\alpha-1}$-Lipschitz continuous for a constant C. Therefore,

$$\mathbb{E}\left\|\sum_{i=1}^N a_i h_1(\boldsymbol{x}_i \circ \beta \boldsymbol{V} + C_i(\beta \boldsymbol{V}))\right\|_2^2 \geq \beta^{-2\alpha}\lambda_R - C'\beta^{-2\alpha-1}C_B \geq (C'C_B)^{-2\alpha}\left(\frac{2}{\lambda_R}\right)^{-2\alpha-1}.$$

for a constant C' when $\beta = \frac{2C'C_B}{\lambda_R}$. We achieve (42) by letting $\boldsymbol{V}' = \beta \boldsymbol{V}$. ■

Appendix C. Rest Proofs

C.1. Proofs of Theorems 4 and 13

We first demonstrate Theorem 13 which shows that the neural feature flow Φ_* is a continuous mapping on $\boldsymbol{\Theta}$ given time t.

Theorem 13 (Property of Φ_*) *Under Assumptions 1 and 3, let Φ_* be the neural feature flow, there exist constants $C, C' \geq 0$ such that for all $t \in [0, T]$, $\boldsymbol{\Theta}_1 \in \operatorname{supp}(p)$ and $\tilde{\boldsymbol{\Theta}}_1 \in \operatorname{supp}(p) \cap \mathcal{B}_\infty(\boldsymbol{\Theta}, 1)$, $\boldsymbol{\Theta}_2 \in \operatorname{supp}(p)$, and $\tilde{\boldsymbol{\Theta}}_2 \in \operatorname{supp}(p) \cap \mathcal{B}_\infty(\boldsymbol{\Theta}_2, 1)$, we have*

$$\begin{aligned}
\left\|\Phi_{*,\ell}^{\boldsymbol{\beta}}(\boldsymbol{\Theta}_1)(t) - \Phi_{*,\ell}^{\boldsymbol{\beta}}(\tilde{\boldsymbol{\Theta}}_1)(t)\right\|_\infty &\leq Ce^{C't}\left(\|\boldsymbol{\Theta}_1\|_\infty + 1\right)\left\|\boldsymbol{\Theta}_1 - \tilde{\boldsymbol{\Theta}}_1\right\|_\infty, \quad \ell \in [L], \\
\left\|\Phi_{*,\ell}^{\boldsymbol{\alpha}}(\boldsymbol{\Theta}_1)(t) - \Phi_{*,\ell}^{\boldsymbol{\alpha}}(\tilde{\boldsymbol{\Theta}}_1)(t)\right\|_\infty &\leq Ce^{C't}\left(\|\boldsymbol{\Theta}_1\|_\infty + 1\right)\left\|\boldsymbol{\Theta}_1 - \tilde{\boldsymbol{\Theta}}_1\right\|_\infty, \quad \ell \in [2:L], \\
\left\|\Phi_{*,1}^{\boldsymbol{v}}(\boldsymbol{\Theta}_1)(t) - \Phi_{*,1}^{\boldsymbol{v}}(\tilde{\boldsymbol{\Theta}}_1)(t)\right\|_\infty &\leq Ce^{C't}\left(\|\boldsymbol{\Theta}_1\|_\infty + 1\right)\left\|\boldsymbol{\Theta}_1 - \tilde{\boldsymbol{\Theta}}_1\right\|_\infty, \\
\left|\Phi_{*,\ell}^{\boldsymbol{v}}(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)(t) - \Phi_{*,\ell}^{\boldsymbol{v}}(\tilde{\boldsymbol{\Theta}}_1, \boldsymbol{\Theta}_2)(t)\right| &\leq Ce^{C't}\left(\|\boldsymbol{\Theta}_1\|_\infty + \|\boldsymbol{\Theta}_2\|_\infty + 1\right)\left\|\boldsymbol{\Theta}_1 - \tilde{\boldsymbol{\Theta}}_1\right\|_\infty, \quad \ell \in [2:L], \\
\left|\Phi_{*,\ell}^{\boldsymbol{v}}(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)(t) - \Phi_{*,\ell}^{\boldsymbol{v}}(\boldsymbol{\Theta}_1, \tilde{\boldsymbol{\Theta}}_2)(t)\right| &\leq Ce^{C't}\left(\|\boldsymbol{\Theta}_1\|_\infty + \|\boldsymbol{\Theta}_2\|_\infty + 1\right)\left\|\boldsymbol{\Theta}_2 - \tilde{\boldsymbol{\Theta}}_2\right\|_\infty, \quad \ell \in [2:L], \\
\left|\Phi_{*,L+1}^{\boldsymbol{v}}(\boldsymbol{\Theta}_1)(t) - \Phi_{*,L+1}^{\boldsymbol{v}}(\tilde{\boldsymbol{\Theta}}_1)(t)\right| &\leq Ce^{C't}\left(\|\boldsymbol{\Theta}_1\|_\infty + 1\right)\left\|\boldsymbol{\Theta}_1 - \tilde{\boldsymbol{\Theta}}_1\right\|_\infty.
\end{aligned}$$

In the proof, we fix the initial continuous Res-Net $(\{v_\ell\}_{\ell=2}^{L+1}, p)$, which is assumed to satisfy the Assumption 3.

C.1.1. PROOF OF THEOREM 4

We first show that our neural feature flow in Definition 2 necessarily satisfies several continuity properties in Lemma 15, which allows us to narrow down the search space for the solution. Then we construct a contraction mapping (also known as Picard iteration) to show the existence of uniqueness of solution in that search space. Recall that a trajectory Φ consists of trajectories of weights $\Phi_\ell^{\boldsymbol{v}}$ for $\ell \in [L+1]$, features $\Phi_\ell^{\boldsymbol{\beta}}$ for $\ell \in [L]$, and residuals $\Phi_\ell^{\boldsymbol{\alpha}}$ for $\ell \in [2:L]$. For $\boldsymbol{\Theta}, \bar{\boldsymbol{\Theta}} \in \operatorname{supp}(p)$, we also abbreviate the notations for individual trajectories as

$$\Phi_\ell^{\boldsymbol{v}}(\boldsymbol{u}_\ell)(t) = v_\ell^t(\boldsymbol{u}_\ell), \quad \Phi_\ell^{\boldsymbol{\beta}}(\boldsymbol{\Theta})(t) = \boldsymbol{\beta}_\ell^t(\boldsymbol{\Theta}), \quad \Phi_\ell^{\boldsymbol{\alpha}}(\boldsymbol{\Theta})(t) = \boldsymbol{\alpha}_\ell^t(\boldsymbol{\Theta}),$$

where $\boldsymbol{u}_\ell$ stands for $\boldsymbol{\Theta}$, $(\boldsymbol{\Theta}, \bar{\boldsymbol{\Theta}})$, $\boldsymbol{\Theta}$ for $\ell = 1$, $2 \leq \ell \leq L$, $\ell = L+1$, respectively.

Throughout the proof, we fix T as a constant. We define the set of continuous restricted trajectories below.

Definition 14 ($\boldsymbol{C}$-Continuous Restricted Trajectory) *Given $\boldsymbol{C} := (\boldsymbol{C}_1, \ldots, \boldsymbol{C}_{L+1}) \in \mathbb{R}_+^{L+1}$, we say Φ is a $\boldsymbol{C}$-continuous restricted trajectory if $\Phi_\ell^{\boldsymbol{v}}(\boldsymbol{u}_\ell)(t)$ is $\boldsymbol{C}_\ell$-Lipschitz continuous in $t \in [0,T]$ for $\ell \in [L+1]$, and $\Phi_\ell^{\boldsymbol{\alpha}}(\boldsymbol{u}_\ell)(t)$ and $\Phi_\ell^{\boldsymbol{\beta}}(\boldsymbol{u}_\ell)(t)$ are determined by the forward-propagation process, i.e., $\boldsymbol{\beta}_1^t(\boldsymbol{\Theta}) = \frac{1}{d}\boldsymbol{X} v_1^t(\boldsymbol{\Theta})$, $\boldsymbol{\alpha}_{\ell+1}^t(\boldsymbol{\Theta}) = \int v_{\ell+1}^t(\boldsymbol{\Theta}, \bar{\boldsymbol{\Theta}}) \dot{h}_1\left(\boldsymbol{\beta}_\ell^t(\bar{\boldsymbol{\Theta}})\right) dp(\bar{\boldsymbol{\Theta}})$, $\boldsymbol{\beta}_{\ell+1}^t(\boldsymbol{\Theta}) = \boldsymbol{\beta}_\ell^t(\boldsymbol{\Theta}) + \dot{h}_2\left(\boldsymbol{\alpha}_{\ell+1}^t(\boldsymbol{\Theta})\right)$ for $\ell \in [L-1]$ and $\boldsymbol{\Theta} \in \mathrm{supp}(p)$. The set of all $\boldsymbol{C}$-continuous restricted trajectories is denoted as $\boldsymbol{\Phi}^{\boldsymbol{C}}$.*

We can find that given the trajectories of weights, the trajectories of residuals and features are determined by the forward-propagation process. Lemma 15 below shows that it suffices to consider a restricted search space.

Lemma 15 *There exists a constant vector $\boldsymbol{C} \in \mathbb{R}_+^{L+1}$ such that every solution Φ of the neural feature flow is a $\boldsymbol{C}$-continuous restricted trajectory.*

In the remaining of the proof we let $\boldsymbol{C}$ be the constant vector in Lemma 15, and let $\boldsymbol{\Phi} := \boldsymbol{\Phi}^{\boldsymbol{C}}$, which will serve as the search space. The solution can be equivalently characterized as the fixed-point of a mapping from $\boldsymbol{\Phi}$ to itself that we introduce next:

Definition 16 *Define $F : \boldsymbol{\Phi} \to \boldsymbol{\Phi}$ as follows: for all $t \in [0,T]$,*

(1) for all $\ell \in [L+1]$ and all $\boldsymbol{u}_\ell$,

$$F(\Phi)_\ell^{\boldsymbol{v}}(\boldsymbol{u}_\ell)(t) = v_\ell(\boldsymbol{u}_\ell) - \int_0^t \underline{\mathcal{G}}_\ell^{\boldsymbol{v}}(\boldsymbol{u}_\ell; \Phi, s) ds,$$

(2) for all $\boldsymbol{\Theta}$,

$$F(\Phi)_1^{\boldsymbol{\beta}}(\boldsymbol{\Theta})(t) = \frac{1}{d}\left[\boldsymbol{X} F(\Phi)_1^{\boldsymbol{v}}(\boldsymbol{\Theta})(t)\right],$$

(3) for all $\ell \in [L-1]$ and $\boldsymbol{\Theta}$,

$$\begin{aligned} F(\Phi)_{\ell+1}^{\boldsymbol{\alpha}}(\boldsymbol{\Theta})(t) &= \int F(\Phi)_{\ell+1}^{\boldsymbol{v}}(\boldsymbol{\Theta}, \bar{\boldsymbol{\Theta}})(t)\, \dot{h}_1\left(\Phi_\ell^{\boldsymbol{\beta}}(\bar{\boldsymbol{\Theta}})(t)\right) dp(\bar{\boldsymbol{\Theta}}), \\ F(\Phi)_{\ell+1}^{\boldsymbol{\beta}}(\boldsymbol{\Theta})(t) &= \dot{h}_2\left(F(\Phi)_{\ell+1}^{\boldsymbol{\alpha}}(\boldsymbol{\Theta})(t)\right) + F(\Phi)_\ell^{\boldsymbol{\beta}}(\boldsymbol{\Theta})(t). \end{aligned}$$

Following the same argument as Lemma 15, we have that the image of $\boldsymbol{\Phi}$ under F is indeed contained in $\boldsymbol{\Phi}$. We then show in Lemma 18 the contraction property of F under an appropriate metric defined below:

Definition 17 *For a pair $\Phi_1, \Phi_2 \in \boldsymbol{\Phi}$, we define the normalized distance between each trajectories over $[0,t]$ as*

$$\mathrm{D}^{[0,t]}(\Phi_1, \Phi_2) := \sup_{s \in [0,t],\ \ell \in [L+1],\ \boldsymbol{u}_\ell} \frac{\|\Phi_{1,\ell}^{\boldsymbol{v}}(\boldsymbol{u}_\ell)(s) - \Phi_{2,\ell}^{\boldsymbol{v}}(\boldsymbol{u}_\ell)(s)\|_\infty}{1 + \|\boldsymbol{u}_\ell\|_\infty}.$$

Lemma 18 *There exists a constant C such that*

$$\mathrm{D}^{[0,t]}(F(\Phi_1), F(\Phi_2)) \le C \int_0^t \mathrm{D}^{[0,s]}(\Phi_1, \Phi_2) ds.$$

Proof [Proof of Theorem 4] Firstly, $\mathbf{\Phi}$ contains the constant trajectory and thus is nonempty. Applying Lemma 18, the proof of existence and uniqueness follows from a similar argument of Picard–Lindelöf theorem. Specifically, iteratively applying Lemma 18 yields that

$$\mathrm{D}^{[0,T]}(F^m(\Phi_1), F^m(\Phi_2)) \leq \frac{(CT)^m}{m!}\mathrm{D}^{[0,T]}(\Phi_1, \Phi_2).$$

Let Φ be the constant trajectory, for any $\tilde{\Phi} \in \mathbf{\Phi}$, by the upper bounds of $\underline{\mathcal{G}}_\ell^{\boldsymbol{v}}$ in Lemma 21 and the Definition of $\mathrm{D}^{[0,T]}$ in Definition 17, there is a constant C such that

$$\mathrm{D}^{[0,T]}(F(\tilde{\Phi}), \Phi) \leq CT < \infty.$$

We first show the uniqueness. For two fixed points of F denoted by Φ_1 and Φ_2, we have

$$\mathrm{D}^{[0,T]}(\Phi_1, \Phi_2) = \mathrm{D}^{[0,T]}(F^m(\Phi_1), F^m(\Phi_2)) \leq \frac{(CT)^{m-1}}{(m-1)!}\mathrm{D}^{[0,T]}(F(\Phi_1), F(\Phi_2)),$$

By the triangle inequality $\mathrm{D}^{[0,T]}(F(\Phi_1), F(\Phi_2)) \leq \mathrm{D}^{[0,T]}(F(\Phi_1), \Phi) + \mathrm{D}^{[0,T]}(F(\Phi_2), \Phi) < \infty$, hence the right-hand side of the above inequality vanishes as m diverges. For the existence, we can consider the sequence $\{F_2^i(\Phi) : i \geq 0\}$ that satisfies

$$\mathrm{D}^{[0,T]}(F_2^{m+1}(\Phi), F_2^m(\Phi)) \leq \frac{(CT)^m}{m!}\mathrm{D}^{[0,T]}(F(\Phi), \Phi),$$

Because $\mathrm{D}^{[0,T]}(F(\Phi), \Phi) < \infty$, $\{F^i(\Phi) : i \geq 0\}$ is a Cauchy sequence. Since $\mathbf{\Phi}$ is complete under $\mathrm{D}^{[0,T]}$ by Lemma 22, the limit point $\Phi_* \in \mathbf{\Phi}$, which is a fixed-point of F. Finally, by dominated convergence theorem, we can directly verify that Φ_* is the solution of neural feature flow. ■

C.1.2. PROOF OF THEOREM 13

Theorem 13 is a Grönwall-type of result. However, it is not straightforward to directly derive a simple differential inequality due to the involved relations among the parameters of deep neural networks. Again we turn to the technique of Picard iterations used in the proof of Theorem 4. This approach has also been used to prove the abstract Grönwall inequality in Turinici (1986).

Recall the set $\mathbf{\Phi}$ in the proof of Theorem 4, and the mapping $F : \mathbf{\Phi} \mapsto \mathbf{\Phi}$ in Definition 16. It is shown that F is a contraction mapping and thus there exists a unique solution $\Phi_* \in \mathbf{\Phi}$. We will construct a closed nonemtpy subset $\tilde{\mathbf{\Phi}} \subseteq \mathbf{\Phi}$ with the desired properties in Theorem 13 such that $F(\tilde{\mathbf{\Phi}}) \subseteq \tilde{\mathbf{\Phi}}$. Then by the same argument as the proof of Theorem 4, the Picard iteration guarantees the solution in $\tilde{\mathbf{\Phi}}$, thereby proving $\Phi_* \in \tilde{\mathbf{\Phi}}$.

We introduce the set of b-locally Lipschitz trajectories with the desired properties in Theorem 13. We use similar notations as in the proof of Theorem 4 by letting $\boldsymbol{u}_\ell$ denote $\mathbf{\Theta}$, $(\mathbf{\Theta}, \bar{\mathbf{\Theta}})$, $\mathbf{\Theta}$ for $\ell = 1$, $2 \leq \ell \leq L$, $\ell = L+1$, respectively.

Definition 19 (b-Locally Lipschitz Trajectory) *Recall the constants C_2 and C_4 in Assumption 3 for the locally Lipschitz continuity at $t = 0$. We say Φ is b-locally Lipschitz if for all $t \in [0, T]$, $\mathbf{\Theta}_1$,*

$\bar{\boldsymbol{\Theta}}_1 \in \mathcal{B}_\infty(\boldsymbol{\Theta}_1, 1)$, $\bar{\boldsymbol{\Theta}}_2$, *and* $\bar{\boldsymbol{\Theta}}_2 \in \mathcal{B}_\infty(\bar{\boldsymbol{\Theta}}_2, 1)$, *we have*

$$\left\|\Phi_1^{\boldsymbol{v}}(\boldsymbol{\Theta}_1)(t) - \Phi_1^{\boldsymbol{v}}(\bar{\boldsymbol{\Theta}}_1)(t)\right\|_\infty \le e^{bt}(\|\boldsymbol{\Theta}_1\|_\infty + 1)\|\boldsymbol{\Theta}_1 - \bar{\boldsymbol{\Theta}}_1\|_\infty, \tag{44a}$$

$$\left|\Phi_{L+1}^{\boldsymbol{v}}(\boldsymbol{\Theta}_1)(t) - \Phi_{L+1}^{\boldsymbol{v}}(\bar{\boldsymbol{\Theta}}_1)(t)\right| \le (1 + C_4)e^{bt}(\|\boldsymbol{\Theta}_1\|_\infty + 1)\|\boldsymbol{\Theta}_1 - \bar{\boldsymbol{\Theta}}_1\|_\infty, \tag{44b}$$

$$\left|\Phi_\ell^{\boldsymbol{v}}(\boldsymbol{u}_\ell)(t) - \Phi_\ell^{\boldsymbol{v}}(\bar{\boldsymbol{\Theta}}_1, \boldsymbol{\Theta}_2)(t)\right| \le (1 + C_2)e^{bt}(\|\boldsymbol{u}_\ell\|_\infty + 1)\|\boldsymbol{\Theta}_1 - \bar{\boldsymbol{\Theta}}_1\|_\infty, \tag{44c}$$

$$\left|\Phi_\ell^{\boldsymbol{v}}(\boldsymbol{u}_\ell)(t) - \Phi_\ell^{\boldsymbol{v}}(\boldsymbol{\Theta}_1, \bar{\boldsymbol{\Theta}}_2)(t)\right| \le (1 + C_2)e^{bt}(\|\boldsymbol{u}_\ell\|_\infty + 1)\|\boldsymbol{\Theta}_2 - \bar{\boldsymbol{\Theta}}_2\|_\infty, \tag{44d}$$

for $\ell \in [2 : L]$. *Denote the set of all* b*-locally Lipschitz trajectories as* $\boldsymbol{\Phi}_b$.

Lemma 20 *There exists a constant* b_* *such that* $F(\boldsymbol{\Phi} \cap \boldsymbol{\Phi}_{b_*}) \subseteq \boldsymbol{\Phi}_{b_*}$.

Proof [Proof of Theorem 13] Let b_* be the constant in Lemma 20 and $\boldsymbol{\Phi}' := \boldsymbol{\Phi} \cap \boldsymbol{\Phi}_{b_*} \subseteq \boldsymbol{\Phi}$, which clearly contains the constant trajectory and thus is nonempty. It follows from Lemma 20 that $F(\boldsymbol{\Phi}') \subseteq \boldsymbol{\Phi}'$. Since F is a contraction mapping by Lemma 18 and $\boldsymbol{\Phi}'$ is a closed set by Lemma 23, by the same argument as the proof of Theorem 4, there exists a unique solution in $\boldsymbol{\Phi}'$, which is necessarily Φ_* by the uniqueness of the solution in Theorem 4.

■

C.1.3. PROOFS OF LEMMAS

Proof [Proof of Lemma 15] We first prove the Lipschitz continuity of Φ for weight. It suffices to show upper bounds of $\underline{\mathcal{G}}_\ell^{\boldsymbol{v}}$ for each layer ℓ. We use the backward equations to inductively upper bound $\underline{\mathcal{D}}_\ell^{\boldsymbol{\beta}}$ and $\underline{\mathcal{D}}_\ell^{\boldsymbol{\alpha}}$, which immediately yield upper bounds $\|\underline{\mathcal{G}}_\ell^{\boldsymbol{v}}\|_\infty \le \boldsymbol{C}_\ell$ for constants $\boldsymbol{C}_\ell$.

For the top layer $\ell = L + 1$, by Assumption 1 that $|\phi_1'| \le L_4$, we have

$$\|\underline{\mathcal{D}}_{L+1}(\Phi, t)\|_\infty \le L_4 := \tilde{\boldsymbol{C}}_{L+1}.$$

At layer $\ell = L$, since $|h_1'| \le L_2$,

$$\left\|\underline{\mathcal{D}}_L^{\boldsymbol{\beta}}(\boldsymbol{\Theta}; \Phi, t)\right\|_\infty \le \underbrace{\left\|\dot{h}_1'\left(\boldsymbol{\beta}_L^t(\boldsymbol{\Theta})\right)\right\|_\infty}_{\le L_2} \underbrace{\left\|\underline{\mathcal{D}}_{L+1}(\Phi, t)\right\|_\infty}_{\le \tilde{\boldsymbol{C}}_{L+1}} \underbrace{\left|v_{L+1}^t(\boldsymbol{\Theta})\right|}_{\le C_3 + T\boldsymbol{C}_{L+1}} \le \tilde{\boldsymbol{C}}_L, \tag{45}$$

where $\tilde{\boldsymbol{C}}_L := (C_3 + \boldsymbol{C}_{L+1}T)L_2\tilde{\boldsymbol{C}}_{L+1}$ and $|v_{L+1}^t| \le C_3 + \boldsymbol{C}_{L+1}T$ by the upper bound of initialization in Assumption 3 and the $\boldsymbol{C}_{L+1} := L_1\tilde{\boldsymbol{C}}_{L+1}$-Lipschitz continuity of v_{L+1}^t in t. For each $\ell = L - 1, \dots, 1$, suppose $\underline{\mathcal{D}}_{\ell+1}^{\boldsymbol{\beta}}$ is uniformly bounded by $\tilde{\boldsymbol{C}}_{\ell+1}$. Then it follows from (11c) and (12b) that

$$\left\|\underline{\mathcal{D}}_{\ell+1}^{\boldsymbol{\alpha}}(\boldsymbol{\Theta}; \Phi, t)\right\|_\infty \le \left\|\underline{\mathcal{D}}_{\ell+1}^{\boldsymbol{\beta}}(\boldsymbol{\Theta}; \Phi, t)\right\|_\infty \left\|\dot{h}_2'\left(\boldsymbol{\alpha}_{\ell+1}^t(\boldsymbol{\Theta})\right)\right\|_\infty \le \tilde{\boldsymbol{C}}_{\ell+1}L_2 := \tilde{\boldsymbol{C}}_{\ell+1}',$$

and

$$\left|\underline{\mathcal{G}}_{\ell+1}^{\boldsymbol{v}}(\boldsymbol{\Theta}, \bar{\boldsymbol{\Theta}}; \Phi, t)\right| \le \left\|\underline{\mathcal{D}}_{\ell+1}^{\boldsymbol{\alpha}}(\boldsymbol{\Theta}; \Phi, t)\right\|_\infty \left\|\dot{h}_1\left(\boldsymbol{\beta}_\ell^t(\boldsymbol{\Theta})\right)\right\|_\infty \le L_1\boldsymbol{C}_{\ell+1}' := \boldsymbol{C}_{\ell+1}.$$

We then similarly apply the upper bounds of initialization in Assumption 3 and the $\boldsymbol{C}_{\ell+1}$-Lipschitz continuity of $v^t_{\ell+1}$ in t and obtain that

$$\begin{aligned}
&\left\| \int v^t_{\ell+1}(\boldsymbol{\Theta}, \bar{\boldsymbol{\Theta}})\, \underline{\mathcal{D}}^{\boldsymbol{\alpha}}_{\ell+1}(\bar{\boldsymbol{\Theta}}; \Phi, t)\, dp(\bar{\boldsymbol{\Theta}}) \right\|_\infty \qquad (46)\\
\leq\ & \tilde{\boldsymbol{C}}'_{\ell+1} \int \left| v^t_{\ell+1}(\boldsymbol{\Theta}, \bar{\boldsymbol{\Theta}}) \right| dp(\bar{\boldsymbol{\Theta}})\\
\leq\ & \tilde{\boldsymbol{C}}'_{\ell+1} \left(\boldsymbol{C}_{\ell+1} t + \int |v_{\ell+1}(\boldsymbol{\Theta}_\ell, \boldsymbol{\Theta}_{\ell+1})| dp_{\ell+1}(\boldsymbol{\Theta}_{\ell+1}) \right)\\
\leq\ & \tilde{\boldsymbol{C}}_{\ell+1}(\boldsymbol{C}_{\ell+1} T + C'),
\end{aligned}$$

for a constant C', where in the last inequality we used the upper bound of $v_{\ell+1}$ in (15), the sub-gaussian property of p, and Corollary 31. Consequently,

$$\begin{aligned}
&\left\| \underline{\mathcal{D}}^{\boldsymbol{\beta}}_{\ell}(\boldsymbol{\Theta}; \Phi, t) \right\|_\infty \qquad (47)\\
\leq\ & \left\| \underline{\mathcal{D}}^{\boldsymbol{\beta}}_{\ell+1}(\boldsymbol{\Theta}; \Phi, t) \right\|_\infty + \left\| \dot{h}'_1\left(\boldsymbol{\beta}^t_\ell(\boldsymbol{\Theta})\right) \circ \int v^t_{\ell+1}(\boldsymbol{\Theta}, \bar{\boldsymbol{\Theta}})\, \underline{\mathcal{D}}^{\boldsymbol{\alpha}}_{\ell+1}(\bar{\boldsymbol{\Theta}}; \Phi, t)\, dp(\bar{\boldsymbol{\Theta}}) \right\|_\infty\\
\leq\ & \tilde{\boldsymbol{C}}_{\ell+1} + (\tilde{\boldsymbol{C}}_{\ell+1}(\boldsymbol{C}_{\ell+1} T + C')) L_2 := \tilde{\boldsymbol{C}}_\ell.
\end{aligned}$$

Finally, denoting the infinity norm of the data $\boldsymbol{X}$ by C_x, (12c) gives the upper bound of $\left|\underline{\mathcal{G}}^{\boldsymbol{v}}_1(\boldsymbol{\Theta}; \Phi, t)\right|$ by $C_x \tilde{\boldsymbol{C}}_1 := \boldsymbol{C}_1$.

Now we turn to the forward steps. We prove that there is a constant C such that for $\ell \in [L]$ and $\boldsymbol{\Theta}$,

$$\left\| \underline{\mathcal{G}}^{\boldsymbol{\beta}}_{\ell}(\boldsymbol{\Theta}; \Phi, t) \right\|_\infty \leq C \left(\|\boldsymbol{\Theta}\|_\infty + 1 \right), \qquad (48)$$

and for all $\ell \in [2:L]$ and $\boldsymbol{\Theta}$,

$$\left\| \underline{\mathcal{G}}^{\boldsymbol{\alpha}}_{\ell}(\boldsymbol{\Theta}; \Phi, t) \right\|_\infty \leq C \left(\|\boldsymbol{\Theta}\|_\infty + 1 \right). \qquad (49)$$

Once we obtain (48) and (49), because p has bounded finite moment (Corollary 31), the dominated convergence theorem directly implies that $\Phi^{\boldsymbol{\alpha}}$ and $\Phi^{\boldsymbol{\beta}}$ satisfy the forward equations in Definition 14, which is our desired result.

For the first layer $\ell = 1$, since $\boldsymbol{X}$ is bounded, it follows from (13a) that

$$\left\| \underline{\mathcal{G}}^{\boldsymbol{\beta}}_{1}(\boldsymbol{\Theta}; \Phi, t) \right\|_\infty \leq C''_1 \leq C''_1 \left(\|\boldsymbol{\Theta}\|_\infty + 1 \right).$$

Suppose that at layer $\ell \in [L-1]$, we have $\left\| \underline{\mathcal{G}}^{\boldsymbol{\beta}}_{\ell}(\boldsymbol{\Theta}; \Phi, t) \right\|_\infty \leq C''_\ell \left(\|\boldsymbol{\Theta}\|_\infty + 1 \right)$. By a similar argument to (47), we have

$$\begin{aligned}
&\left\| \int v^t_{\ell+1}(\boldsymbol{\Theta}, \bar{\boldsymbol{\Theta}})\, \underline{\mathcal{G}}^{\boldsymbol{\beta}}_{\ell}(\bar{\boldsymbol{\Theta}}; \Phi, t)\, dp(\bar{\boldsymbol{\Theta}}) \right\|_\infty\\
\leq\ & \boldsymbol{C}''_\ell \int \left(\boldsymbol{C}_{\ell+1} T + \left| v_{\ell+1}(\boldsymbol{\Theta}, \bar{\boldsymbol{\Theta}}) \right| \right) \left(\|\bar{\boldsymbol{\Theta}}\|_\infty + 1 \right) dp(\bar{\boldsymbol{\Theta}})\\
\overset{(15)}{\leq}\ & \boldsymbol{C}''_\ell (\boldsymbol{C}_{\ell+1} T + C_1(\|\boldsymbol{\Theta}\|_\infty + 1)) \int (\|\bar{\boldsymbol{\Theta}}\|_\infty + 1) dp(\bar{\boldsymbol{\Theta}})\\
\leq\ & \tilde{\boldsymbol{C}}''_{\ell+1}(\|\boldsymbol{\Theta}\|_\infty + 1),
\end{aligned}$$

for some constant $\tilde{C}''_{\ell+1}$. Therefore, applying (13b) yields that

$$\begin{aligned}\left\|\underline{\mathcal{G}}^{\boldsymbol{\alpha}}_{\ell+1}(\boldsymbol{\Theta};\Phi,t)\right\|_\infty &\le L_2\tilde{C}''_{\ell+1}(\|\boldsymbol{\Theta}\|_\infty+1)+\int \underbrace{\left\|\dot{h}_1\left(\boldsymbol{\beta}^t_\ell\right)\right\|_\infty}_{\le L_1}\underbrace{\left|\underline{\mathcal{G}}^{v}_{\ell+1}(\boldsymbol{\Theta},\bar{\boldsymbol{\Theta}};\Phi,t)\right|}_{\le \boldsymbol{C}_\ell} dp(\bar{\boldsymbol{\Theta}})\\ &\le \bar{\boldsymbol{C}}_{\ell+1}(\|\boldsymbol{\Theta}\|_\infty+1), \end{aligned} \tag{50}$$

for some constant $\bar{\boldsymbol{C}}_{\ell+1}$. We obtain

$$\begin{aligned}&\left\|\underline{\mathcal{G}}^{\boldsymbol{\beta}}_{\ell+1}(\boldsymbol{\Theta};\Phi,t)\right\|_\infty\\ \le\ &\left\|\dot{h}'_2\left(\boldsymbol{\alpha}^t_{\ell+1}(\boldsymbol{\Theta})\right)\circ\underline{\mathcal{G}}^{\boldsymbol{\alpha}}_{\ell+1}(\boldsymbol{\Theta};\Phi,t)\right\|_\infty+\left\|\underline{\mathcal{G}}^{\boldsymbol{\beta}}_{\ell}(\boldsymbol{\Theta};\Phi,t)\right\|_\infty\\ \le\ &\left(L_2\bar{\boldsymbol{C}}_{\ell+1}+\boldsymbol{C}''_\ell\right)(\|\boldsymbol{\Theta}\|_\infty+1),\end{aligned}$$

which suggests to pick $\boldsymbol{C}''_{\ell+1}=(L_2\bar{\boldsymbol{C}}_{\ell+1}+\boldsymbol{C}''_\ell)$. We achieve Lemma 15. ■

Before proving Lemma 18, we first present in Lemma 21 properties of $\Phi\in\boldsymbol{\Phi}$ that will be used to prove the contraction lemma. The proof is exactly the same as Lemma 15 and is omitted.

Lemma 21 (Property of $\boldsymbol{\Phi}$) *There exists a generic constant C such that, for any $\Phi\in\boldsymbol{\Phi}$, we have*

- $\|\underline{\mathcal{D}}_{L+1}(\Phi,t)\|_\infty\le C$ *and* $\|\underline{\mathcal{D}}^{\boldsymbol{\beta}}_\ell(\boldsymbol{\Theta};\Phi,t)\|_\infty\le C$ *for* $\ell\in[L]$;
- $\|\underline{\mathcal{D}}^{\boldsymbol{\alpha}}_\ell(\boldsymbol{\Theta};\Phi,t)\|_\infty\le C$ *for* $\ell\in[2:L]$;
- $\|\underline{\mathcal{G}}^{\boldsymbol{v}}_\ell(\boldsymbol{u}_\ell;\Phi,t)\|_\infty\le C$ *and* $\|v^t_\ell(\boldsymbol{u}_\ell)\|_\infty\le\|v^0_\ell(\boldsymbol{u}_\ell)\|_\infty+C\,t$ *for* $\ell\in[L+1]$;
- $\|\underline{\mathcal{G}}^{\boldsymbol{\beta}}_\ell(\boldsymbol{\Theta};\Phi,t)\|_\infty\le C\,(\|\boldsymbol{\Theta}\|_\infty+1)$ *for* $\ell\in[L]$;
- $\|\underline{\mathcal{G}}^{\boldsymbol{\alpha}}_\ell(\boldsymbol{\Theta};\Phi,t)\|_\infty\le C\,(\|\boldsymbol{\Theta}\|_\infty+1)$ *for* $\ell\in[2:L]$.

Proof [Proof of Lemma 18] The proof entails upper bounds of the gradient differences $\|\underline{\mathcal{G}}^{\boldsymbol{v}}_\ell(\boldsymbol{u}_\ell;\Phi_1,t)-\underline{\mathcal{G}}^{\boldsymbol{v}}_\ell(\boldsymbol{u}_\ell;\Phi_2,t)\|_\infty$ in terms of the differences $|v^t_{1,\ell}-v^t_{2,\ell}|$ for $\ell\in[L+1]$, which can be further upper bounded in terms of $d_t:=\mathrm{D}^{[0,t]}(\Phi_1,\Phi_2)$, that is by Definition 17:

$$\left\|v^t_{1,\ell}(\boldsymbol{u}_\ell)-v^t_{2,\ell}(\boldsymbol{u}_\ell)\right\|_\infty\le(\|\boldsymbol{u}_\ell\|_\infty+1)d_t,\quad \ell\in[L+1]. \tag{51}$$

We will use the forward equations to inductively upper bound the differences between $\boldsymbol{\beta}_\ell$ and $\boldsymbol{\alpha}_\ell$. Especially, for some constant C, we prove

$$\left\|\boldsymbol{\beta}^t_{1,\ell}(\boldsymbol{\Theta})-\boldsymbol{\beta}^t_{2,\ell}(\boldsymbol{\Theta})\right\|_\infty\le C\,(\|\boldsymbol{\Theta}\|_\infty+1)\,d_t,\quad \ell\in[L], \tag{52}$$

$$\left\|\boldsymbol{\alpha}^t_{1,\ell}(\boldsymbol{\Theta})-\boldsymbol{\alpha}^t_{2,\ell}(\boldsymbol{\Theta})\right\|_\infty\le C\,(\|\boldsymbol{\Theta}\|_\infty+1)\,d_t,\quad \ell\in[2:L]. \tag{53}$$

We then use the backward equations to upper bound the difference between $\underline{\mathcal{D}}^{\boldsymbol{\beta}}_\ell$, $\underline{\mathcal{D}}^{\boldsymbol{\alpha}}_\ell$, $\underline{\mathcal{G}}^{\boldsymbol{v}}_\ell$. Namely, we prove

$$\left\|\underline{\mathcal{D}}^{\boldsymbol{\beta}}_\ell(\boldsymbol{\Theta};\Phi_1,t)-\underline{\mathcal{D}}^{\boldsymbol{\beta}}_\ell(\boldsymbol{\Theta};\Phi_2,t)\right\|_\infty\le C(1+\|\boldsymbol{\Theta}\|_\infty)d_t,\quad \ell\in[L], \tag{54}$$

$$\|\underline{\mathcal{D}}^{\boldsymbol{\alpha}}_\ell(\boldsymbol{\Theta};\Phi_1,t)-\underline{\mathcal{D}}^{\boldsymbol{\alpha}}_\ell(\boldsymbol{\Theta};\Phi_2,t)\|_\infty\le C(1+\|\boldsymbol{\Theta}\|_\infty)d_t,\quad \ell\in[2:L], \tag{55}$$

$$\left\|\underline{\mathcal{G}}^{\boldsymbol{v}}_\ell(\boldsymbol{u}_\ell;\Phi_1,t)-\underline{\mathcal{G}}^{\boldsymbol{v}}_\ell(\boldsymbol{u}_\ell;\Phi_2,t)\right\|_\infty\le C(1+\|\boldsymbol{u}_\ell\|_\infty)d_t,\quad \ell\in[L+1], \tag{56}$$

Then the conclusion,i.e., Lemma 18 directly follows from (56), and the definitions of F and $\mathrm{D}^{[0,t]}$ in Definitions 16 and 17, respectively.

We consider forward steps. When $\ell = 1$, because $\boldsymbol{X}$ is bounded, by the Definition 16 (2) and (51), we have $\left\|\boldsymbol{\beta}_{1,1}^t(\boldsymbol{\Theta}) - \boldsymbol{\beta}_{2,1}^t(\boldsymbol{\Theta})\right\|_\infty \leq C\left(\|\boldsymbol{\Theta}\|_\infty + 1\right) d_t$. Consider at layer $\ell \in [L-1]$, we obtain $\left\|\boldsymbol{\beta}_{1,\ell}^t(\boldsymbol{\Theta}) - \boldsymbol{\beta}_{2,\ell}^t(\boldsymbol{\Theta})\right\|_\infty \leq C\left(\|\boldsymbol{\Theta}\|_\infty + 1\right) d_t$. For layer $\ell+1$, by the boundedness and Lipschitz continuity of h_1, we have

$$
\begin{aligned}
& \left\|\boldsymbol{\alpha}_{1,\ell+1}^t(\boldsymbol{\Theta}) - \boldsymbol{\alpha}_{2,\ell+1}^t(\boldsymbol{\Theta})\right\|_\infty \\
\leq\ & \left\| \int \dot{h}_1\big(\boldsymbol{\beta}_{1,\ell}^t(\bar{\boldsymbol{\Theta}})\big) v_{1,\ell+1}^t(\bar{\boldsymbol{\Theta}}, \boldsymbol{\Theta}) \ - \dot{h}_1\big(\boldsymbol{\beta}_{2,\ell}^t(\bar{\boldsymbol{\Theta}})\big)\ v_{2,\ell+1}^t(\bar{\boldsymbol{\Theta}}, \boldsymbol{\Theta}) dp(\bar{\boldsymbol{\Theta}}) \right\|_\infty \\
\leq\ & \int \underbrace{\left\| \dot{h}_1\big(\boldsymbol{\beta}_{1,\ell}^t(\bar{\boldsymbol{\Theta}})\big) - \dot{h}_1\big(\boldsymbol{\beta}_{2,\ell}^t(\bar{\boldsymbol{\Theta}})\big) \right\|_\infty}_{\leq L_2 C\left(\|\bar{\boldsymbol{\Theta}}\|+1\right) d_t} \left| v_{1,\ell+1}^t(\bar{\boldsymbol{\Theta}}, \boldsymbol{\Theta}) \right| dp(\bar{\boldsymbol{\Theta}}) \\
& \qquad + \int \underbrace{\left\| \dot{h}_1\big(\boldsymbol{\beta}_{2,\ell}^t(\bar{\boldsymbol{\Theta}})\big) \right\|_\infty}_{\leq L_1} \underbrace{\left| v_{1,\ell+1}^t(\bar{\boldsymbol{\Theta}}, \boldsymbol{\Theta}) - v_{2,\ell+1}^t(\bar{\boldsymbol{\Theta}}, \boldsymbol{\Theta}) \right|}_{\leq \left(\|\bar{\boldsymbol{\Theta}}\|_\infty + \|\boldsymbol{\Theta}\|_\infty + 1\right) d_t} dp(\bar{\boldsymbol{\Theta}}) \\
\leq\ & C'\left(\|\boldsymbol{\Theta}\|_\infty + 1\right) d_t,
\end{aligned}
$$

for a constant C', where the last step is due to the sub-gaussianness of p, Corollary 31, and the upper bound of $v_{\ell+1}^t$ in Lemma 21. Consequently, the Lipschitz continuity of L_2 gives that

$$
\begin{aligned}
& \left\|\boldsymbol{\beta}_{1,\ell+1}^t(\boldsymbol{\Theta}) - \boldsymbol{\beta}_{2,\ell+1}^t(\boldsymbol{\Theta})\right\|_\infty \\
\leq\ & \left\| \dot{h}_2\big(\boldsymbol{\alpha}_{1,\ell+1}^t(\boldsymbol{\Theta})\big) - \dot{h}_2\big(\boldsymbol{\alpha}_{2,\ell+1}^t(\boldsymbol{\Theta})\big) \right\|_\infty + \left\|\boldsymbol{\beta}_{1,\ell}^t(\boldsymbol{\Theta}) - \boldsymbol{\beta}_{2,\ell}^t(\boldsymbol{\Theta})\right\|_\infty \\
\leq\ & C''\left(\|\boldsymbol{\Theta}\|_\infty + 1\right) d_t
\end{aligned}
$$

for a constant C''. We achieve (52) and (53).

We turn to the backward steps. We focus on the upper bound of the difference between $\underline{\mathcal{D}}_\ell^{\boldsymbol{\alpha}}$ and $\underline{\mathcal{D}}_\ell^{\boldsymbol{\beta}}$. Since both h_1 and $\boldsymbol{X}$ are bounded, h_1 is Lipschitz continuous, and $\underline{\mathcal{D}}_\ell^{\boldsymbol{\alpha}}$ is bounded by Lemma 21, the upper bound of the difference between $\underline{\mathcal{G}}_\ell^{\boldsymbol{v}}$ follows immediately. To begin with, we introduce

$$
\underline{\mathcal{D}}_\ell^{\gamma}(\boldsymbol{\Theta}; \Phi, t) := \int v_\ell^t(\boldsymbol{\Theta}, \bar{\boldsymbol{\Theta}})\, \underline{\mathcal{D}}_\ell^{\boldsymbol{\alpha}}(\bar{\boldsymbol{\Theta}}; \Phi, t) dp(\bar{\boldsymbol{\Theta}}) \circ \dot{h}_1'\big(\boldsymbol{\beta}_{\ell-1}^t(\boldsymbol{\Theta})\big), \quad \ell \in [2:L]. \tag{57}
$$

Then we have

$$
\underline{\mathcal{D}}_{\ell-1}^{\boldsymbol{\beta}}(\boldsymbol{\Theta}; \Phi, t) = \underline{\mathcal{D}}_\ell^{\boldsymbol{\beta}}(\boldsymbol{\Theta}; \Phi, t) + \underline{\mathcal{D}}_\ell^{\gamma}(\boldsymbol{\Theta}; \Phi, t), \quad \ell \in [2:L].
$$

For the top layer $\ell = L+1$, the Lipschitz continuity of ϕ_1' in Assumption 1 implies that,

$$
\begin{aligned}
& \|\underline{\mathcal{D}}_{L+1}(\Phi_1, t) - \underline{\mathcal{D}}_{L+1}(\Phi_2, t)\|_\infty \\
\leq\ & L_5 \|\boldsymbol{\beta}_{1,L+1}^t - \boldsymbol{\beta}_{2,L+1}^t\|_\infty \\
\leq\ & L_5 \int \|h_1(\boldsymbol{\beta}_{1,L}^t) v_{1,L+1}^t(\boldsymbol{\Theta}) - h_1(\boldsymbol{\beta}_{2,L}^t) v_{2,L+1}^t(\boldsymbol{\Theta})\|_\infty dp(\boldsymbol{\Theta}).
\end{aligned} \tag{58}
$$

Since h_1 is bounded and Lipschitz continuous, $v^t_{i,L+1}$ is bounded for $t \le T$ by Lemma 21, and (52), we have

$$\|\underline{\mathcal{D}}_{L+1}(\Phi_1, t) - \underline{\mathcal{D}}_{L+1}(\Phi_2, t)\|_\infty \le C d_t,$$

for a constant C. At layer $\ell = L$, recall that

$$\underline{\mathcal{D}}_L(\boldsymbol{\Theta}; \Phi, t) = v^t_{L+1}(\boldsymbol{\Theta})\, \underline{\mathcal{D}}_{L+1}(\Phi, t) \circ \dot{h}'_1\left(\boldsymbol{\beta}^t_L\right).$$

Since the three terms in the product are all bounded, and h'_1 is L_3-Lipschitz continuous, we have

$$\|\underline{\mathcal{D}}^{\boldsymbol{\beta}}_L(\boldsymbol{\Theta}; \Phi_1, t) - \underline{\mathcal{D}}^{\boldsymbol{\beta}}_L(\boldsymbol{\Theta}; \Phi_2, t)\|_\infty \le C'(1 + \|\boldsymbol{\Theta}\|_\infty) d_t, \tag{59}$$

for a constant C'. For each $\ell = L-1, \ldots, 1$, suppose there is a constant C, such that

$$\left\|\underline{\mathcal{D}}^{\boldsymbol{\beta}}_{\ell+1}(\boldsymbol{\Theta}; \Phi_1, t) - \underline{\mathcal{D}}^{\boldsymbol{\beta}}_{\ell+1}(\boldsymbol{\Theta}; \Phi_2, t)\right\|_\infty \le C\left(\|\boldsymbol{\Theta}\|_\infty + 1\right) d_t.$$

As h'_2 is L_3-Lipschitz continuous, the boundednesses of h_2 and $\underline{\mathcal{D}}^{\boldsymbol{\beta}}_{\ell+1}$ in Lemma 21 and (53) give that

$$\left\|\underline{\mathcal{D}}^{\boldsymbol{\alpha}}_{\ell+1}(\boldsymbol{\Theta}; \Phi_1, t) - \underline{\mathcal{D}}^{\boldsymbol{\alpha}}_{\ell+1}(\boldsymbol{\Theta}; \Phi_2, t)\right\|_\infty \le C'\left(\|\boldsymbol{\Theta}\|_\infty + 1\right) d_t,$$

for a constant C'. Moreover,

$$\begin{aligned}
& \int \|v^t_{1,\ell+1}(\boldsymbol{\Theta}, \bar{\boldsymbol{\Theta}})\, \underline{\mathcal{D}}^{\boldsymbol{\alpha}}_{\ell+1}(\bar{\boldsymbol{\Theta}}; \Phi_1, t) - v^t_{2,\ell+1}(\boldsymbol{\Theta}, \bar{\boldsymbol{\Theta}})\, \underline{\mathcal{D}}^{\boldsymbol{\alpha}}_{\ell+1}(\bar{\boldsymbol{\Theta}}; \Phi_2, t)\|_\infty dp(\bar{\boldsymbol{\Theta}}) \\
\le & \int \underbrace{\left|v^t_{1,\ell+1}(\boldsymbol{\Theta}, \bar{\boldsymbol{\Theta}}) - v^t_{2,\ell+1}(\boldsymbol{\Theta}, \bar{\boldsymbol{\Theta}})\right|}_{\le (\|\boldsymbol{\Theta}\|_\infty + \|\bar{\boldsymbol{\Theta}}\|_\infty + 1) d_t} \left\|\underline{\mathcal{D}}^{\boldsymbol{\alpha}}_{\ell+1}(\bar{\boldsymbol{\Theta}}; \Phi_1, t)\right\|_\infty \\
& \qquad + \left|v^t_{2,\ell+1}(\boldsymbol{\Theta}, \bar{\boldsymbol{\Theta}})\right| \underbrace{\left\|\underline{\mathcal{D}}^{\boldsymbol{\alpha}}_{\ell+1}(\bar{\boldsymbol{\Theta}}; \Phi_1, t) - \underline{\mathcal{D}}^{\boldsymbol{\alpha}}_{\ell+1}(\bar{\boldsymbol{\Theta}}; \Phi_2, t)\right\|_\infty}_{\le \tilde{C}'(\|\bar{\boldsymbol{\Theta}}\|_\infty + 1) d_t} dp(\bar{\boldsymbol{\Theta}}) \\
\le & C''(\|\boldsymbol{\Theta}\|_\infty + 1) d_t,
\end{aligned} \tag{60}$$

for a constant C'', where the last step is due to the sub-gaussianness of p, Corollary 31, and the upper bounds of $\underline{\mathcal{D}}^{\boldsymbol{\alpha}}_{\ell+1}$ and $v^t_{\ell+1}$ in Lemma 21. Consequently,

$$\left\|\underline{\mathcal{D}}^{\boldsymbol{\gamma}}_{\ell+1}(\boldsymbol{\Theta}; \Phi_1, t) - \underline{\mathcal{D}}^{\boldsymbol{\gamma}}_{\ell+1}(\boldsymbol{\Theta}; \Phi_2, t)\right\|_\infty \le \bar{C}\left(\|\boldsymbol{\Theta}\|_\infty + 1\right) d_t,$$

for a constant $\bar{C}$, which further implies

$$\begin{aligned}
& \left\|\underline{\mathcal{D}}^{\boldsymbol{\beta}}_{\ell}(\boldsymbol{\Theta}; \Phi_1, t) - \underline{\mathcal{D}}^{\boldsymbol{\beta}}_{\ell}(\boldsymbol{\Theta}; \Phi_2, t)\right\|_\infty \\
\le & \left\|\underline{\mathcal{D}}^{\boldsymbol{\beta}}_{\ell+1}(\boldsymbol{\Theta}; \Phi_1, t) - \underline{\mathcal{D}}^{\boldsymbol{\beta}}_{\ell+1}(\boldsymbol{\Theta}; \Phi_2, t)\right\|_\infty + \left\|\underline{\mathcal{D}}^{\boldsymbol{\gamma}}_{\ell+1}(\boldsymbol{\Theta}; \Phi_1, t) - \underline{\mathcal{D}}^{\boldsymbol{\gamma}}_{\ell+1}(\boldsymbol{\Theta}; \Phi_2, t)\right\|_\infty \\
\le & \bar{C}'\left(\|\boldsymbol{\Theta}\|_\infty + 1\right) d_t,
\end{aligned}$$

for a constant $\bar{C}'$. We finish the proof. ■

Lemma 22 $\boldsymbol{\Phi}$ *is complete under* $\mathrm{D}^{[0,T]}$.

Proof Let $\{\Phi_n : n \geq 0\}$ be a Cauchy sequence under $\mathrm{D}^{[0,T]}$. Then $\frac{\Phi^{\boldsymbol{v}}_{n,\ell}(\boldsymbol{u}_\ell)(t)}{1+\|\boldsymbol{u}_\ell\|_\infty}$ converges uniformly under the ℓ_∞-norm. Let $\Phi^{\boldsymbol{v}}_{*,\ell}(\boldsymbol{u}_\ell)(t) = \lim_{n\to\infty} \Phi^{\boldsymbol{v}}_{n,\ell}(\boldsymbol{u}_\ell)(t)$ for $\ell \in [L+1]$. Since the Lipschitz continuity is preserved under the pointwise convergence, we have $\Phi^{\boldsymbol{v}}_{*,\ell}$ is $\boldsymbol{C}$-Lipschitz continuous in t. Let

$$\begin{aligned}
\Phi^{\boldsymbol{\beta}}_{*,1}(\boldsymbol{\Theta})(t) &= \frac{1}{d}\boldsymbol{X}\Phi^{\boldsymbol{v}}_{*,1}(\boldsymbol{\Theta})(t),\\
\Phi^{\boldsymbol{\alpha}}_{*,\ell+1}(\boldsymbol{\Theta})(t) &= \int \Phi^{\boldsymbol{v}}_{*,\ell+1}(\boldsymbol{\Theta},\bar{\boldsymbol{\Theta}})(t)\, \dot{h}_1\left(\Phi^{\boldsymbol{\beta}}_{*,\ell}(\bar{\boldsymbol{\Theta}})(t)\right) dp(\bar{\boldsymbol{\Theta}}), \quad \ell \in [L],\\
\Phi^{\boldsymbol{\beta}}_{*,\ell+1}(\boldsymbol{\Theta})(t) &= \Phi^{\boldsymbol{\beta}}_{*,\ell}(\boldsymbol{\Theta})(t) + \dot{h}_2\left(\Phi^{\boldsymbol{\alpha}}_{*,\ell+1}(\boldsymbol{\Theta})(t)\right), \quad \ell \in [L].
\end{aligned}$$

By the dominated convergence theorem, we have $\Phi^{\boldsymbol{\beta}}_{*,\ell}(\boldsymbol{\Theta})(t) = \lim_{n\to\infty} \Phi^{\boldsymbol{\beta}}_{n,\ell}(\boldsymbol{\Theta})(t)$ and $\Phi^{\boldsymbol{\alpha}}_{*,\ell}(\boldsymbol{\Theta})(t) = \lim_{n\to\infty} \Phi^{\boldsymbol{\alpha}}_{n,\ell}(\boldsymbol{\Theta})(t)$. Then Φ_* is a limit point of $\{\Phi_n : n \geq 0\}$ under $\mathrm{D}^{[0,T]}$ and $\Phi_* \in \boldsymbol{\Phi}$. ■

Next we prove lemmas for Theorem 13.

Proof [Proof of Lemma 20] Analogous to the notation of $\boldsymbol{u}_\ell$, for the convenience of presenting continuity of $\Phi^{\boldsymbol{v}}_\ell$, we introduce notations $\bar{\boldsymbol{u}}$ and $\bar{\boldsymbol{u}}'_\ell$ by letting

$$\bar{\boldsymbol{u}}_\ell = \begin{cases} \bar{\boldsymbol{\Theta}}_1, \\ (\bar{\boldsymbol{\Theta}}_1, \boldsymbol{\Theta}_2), \\ \bar{\boldsymbol{\Theta}}_1 \end{cases} \qquad \bar{\boldsymbol{u}}'_\ell = \begin{cases} \bar{\boldsymbol{\Theta}}_1, & \ell = 1, \\ (\boldsymbol{\Theta}_1, \bar{\boldsymbol{\Theta}}_2), & \ell \in [2:L], \\ \bar{\boldsymbol{\Theta}}_1 & \ell = L+1. \end{cases}$$

We also abbreviate the notations for the individual trajectories as:

$$\Phi^{\boldsymbol{v}}_\ell(\boldsymbol{u}_\ell)(t) = v^t_\ell(\boldsymbol{u}_\ell), \qquad \Phi^{\boldsymbol{v}}_\ell(\bar{\boldsymbol{u}}_\ell)(t) = v^t_\ell(\bar{\boldsymbol{u}}_\ell), \qquad \Phi^{\boldsymbol{v}}_\ell(\bar{\boldsymbol{u}}'_\ell)(t) = v^t_\ell(\bar{\boldsymbol{u}}'_\ell),$$

$\Phi^{\boldsymbol{\beta}}_{\ell_1}(\boldsymbol{\Theta}_1)(t) = \boldsymbol{\beta}^t_{\ell_1}(\boldsymbol{\Theta}_1)$, $\Phi^{\boldsymbol{\beta}}_{\ell_1}(\bar{\boldsymbol{\Theta}}_1)(t) = \boldsymbol{\beta}^t_{\ell_1}(\bar{\boldsymbol{\Theta}}_1)$, $\Phi^{\boldsymbol{\alpha}}_{\ell_2}(\boldsymbol{\Theta}_1)(t) = \boldsymbol{\alpha}^t_{\ell_2}(\boldsymbol{\Theta}_1)$, $\Phi^{\boldsymbol{\alpha}}_{\ell_2}(\bar{\boldsymbol{\Theta}}_1)(t) = \boldsymbol{\alpha}^t_{\ell_2}(\bar{\boldsymbol{\Theta}}_1)$, for $\ell \in [L+1]$, $\ell_1 \in [L]$, and $\ell_2 \in [2:L]$, respectively.

We first investigate the set $F(\boldsymbol{\Phi} \cap \boldsymbol{\Phi}_b)$ for a general b. We follow similar steps as the proof of Lemma 18. We first consider forward steps and inductively show upper bound for the differences between $\boldsymbol{\beta}_\ell$ and $\boldsymbol{\alpha}_\ell$. Namely, we prove for any $\Phi \in \boldsymbol{\Phi} \cap \boldsymbol{\Phi}_b$,

$$\left\|\boldsymbol{\beta}^t_\ell(\boldsymbol{\Theta}_1) - \boldsymbol{\beta}^t_\ell(\bar{\boldsymbol{\Theta}}_1)\right\|_\infty \leq Ce^{bt}\left(\|\boldsymbol{\Theta}_1\|_\infty + 1\right)\|\boldsymbol{\Theta}_1 - \bar{\boldsymbol{\Theta}}_1\|_\infty, \quad \ell \in [L], \tag{61}$$

$$\left\|\boldsymbol{\alpha}^t_\ell(\boldsymbol{\Theta}_1) - \boldsymbol{\alpha}^t_\ell(\bar{\boldsymbol{\Theta}}_1)\right\|_\infty \leq Ce^{bt}\left(\|\boldsymbol{\Theta}_1\|_\infty + 1\right)\|\boldsymbol{\Theta}_1 - \bar{\boldsymbol{\Theta}}_1\|_\infty, \quad \ell \in [2:L], \tag{62}$$

for a constant C. Then we study the backward steps, and prove that there is a constant $\tilde{C}$ independent of b such that for any $\Phi \in \boldsymbol{\Phi} \cap \boldsymbol{\Phi}_b$,

$$\|\underline{\mathcal{D}}^{\boldsymbol{\beta}}_\ell(\boldsymbol{\Theta}_1;\Phi,t) - \underline{\mathcal{D}}^{\boldsymbol{\beta}}_\ell(\bar{\boldsymbol{\Theta}}_1;\Phi,t)\|_\infty \leq \tilde{C}e^{bt}\left(\|\boldsymbol{\Theta}_1\|_\infty + 1\right)\|\boldsymbol{\Theta}_1 - \bar{\boldsymbol{\Theta}}_1\|_\infty, \qquad \ell \in [L], \tag{63}$$

$$\|\underline{\mathcal{D}}^{\boldsymbol{\alpha}}_\ell(\boldsymbol{\Theta}_1;\Phi,t) - \underline{\mathcal{D}}^{\boldsymbol{\alpha}}_\ell(\bar{\boldsymbol{\Theta}}_1;\Phi,t)\|_\infty \leq \tilde{C}e^{bt}\left(\|\boldsymbol{\Theta}_1\|_\infty + 1\right)\|\boldsymbol{\Theta}_1 - \bar{\boldsymbol{\Theta}}_1\|_\infty, \qquad \ell \in [2:L], \tag{64}$$

$$\|\underline{\mathcal{G}}^{\boldsymbol{v}}_\ell(\boldsymbol{u}_\ell;\Phi,t) - \underline{\mathcal{G}}^{\boldsymbol{v}}_\ell(\bar{\boldsymbol{u}}_\ell;\Phi,t)\|_\infty \leq \tilde{C}e^{bt}\left(1 + \|\boldsymbol{u}_\ell\|_\infty\right)\|\boldsymbol{u}_\ell - \bar{\boldsymbol{u}}_\ell\|_\infty, \qquad \ell \in [L+1], \tag{65}$$

$$\|\underline{\mathcal{G}}^{\boldsymbol{v}}_\ell(\boldsymbol{u}_\ell;\Phi,t) - \underline{\mathcal{G}}^{\boldsymbol{v}}_\ell(\bar{\boldsymbol{u}}'_\ell;\Phi,t)\|_\infty \leq \tilde{C}e^{bt}\left(1 + \|\boldsymbol{u}_\ell\|_\infty\right)\|\boldsymbol{u}_\ell - \bar{\boldsymbol{u}}'_\ell\|_\infty, \qquad \ell \in [L+1]. \tag{66}$$

In forward steps, we prove (61) and (62). For the 1-st layer, because $\boldsymbol{X}$ is bounded, applying (44a) with the formula of $\boldsymbol{\beta}^t_1$ as Definition 16 (2) yields (61) with $\ell = 1$. Suppose at layer $\ell \in [L-1]$, we have

$$\left\|\boldsymbol{\beta}^t_\ell(\boldsymbol{\Theta}_1) - \boldsymbol{\beta}^t_\ell(\bar{\boldsymbol{\Theta}}_1)\right\|_\infty \leq C'e^{bt}\left(\|\boldsymbol{\Theta}_1\|_\infty + 1\right)\|\boldsymbol{\Theta}_1 - \bar{\boldsymbol{\Theta}}_1\|_\infty$$

for a constant C'. It follows that

$$\begin{aligned}\left\|\boldsymbol{\alpha}_{\ell+1}^t(\boldsymbol{\Theta}_2)-\boldsymbol{\alpha}_{\ell+1}^t(\bar{\boldsymbol{\Theta}}_2)\right\|_\infty &\le \int \underbrace{\left\|\dot{h}_1\left(\boldsymbol{\beta}_\ell^t(\boldsymbol{\Theta}_1)\right)\right\|_\infty}_{\le L_1} \underbrace{\left|v_{\ell+1}^t(\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2)-v_{\ell+1}^t(\boldsymbol{\Theta}_1,\bar{\boldsymbol{\Theta}}_2)\right|}_{(1+C_2)e^{bt}(\|\boldsymbol{\Theta}_1\|_\infty+\|\boldsymbol{\Theta}_2\|_\infty+1)\|\boldsymbol{\Theta}_2-\bar{\boldsymbol{\Theta}}_2\|_\infty} dp(\boldsymbol{\Theta}_1)\\ &\le C''e^{bt}(\|\boldsymbol{\Theta}_2\|_\infty+1)\|\boldsymbol{\Theta}_2-\bar{\boldsymbol{\Theta}}_2\|_\infty,\end{aligned}$$

for a constant C'', where we use sub-gaussianness of p and Corollary 31. We conclude that

$$\begin{aligned}\left\|\boldsymbol{\beta}_{\ell+1}^t(\boldsymbol{\Theta}_1)-\boldsymbol{\beta}_{\ell+1}^t(\bar{\boldsymbol{\Theta}}_1)\right\|_\infty &\le \left\|\dot{h}_2\left(\boldsymbol{\alpha}_{\ell+1}^t(\boldsymbol{\Theta}_1)\right)-\dot{h}_2\left(\boldsymbol{\alpha}_{\ell+1}^t(\bar{\boldsymbol{\Theta}}_1)\right)\right\|_\infty+\left\|\boldsymbol{\beta}_\ell^t(\boldsymbol{\Theta}_1)-\boldsymbol{\beta}_\ell^t(\bar{\boldsymbol{\Theta}}_1)\right\|_\infty\\ &\le C'''e^{bt}(\|\boldsymbol{\Theta}_1\|_\infty+1)\|\boldsymbol{\Theta}_1-\bar{\boldsymbol{\Theta}}_1\|_\infty.\end{aligned}$$

Therefore, (61) and (62) is achieved.

We turn to backward process. Again, we focus on the difference between $\underline{\mathcal{D}}_\ell^{\boldsymbol{\beta}}$ and $\underline{\mathcal{D}}_\ell^{\boldsymbol{\alpha}}$, i.e., (63) and (64). Then the upper bound for the difference between $\underline{\mathcal{G}}_\ell^{\boldsymbol{v}}$, i.e., (65) and (66) follows immediately. For example, for the top layer $\ell=L+1$, because of the boundedness of h_1, $\underline{\mathcal{D}}_{L+1}$, applying (63) with the formula of $\underline{\mathcal{G}}_{L+1}^{\boldsymbol{v}}$ (12a) gives that

$$\left|\underline{\mathcal{G}}_{L+1}^{\boldsymbol{v}}(\boldsymbol{\Theta}_1;\Phi,t)-\underline{\mathcal{G}}_{L+1}^{\boldsymbol{v}}(\bar{\boldsymbol{\Theta}}_1;\Phi,t)\right|\le Ce^{bt}(\|\boldsymbol{\Theta}_1\|_\infty+1)\|\boldsymbol{\Theta}_1-\bar{\boldsymbol{\Theta}}_1\|_\infty.$$

Other layers can be analogously obtained. At layer $\ell=L$, recall that

$$\underline{\mathcal{D}}_L(\boldsymbol{\Theta};\Phi,t)=v_{L+1}^t(\boldsymbol{\Theta})\,\underline{\mathcal{D}}_{L+1}(\Phi,t)\circ\dot{h}_1'\left(\boldsymbol{\beta}_L^t\right).$$

For, $\Phi\in\boldsymbol{\Phi}_b$, the upper bound for $|v_{L+1}^t(\boldsymbol{\Theta}_1)-v_{L+1}^t(\bar{\boldsymbol{\Theta}}_1)|$ is given in (44b). Then applying the Lipschitz continuity of h_1', (63), and the boundednesses of v_{L+1}^t, $\underline{\mathcal{D}}_{L+1}$ in Lemma 21 and h_1' yields

$$\left\|\underline{\mathcal{D}}_L^{\boldsymbol{\beta}}(\boldsymbol{\Theta}_1;\Phi,t)-\underline{\mathcal{D}}_L^{\boldsymbol{\beta}}(\bar{\boldsymbol{\Theta}}_1;\Phi,t)\right\|_\infty\le C'e^{bt}(\|\boldsymbol{\Theta}_1\|_\infty+1)\|\boldsymbol{\Theta}_1-\bar{\boldsymbol{\Theta}}_1\|_\infty,\tag{67}$$

for a constant C'.

For each $\ell=L-1,L-2,\cdots,1$, suppose we have

$$\left\|\underline{\mathcal{D}}_{\ell+1}^{\boldsymbol{\beta}}(\boldsymbol{\Theta}_1;\Phi,t)-\underline{\mathcal{D}}_{\ell+1}^{\boldsymbol{\beta}}(\bar{\boldsymbol{\Theta}}_1;\Phi,t)\right\|_\infty\le Ce^{bt}\left(\|\boldsymbol{\Theta}_1\|_\infty+1\right)\|\boldsymbol{\Theta}_1-\bar{\boldsymbol{\Theta}}_1\|_\infty,$$

for a constant C. Combining the Lipschitz continuity of h_2' with the boundednesses of h_2', $\underline{\mathcal{D}}_{\ell+1}^{\boldsymbol{\beta}}$ in Lemma 21 gives that

$$\begin{aligned}&\left\|\underline{\mathcal{D}}_{\ell+1}^{\boldsymbol{\alpha}}(\boldsymbol{\Theta}_1;\Phi,t)-\underline{\mathcal{D}}_{\ell+1}^{\boldsymbol{\alpha}}(\bar{\boldsymbol{\Theta}}_1;\Phi,t)\right\|_\infty\\ =\;&\left\|\dot{h}_2'\left(\boldsymbol{\alpha}_{\ell+1}^t(\boldsymbol{\Theta}_1)\right)\circ\underline{\mathcal{D}}_{\ell+1}^{\boldsymbol{\beta}}(\boldsymbol{\Theta}_1;\Phi,t)-\dot{h}_2'\left(\boldsymbol{\alpha}_{\ell+1}^t(\bar{\boldsymbol{\Theta}}_1)\right)\circ\underline{\mathcal{D}}_{\ell+1}^{\boldsymbol{\beta}}(\bar{\boldsymbol{\Theta}}_1;\Phi,t)\right\|_\infty\\ \le\;&C'e^{bt}(\|\boldsymbol{\Theta}_1\|_\infty+1)\|\boldsymbol{\Theta}_1-\bar{\boldsymbol{\Theta}}_1\|_\infty,\end{aligned}$$

for a constant C'. Moreover,

$$\begin{aligned}&\int \underbrace{\left|v_{\ell+1}^t(\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2)-v_{\ell+1}^t(\bar{\boldsymbol{\Theta}}_1,\boldsymbol{\Theta}_2)\right|}_{\le Ce^{bt}(\|\boldsymbol{\Theta}_1\|_\infty+\|\boldsymbol{\Theta}_2\|_\infty+1)\|\boldsymbol{\Theta}_1-\bar{\boldsymbol{\Theta}}_1\|_\infty}\left\|\underline{\mathcal{D}}_{\ell+1}^{\boldsymbol{\alpha}}(\boldsymbol{\Theta}_2;\Phi,t)\right\|_\infty dp(\boldsymbol{\Theta}_2)\\ \le\;&C'e^{bt}(\|\boldsymbol{\Theta}_1\|_\infty+1)\|\boldsymbol{\Theta}_1-\bar{\boldsymbol{\Theta}}_1\|_\infty,\end{aligned}\tag{68}$$

where in the last step we used the sub-gaussianness of p, Corollary 31, and upper bound of $\overline{\mathcal{D}}^{\boldsymbol{\alpha}}_{\ell+1}$ in Lemma 21. Then, by the upper bound in (46), boundedness and Lipschitz continuity of h_1', we obtain from (57) that

$$\left\|\underline{\mathcal{D}}^{\gamma}_{\ell+1}(\boldsymbol{\Theta}_1;\Phi,t)-\underline{\mathcal{D}}^{\gamma}_{\ell+1}(\bar{\boldsymbol{\Theta}}_1;\Phi,t)\right\|_\infty \le C''e^{bt}(\|\boldsymbol{\Theta}_1\|_\infty+1)\|\boldsymbol{\Theta}_1-\bar{\boldsymbol{\Theta}}_1\|_\infty,$$

for a constant C'', which further yields

$$\left\|\underline{\mathcal{D}}^{\boldsymbol{\beta}}_{\ell}(\boldsymbol{\Theta}_1;\Phi,t)-\underline{\mathcal{D}}^{\boldsymbol{\beta}}_{\ell}(\bar{\boldsymbol{\Theta}}_1;\Phi,t)\right\|_\infty \le C'''e^{bt}(\|\boldsymbol{\Theta}_1\|_\infty+1)\|\boldsymbol{\Theta}_1-\bar{\boldsymbol{\Theta}}_1\|_\infty.$$

We achieve (63) and (64).

Finally, let $b_*=\tilde{C}$ in (65) and (66). It remains to verify that $F(\Phi)\in\boldsymbol{\Phi}_{b_*}$ for any $\Phi\in\boldsymbol{\Phi}\cap\boldsymbol{\Phi}_{b_*}$, that is, to verify the conditions in Definition 19. For $F(\Phi)^{\boldsymbol{v}}_1$, we have

$$\begin{aligned}
&\left\|F(\Phi)^{\boldsymbol{v}}_1(\boldsymbol{\Theta}_1)(t)-F(\Phi)^{\boldsymbol{v}}_1(\bar{\boldsymbol{\Theta}}_1)(t)\right\|_\infty\\
\le\ &\left\|\boldsymbol{\Theta}_1-\bar{\boldsymbol{\Theta}}_1\right\|_\infty+\int_0^t\left\|\underline{\mathcal{G}}^{\boldsymbol{v}}_1(\boldsymbol{\Theta}_1;\Phi,s)-\underline{\mathcal{G}}^{\boldsymbol{v}}_1(\bar{\boldsymbol{\Theta}}_1;\Phi,s)\right\|_\infty ds\\
\le\ &\left\|\boldsymbol{\Theta}_1-\bar{\boldsymbol{\Theta}}_1\right\|_\infty+\int_0^t\tilde{C}(\|\boldsymbol{\Theta}_1\|_\infty+1)e^{b_*s}\left\|\boldsymbol{\Theta}_1-\bar{\boldsymbol{\Theta}}_1\right\|_\infty ds\\
\le\ &e^{b_*s}(\|\boldsymbol{\Theta}_1\|_\infty+1)\|\boldsymbol{\Theta}_1-\bar{\boldsymbol{\Theta}}_1\|_\infty. \qquad (69)
\end{aligned}$$

The verification of other cases are entirely analogous and is omitted. Thus we obtain Lemma 20. ■

Lemma 23 *$\boldsymbol{\Phi}\cap\boldsymbol{\Phi}_b$ is a closed set.*

Proof Given a convergent sequence $\{\Phi_n:n\ge0\}\subseteq\boldsymbol{\Phi}\cap\boldsymbol{\Phi}_b$, it follows from Lemma 22 that the limit point $\Phi_*\in\boldsymbol{\Phi}$. Since Lipschitz property is preserved under pointwise convergence, we also have $\Phi_*\in\boldsymbol{\Phi}_b$. ■

C.2. Proof of Theorem 6

In the proof, we fix Φ_* and the initialization $\{\bar{\boldsymbol{\Theta}}_i\}_{i=1}^m$. We first write down the exact formula for the gradient of the actual discrete DNN, i.e. $\hat{\mathcal{G}}^k_{\ell,i,j}$. Especially, define intermediate variables in the back-propagation as

$$\begin{aligned}
\hat{\mathcal{D}}^k_{L+1,1} &:= N\frac{\partial\hat{\mathcal{L}}^k}{\partial\hat{\boldsymbol{\beta}}_{L+1}}=\left[\phi_1'\left(\hat{\boldsymbol{\beta}}^k_{L+1}(1),y^1\right),\phi_1'\left(\hat{\boldsymbol{\beta}}^k_{L+1}(2),y^2\right),\ldots,\phi_1'\left(\hat{\boldsymbol{\beta}}^k_{L+1}(N),y^N\right)\right]^\top,\\
\hat{\mathcal{D}}^{\boldsymbol{\beta},k}_{L,i} &:= N\frac{\partial\hat{\mathcal{L}}^k}{\partial\hat{\boldsymbol{\beta}}_{L,i}}=\frac{1}{m}\left[\hat{v}^k_{L+1,i,1}\,\hat{\mathcal{D}}^k_{L+1,1}\right]\circ\dot{h}_1'\left(\hat{\boldsymbol{\beta}}^k_{L,i}\right),\quad i\in[m],\\
\hat{\mathcal{D}}^{\boldsymbol{\alpha},k}_{L,i} &:= N\frac{\partial\hat{\mathcal{L}}^k}{\partial\hat{\boldsymbol{\alpha}}_{L,i}}=\hat{\mathcal{D}}^{\boldsymbol{\beta},k}_{L,i}\circ\dot{h}_2'\left(\hat{\boldsymbol{\alpha}}^k_{L,i}\right),\quad i\in[m],\\
\hat{\mathcal{D}}^{\boldsymbol{\beta},k}_{\ell,i} &:= N\frac{\partial\hat{\mathcal{L}}^k}{\partial\hat{\boldsymbol{\beta}}_{\ell,i}}=\frac{1}{m}\left[\sum_{j=1}^m\hat{v}^k_{\ell+1,i,j}\,\hat{\mathcal{D}}^{\boldsymbol{\alpha},k}_{\ell+1,j}\right]\circ\dot{h}_1'\left(\hat{\boldsymbol{\beta}}^k_{\ell,i}\right)+\hat{\mathcal{D}}^{\boldsymbol{\beta},k}_{\ell+1,i},\quad \ell\in[L-1],\ i\in[m],\\
\hat{\mathcal{D}}^{\boldsymbol{\alpha},k}_{\ell,i} &:= N\frac{\partial\hat{\mathcal{L}}^k}{\partial\hat{\boldsymbol{\alpha}}_{\ell,i}}=\hat{\mathcal{D}}^{\boldsymbol{\beta},k}_{\ell,i}\circ\dot{h}_2'\left(\hat{\boldsymbol{\alpha}}^k_{\ell,i}\right),\quad \ell\in[2:L-1],\ i\in[m].
\end{aligned}$$

We have

$$\begin{aligned}\hat{\mathcal{G}}^k_{L+1,i,1} &= \frac{1}{Nm}\left[\hat{\mathcal{D}}^k_{L+1}\right]^\top \dot{h}_1\left(\hat{\boldsymbol{\beta}}^k_{\ell,i}\right), \quad i\in[m],\\ \hat{\mathcal{G}}^k_{\ell+1,i,j} &= \frac{1}{Nm}\left[\hat{\mathcal{D}}^{\boldsymbol{\alpha},k}_{\ell+1,j}\right]^\top \dot{h}_1\left(\hat{\boldsymbol{\beta}}^k_{\ell,i}\right), \quad \ell\in[L-1],\ i,j\in[m],\\ \hat{\mathcal{G}}^k_{1,i,j} &= \frac{1}{Nd}\left[\hat{\mathcal{D}}^{\boldsymbol{\beta},k}_{1,j}\right]^\top \hat{\boldsymbol{\beta}}_{0,i}, \quad i\in[d],\ j\in[m].\end{aligned}$$

To compare the discrete and continuous trajectories on the same time scale, we normalize discrete gradients by

$$\widehat{\mathcal{ND}}^{\boldsymbol{\alpha},k}_{\ell,i} = [m]\,\hat{\mathcal{D}}^{\boldsymbol{\alpha},k}_{\ell,i},\ \ \ell\in[2:L], \qquad \widehat{\mathcal{ND}}^{\boldsymbol{\beta},k}_{\ell,i} = [m]\,\hat{\mathcal{D}}^{\boldsymbol{\beta},k}_{\ell,i},\ \ \ell\in[L]$$

and

$$\widehat{\mathcal{NG}}^k_{\ell,i,j} = [m_{\ell-1}m_\ell]\,\hat{\mathcal{G}}^k_{\ell,i,j}, \quad \ell\in[L+1].$$

Moreover, recalling the definition of $\underline{\mathcal{D}}^\gamma_\ell$ in (57), we also introduce

$$\widehat{\mathcal{ND}}^{\gamma,k}_{\ell,i} = \frac{1}{m}\sum_{j=1}^m\left[\hat{v}^k_{\ell,i,j}\,\widehat{\mathcal{ND}}^{k,\boldsymbol{\alpha}}_{\ell,i}\right]\circ\dot{h}'_1\left(\hat{\boldsymbol{\beta}}^k_{\ell-1,i}\right), \quad \ell\in[2:L],$$

For the ideal process, similar to the notation $\boldsymbol{u}_\ell$ in the proof of Theorem 4, we introduce the notations $\bar{\boldsymbol{u}}_{\ell,i,j}$ that stands for $\bar{\boldsymbol{\Theta}}_j$, $(\bar{\boldsymbol{\Theta}}_i,\bar{\boldsymbol{\Theta}}_i)$, $\bar{\boldsymbol{\Theta}}_i$ for $\ell=1$, $2\le\ell\le L$, $\ell=L+1$, respectively.

We also abbreviate the gradients of the ideal process as

$$\begin{aligned}\underline{\mathcal{D}}^{\boldsymbol{\beta},t}_{\ell,i} &:= \underline{\mathcal{D}}^{\boldsymbol{\beta}}_\ell(\bar{\boldsymbol{\Theta}}_i,\Phi_*,t), \quad \underline{\mathcal{D}}^{\boldsymbol{\alpha},t}_{\ell,i} := \underline{\mathcal{D}}^{\boldsymbol{\alpha}}_\ell(\bar{\boldsymbol{\Theta}}_i,\Phi_*,t),\\ \underline{\mathcal{D}}^{\gamma,t}_{\ell,i} &:= \underline{\mathcal{D}}^{\gamma}_\ell(\bar{\boldsymbol{\Theta}}_i,\Phi_*,t), \quad \underline{\mathcal{G}}^{\boldsymbol{v}}_{\ell,i,j} = \underline{\mathcal{G}}^{\boldsymbol{v}}_\ell(\bar{\boldsymbol{u}}_{\ell,i,j};\Phi_*,t).\end{aligned}$$

We use a common notation $\bar{v}^t_{\ell,i,j}$ to the weights at layer ℓ; for $\ell=1$ let $\bar{v}^t_{1,i,j}=\bar{\boldsymbol{v}}^t_{1,j}$.

When m is finite, the forward and backward propagation for the ideal process is no long exact. Nevertheless, for sufficiently large m, those propagations relations approximately holds by the following events that happen with high probability:

$$\left\|\frac{1}{m}\sum_{i=1}^m\left[\bar{v}^{k\eta}_{\ell+1,i,j}\,\dot{h}_1\left(\bar{\boldsymbol{\beta}}^{k\eta}_{\ell,i}\right)\right]-\bar{\boldsymbol{\alpha}}^{k\eta}_{\ell+1,j}\right\|_\infty \le \left(\|\bar{\boldsymbol{\Theta}}_j\|_\infty+1\right)\varepsilon_1,\ \ell\in[L-1],\ k\in[0:K],\ j\in[m], \tag{70}$$

$$\left\|\frac{1}{m}\sum_{j=1}^m\left[\bar{v}^{k\eta}_{\ell,i,j}\,\underline{\mathcal{D}}^{\boldsymbol{\alpha},k\eta}_{\ell,j}\right]\circ\dot{h}'_1\left(\bar{\boldsymbol{\beta}}^{k\eta}_{\ell-1,i}\right)-\underline{\mathcal{D}}^{\gamma,k\eta}_{\ell,i}\right\|_\infty \le \varepsilon_1,\ \ell\in[2:L],\ k\in[0:K],\ i\in[m], \tag{71}$$

$$\left\|\bar{\boldsymbol{\Theta}}_i\right\|_\infty \le C\sqrt{\log(\frac{m}{\delta})}, \quad i\in[m], \tag{72}$$

$$\frac{1}{m}\sum_{i=1}^m\|\bar{\boldsymbol{\Theta}}_i\|^j_\infty \le C, \quad j\in[2], \tag{73}$$

for a constant C. In the proofs of this section, we condition on those events.

Lemma 24 *The events* (70) – (73) *happen with probability* $1-\delta$.

The proof consists of the deviation of the actual discrete trajectory from the ideal trajectory over the iteration $k \in [0:K]$. We will upper bound the deviation by induction on k. For $k=0$, we have the deviation of weights $\|\bar{v}^0_{\ell,i,j} - \hat{v}^0_{\ell,i,j}\|_\infty$ from the initial conditions in Definition 5. The induction proceeds as follows. In Lemma 25, we first upper bound the deviation of features using the forward propagation, and then upper bound the deviation of gradients using the backward propagation. Note that

$$\left\|\bar{v}^{(k+1)\eta}_{\ell,i,j} - \hat{v}^{k+1}_{\ell,i,j}\right\|_\infty \le \|\bar{v}^{k\eta}_{\ell,i,j} - \hat{v}^{k}_{\ell,i,j}\|_\infty + \int_{k\eta}^{(k+1)\eta} \left\|\underline{\mathcal{G}}^s_{\ell,i,j} - \widehat{\mathcal{NG}}^k_{\ell,i,j}\right\|_\infty ds. \tag{74}$$

Combining with the Lipschitz continuity of $\underline{\mathcal{G}}^t_{\ell,i,j}$ in Lemma 26, we complete the inductive step.

Lemma 25 *Given* $k \in [0:K]$ *and* $\varepsilon < 1$. *Suppose*

$$\left\|\bar{v}^{k\eta}_{\ell,i,j} - \hat{v}^{k}_{\ell,i,j}\right\|_\infty \le (\|\bar{\boldsymbol{u}}_{\ell,i,j}\|_\infty + 1)\varepsilon, \qquad \forall\, \ell \in [L+1],\ i \in [m_{\ell-1}],\ j \in [m_\ell]. \tag{75}$$

Then there exists a constant C *such that*

$$\left\|\bar{\boldsymbol{\beta}}^{k\eta}_{L+1,1} - \hat{\boldsymbol{\beta}}^{k}_{L+1,1}\right\|_\infty \le C\left(\varepsilon + \varepsilon_1\right), \tag{76}$$

$$\left\|\bar{\boldsymbol{\beta}}^{k\eta}_{\ell,i} - \hat{\boldsymbol{\beta}}^{k}_{\ell,i}\right\|_\infty \le C\left(\left\|\bar{\boldsymbol{\Theta}}_i\right\|_\infty + 1\right)\left(\varepsilon + \varepsilon_1\right),\ \forall\, \ell \in [L],\ i \in [m], \tag{77}$$

$$\left\|\bar{\boldsymbol{\alpha}}^{k\eta}_{\ell,i} - \hat{\boldsymbol{\alpha}}^{k}_{\ell,i}\right\|_\infty \le C\left(\left\|\bar{\boldsymbol{\Theta}}_i\right\|_\infty + 1\right)\left(\varepsilon + \varepsilon_1\right),\ \forall\, \ell \in [2:L],\ i \in [m], \tag{78}$$

$$\left\|\underline{\mathcal{G}}^{k\eta}_{\ell,i,j} - \widehat{\mathcal{NG}}^{k}_{\ell,i,j}\right\|_\infty \le C\left(\|\bar{\boldsymbol{u}}_{\ell,i,j}\|_\infty + 1\right)\left(\varepsilon + \varepsilon_1\right),\ \forall\, \ell \in [L+1],\ i \in [m_{\ell-1}],\ j \in [m_\ell]. \tag{79}$$

Lemma 26 *There exists a constant* C *such that, for all* $\ell \in [L+1]$, $t_1, t_2 \in [0,T]$, *and* $\boldsymbol{u}_\ell$,

$$\left\|\underline{\mathcal{G}}^{t_1}_{\ell,i,j} - \underline{\mathcal{G}}^{t_2}_{\ell,i,j}\right\|_\infty \le C(\|\bar{\boldsymbol{u}}_{\ell,i,j}\|_\infty + 1)|t_1 - t_2|.$$

Proof [Proof of Theorem 6] By Lemma 24, the events in (70) – (73) happen with probability $1-\delta$. Conditioned on those events, we prove by induction on $k \in [0:K]$ that

$$\left\|\bar{v}^{k\eta}_{\ell,i,j} - \hat{v}^{k}_{\ell,i,j}\right\|_\infty \le (\|\bar{\boldsymbol{u}}_{\ell,i,j}\|_\infty + 1)\, e^{Ck\eta}\varepsilon_1, \quad \forall\, \ell \in [L+1],\ \ i \in [m_{\ell-1}],\ j \in [m_\ell], \tag{80}$$

for some constant C to be specified. The base case $k=0$ follows from Definition 5. Suppose that (80) holds for $k \in [0:K-1]$. By Lemmas 25 and 26, for $s \in [k\eta, (k+1)\eta]$,

$$\left\|\underline{\mathcal{G}}^{s}_{\ell,i,j} - \widehat{\mathcal{NG}}^{k}_{\ell,i,j}\right\|_\infty \le C'\left(\|\bar{\boldsymbol{u}}_{\ell,i,j}\|_\infty + 1\right)\left(e^{Ck\eta}\varepsilon_1 + \varepsilon_1 + s - k\eta\right).$$

Applying (74) yields that

$$\begin{aligned}
\left\|\bar{v}^{(k+1)\eta}_{\ell,i,j} - \hat{v}^{k+1}_{\ell,i,j}\right\|_\infty &\le \left\|\bar{v}^{k\eta}_{\ell,i,j} - \hat{v}^{k}_{\ell,i,j}\right\|_\infty + \int_{k\eta}^{(k+1)\eta} \left\|\underline{\mathcal{G}}^s_{\ell,i,j} - \widehat{\mathcal{NG}}^k_{\ell,i,j}\right\|_\infty ds \\
&\le (\|\bar{\boldsymbol{u}}_{\ell,i,j}\|_\infty + 1)\left(e^{Ck\eta}\varepsilon_1 + 2C' e^{Ck\eta}\varepsilon_1\eta + C'\frac{\eta^2}{2}\right) \\
&\le (\|\bar{\boldsymbol{u}}_{\ell,i,j}\|_\infty + 1)\, e^{Ck\eta}\varepsilon_1(1 + C''\eta),
\end{aligned}$$

for a constant C''. By letting $C = C''$, we arrive at (80) for $k+1$ using $1 + C\eta \le e^{C\eta}$. Note that $k\eta \le T$ for $k \in [0:K]$, $\varepsilon_1 \le \tilde{\mathcal{O}}(1/\sqrt{m})$, and $\|\bar{\boldsymbol{u}}_{\ell,i,j}\|_\infty \le \mathcal{O}(\log(m))$ from (72). The conclusion follows from Lemma 25 and the Lipschitz continuity of ϕ.. ■

C.2.1. PROOFS OF LEMMAS

Proof [Proof of Lemma 24] We prove each of the four events happens with probability $1 - \frac{\delta}{4}$ by standard concentration inequalities. Both (72) and (73) happen with probability $1 - \frac{\delta}{4}$ by the concentration of sub-gaussian random variables; in particular, (72) follows from Lemma 27 and (73) follows from Lemmas 28 and 29.

For (70) with a given k, ℓ, j, n, consider random vectors

$$\zeta_i := \frac{\bar{v}_{\ell,i,j}^{k\eta} h_1\left(\bar{\boldsymbol{\beta}}_{\ell-1,i}^{k\eta}(n)\right)}{\left\|\bar{\boldsymbol{\Theta}}_j\right\|_\infty + 1},$$

which are bounded by a constant C' due to the upper bound of $\bar{v}_\ell$ in Lemma 21. Conditioned on $\bar{\boldsymbol{\Theta}}_j$, when $i \neq j$, ζ_i are independent and $\mathbb{E}[\zeta_i | \bar{\boldsymbol{\Theta}}_j] = \frac{\bar{\boldsymbol{\alpha}}_j^{k\eta}(n)}{\|\bar{\boldsymbol{\Theta}}_j\|_\infty + 1}$. By Hoeffding's inequality, we have

$$\left| \frac{1}{m-1} \sum_{i=1,\, i\neq j}^{m} \zeta_i - \frac{\bar{\boldsymbol{\alpha}}_{\ell,j}^{k\eta}(n)}{\|\bar{\boldsymbol{\Theta}}_j\|_\infty + 1} \right| < \varepsilon_1/2,$$

with probability $1 - \frac{\delta}{4mL(K+1)N}$. On the other hand, when $i = j$, we also have

$$\frac{1}{m} \left| \zeta_j - \frac{\bar{\boldsymbol{\alpha}}_{\ell,j}^{k\eta}(n)}{\|\bar{\boldsymbol{\Theta}}_j\|_\infty + 1} \right| \leq \tilde{C}' \varepsilon_1^2 \leq \varepsilon_1/2,$$

where we use the upper bound of $\bar{\boldsymbol{\alpha}}_\ell$ in Lemma 21. Therefore, applying the union bound over $k \in [0:K],\ \ell \in [L],\ j \in [m]$ and $n \in [N]$, we have (70) with probability $1 - \frac{\delta}{4}$.

For (71) with a given k, ℓ, i, n, consider the random vectors

$$\zeta_j' := [\bar{v}_{\ell+1,i,j}^{k\eta}\ \underline{\mathcal{D}}_{\ell+1,j}^{\boldsymbol{\alpha},k\eta}(n)]\ h_1'\left(\boldsymbol{\beta}_{\ell,i}^{k\eta}(n)\right).$$

Conditioned on $\boldsymbol{\Theta}_i$, when $i \neq j$, ζ_j' are independent and $\mathbb{E}[\zeta_j' | \bar{\boldsymbol{\Theta}}_i] = \underline{\mathcal{D}}_{\ell+1,i}^{\gamma,k\eta}(n)$. By the boundedness of h_1' and the upper bound of $\underline{\mathcal{D}}_{\ell+1}^{\boldsymbol{\alpha}}$ in Lemma 21, we have ζ_j'

$$|\zeta_j'| \leq C' |\bar{v}_{\ell+1,i,j}^{k\eta}| \leq C(1 + \|\boldsymbol{\Theta}_j\|_\infty),$$

and thus ξ_j' is sub-gaussian. Applying Lemma 28, we obtain that

$$\left| \frac{1}{m-1} \sum_{j=1,\, j\neq i}^{m} \zeta_j' - \underline{\mathcal{D}}_{\ell+1,i}^{\gamma,k\eta}(n) \right| < \varepsilon_1/2,$$

with probability $1 - \frac{\delta}{4mL(K+1)N}$. On the other hand, under event (72), we have

$$\frac{1}{m} \left| \zeta_i' - \underline{\mathcal{D}}_{\ell+1,i}^{\gamma,k\eta}(n) \right| \leq \tilde{\mathcal{O}}(\varepsilon_1^2) \leq \varepsilon_1/2$$

Therefore, applying the union bound again over $k \in [0:K],\ \ell \in [L],\ j \in [m]$, and $n \in [N]$, we have (71) with probability $1 - \frac{\delta}{4}$.

■

Proof [Proof of Lemma 25] We first consider the forward propagation and prove (76), (77) and (78). For $\ell = 1$, since $\boldsymbol{X}$ is bounded,

$$\left\|\bar{\boldsymbol{\beta}}_{1,i}^{k\eta} - \hat{\boldsymbol{\beta}}_{1,i}^{k}\right\|_\infty \le C\|\bar{\boldsymbol{v}}_{1,i}^{k\eta} - \hat{\boldsymbol{v}}_{1,i}^{k}\|_\infty \le C(\|\bar{\boldsymbol{\Theta}}_i\|_\infty + 1)\varepsilon.$$

For $\ell \in [2:L]$, by the triangle inequality,

$$\begin{aligned} &\left\|\bar{\boldsymbol{\alpha}}_{\ell+1,j}^{k\eta} - \hat{\boldsymbol{\alpha}}_{\ell+1,j}^{k}\right\|_\infty \qquad (81)\\ \le &\left\|\bar{\boldsymbol{\alpha}}_{\ell+1,j}^{k\eta} - \frac{1}{m}\sum_{i=1}^m \left[\bar{v}_{\ell+1,i,j}^{k\eta}\, \dot{h}_1\left(\bar{\boldsymbol{\beta}}_{\ell,i}^{k\eta}\right)\right]\right\|_\infty + \left\|\frac{1}{m}\sum_{i=1}^m \left[\bar{v}_{\ell+1,i,j}^{k\eta}\, \dot{h}_1\left(\bar{\boldsymbol{\beta}}_{\ell,i}^{k\eta}\right) - \hat{v}_{\ell+1,i,j}^{k}\, \dot{h}_1\left(\hat{\boldsymbol{\beta}}_{\ell,i}^{k}\right)\right]\right\|_\infty. \end{aligned}$$

The first term is approximately the forward propagation that is at most $(\|\bar{\boldsymbol{\Theta}}\|_\infty + 1)\varepsilon_1$ by (70). For the second term, since h_1 is bounded and Lipschitz continuous and the weights $\bar{v}_{\ell,i,j}$ are upper bounded by Lemma 21 and Assumption 3, we have a further upper bound

$$\begin{aligned} &\frac{1}{m}\sum_{i=1}^m \underbrace{\left|\bar{v}_{\ell+1,i,j}^{k\eta}\right|}_{\le C(\|\bar{\boldsymbol{\Theta}}_j\|_\infty+1)} \left\|\dot{h}_1\left(\bar{\boldsymbol{\beta}}_{\ell,i}^{k\eta}\right) - \dot{h}_1\left(\hat{\boldsymbol{\beta}}_{\ell,i}^{k}\right)\right\|_\infty + \frac{1}{m}\sum_{i=1}^m \underbrace{\left|\bar{v}_{\ell+1,i,j}^{k\eta} - \hat{v}_{\ell+1,i,j}^{k}\right|}_{(75)} \left\|\dot{h}_1(\hat{\boldsymbol{\beta}}_{\ell,i}^{k})\right\|_\infty \\ \le\ & C(\|\bar{\boldsymbol{\Theta}}_j\|_\infty + 1)(\varepsilon + \varepsilon_1), \end{aligned}$$

where in the last step we used (73). We have

$$\left\|\bar{\boldsymbol{\alpha}}_{\ell+1,i}^{k\eta} - \hat{\boldsymbol{\alpha}}_{\ell+1,i}^{k}\right\|_\infty \le C'(\|\bar{\boldsymbol{\Theta}}_i\|_\infty + 1)\varepsilon,$$

for a constant C', which gives that

$$\begin{aligned} &\left\|\bar{\boldsymbol{\beta}}_{\ell+1,i}^{k\eta} - \hat{\boldsymbol{\beta}}_{\ell+1,i}^{k}\right\|_\infty \\ \le\ &\left\|\bar{\boldsymbol{\beta}}_{\ell,i}^{k\eta} - \hat{\boldsymbol{\beta}}_{\ell,i}^{k}\right\|_\infty + \left\|\dot{h}_2\left(\bar{\boldsymbol{\alpha}}_{\ell+1,i}^{k\eta}\right) - \dot{h}_2\left(\hat{\boldsymbol{\alpha}}_{\ell+1,i}^{k}\right)\right\|_\infty \\ \le\ & C''(\|\bar{\boldsymbol{\Theta}}_i\|_\infty + 1)\varepsilon. \end{aligned}$$

The output layer $\ell = L+1$ is similar by applying the upper bound of v_{L+1} in Assumption 3.

Next we consider the backward propagation and prove (79). Since $\boldsymbol{X}$ is bounded, h_1 is bounded and Lipschitz continuous, and $\underline{\mathcal{D}}_\ell^{\boldsymbol{\alpha}}$ is bounded by Lemma 21, it suffices to prove that

$$\begin{aligned} &\left\|\underline{\mathcal{D}}_{L+1}(\Phi_*, k\eta) - \hat{\mathcal{D}}_{L+1,1}^{k}\right\|_\infty \le C\varepsilon, \\ &\left\|\underline{\mathcal{D}}_{\ell,i}^{\boldsymbol{\beta},k\eta} - \widehat{\mathcal{ND}}_{\ell,i}^{\boldsymbol{\beta},k}\right\|_\infty \le C\left(1 + \left\|\bar{\boldsymbol{\Theta}}_i\right\|_\infty\right)\varepsilon, \quad \ell \in [L],\ i \in [m], \qquad (82)\\ &\left\|\underline{\mathcal{D}}_{\ell,i}^{\boldsymbol{\alpha},k\eta} - \widehat{\mathcal{ND}}_{\ell,i}^{\boldsymbol{\alpha},k}\right\|_\infty \le C\left(1 + \left\|\bar{\boldsymbol{\Theta}}_i\right\|_\infty\right)\varepsilon, \quad \ell \in [2:L],\ i \in [m], \end{aligned}$$

for a constant C.

At the output layer $\ell = L+1$, since ϕ_1' is Lipschitz continuous on the first argument,

$$\left\|\underline{\mathcal{D}}_{L+1}(\Phi_*, k\eta) - \widehat{\mathcal{ND}}^k_{L+1,1}\right\|_\infty \le L_5 \left\|\bar{\boldsymbol{\beta}}^{k\eta}_{L+1,1} - \hat{\boldsymbol{\beta}}^k_{L+1,1}\right\|_\infty \le C'(\varepsilon + \varepsilon_1),$$

for a constant C'. At layer $\ell = L$, since h_1' is bounded and Lipschitz continuous and $\underline{\mathcal{D}}_{L+1}$ is bounded by Lemma 21, applying (77) yields that

$$\left\|\underline{\mathcal{D}}^{k\eta}_{L+1,1}\, h_1'\left(\bar{\boldsymbol{\beta}}^{k\eta}_{L,i}\right) - \widehat{\mathcal{ND}}^k_{L+1,1}\, h_1'\left(\hat{\boldsymbol{\beta}}^{k\eta}_{L,i}\right)\right\|_\infty \le C\left(\left\|\bar{\boldsymbol{\Theta}}_i\right\|_\infty + 1\right)(\varepsilon + \varepsilon_1).$$

Moreover, applying (75) and the upper bound of $\bar{v}_{L,i}$ in Lemma 21, we obtain that

$$\left\|\underline{\mathcal{D}}^{\boldsymbol{\beta},k\eta}_{L,i} - \widehat{\mathcal{ND}}^{\boldsymbol{\beta},k}_{L,i}\right\|_\infty \le C\left(1 + \left\|\bar{\boldsymbol{\Theta}}_i\right\|_\infty\right)\varepsilon,$$

for a constant C.

For each layer ℓ from $L-1$ to 1, suppose we have

$$\left\|\underline{\mathcal{D}}^{\boldsymbol{\beta},k\eta}_{\ell+1,i} - \widehat{\mathcal{ND}}^{\boldsymbol{\beta},k}_{\ell+1,i}\right\|_\infty \le C\left(1 + \left\|\bar{\boldsymbol{\Theta}}_i\right\|_\infty\right)\varepsilon.$$

Since h_2' is bounded and Lipschitz continuous and $\underline{\mathcal{D}}^{\boldsymbol{\beta}}_{\ell+1}$ is bounded by Lemma 21, by (78), we further have

$$\begin{aligned} &\left\|\underline{\mathcal{D}}^{\boldsymbol{\alpha},k\eta}_{\ell+1,i} - \widehat{\mathcal{ND}}^{\boldsymbol{\alpha},k}_{\ell+1,i}\right\|_\infty \\ \le\ &\left\|\underline{\mathcal{D}}^{\boldsymbol{\beta},k\eta}_{\ell+1,i} \circ \dot{h}_2'\left(\bar{\boldsymbol{\alpha}}^{k\eta}_{\ell+1,i}\right) - \widehat{\mathcal{ND}}^{\boldsymbol{\beta},k}_{\ell+1,i} \circ \dot{h}_2'\left(\hat{\boldsymbol{\alpha}}^k_{\ell+1,i}\right)\right\|_\infty \\ \le\ &C'\left(1 + \left\|\bar{\boldsymbol{\Theta}}_i\right\|_\infty\right)\varepsilon, \end{aligned} \tag{83}$$

for a constant C'. By the triangle inequality,

$$\begin{aligned} &\left\|\underline{\mathcal{D}}^{\gamma,k\eta}_{\ell+1,i} - \widehat{\mathcal{ND}}^{\gamma,k}_{\ell+1,i}\right\|_\infty \\ \le &\left\|\underline{\mathcal{D}}^{\gamma,k\eta}_{\ell+1,i} - \frac{1}{m}\sum_{j=1}^m \bar{v}^{k\eta}_{\ell+1,i,j}\, \underline{\mathcal{D}}^{\boldsymbol{\alpha},k\eta}_{\ell+1,j} \circ \dot{h}_1'\left(\bar{\boldsymbol{\beta}}^{k\eta}_{\ell+1,i}\right)\right\|_\infty \\ + &\left\|\frac{1}{m}\sum_{j=1}^m \bar{v}^{k\eta}_{\ell+1,i,j}\left[\underline{\mathcal{D}}^{\boldsymbol{\alpha},k\eta}_{\ell+1,j} \circ \dot{h}_1'\left(\bar{\boldsymbol{\beta}}^{k\eta}_{\ell,i}\right)\right] - \frac{1}{m}\sum_{j=1}^m \hat{v}^k_{\ell+1,i,j}\left[\widehat{\mathcal{ND}}^{\boldsymbol{\alpha},k}_{\ell+1,j} \circ \dot{h}_1'\left(\hat{\boldsymbol{\beta}}^{k\eta}_{\ell,i}\right)\right]\right\|_\infty. \end{aligned} \tag{84}$$

The first term is approximately backward propagation and is at most ε_1 by (71). For the second term, note that h_1' is bounded and Lipschitz continuous, $\bar{v}_{\ell+1,i,j}$ and $\underline{\mathcal{D}}^{\boldsymbol{\alpha}}_{\ell,j}$ are upper bounded by Lemma 21. Applying (75), (77), and (83) at layer $\ell+1$ yields that

$$\begin{aligned} &\left\|\underline{\mathcal{D}}^{\boldsymbol{\alpha},k\eta}_{\ell+1,j} \circ \left[\bar{v}^{k\eta}_{\ell+1,i,j}\, \dot{h}'\left(\bar{\boldsymbol{\beta}}^{k\eta}_{\ell,i}\right)\right] - \widehat{\mathcal{ND}}_{\ell+1,j} \circ \left[\hat{v}^k_{\ell+1,i,j}\, \dot{h}_1'\left(\hat{\boldsymbol{\beta}}^{k\eta}_{\ell,i}\right)\right]\right\|_\infty \\ \le\ &C(\|\bar{\boldsymbol{\Theta}}_j\|_\infty + 1)\left(\|\bar{\boldsymbol{\Theta}}_j\|_\infty + \|\bar{\boldsymbol{\Theta}}_i\|_\infty + 1\right)(\varepsilon + \varepsilon_1). \end{aligned}$$

Therefore, by (73), we have

$$\left\|\underline{\mathcal{D}}^{\gamma,k\eta}_{\ell+1,i} - \widehat{\mathcal{ND}}^{\gamma,k}_{\ell+1,i}\right\|_\infty \le C''\left(1 + \left\|\bar{\boldsymbol{\Theta}}_i\right\|_\infty\right)\varepsilon,$$

for a constant C''. We conclude

$$\begin{aligned} & \left\|\underline{\mathcal{D}}_{\ell,i}^{\boldsymbol{\beta},k\eta} - \widehat{\mathcal{ND}}_{\ell,i}^{\boldsymbol{\beta},k}\right\|_\infty \\ \leq & \left\|\underline{\mathcal{D}}_{\ell+1,i}^{\boldsymbol{\gamma},k\eta} - \widehat{\mathcal{ND}}_{\ell+1,i}^{\boldsymbol{\gamma},k}\right\|_\infty + \left\|\underline{\mathcal{D}}_{\ell+1,i}^{\boldsymbol{\beta},k\eta} - \widehat{\mathcal{ND}}_{\ell+1,i}^{\boldsymbol{\beta},k}\right\|_\infty \\ \leq & C'''\left(1+\left\|\bar{\boldsymbol{\Theta}}_i\right\|_\infty\right)\varepsilon, \end{aligned}$$

for a constant C''' and obtain (82). ■

Proof [Proof of Lemma 26] It suffices to prove $\underline{\mathcal{G}}_\ell^{\boldsymbol{v}}(\boldsymbol{u}_\ell;\Phi_*,t)$ is continuous on t for all $\boldsymbol{u}_\ell$. The proof is similar to the backward steps in Lemma 25. Recalling boundedness of $\underline{\mathcal{G}}_\ell^{\boldsymbol{\beta}}$ and $\underline{\mathcal{G}}_\ell^{\boldsymbol{\alpha}}$ in Lemma 21, for any fix $\boldsymbol{\Theta}$, we have

$$\begin{aligned} \left\|\boldsymbol{\beta}_\ell^{t_1} - \boldsymbol{\beta}_\ell^{t_2}\right\| &\leq C\left(\|\boldsymbol{\Theta}\|_\infty+1\right)|t_1-t_2|, \quad \ell\in[L], \\ \left\|\boldsymbol{\alpha}_\ell^{t_1} - \boldsymbol{\alpha}_\ell^{t_2}\right\| &\leq C\left(\|\boldsymbol{\Theta}\|_\infty+1\right)|t_1-t_2|, \quad \ell\in[2:L], \end{aligned} \tag{85}$$

where $\boldsymbol{\beta}_\ell^t = \Phi_{*,\ell}^{\boldsymbol{\beta}}(\boldsymbol{\Theta})(t)$ and $\boldsymbol{\alpha}_\ell^t = \Phi_{*,\ell}^{\boldsymbol{\alpha}}(\boldsymbol{\Theta})(t)$.

By the boundedness of h_1 and $\boldsymbol{X}$, it suffices to prove the Lipschitz continuity for $\underline{\mathcal{D}}_\ell^{\boldsymbol{\beta}}$ for $\ell=1$ and $\underline{\mathcal{D}}_\ell^{\boldsymbol{\alpha}}$ for $\ell\in[2:L]$. We prove

$$\left\|\underline{\mathcal{D}}_{L+1}(\Phi_*,t_1) - \underline{\mathcal{D}}_{L+1}(\Phi_*,t_2)\right\|_\infty \leq C'|t_1-t_2|, \tag{86}$$

$$\left\|\underline{\mathcal{D}}_\ell^{\boldsymbol{\beta}}(\boldsymbol{\Theta};\Phi_*,t_1) - \underline{\mathcal{D}}_\ell^{\boldsymbol{\beta}}(\boldsymbol{\Theta};\Phi_*,t_2)\right\|_\infty \leq C'\left(1+\|\boldsymbol{\Theta}\|_\infty\right)|t_1-t_2|, \quad \ell\in[L], \tag{87}$$

$$\left\|\underline{\mathcal{D}}_\ell^{\boldsymbol{\alpha}}(\boldsymbol{\Theta};\Phi_*,t_1) - \underline{\mathcal{D}}_\ell^{\boldsymbol{\alpha}}(\boldsymbol{\Theta};\Phi_*,t_2)\right\|_\infty \leq C'\left(1+\|\boldsymbol{\Theta}\|_\infty\right)|t_1-t_2|, \quad \ell\in[2:L]. \tag{88}$$

At the output layer $\ell=L+1$, by the Lipschitz continuity of ϕ', we have

$$\begin{aligned} & \left\|\underline{\mathcal{D}}_{L+1}\left(\Phi_*,t_1\right) - \underline{\mathcal{D}}_{L+1}\left(\Phi_*,t_2\right)\right\|_\infty \\ \leq & L_5\left\|\boldsymbol{\beta}_{L+1}^{t_1} - \boldsymbol{\beta}_{L+1}^{t_2}\right\|_\infty \\ \leq & L_5\left\|\int v_{L+1}^{t_1}\,\dot{h}_1\left(\boldsymbol{\beta}_L^{t_1}\right) - v_{L+1}^{t_2}\,\dot{h}_1\left(\boldsymbol{\beta}_L^{t_2}\right) dp(\boldsymbol{\Theta})\right\|_\infty. \end{aligned} \tag{89}$$

By the upper bound and Lipschitz continuity of v_{L+1} in Lemma 21, we obtain (86). At layer $\ell=L$, using (11b), we obtain (87) from the upper bounds and the Lipschitz continuity of $\underline{\mathcal{D}}_{L+1}(\boldsymbol{\Theta},\Phi_*,t)$, $v_{L+1}^t(\boldsymbol{\Theta})$, and $\dot{h}_1'\left(\boldsymbol{\beta}_L^t\right)$.

For each layer ℓ from $L-1$ to 1, suppose we have (87) at layer $\ell+1$. From the upper bounds and the Lipschitz continuity of $\underline{\mathcal{D}}_{\ell+1}^{\boldsymbol{\beta}}$ and $\dot{h}_2'(\bar{\boldsymbol{\alpha}}_{\ell+1}^t)$, we have

$$\left\|\underline{\mathcal{D}}_{\ell+1}^{\boldsymbol{\alpha}}(\boldsymbol{\Theta};\Phi_*,t_1) - \underline{\mathcal{D}}_{\ell+1}^{\boldsymbol{\alpha}}(\boldsymbol{\Theta};\Phi_*,t_2)\right\|_\infty \leq C''\left(1+\|\boldsymbol{\Theta}\|_\infty\right)|t_1-t_2|, \tag{90}$$

for a constant C''. Moreover,

$$\begin{aligned}
&\int \left\| v_{\ell+1}^{t_1}(\boldsymbol{\Theta}, \bar{\boldsymbol{\Theta}})\, \underline{\mathcal{D}}_{\ell+1}^{\boldsymbol{\alpha}}(\bar{\boldsymbol{\Theta}}; \Phi_*, t_1) - v_{\ell+1}^{t_2}(\boldsymbol{\Theta}, \bar{\boldsymbol{\Theta}})\, \underline{\mathcal{D}}_{\ell+1}^{\boldsymbol{\alpha}}(\bar{\boldsymbol{\Theta}}; \Phi_*, t_2) \right\|_\infty dp(\bar{\boldsymbol{\Theta}}) \\
&\leq \int \underbrace{\left| v_{\ell+1}^{t_1}(\boldsymbol{\Theta}, \bar{\boldsymbol{\Theta}}) - v_{\ell+1}^{t_2}(\boldsymbol{\Theta}, \bar{\boldsymbol{\Theta}}) \right|}_{\leq C|t_1 - t_2|} \left\| \underline{\mathcal{D}}_{\ell+1}^{\boldsymbol{\alpha}}(\bar{\boldsymbol{\Theta}}; \Phi_*, t) \right\|_\infty dp(\bar{\boldsymbol{\Theta}}) \\
&\qquad + \int \left| v_{\ell+1}^{t_2}(\boldsymbol{\Theta}, \bar{\boldsymbol{\Theta}}) \right| \underbrace{\left\| \underline{\mathcal{D}}_{\ell+1}^{\boldsymbol{\alpha}}(\bar{\boldsymbol{\Theta}}; \Phi_*, t_1) - \underline{\mathcal{D}}_{\ell+1}(\bar{\boldsymbol{\Theta}}; \Phi_*, t_2) \right\|_\infty}_{\leq C\left(\|\bar{\boldsymbol{\Theta}}\|_\infty + 1\right)\, |t_1 - t_2|} dp(\bar{\boldsymbol{\Theta}}) \\
&\leq C'|t_1 - t_2|, \qquad (91)
\end{aligned}$$

where in the last step we used the upper bounds of $v_{\ell+1}$ and $\underline{\mathcal{D}}_{\ell+1}^{\boldsymbol{\alpha}}$ in Lemma 21, sub-gaussianness of p, and Corollary 31. Combing the above results with

$$\left\| \dot{h}_1(\boldsymbol{\beta}_\ell^{t_1}) - \dot{h}_1(\boldsymbol{\beta}_\ell^{t_2}) \right\| \leq C\left(\|\boldsymbol{\Theta}\|_\infty + 1\right)|t_1 - t_2|, \quad \ell \in [L]$$

from (85), we obtain

$$\left\| \underline{\mathcal{D}}_{\ell+1}^{\gamma}(\boldsymbol{\Theta}; \Phi_*, t_1) - \underline{\mathcal{D}}_{\ell+1}^{\gamma}(\boldsymbol{\Theta}; \Phi_*, t_2) \right\|_\infty \leq C'''\left(1 + \|\boldsymbol{\Theta}\|_\infty\right)|t_1 - t_2|, \qquad (92)$$

for a constant C'''. Combining (90) and (92), we can achieve (87) at ℓ. ■

C.3. Proof of Proposition 1

Proof [Proof of Proposition 1] Explicitly shown in (Nguyen and Pham, 2020, Corollary 25), in the mean-field limit that $m \to \infty$, the weights remain mutually independent and follow a common distribution that only depends on time t in the intermediate layers. Therefore, by the law of large numbers, the features are the same. We have Proposition 1. ■

Appendix D. Preliminary of Proofs

In this paper, we adopt the definition of sub-gaussian distributions in Vershynin (2010). Below we present properties of sub-gaussian distributions. The equivalence among those properties are given in (Vershynin, 2010, Lemma 5.5).

Lemma 27 *Let ξ be an univariate random variable that follows a σ-sub-gaussian distribution. Then there exists an absolute constant C such that*

(1) Tails $\mathbb{P}(|\xi| > t) \leq \exp(1 - t^2/(C\sigma)^2)$ for all $t \geq 0$;

(2) Moments: $(\mathbb{E}|\xi|^q)^{1/q} \leq C\sigma\sqrt{q}$ for all $q \geq 1$;

(3) If $\mathbb{E}[\xi] = 0$, then $\mathbb{E}[\exp(t\xi)] \leq \exp(t^2(C\sigma)^2)$ for all $t \in \mathbb{R}$.

Lemma 28 (Concentration Inequality for Sub-gaussian Distributions (Vershynin, 2010, Proposition 5.10)) *Let $\{\xi_i\}_{i=1}^m$ be independent centered σ-sub-gaussian random variables. Then, for an absolute constant C,*

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^m \xi_i\right| \geq \varepsilon\right) \leq 2\exp\left(-\frac{m\varepsilon^2}{4(C\sigma)^2}\right).$$

We say a random variable ξ is sub-exponential if

$$\sup_{q\geq 1} q^{-1}(\mathbb{E}|\xi|^q)^{1/q} < \infty.$$

A sub-exponential random variable is equivalent to the a squared sub-gaussian random variable (Vershynin, 2010, Lemma 5.14). It satisfies the following concentration inequality:

Lemma 29 (Bernstein's Inequality for Sub-Exponential Distributions (Vershynin, 2010, Corollary 5.17)) *Let $\{\xi_i\}_{i=1}^m$ be independent centered sub-exponential random variables such that $(\mathbb{E}|\xi|^q)^{1/q} \leq Kq$ for all $q \geq 1$. Then, for an absolute constant c,*

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^m z_i\right| \geq \varepsilon\right) \leq 2\exp\left[-cm\min\left(\frac{\varepsilon^2}{K^2}, \frac{\varepsilon}{K}\right)\right].$$

Lemma 30 *For a d-dimensional random vector $\boldsymbol{\xi} \in \mathbb{R}^d$, we have:*

(1) $\|\boldsymbol{\xi}\|_\infty$ is σ-sub-gaussian $\Longrightarrow \boldsymbol{\xi}$ is $(\sigma\sqrt{d})$-sub-gaussian;

(2) $\boldsymbol{\xi}$ is σ-sub-gaussian $\Longrightarrow \|\boldsymbol{\xi}\|_\infty$ is $\mathcal{O}(\sigma\sqrt{\log d})$-sub-gaussian.

Proof (1) For any $\boldsymbol{u} \in \mathbb{S}^{d-1}$, we have $|\boldsymbol{u}^\top \boldsymbol{\xi}| \leq \sqrt{d}\|\boldsymbol{\xi}\|_\infty$.

(2) Note that $\|\boldsymbol{\xi}\|_\infty = \max_{i\in[d]} |\boldsymbol{e}_i^\top \boldsymbol{\xi}|$, where $\boldsymbol{e}_i$ denotes the unit vector with i-th coordinate being one. Applying Lemma 27 (1) and the union bound yields that

$$\mathbb{P}[\|\boldsymbol{\xi}\|_\infty > t] \leq \min\left\{de^{1-\frac{t^2}{(C\sigma)^2}}, 1\right\} = e^{-\left(\frac{t^2}{(C\sigma)^2}-\log(ed)\right)_+} \leq e^{1-\frac{t^2}{(C\sigma\sqrt{\log(ed)})^2}},$$

where we used the fact that $(\frac{t^2}{a} - b)_+ \geq \frac{t^2}{ab} - 1$ when $b \geq 1$. Therefore, $\|\boldsymbol{\xi}\|_\infty$ is $\mathcal{O}(\sigma\sqrt{\log d})$-sub-gaussian by the equivalent definition of sub-gaussian distributions in (Vershynin, 2010, Lemma 5.5). ■

Corollary 31 *For a σ-sub-gaussian random vector $\boldsymbol{\xi} \in \mathbb{R}^d$, we have*

$$(\mathbb{E}\|\boldsymbol{\xi}\|_\infty^q)^{1/q} \leq \mathcal{O}(\sigma\sqrt{q\log d}), \quad q \geq 1.$$

Proof From Lemma 30, $\|\boldsymbol{\xi}\|_\infty$ is $\mathcal{O}(\sigma\sqrt{\log d})$-sub-gaussian. Applying Lemma 27 (2), we achieve the desired result. ■

Appendix E. Extensions on Fully-connected DNNs

E.1. Discrete Fully-connected DNN

We first introduce the standard DNN. Let m_ℓ denote the number of units at layer ℓ for $\ell \in [0 : L+1]$. Let $m_0 = d$ and node i outputs the value of i-th coordinate of the training data for $i \in [d]$. Let $m_{L+1} = 1$ that is the unit of the final network output. For $\ell \in [L+1]$, the output, i.e. features, of node i in layer ℓ is denoted by $\hat{\boldsymbol{\theta}}_{\ell,i} \in \mathbb{R}^N$; the weight that connects the node i for the N training samples at layer $\ell - 1$ to node j at layer ℓ is denoted by $\hat{w}_{\ell,i,j} \in \mathbb{R}$.

(1) At the input layer, for $i \in [d]$, let

$$\hat{\boldsymbol{\theta}}_{0,i} := \left[\boldsymbol{x}^1(i), \boldsymbol{x}^2(i), \ldots, \boldsymbol{x}^N(i)\right]^\top. \tag{93}$$

(2) We recursively define the upper layers ($\ell \in [L]$) as below.

$$\hat{\boldsymbol{\theta}}_{\ell,j} := \begin{cases} \frac{1}{m_0} \sum_{i=1}^{m_0} \hat{w}_{1,i,j}\, \hat{\boldsymbol{\theta}}_{0,i}, & j \in [m_1],\ \ \ell = 1, \\ \frac{1}{m_{\ell-1}} \sum_{i=1}^{m_{\ell-1}} \hat{w}_{\ell,i,j}\, \dot{h}\left(\hat{\boldsymbol{\theta}}_{\ell-1,i}\right), & j \in [m_\ell],\ \ell \in [2 : L], \end{cases} \tag{94}$$

where h is the activation function.

(3) At the output layer,

$$\hat{\boldsymbol{\theta}}_{L+1,1} := \frac{1}{m_L} \sum_{i=1}^{m_L} \hat{w}_{L+1,i}\, \dot{h}\left(\hat{\boldsymbol{\theta}}_{L,i}\right). \tag{95}$$

We collect the weights at the ℓ-th layer ($\ell \in [L+1]$) into a single vector denoted by $\hat{\boldsymbol{w}}_\ell$ and all the weights into a single vector denoted by $\hat{\boldsymbol{w}}$. Similarly, we aggregate features at ℓ-th layer ($\ell \in [L]$) into a single vector denoted by $\hat{\boldsymbol{\theta}}_\ell$ and all the features into a single vector denoted by $\hat{\boldsymbol{\theta}}$.

E.2. Continuous Fully-connected DNN

We introduce our continuous DNN formulation using similar forward propagation of the the discrete DNN in Section E.1.

(1) At the input layer, let $\boldsymbol{X} = \left[\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^N\right]^\top \in \mathbb{R}^{N \times d}$.

(2) At the first layer, each hidden node (before the activation function) is computed by a linear mapping of the input data, so each node can be indexed by the weights connecting it to the input. We introduce a probability measure $p_1(\boldsymbol{w}_1) \in \mathcal{P}\left(\mathbb{R}^d\right)$ for the weights to describe the states of first layer and let[6]

$$\boldsymbol{\theta}_1(\boldsymbol{w}_1) := \frac{1}{d}(\boldsymbol{X}\boldsymbol{w}_1). \tag{96}$$

6. The state of the first layer can be equivalently characterized by either the output or the weight that are related by a linear mapping.

Algorithm 3 Scaled Gradient Descent for Training a DNN

1: Input the data $\{\boldsymbol{x}^i, y^i\}_{i=1}^N$, step size η, and initial weights $\hat{\boldsymbol{w}}^0$.
2: **for** $k = 0, 1, \ldots, K-1$ **do**
3: Perform forward-propagation (94) and (95) to compute $\hat{\boldsymbol{\theta}}_{L+1,1}^k$.
4: Perform backward-propagation to compute the gradient $\hat{\mathcal{G}}_{\ell,i,j}^k = \frac{\partial \hat{\mathcal{L}}_N}{\partial \hat{w}_{\ell,i,j}^k}$.
5: Perform scaled Gradient Descent:

$$\hat{w}_{\ell,i,j}^{k+1} = \hat{w}_{\ell,i,j}^k - \left[\eta m_{\ell-1} m_\ell\right] \hat{\mathcal{G}}_{\ell,i,j}^k, \quad \ell \in [L+1],\ i \in [m_{\ell-1}],\ j \in [m_\ell].$$

6: **end for**
7: Output the weights $\hat{\boldsymbol{w}}^K$.

(3) At the second layer, recall that the output of each node, i.e., the feature, for the training samples is a N-dimensional vector. We use the features $\boldsymbol{\theta}_2 \in \mathbb{R}^N$ to index those nodes. We introduce a probability measure $p_2(\boldsymbol{\theta}_2) \in \mathcal{P}(\mathbb{R}^N)$ to describe the overall states of the second layer and function $w_2 : \mathrm{supp}(p_1) \times \mathrm{supp}(p_2) \to \mathbb{R}$ to denote the weights on the connections from layer 1 to 2. We have for any $\boldsymbol{\theta}_2 \in \mathrm{supp}(p_2)$, we have the constraint for w_2 and p_1 that

$$\int w_2\left(\boldsymbol{w}_1, \boldsymbol{\theta}_2\right)\ \dot{h}\left(\boldsymbol{\theta}_1(\boldsymbol{w}_1)\right) dp_1\left(\boldsymbol{w}_1\right) = \boldsymbol{\theta}_2. \tag{97}$$

(4) Similarly, for $\ell \in [3:L]$, let $\boldsymbol{\theta}_\ell \in \mathbb{R}^N$ be the index of nodes according to the features. We introduce a probability measure $p_\ell(\boldsymbol{\theta}_\ell) \in \mathcal{P}(\mathbb{R}^N)$ to describe the states the ℓ-th layer and function $w_\ell : \mathrm{supp}(p_{\ell-1}) \times \mathrm{supp}(p_\ell) \to \mathbb{R}$ to denote the weights on the connections from layer $\ell-1$ to ℓ. We have any all $\boldsymbol{\theta}_\ell \in \mathrm{supp}(p_\ell)$, we have the constraint for w_ℓ and p_{ℓ_1} that:

$$\int w_\ell\left(\boldsymbol{\theta}_{\ell-1}, \boldsymbol{\theta}_\ell\right)\ \dot{h}\left(\boldsymbol{\theta}_{\ell-1}\right) dp_{\ell-1}\left(\boldsymbol{\theta}_{\ell-1}\right) = \boldsymbol{\theta}_\ell. \tag{98}$$

(5) Finally, let $w_{L+1} : \mathrm{supp}(p) \to \mathbb{R}$ be the weights in the layer $L+1$ and $\boldsymbol{\theta}_{L+1}$ be the final output, and we have the constraint for w_{L+1} and p_L that

$$\int w_{L+1}\left(\boldsymbol{\theta}_L\right)\ \dot{h}\left(\boldsymbol{\theta}_L\right) dp_L\left(\boldsymbol{\theta}_L\right) = \boldsymbol{\theta}_{L+1}. \tag{99}$$

Finally, a loss is imposed on the top layer and our target is to minimize the objective $\frac{1}{N}\sum_{n=1}^N \phi\left(\boldsymbol{\theta}_{L+1}(n), y^n\right)$. We can find that a static continuous DNN in our formulation is characterized by $\{w_\ell\}_{\ell=2}^{L+1}$ and $\{p_\ell\}_{\ell=1}^L$.

E.3. Scaled Gradient Descent for Training Fully-connected DNN

We still consider the scaled Gradient Descent algorithm to optimize the DNN. Given an initial weights $\hat{\boldsymbol{w}}^0$, the meta algorithm of the scaled Gradient Descent is shown in Algorithm 3, where the gradients $\hat{\mathcal{G}}_{\ell,i,j}^k$ can be obtained by the standard backward-propagation algorithm.

E.4. Neural Feature Flow for Training Continuous Fully-connected DNN

We derive the evolution of the Gradient Descent algorithm on a continuous DNN $\left(\{w_\ell\}_{\ell=2}^{L+1}, \{p_\ell\}_{\ell=1}^{L}\right)$. We first introduce the notations for the trajectories of $\boldsymbol{w}_1$, $\{w_\ell\}_{\ell=2}^{L}$, and $\{\boldsymbol{\theta}_\ell\}_{\ell=2}^{L}$:

- $\Psi_\ell^{\boldsymbol{\theta}} : \mathrm{supp}(p_\ell) \to C\left([0,T], \mathbb{R}^N\right)$ is the trajectory of $\boldsymbol{\theta}_\ell$ for $\ell \in [2:L]$;
- $\Psi_1^{\boldsymbol{w}} : \mathrm{supp}(p_1) \to C([0,T], \mathbb{R}^d)$ and $\Psi_{L+1}^{\boldsymbol{w}} : \mathrm{supp}(p_L) \to C([0,T], \mathbb{R})$ are the trajectories of $\boldsymbol{w}_1$ and w_{L+1}, respectively;
- $\Psi_\ell^{\boldsymbol{w}} : \mathrm{supp}(p_{\ell-1}) \times \mathrm{supp}(p_\ell) \to C([0,T], \mathbb{R})$ is the trajectory of w_ℓ for $\ell \in [2:L]$;
- Let Ψ be the collection of these trajectories.

The continuous gradient for the weight can be obtained from the backward-propagation algorithm. Especially, we define

$$\boldsymbol{\theta}_{L+1}(\Psi,t) := \int \Psi_{L+1}^{\boldsymbol{w}}(\boldsymbol{\theta}_L)(t)\, \dot{h}\left(\Psi_L^{\boldsymbol{\theta}}(\boldsymbol{\theta}_L)(t)\right) dp_L(\boldsymbol{\theta}_L), \tag{100a}$$

$$\begin{aligned}\overline{\mathcal{D}}_{L+1}(\Psi,t) &:= \left[\phi_1'\left(\boldsymbol{\theta}_{L+1}^t(1), y^1\right), \phi_1'\left(\boldsymbol{\theta}_{L+1}^t(2), y^2\right), \ldots, \phi_1'\left(\boldsymbol{\theta}_{L+1}^t(N), y^N\right)\right]^\top,\\ \overline{\mathcal{D}}_L(\boldsymbol{\theta}_L;\Psi,t) &:= \left[\Psi_{L+1}^{\boldsymbol{w}}(\boldsymbol{\theta}_L)(t)\ \overline{\mathcal{D}}_{L+1}(\Psi,t)\right] \circ \dot{h}'\left(\Psi_L^{\boldsymbol{\theta}}(\boldsymbol{\theta}_L)(t)\right),\end{aligned} \tag{100b}$$

$$\overline{\mathcal{D}}_\ell(\boldsymbol{\theta}_\ell;\Psi,t) := \left[\int \Psi_{\ell+1}^{\boldsymbol{w}}(\boldsymbol{\theta}_\ell,\boldsymbol{\theta}_{\ell+1})(t)\, \overline{\mathcal{D}}_{\ell+1}(\boldsymbol{\theta}_{\ell+1};\Psi,t)\, dp_{\ell+1}(\boldsymbol{\theta}_{\ell+1})\right] \circ \dot{h}'\left(\Psi_\ell^{\boldsymbol{\theta}}(\boldsymbol{\theta}_\ell)(t)\right), \tag{100c}$$

$$\overline{\mathcal{D}}_1(\boldsymbol{w}_1;\Psi,t) := \left[\int \Psi_2^{\boldsymbol{w}}(\boldsymbol{w}_1,\boldsymbol{\theta}_2)(t)\, \overline{\mathcal{D}}_2(\boldsymbol{\theta}_2;\Psi,t)\, dp_2(\boldsymbol{\theta}_2)\right] \circ \dot{h}'\left(\boldsymbol{\theta}_1\left(\Psi_1^{\boldsymbol{w}}(\boldsymbol{w}_1)(t)\right)\right), \tag{100d}$$

where in (100b), $\boldsymbol{\theta}_L \in \mathrm{supp}(p_L)$, in (100c), $\ell \in [2:L-1]$ and $\boldsymbol{\theta}_\ell \in \mathrm{supp}(p_\ell)$, and in (100d), $\boldsymbol{w}_1 \in \mathrm{supp}(p_1)$ and $\boldsymbol{\theta}_1(\cdot)$ is defined by (96). Then the gradient of the weights can be written as below.

$$\overline{\mathcal{G}}_{L+1}^{\boldsymbol{w}}(\boldsymbol{\theta}_L;\Psi,t) := \frac{1}{N}\left[\overline{\mathcal{D}}_{L+1}(\Psi,t)\right]^\top \dot{h}\left(\Psi_L^{\boldsymbol{\theta}}(\boldsymbol{\theta}_L)(t)\right), \tag{101a}$$

$$\overline{\mathcal{G}}_\ell^{\boldsymbol{w}}(\boldsymbol{\theta}_{\ell-1},\boldsymbol{\theta}_\ell;\Psi,t) := \frac{1}{N}\left[\overline{\mathcal{D}}_\ell(\boldsymbol{\theta}_\ell;\Psi,t)\right]^\top \dot{h}\left(\Psi_{\ell-1}^{\boldsymbol{\theta}}(\boldsymbol{\theta}_{\ell-1})(t)\right), \tag{101b}$$

$$\overline{\mathcal{G}}_2^{\boldsymbol{w}}(\boldsymbol{w}_1,\boldsymbol{\theta}_2;\Psi,t) := \frac{1}{N}\left[\overline{\mathcal{D}}_2(\boldsymbol{\theta}_2;\Psi,t)\right]^\top \dot{h}\left(\boldsymbol{\theta}_1\left(\Psi_1^{\boldsymbol{w}}(\boldsymbol{w}_1)(t)\right)\right), \tag{101c}$$

$$\overline{\mathcal{G}}_1^{\boldsymbol{w}}(\boldsymbol{w}_1;\Psi,t) := \frac{1}{N}\boldsymbol{X}^\top\left[\overline{\mathcal{D}}_1(\boldsymbol{w}_1;\Psi,t)\right], \tag{101d}$$

where in (101a), $\boldsymbol{\theta}_L \in \mathrm{supp}(p_L)$, in (101b), $\ell \in [3:L]$, $\boldsymbol{\theta}_{\ell-1} \in \mathrm{supp}(p_{\ell-1})$, and $\boldsymbol{\theta}_\ell \in \mathrm{supp}(p_\ell)$, in (101c), $\boldsymbol{w}_1 \in \mathrm{supp}(p_1)$ and $\boldsymbol{\theta}_2 \in \mathrm{supp}(p_2)$, and in (101d), $\boldsymbol{w}_1 \in \mathrm{supp}(p_1)$.

Moreover, we expect that the features satisfy the constraints:

$$\begin{aligned}\int \Psi_2^{\boldsymbol{w}}(\boldsymbol{w}_1,\boldsymbol{\theta}_2)(t)\, \dot{h}\left(\boldsymbol{\theta}_1\left(\Psi_1^{\boldsymbol{w}}(\boldsymbol{w}_1)(t)\right)\right) dp_1(\boldsymbol{w}_1) &= \Psi_2^{\boldsymbol{\theta}}(\boldsymbol{\theta}_2)(t), \quad \boldsymbol{\theta}_2 \in \mathrm{supp}(p_2),\\ \int \Psi_\ell^{\boldsymbol{w}}(\boldsymbol{\theta}_{\ell-1},\boldsymbol{\theta}_\ell)(t)\, \dot{h}\left(\Psi_{\ell-1}^{\boldsymbol{\theta}}(\boldsymbol{\theta}_{\ell-1})(t)\right) dp_{\ell-1}(\boldsymbol{\theta}_{\ell-1}) &= \Psi_\ell^{\boldsymbol{\theta}}(\boldsymbol{\theta}_\ell)(t), \quad \ell \in [3:L],\ \boldsymbol{\theta}_\ell \in \mathrm{supp}(p_\ell).\end{aligned}$$

So the drift term for the features can be obtained by the chain rule:

$$\overline{\mathcal{G}}_1^{\boldsymbol{\theta}}\left(\boldsymbol{w}_1;\Psi,t\right):=\frac{1}{d}\left[\boldsymbol{X}\,\overline{\mathcal{G}}_1^{\boldsymbol{w}}\left(\boldsymbol{w}_1;\Psi,t\right)\right], \tag{102a}$$

$$\begin{aligned}\overline{\mathcal{G}}_2^{\boldsymbol{\theta}}\left(\boldsymbol{\theta}_2;\Psi,t\right):=&\int\Psi_2^{\boldsymbol{w}}\left(\boldsymbol{w}_1,\boldsymbol{\theta}_2\right)(t)\left[\dot{h}'\Big(\boldsymbol{\theta}_1\big(\Psi_1^{\boldsymbol{w}}(\boldsymbol{w}_1)(t)\big)\Big)\circ\overline{\mathcal{G}}_1^{\boldsymbol{\theta}}\left(\boldsymbol{w}_1;\Psi,t\right)\right]dp_1\left(\boldsymbol{w}_1\right)\\&+\int\dot{h}\Big(\boldsymbol{\theta}_1\big(\Psi_1^{\boldsymbol{w}}(\boldsymbol{w}_1)(t)\big)\Big)\circ\overline{\mathcal{G}}_2^{\boldsymbol{w}}\left(\boldsymbol{w}_1,\boldsymbol{\theta}_2;\Psi,t\right)dp_1\left(\boldsymbol{w}_1\right),\end{aligned} \tag{102b}$$

$$\begin{aligned}\overline{\mathcal{G}}_\ell^{\boldsymbol{\theta}}\left(\boldsymbol{\theta}_\ell;\Psi,t\right):=&\int\Psi_\ell^{\boldsymbol{w}}\left(\boldsymbol{\theta}_{\ell-1},\boldsymbol{\theta}_\ell\right)(t)\left[\dot{h}'\left(\Psi_{\ell-1}^{\boldsymbol{\theta}}\left(\boldsymbol{\theta}_{\ell-1}\right)(t)\right)\circ\overline{\mathcal{G}}_{\ell-1}^{\boldsymbol{\theta}}\left(\boldsymbol{\theta}_{\ell-1};\Psi,t\right)\right]dp_{\ell-1}\left(\boldsymbol{\theta}_{\ell-1}\right)\\&+\int\dot{h}\left(\Psi_{\ell-1}^{\boldsymbol{\theta}}\left(\boldsymbol{\theta}_{\ell-1}\right)(t)\right)\circ\overline{\mathcal{G}}_\ell^{\boldsymbol{w}}\left(\boldsymbol{\theta}_{\ell-1},\boldsymbol{\theta}_\ell;\Psi,t\right)dp_{\ell-1}\left(\boldsymbol{\theta}_{\ell-1}\right),\end{aligned} \tag{102c}$$

where in (102a), $\boldsymbol{w}_1\in\operatorname{supp}(p_1)$, in (102b), $\boldsymbol{\theta}_2\in\operatorname{supp}(p_2)$, and in (102c), $\ell\in[3:L]$ and $\boldsymbol{\theta}_\ell\in\operatorname{supp}(p_\ell)$. Now we define the process of a continuous DNN trained by Gradient Descent, i.e., neural feature flow. It characterizes the evolution of both weights and features.

Definition 32 (Neural Feature Flow for DNN) *Given an initial continuous DNN* $\left(\{w_\ell\}_{\ell=2}^{L+1},\{p_\ell\}_{\ell=1}^{L}\right)$ *that satisfies* (96) – (99) *and* $T<\infty$*, we say a trajectory* Ψ_* *is a neural feature flow if for all* $t\in[0,T]$*,*

(1) for all $\ell\in[2:L]$ *and* $\boldsymbol{\theta}_\ell\in\operatorname{supp}(p_\ell)$*,*

$$\Psi_{*,\ell}^{\boldsymbol{\theta}}\left(\boldsymbol{\theta}_\ell\right)(t)=\boldsymbol{\theta}_\ell-\int_{s=0}^{t}\overline{\mathcal{G}}_\ell^{\boldsymbol{\theta}}\left(\boldsymbol{\theta}_\ell;\Psi_*,s\right)ds,$$

(2) for all $\boldsymbol{w}_1\in\operatorname{supp}(p_1)$*,*

$$\Psi_{*,1}^{\boldsymbol{w}}\left(\boldsymbol{w}_1\right)(t)=\boldsymbol{w}_1-\int_{s=0}^{t}\overline{\mathcal{G}}_1^{\boldsymbol{w}}\left(\boldsymbol{w}_1;\Psi_*,s\right)ds,$$

(3) for all $\boldsymbol{w}_1\in\operatorname{supp}(p_1)$ *and* $\boldsymbol{\theta}_2\in\operatorname{supp}(p_2)$*,*

$$\Psi_{*,2}^{\boldsymbol{w}}\left(\boldsymbol{w}_1,\boldsymbol{\theta}_2\right)(t)=w_2\left(\boldsymbol{w}_1,\boldsymbol{\theta}_2\right)-\int_{s=0}^{t}\overline{\mathcal{G}}_2^{\boldsymbol{w}}\left(\boldsymbol{w}_1,\boldsymbol{\theta}_2;\Psi_*,s\right)ds,$$

(4) for all $\ell\in[2:L-1]$*,* $\boldsymbol{\theta}_\ell\in\operatorname{supp}(p_\ell)$*, and* $\boldsymbol{\theta}_{\ell+1}\in\operatorname{supp}(p_{\ell+1})$*,*

$$\Psi_{*,\ell+1}^{\boldsymbol{w}}\left(\boldsymbol{\theta}_\ell,\boldsymbol{\theta}_{\ell+1}\right)(t)=w_{\ell+1}\left(\boldsymbol{\theta}_\ell,\boldsymbol{\theta}_{\ell+1}\right)-\int_{s=0}^{t}\overline{\mathcal{G}}_{\ell+1}^{\boldsymbol{w}}\left(\boldsymbol{\theta}_\ell,\boldsymbol{\theta}_{\ell+1};\Psi_*,s\right)ds,$$

(5) for all $\boldsymbol{\theta}_L\in\operatorname{supp}(p_L)$*,*

$$\Psi_{*,L+1}^{\boldsymbol{w}}\left(\boldsymbol{\theta}_L\right)(t)=w_{L+1}\left(\boldsymbol{\theta}_L\right)-\int_{s=0}^{t}\overline{\mathcal{G}}_{L+1}^{\boldsymbol{w}}\left(\boldsymbol{\theta}_L;\Psi_*,s\right)ds.$$

Algorithm 4 Initializing a Discrete DNN.

Input the data $\{\hat{\boldsymbol{\theta}}_{0,i}\}_{i=1}^{d}$ in (93), variance $\sigma_1 > 0$, and a constant C_3.
Independently draw $\hat{w}_{1,i,j} \sim p_0 = \mathcal{N}\left(0, d\sigma_1^2\right)$ for $i \in [d]$ and $j \in [m]$.
Set $\hat{\boldsymbol{\theta}}_{1,j} = \frac{1}{d}\sum_{i=1}^{d} \hat{w}_{1,i,j}\, \hat{\boldsymbol{\theta}}_{0,i}$ where $j \in [m]$. ◇ Standard Initialization for layer 1
for $\ell = 2, \ldots, L$ **do**
  Independently draw $\tilde{w}_{\ell,i,j} \sim \mathcal{N}\left(0, m\sigma_1^2\right)$ for $i, j \in [m]$.
  Set $\hat{\boldsymbol{\theta}}_{\ell,j} = \frac{1}{m}\sum_{i=1}^{m} \tilde{w}_{\ell,i,j}\, \dot{h}(\hat{\boldsymbol{\theta}}_{\ell-1,i})$ where $j \in [m]$. ◇ Standard Initialization for layer ℓ
end for
Set $\hat{w}_{L+1,i,1} = C$ where $i \in [m]$. ◇ Simply initialize $\{\hat{w}_{L+1,i,1}\}_{i=1}^{m}$ by a constant
for $\ell = 2, \ldots, L$ **do**
  for $j = 1, \ldots, m$ **do**
    Solve convex optimization problem: ◇ Perform ℓ_2-regression to reduce redundancy

$$\min_{\{\hat{w}_{\ell,i,j}\}_{i=1}^{m}} \frac{1}{m}\sum_{i=1}^{m} (\hat{w}_{\ell,i,j})^2, \quad \text{s.t. } \hat{\boldsymbol{\theta}}_{\ell,j} = \frac{1}{m}\sum_{i=1}^{m} \hat{w}_{\ell,i,j}\, \dot{h}(\hat{\boldsymbol{\theta}}_{\ell-1,i}).$$

  end for
end for
Output the discrete DNN parameters $(\hat{\boldsymbol{w}}, \hat{\boldsymbol{\theta}})$.

E.5. Informal Result

We still consider the DNN initialized by a standard initialization with an additional regression procedure, shown in Algorithm 4. We show in Theorem 6 that there is a neural feature flow that can capture the evolution of a DNN that is initialized by Algorithm 4 and trained by Gradient Descent, i.e, Algorithm 3.

Theorem 33 (Informal) *Under suitable conditions, there is an initialization* $(\{w_\ell\}_{\ell=2}^{L+1}, \{p_\ell\}_{\ell=1}^{L})$ *such that the continuous DNN has the following properties.*

(1) For any $T < \infty$*, there exists an unique neural feature flow* Ψ_* *satisfying Definition 32.*

(2) Suppose $\varepsilon \le \tilde{\mathcal{O}}(1)$*,* $\delta \le 1$*,* $m \ge \tilde{\Omega}(\varepsilon^{-2})$*, the step size* $\eta \le \tilde{\mathcal{O}}(\varepsilon)$*. Let* T *be a constant and* $K := \lfloor T/\eta \rfloor$*. Let* $\hat{\mathcal{L}}_N^k := \frac{1}{n}\sum_{n=1}^{N} \phi(\hat{\boldsymbol{\theta}}_{L+1,1}^k(n), y^n)$ *be the loss of running scaled Gradient Descent Algorithm 3 on a DNN initialized by Algorithm 4 at* k*-th step, and* $\mathcal{L}_N^t := \frac{1}{N}\sum_{n=1}^{n} \phi\left(\boldsymbol{\theta}_{L+1}(\Psi_*, t)(n), y^n\right)$ *be the loss of neural feature flow at time* t*. Then, with probability* $1-\delta$*,*

$$\sup_{k\in[0:K]} \left| \hat{\mathcal{L}}_N^k - \mathcal{L}_N^{k\eta} \right| \le \tilde{\mathcal{O}}(\varepsilon).$$

Appendix F. Simulations

In this section, we perform a toy simulation to validate our theory. We consider a synthetic 1-D regression task: $f(x) = \sin(x)$. We randomly generate $N = 100$ training samples uniformly from $[-\pi, \pi]$ and experiment on a four-hidden-layer ($L = 4$) NN. We choose the activation function h_1 and h_2 as $\tanh(x)$ and apply ℓ_2 loss, i.e., $\phi(y', y) = |y' - y|^2$. The weights are initialized by a

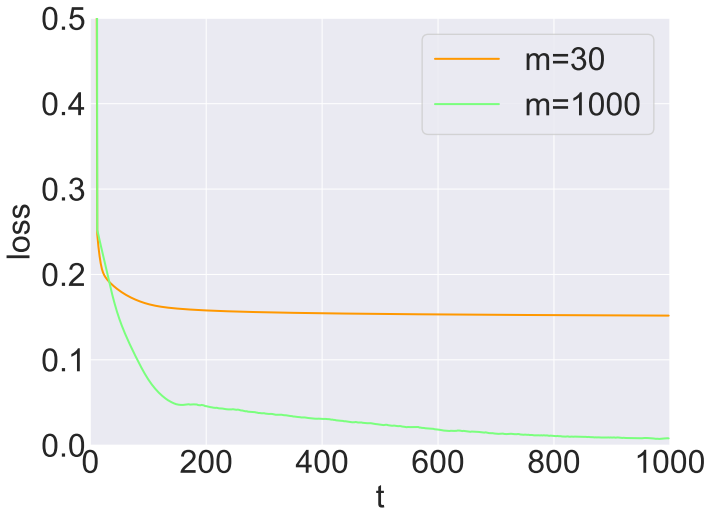

Figure 1: The experimental results on the synthetic data. m denotes the number of hidden units.

standard strategy (Glorot and Bengio, 2010) with an additional ℓ_2 regression in Algorithm 2. We see that the scaled GD in Algorithm 3 from the initialization achieves the global optimal solution for overparamterized DNNs.

The experimental result is shown Fig. 1. From the experiments, we can conclude that (1) the initialization and scaled GD are workable; (2) with the growth of the number of hidden units m, scaled GD achieves the global minimum.