MODELING FROM FEATURES: A MEAN-FIELD FRAMEWORK FOR OVER-PARAMETERIZED DEEP NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

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. Our analysis leads to the first global convergence proof for over-parameterized neural network training with more than 3 layers in the mean-field regime.

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 & Liang, 2018; Du et al., 2019a; Arora et al., 2019; Du et al., 2019b; Allen-Zhu et al., 2018; Allen-Zhu & 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 & Shamir (2019).

(ii) Another line of research applies the mean-field analysis for NNs (Mei et al., 2018; Chizat & Bach, 2018; Sirignano & Spiliopoulos, 2019a; Rotskoff & Vanden-Eijnden, 2018; Mei et al., 2019; Dou & Liang, 2019; Wei et al., 2018; Sirignano & Spiliopoulos, 2019b; Fang et al., 2019; Araujo et al., 2019; Nguyen & Pham, 2020; Chen et al., 2020). Learning a two-level 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 of McKean-Vlasov type (Sznitman, 1991). In the mean-field limit, the Wasserstein gradient flow converges to the globally optimal solution for two-level NNs (Mei et al., 2018; Chizat & 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, 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. Furthermore, understandings for the global convergence of DNNs are still limited
in the mean-field regime. Beyond two layers, the only result to the best of our knowledge came from \cite{Nguyen2020}, recently, in which they proved the global convergence for three-level DNNs under restrictive conditions. It is not clear how to extend their analysis to deeper NNs.

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 \cite{Araujo2019,Nguyen2020}. We also 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 \cite{He2016}. 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 \cite{Glorot2010,He2015}, the NN weights scale to infinity with the growth of the number of hidden units. There are empirical studies, e.g. \cite{Zhang2019}, which show that properly rescaling the standard initialization stabilizes training. We introduce a simple $\ell_2$-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. More or less surprisingly, we show that when the neural feature flow process converges, it reaches a globally optimal solution under suitable conditions. To the best of our knowledge, our analysis leads to the first proof for the global convergence of training over-parameterized DNNs with more than 3 layers in the mean-field regime. We conclude 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.

We also note that this paper presents the motivations and the primary outcomes of our new mean-field framework. We provide the key proofs in the appendix. In the supplementary material, we also attach the extended version of this paper, which similarly analyzes the fully-connected DNNs as well. We also consider general initialization conditions for both DNNs and Res-Nets and provides all proofs and more discussions.

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 $A \in \mathbb{R}^{n \times m}$, let $\|A\|_2$, $\|A\|_F$, and $\|A\|_\infty$ denote its operator, Frobenius, max norms, respectively. If $A$ is symmetric, let $\lambda_{\min}(A)$ be its smallest eigenvalue. Vectors are treated as columns. For a vector $a \in \mathbb{R}^n$, let $\|a\|_2$ and $\|a\|_\infty$ denote its $\ell_2$ and $\ell_\infty$ norms, respectively. The $i$-th coordinate is denoted by $a(i)$. For $a, b \in \mathbb{R}^n$, denote the entrywise product by $a \cdot b$ that $[a \cdot b](i) := a(i) \cdot b(i)$ for $i \in [n]$. For an unary function $f : \mathbb{R} \to \mathbb{R}$, define $\hat{f} : \mathbb{R}^n \to \mathbb{R}^n$ as the entrywise operation that $\hat{f}(a)(i) = f(a(i))$ for $i \in [n]$ and $a \in \mathbb{R}^n$. Denote $n$-dimensional identity matrix by $I^n$. Denote $m$-by-$n$ zero matrix and $n$-dimensional zero vector by $0^{m \times n}$ and $0^n$, respectively. For two positive sequences $\{p_n\}$ and $\{q_n\}$, $p_n = \mathcal{O}(q_n)$ if $p_n \leq C q_n$ for some positive constant $C$, and $p_n = \Omega(q_n)$ if $q_n = \mathcal{O}(p_n)$. Moreover, $p_n = \Omega(q_n) \text{ if } p_n = \mathcal{O}(q_n \log^{k} q_n)$ for some $k > 0$, and $p_n = \Omega(q_n)$ if $q_n = \mathcal{O}(p_n)$.

2 Challenges on Mean-field Theory for DNNs

We discuss related mean-field studies and point out the challenges in modeling DNNs. For two-level NNs, most of the existing works \cite{Mei2018,Chizat2018,Sirignano2019,Rotskoff2018,Eijnden2018} formulate the continuous limit as

$$f(x;p) = \int w_2 \, h \left( w_1^T x \right) \, dp \left( w_2, w_1 \right),$$
where \( p \) is the probability distribution over the pair of weights \((w_2, w_1)\). The weights of the second layer \( w_2 \) can be viewed as functions of \( w_1 \), which is a \( d \)-dimensional vector. However, this approach indexes higher-layer weights, say \( w_3 \), by functions over features of the hidden layer, with a diverging dimensionality in the mean-field limit. For 3-level NNs, \( w_3 \) as the last hidden layer is indexed by the connection to the output units in Nguyen & Pham (2020), which is not generalizable when middle layers present. An alternative approach is to model DNNs with nested measures (also known as multi-level 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 & Spiliopoulos, 2019b, Section 4.3).

The continuous limit of DNNs is investigated by Araújo et al. (2019); Nguyen & 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 \( \hat{\beta} \) of a middle-layer neuron connecting to \( m \) hidden neurons in the previous layer:

\[
\hat{\beta} = \frac{1}{m} \sum_{i=1}^{m} h(\hat{\beta}_i^\ell) \; w_i,
\]

where \( \hat{\beta}_i^\ell \) 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 \( \text{var}[\hat{\beta}] \to 0 \) as \( m \to \infty \). 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 is characterized in 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 equation (2.1) are carried out by Glorot & Bengio (2010); He et al. (2015) and motivate the popular initialization strategy with \( \mathcal{N}(0, \mathcal{O}(m)) \) such that the variance of \( \beta \) 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 \( \hat{\beta}_{\ell, i} \) denote the output of \( i \)-th hidden neuron at \( \ell \)-th layer and \( k \)-th iteration, and define \( \Delta_{\ell, m} := \max_{i \neq j, k \in [K]} \| \hat{\beta}_{\ell, i}^k - \hat{\beta}_{\ell, j}^k \|_\infty \). Then, for every \( \ell \in [2 : L - 1] \),

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

### 3 Formulation of Continuous Res-Nets

We consider the empirical minimization problem over \( N \) training samples \( \{x^i, y^i\}_{i=1}^N \), where \( x^i \in \mathbb{R}^d \) 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 \( m_\ell \) denote the number of units at layer \( \ell \) 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 \( \ell \) is denoted by \( \hat{\beta}_{\ell, i} \in \mathbb{R}^N \); the weight that connects the node \( i \) at layer \( \ell - 1 \) to node \( j \) at layer \( \ell \) is denoted by \( \hat{v}_{\ell, i, j} \in \mathbb{R} \).

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

\[
\hat{\beta}_{0, i} := (x^1(i), x^2(i), \ldots, x^N(i))^T.
\]

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

\[
\hat{\beta}_{1, j} = \frac{1}{m_0} \sum_{i=1}^{m_0} \hat{v}_{1, i, j} \hat{\beta}_{0, i}.
\]
respectively, where $h_1 : \mathbb{R} \rightarrow \mathbb{R}$ is the activation function and $\hat{h}_1 : \mathbb{R}^N \rightarrow \mathbb{R}^N$ is the entrywise operation for $h_1$, which satisfies $h_1(a)(i) = h_1(a(i))$ for $i \in [N]$ and $a \in \mathbb{R}^N$. Furthermore, we consider the following coupling between the residual and the previous feature:

$$
\hat{\beta}_{t,j} = \hat{h}_2(\hat{\alpha}_{t,j}) + \hat{\beta}_{t-1,j}, \quad j \in [m].
$$

(3.4)

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

(4) At the output layer,

$$
\hat{\beta}_{L+1,1} = \frac{1}{m} \sum_{i=1}^m \hat{v}_{L+1,i} \hat{h}_1(\hat{\beta}_{L,i}).
$$

(3.5)

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

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

(3.6)

where $(\hat{v}, \hat{\alpha}, \hat{\beta})$ satisfies equation 3.2 – equation 3.5 and $\phi : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ denotes the loss function. One noteworthy feature in the architecture is equation 3.4 where we introduce a mapping $h_2$ on the residual $\hat{\alpha}_{t,j}$ before fusing it with $\hat{\beta}_{t-1,j}$. We assume that $h_2$ is bounded by a constant $L_1$, and hence $\|\hat{\beta}_{t,j} - \hat{\beta}_{t-1,j}\|_\infty \leq L_1$. Therefore, the high-level features can be regarded as perturbations of the low-level ones. Similar ideas have also appeared in [Du et al., 2019a; Hardt & Ma, 2016], but realized in a different way. For example, in the lazy training regime, [Du et al., 2019a] achieved it by scaling $\hat{\alpha}_{t,j}$ with a vanishing $O\left(\frac{1}{\sqrt{m}}\right)$ factor.

3.2 CONTINUOUS RES-NETS FORMULATION

We propose our formulation for the continuous Res-Nets. The key structure of Res-Nets is the skip connections in equation 3.4. However, 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 $\Theta = (v_1, \alpha_2, \ldots, \alpha_L) \in \mathbb{R}^D$ for $D = d + (N - 1)L$ to parametrize the skip connections that are described in equation 3.5 and equation 3.9 below. Each $\Theta$ consists of $v_1, \alpha_2, \ldots, \alpha_L$ that can be regarded as an input-output path $v_1 \rightarrow \alpha_2 \rightarrow \ldots \rightarrow \alpha_L$, 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 $\beta_\ell$ in the hidden layer $\ell \in [L]$ as functions of $\Theta$ that we introduce next:

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

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

$$
\beta_1(\Theta) = \frac{1}{d} (Xv_1).
$$

(3.7)

(3) At layer $\ell \in [2 : L]$, let $v_\ell : \text{supp}(p) \times \text{supp}(p) \rightarrow \mathbb{R}$ denote the weights on the connections from layer $\ell - 1$ to $\ell$, then for all $\Theta = (v_1, \alpha_2, \ldots, \alpha_L) \in \mathbb{R}^D$, we have the forward-propagation constraint for $v_\ell$ and $p$:

$$
\alpha_\ell = \int v_\ell(\Theta, \tilde{\Theta}) \hat{h}_1(\beta_{\ell-1}(\tilde{\Theta})) \, dp(\tilde{\Theta}),
$$

(3.8)

$$
\beta_\ell(\Theta) = \hat{h}_2(\alpha_\ell) + \beta_{\ell-1}(\Theta).
$$

(3.9)

Here, $\Theta$ takes on values in $\mathbb{R}^D$ and for each $\Theta$, $\alpha_\ell$ is one part of $\Theta$ and $\beta_\ell$ is a function of $\Theta$. $\alpha_\ell$ represents the residual at layer $\ell$ on the skip connected path described by $\Theta$. And $\beta_\ell(\Theta)$ represents the corresponding feature.
At the output layer, let $v_{L+1}$ be the weights in the layer $L + 1$, and we have

$$
\beta_{L+1} = \int v_{L+1} (\Theta) \dot{h}_1 (\beta_L (\Theta)) dp (\Theta).
$$

Now we describe the continuous limit of the Res-Net trained by Algorithm 1 by the continuous trajectories of the Res-Nets. A trajectory is denoted by $\Phi$ that maps the initial Res-Nets at $t = 0$ to a Res-Net process over $[0, T]$. Specifically, it consists of the following parts:

- $\Phi^\beta : \text{supp}(p) \to C([0, T], \mathbb{R}^N)$ is the trajectory of $\beta\ell$ for $\ell \in [L]$;
- $\Phi^\alpha : \text{supp}(p) \to C([0, T], \mathbb{R}^N)$ is the trajectory of $\alpha\ell$ for $\ell \in [2 : L]$;
- $\Phi^\gamma : \text{supp}(p) \to C([0, T], \mathbb{R}^D)$ and $\Phi^\gamma_{L+1} : \text{supp}(p) \to C([0, T], \mathbb{R})$ are the trajectories of $v_1$ and $v_{L+1}$, respectively;
- $\Phi^\gamma : \text{supp}(p) \times \text{supp}(p) \to C([0, T], \mathbb{R})$ is the trajectory of $v_{L+1}$ for $\ell \in [2 : L]$.

The continuous gradient for the weight can be obtained from the backward-propagation algorithm. For a given trajectory $\Phi$, 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 $\Theta = (v_1, \alpha_2, \ldots, \alpha_L) \in \text{supp}(p)$, $t \in [0, T]$, and $\ell \in [2 : L]$, let

$$
\beta_{L+1} (\Phi, t) := \int \Phi_{L+1}^\gamma (\Theta) (t) \dot{h}_1 (\Phi_L^\alpha (\Theta) (t)) dp (\Theta),
$$

$$
\partial_{\ell+1} (\Phi, t) := \{ \phi' (\beta_{L+1} (\Phi, t) (n), y^n) : n \in [N] \},
$$

$$
\partial_{\ell}^\alpha (\Theta; \Phi, t) := [\Phi_{L+1}^\gamma (\Theta) (t) \partial_{\ell+1} (\Phi, t)] : \dot{h}' (\Phi_L^\alpha (\Theta) (t)),$n

$$
\partial_{\ell}^\alpha (\Theta; \Phi, t) := \partial_{\ell}^\alpha (\Theta; \Phi, t) \dot{h}' (\Phi_L^\alpha (\Theta) (t)) + \left[ \Phi_{L+1}^\gamma (\Theta, \Theta) (t) \partial_{\ell}^\alpha (\Theta, \Phi, t) dp (\Theta) \right] : \dot{h}' (\Phi_{\ell-1}^\gamma (\Theta) (t)).
$$

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

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

$$G_{l+1}^\alpha (\theta; \Phi, t) := \frac{1}{N} [D_{l+1} (\Phi, t)]^T h_1 \left( \Phi_{l}^\alpha (\theta) (t) \right) ,$$

$$G_{l}^\alpha (\theta, \Theta, \Phi, t) := \frac{1}{N} [D_{l}^\alpha (\theta, \Theta, \Phi, t)]^T h_1 \left( \Phi_{l-1}^\alpha (\theta) (t) \right) , \quad \ell \in [2 : L],$$

$$G_{l}^\alpha (\theta; \Phi, t) := \frac{1}{N} X D_{l}^\alpha (\theta; \Phi, t) .$$

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 all $\ell \in [L - 1]$ and $\Theta \in \text{supp}(p)$,

$$G_{l}^\beta (\theta; \Phi, t) := \frac{1}{d} [X G_{l+1}^\gamma (\theta; \Phi, t)] ,$$

$$G_{l+1}^\gamma (\theta; \Theta, \Phi, t) := \int \Phi_{l+1}^\gamma (\theta, \Theta) (t) \left[ h_1 \left( \Phi_{l}^\beta (\theta) (t) \right) \cdot G_{l}^\alpha (\theta, \Theta, \Phi, t) \right] dp (\Theta) \quad +$$

$$+ \int h_1 \left( \Phi_{l}^\beta (\theta) (t) \right) \cdot G_{l+1}^\gamma (\theta, \Theta, \Phi, t) dp (\Theta) ,$$

$$G_{l+1}^\gamma (\theta, \Phi, t) := G_{l}^\gamma (\theta, \Phi, t) + G_{l+1}^\alpha (\theta, \Phi, t) h_2 \left( \Phi_{l+1}^\alpha (\theta) (t) \right) .$$

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

**Definition 1** (Neural Feature Flow for Res-Net). Given an initial continuous Res-Net $(\{v_k\}_{k=1}^{L+1}, p)$ that satisfies the equation 3.7—equation 3.10 and $T < \infty$, we say a trajectory $\Phi_*$ is a neural feature flow if for all $\theta = (\alpha_1, \alpha_2, \ldots, \alpha_L) \in \text{supp}(p), \Theta \in \text{supp}(p)$, and $t \in [0, T]$, $\Phi_\alpha (\theta ; t) = v_1 - \int_0^t G_{l}^\alpha (\theta; \Phi_*, s) ds,$

$$\Phi_\alpha (\theta ; t) = \alpha L - \int_0^t G_{l}^\alpha (\theta; \Phi_*, s) ds,$$

$$\Phi_\alpha (\theta ; t) = v_1 - \int_0^t G_{l}^\alpha (\theta; \Phi_*, s) ds,$$

$$\Phi_\alpha (\theta ; t) = v_1 (\theta; \Theta) - \int_0^t G_{l}^\alpha (\theta, \Theta, \Phi_*, s) ds,$$

$$\Phi_\alpha (\theta ; t) = v_1 (\theta, \Theta) - \int_0^t G_{l}^\alpha (\theta, \Theta, \Phi_*, s) ds.$$

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

5 Main Results

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 $x \in \mathbb{R},$

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

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

$$|h_2 (x) - h_2 (y)| \leq L_3 |x - y|, \quad |h_2 (x) - h_2 (y)| \leq L_3 |x - y|.$$  

For the loss function, we assume that there exist constants $L_4, L_5, L_6 > 0$ such that, for all $y \in \mathcal{Y}, x_1 \in \mathbb{R},$ and $x_2 \in \mathbb{R},$

$$|\phi_1 (x_1, y)| \leq L_4, \quad |\phi_1 (x_1, y) - \phi_1 (x_2, y)| \leq L_5 |x_1 - x_2|.$$  

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].
Algorithm 2: Initialize a Discrete Res-Net.

1: Input the data $X$, variance $\sigma_I$, and a constant $C_7$.
2: Independently draw $\tilde{v}_{1,i,j} \sim p_0 = \mathcal{N}(0, d\sigma_I^2)$ for $i \in [d]$ and $j \in [m]$.
3: Set $\beta_{1,j} = \frac{1}{d} \sum_{i=1}^{d} \tilde{v}_{1,i,j} \beta_{0,i}$ where $j \in [m]$. $\triangleleft$ Standard Initialization for layer 1
4: for $\ell = 2, \ldots, L$ do
5: Independently draw $\tilde{v}_{\ell,i,j} \sim \mathcal{N}(0, m\sigma_I^2)$ for $i, j \in [m]$.
6: Set $\bar{\alpha}_{\ell,j} = \frac{1}{m} \sum_{i=1}^{m} \tilde{v}_{\ell,i,j} h_1(\beta_{\ell-1,i})$ where $j \in [m]$. $\triangleleft$ Standard Initialization for layer $\ell$
7: Set $\beta_{\ell,j} = \beta_{\ell-1,j} + \bar{h}_2(\bar{\alpha}_{\ell,j})$ for $j \in [m]$.
8: end for
9: Set $\tilde{v}_{L+1,i,1} = C_7$ where $i \in [m]$. $\triangleleft$ Simply initialize $\{\tilde{v}_{L+1,i,1}\}_{i=1}^{m}$ by a constant
10: for $\ell = 2, \ldots, L$ do
11: for $j = 1, \ldots, m$ do
12: Solve convex optimization problem: $\min_{\{\tilde{v}_{\ell,i,j}\}_{i=1}^{m}} \frac{1}{m} \sum_{i=1}^{m} (\tilde{v}_{\ell,i,j})^2$, s.t. $\bar{\alpha}_{\ell,j} = \frac{1}{m} \sum_{i=1}^{m} \tilde{v}_{\ell,i,j} h_1(\beta_{\ell-1,i})$. (5.2)
13: end for
14: end for
15: Output the discrete Res-Net parameters $(\vec{v}, \bar{\alpha}, \bar{\beta})$.

Assumption 2 (Strong Universal Approximation Property). Assume that for any function $f_2 : \mathbb{R}^d \rightarrow \mathbb{R}^N$ that is bounded by $C_B$, i.e., for all $v_1 \in \mathbb{R}^d$, $\|f_2(v_1)\|_{\infty} \leq C_B$, we have

$$\min \left[ \int h_1 \left( \frac{1}{d} Xv_1 + f_2(v_1) \right) \right] \left[ h_1 \left( \frac{1}{d} Xv_1 + f_2(v_1) \right) \right]^T dp_1 (v_1) \geq \tilde{\lambda} > 0. \quad (5.1)$$

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

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. 2019) Lemma F.1 that the assumption holds for all analytic non-polynomial $h_1$. Lemma 1 affords many examples that satisfy the assumption for constant $C_B$.

Lemma 1. Suppose that the data is non-parallel, i.e., $x_i \notin \text{Span}(x_j)$ for all $i \neq j$.

(i) If $g : \mathbb{R} \rightarrow \mathbb{R}$ is a non-polynomial function, 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 $h_2(x) = c|x|^{-\alpha}$ for $|x| > c'$, where $c, c', \alpha > 0$, then $h_1$ satisfies Assumption 2.

We consider the Res-Net initialized by Algorithm 2 which is composed of a standard initialization (Glorot & Bengio 2010; He et al. 2015) and an additional regression procedure while preserving all initial features. The standard initialization strategy scales the weights to $\sqrt{m}$, which diverges in the mean-field limit. We perform the simple $\ell_2$-regression to reduce the redundancy of the weights.

Now we start the analysis. First, we show in Theorem 1 that there is a neural feature flow that can capture the evolution of a Res-Net that is initialized by Algorithm 2 and trained by Gradient Descent, i.e., Algorithm 1.

Theorem 1. Under Assumptions 1 and 2, there is an initialization $(\{v_\ell\}_{\ell=2}^{L+1}, p)$ such that the continuous Res-Net has the following properties.

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 $\ell_2$-regression. The $\ell_2$-regression equation (5.2) can be replaced by a soft version

$$\min_{\{\tilde{v}_{\ell,i,j}\}_{i=1}^{m}} \left( \lambda \frac{m}{m} \sum_{i=1}^{m} (\tilde{v}_{\ell,i,j})^2 + \left\| \bar{\alpha}_{\ell,j} - \frac{1}{m} \sum_{i=1}^{m} \tilde{v}_{\ell,i,j} h_1(\beta_{\ell-1,i}) \right\|^2 \right). \quad (5.3)$$
(1) For any $T < \infty$, there exists an unique neural feature flow $\Phi_*$ satisfying Definition 1.

(2) Suppose $\varepsilon \leq \tilde{O}(1)$, $\delta \leq 1$, $m \geq \Omega(\varepsilon^{-2})$, the step size $\eta \leq \tilde{O}(\varepsilon)$. Let $T$ be a constant and $K := \lfloor T/\eta \rfloor$. Let $\tilde{L}^k := \frac{1}{N} \sum_{n=1}^{N} \phi(\tilde{\beta}_{L+1}^k(n), y^n)$ be the loss of running scaled Gradient Descent Algorithm 1 on a Res-Net initialized by Algorithm 1 at $k$-th step, and $L^\ell := \frac{1}{N} \sum_{n=1}^{N} \phi(\beta_{L+1}(\Phi_*, t)(n), y^n)$ be the loss of neural feature flow at time $t$. Then, with probability $\sup_{k \in [0, K]} |\tilde{L}^k - L^{k\eta}| \leq O(\varepsilon)$.

The existence and uniqueness of $\Phi_*$ follows from 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 second result shows that $\Phi_*$ can approximate the training process of a sufficiently over-parametrized Res-Net. In the discrete DNNs, although the connecting weights are independently initialized, the features $\theta_{l,j}$ are not mutually independent since they all depend on a common set of random outputs from the previous layer. Our key observation is that the skip-connected paths of the discrete Res-Net $\{v_{1,1}, \alpha_{2,1}, \ldots, \alpha_{L,1}\}_{i=1}^{m}$, from the standard initialization are nearly independent when $m$ is sufficiently large, 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 compare an ideal process determined by $\Phi_*$ with the discrete one from Algorithm 1.

Finally, we consider the global convergence of the neural feature flow. We show in Theorem 2 that the neural feature flow always finds a globally optimal solution when it converges.

**Theorem 2.** Under Assumptions 1 and 2, assume that the loss function $\phi$ is convex in the first argument. Let $\{v_{1,1}, \alpha_{2,1}, \ldots, \alpha_{L,1}\}_{i=1}^{m}$ be the initial continuous Res-Net in Theorem 1, and $\Phi_*$ and $L^\ell$ be the solution and loss of the neural feature flow, respectively. If $\Phi_{*,L}^\Theta(t)$ converges in $\ell_\infty(p)$ and $\Phi_{*,L+1}^\Theta(t)$ converges in $\ell_1(p)$ as $t \to \infty$, where $\Theta \sim p$, then we have

$$\lim_{t \to \infty} L^\ell = \sum_{n=1}^{N} \min_{y'} \phi(y', y^n).$$

Theorem 2 is an important application of our mean-field framework, which shows that neural feature flow can find a global minimizer after it converges. 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 & Bach [2018], we show that all bad local minima are unstable. Our global convergence holds for Res-Nets with arbitrary (finite) depth. Before us, the global convergence result was proved only for two-level NNs [Mei et al. 2018; Chizat & Bach 2018], and more recently for three-level ones [Nguyen & Pham 2020] under a similar convergence assumption on the weights in the second layer.

**6 Conclusions**

This paper proposed a new mean-field framework for DNNs where features in hidden layers have non-vanishing variances. 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 this framework to be further investigated:

(A) It is not clear whether the dynamics of DNNs trained by Gradient Descent can be characterized by PDEs of Mckean-Vlasov type. Recently [Araujo et al. 2019] pointed out the difficulty lied in the potential discontinuity of the conditional distribution under Wasserstein metric. From the viewpoint of our framework, the features of the hidden units potentially collide with others along the evolution.

(B) It is not answered in this paper how to analyze the evolution of DNN with special regularizers such as relative entropy regularizer. Can we prove that Gradient Descent find a global minimum under such regularizers? More discussions about fusing suitable regularizers is shown in Section F of supplementary material.

(C) This paper only studies the optimization aspect of DNNs. It is interesting to study the generalization property. There are two potential directions: we may incorporate regularizers on the DNNs to control the model complexity; implicit regularization is often observed in practice, which is hopefully preserved in the neural feature flow.

We encourage the readers to see more discussions in the full version of this paper in the supplementary material.
REFERENCES


A Appendix

In this appendix, we present the key proofs for the results in the main text, and we defer the technical proofs to the full version of this paper in the supplementary material. In the proof, we fix a set of non-parallel 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.

The appendix is organized as follows: Section 1 introduces more notations that will be used in our proof. Section 2 provides the key proof for Theorem 1 in which the proofs of Theorems 2 – 5 are deferred to supplementary material. Section 3 proves Theorem 2 in which an additional lemma (Lemma 16 in the supplementary material) is used. Finally, in Section 4 we provide the proofs for Lemma 1 and Proposition 1.

A.1 Additional Notations

We introduce more notations which shall be used in our proof. For some argument. In detail, there is a constant $C$.

Moreover, we assume that $C$.

Theorem 2, in which an additional lemma (Lemma 16 in the supplementary material) is used. Finally, in Section A.4 we provide the key proof for Theorem 1, in which the proofs of Theorems 3 – 5 are deferred to supplementary material. Section A.3 proves the initial Res-Net with mutually independent features such that the distance to the actual discrete Res-Net is vanishing with $m$.

Road Map: In Section A.2.1 we propose the general initial conditions for continuous and discrete Res-Nets, respectively. In Section A.2.2 we analyze the behaviors of the Res-Nets under those general initial conditions. We first study the continuous Res-Net and the continuous Res-Net.

A.2 Proof of Theorem 1

In Theorem 1 the well-posedness of the neural feature flow in Definition 1 can be established under mild conditions on the initialization including the tail of initial distribution $p$ and the continuity of the initial weights (see Theorem 3). To prove Theorem 1 it remains to construct an initial continuous Res-Net satisfying those conditions and show the corresponding neural feature flow is close to the training trajectory of the scaled Gradient Descent on a sufficiently overparametrized discrete Res-Net initialized by Algorithm 2.

However, in the discrete Res-Nets, although the connecting weights are independently initialized, the features are not mutually independent since they all depend on a common set of random outputs from the previous layer. Our crucial observation is that those features are almost independent when the width $m$ of the hidden layers are sufficiently large; namely, there exists an ideal initial Res-Net with mutually independent features such that the distance to the actual discrete Res-Net is vanishing with $m$.

A precise definition is given in Definition 2. This allows us to construct an ideal process to bridge the trajectories of the discrete Res-Net and the continuous Res-Net.

Road Map: In Section A.2.1 we propose the general initial conditions for continuous and discrete Res-Nets, respectively. In Section A.2.2 we analyze the behaviors of the Res-Nets under those general initial conditions. We first study the continuous Res-Net and the continuous Res-Net.

A.2.1 Definitions

We first propose the initial condition for the continuous Res-Nets. In our proof, unless otherwise specified, any initial continuous Res-Net $(v_1, v_2, \ldots, v_{L+1})$ is needed to satisfies the forward propagation constraints, i.e., equation 5.7 – equation 5.10.

Assumption 3 (Initialization for Continuous-Net). We assume that $p$ is $\sigma$-sub-gaussian distribution. We assume that, for all $\ell \in \{1 \leq L\}$, $v_\ell(\cdot, \cdot)$ has sublinear growth on the second argument, that is, there is a constant $C_5$ such that

$$
|v_\ell(\Theta_1, \Theta_2)| \leq C_5 (1 + \|\Theta_2\|_\infty), \quad \text{for all } \Theta_1, \Theta_2 \in \text{supp}(p), \ell \in \{1 \leq L\}.
$$

Moreover, we assume that $v_\ell(\cdot, \cdot)$ are locally Lipschitz continuous where the Lipschitz constant has sub-linear growth on the second argument. In detail, there is a constant $C_6$, such that for $\Theta_1 \in \text{supp}(p), \Theta_1 \in \text{supp}(p) \cap B_\infty(\Theta_1, 1), \Theta_2 \in \text{supp}(p)$, and $\Theta_2 \in \text{supp}(p) \cap B_\infty(\Theta_2, 1)$, we have

$$
|v_\ell(\Theta_1, \Theta_2) - v_\ell(\tilde{\Theta}_1, \tilde{\Theta}_2)| \leq C_6 (1 + \|\Theta_2\|_\infty) \left(\|\Theta_1 - \tilde{\Theta}_1\|_\infty + \|\Theta_2 - \tilde{\Theta}_2\|_\infty\right).
$$

For the last layer, there exist constants $C_7$ and $C_8$, such that for all $\Theta, \Theta \in \text{supp}(p)$, we have

$$
|v_{L+1}(\Theta)| \leq C_7 \quad \text{and} \quad |v_{L+1}(\Theta) - v_{L+1}(\tilde{\Theta})| \leq C_8 \|\Theta - \tilde{\Theta}\|_\infty.
$$

Based on the initial condition for the continuous Res-Net, we then introduce the initial condition for discrete Res-Net.

Definition 2 $(\varepsilon_1$-Independent Initial Discrete Res-Net). We say an initial discrete Res-Net $(\tilde{v}, \tilde{\alpha}, \tilde{\beta})$ is $\varepsilon_1$-independent if there exist a continuous initial Res-Net $(\{v_\ell\}_{\ell=2}^{L+1}, p)$ satisfying Assumption 3 and $(\tilde{v}, \tilde{\alpha}, \tilde{\beta})$ such that

$\varepsilon_1$ Here the value $\varepsilon$ can be replaced by any number greater than one. See Vershynin [2010] Remark 5.6.
(1) $\tilde{\Theta}_i = (\tilde{v}_{1,i}, \alpha_{2,i}, \ldots, \alpha_{L,i}) \iid p$.

(2) For $\tilde{\beta}$ and $\Phi$,

- $\tilde{\beta}_{\ell,i} = \frac{1}{\alpha_\ell} (Xv_{1,i}) + \sum_{t=2}^{L} \tilde{h}_2(\alpha_{t,i})$ for $\ell \in [L]$ and $i \in [m]$;
- $v_{\ell,i,j} = v_{\ell}(\tilde{\Theta}_i, \tilde{\Theta}_j)$ for $\ell \in [2 : L]$, $i, j \in [m]$;
- $\bar{v}_{L+1,i,1} = v_{L+1}(\tilde{\Theta}_i)$ for $i \in [m]$;

(3) $\varepsilon_1$-closeness:

- $\|\tilde{v}_{1,i} - \bar{v}_{1,i}\|_{\infty} \leq (1 + \|\bar{\Theta}_i\|_{\infty}) \varepsilon_1$ for $i \in [m]$;
- $|\bar{v}_{\ell+1,i,j} - \bar{v}_{\ell+1,i,j}| \leq (1 + \|\bar{\Theta}_i\|_{\infty} + \|v_{\ell,i,j}\|_{\infty}) \varepsilon_1$ for $\ell \in [L-1]$, $i, j \in [m]$;
- $|\bar{v}_{L+1,i,1} - \bar{v}_{L+1,i,1}| \leq (1 + \|\bar{\Theta}_i\|_{\infty}) \varepsilon_1$ for $i \in [m]$.

A.2.2 General Analysis

We analyze the neural feature flow for the continuous Res-Net under the general conditions in Section A.2.1. The following theorem guarantees the existence and uniqueness.

Theorem 3 (Existence and Uniqueness of Neural Feature Flow on Res-Net). Under Assumptions 7 and 2 for any $T < \infty$, there exists an unique neural feature flow $\Phi_\ast$.

Moreover, we show that $\Phi_\ast$ is a continuous mapping on $\Theta$ given time $t$, which is stated as follows. Theorem 4 will be used in the proof of global convergence in Theorem 2.

Theorem 4 (Property of $\Phi_\ast$). Under Assumptions 7 and 2 let $\Phi_\ast$ be the neural feature flow, there exist constants $C, C' \geq 0$ such that for all $t \in [0, T]$, $\Theta_1 \in \text{supp}(p)$ and $\tilde{\Theta}_1 \in \text{supp}(p) \cap B_\infty (\Theta_1, 1)$, $\Theta_2 \in \text{supp}(p)$, and $\tilde{\Theta}_2 \in \text{supp}(p) \cap B_\infty (\Theta_2, 1)$, we have

$$
\|\Phi_\ast^\ell, t(\tilde{\Theta}_1) - \Phi_\ast^\ell, t(\Theta_1)\|_{\infty} \leq C e^{C' t} (\|\Theta_1\|_{\infty} + 1) \|\tilde{\Theta}_1 - \Theta_1\|_{\infty}, \quad \ell \in [L],
$$

$$
\|\Phi_\ast^\ell, t(\tilde{\Theta}_1) - \Phi_\ast^\ell, t(\tilde{\Theta}_2)\|_{\infty} \leq C e^{C' t} (\|\Theta_1\|_{\infty} + 1) \|\tilde{\Theta}_1 - \tilde{\Theta}_2\|_{\infty}, \quad \ell \in [2 : L],
$$

$$
\|\Phi_\ast^\ell, t(\tilde{\Theta}_1) - \Phi_\ast^\ell, t(\tilde{\Theta}_2)\|_{\infty} \leq C e^{C' t} (\|\Theta_1\|_{\infty} + 1) \|\tilde{\Theta}_1 - \tilde{\Theta}_2\|_{\infty}, \quad \ell \in [2 : L],
$$

$$
\|\Phi_\ast^\ell, t(\tilde{\Theta}_1) - \Phi_\ast^\ell, t(\tilde{\Theta}_2)\|_{\infty} \leq C e^{C' t} (\|\Theta_1\|_{\infty} + 1) \|\tilde{\Theta}_1 - \tilde{\Theta}_2\|_{\infty}, \quad \ell \in [2 : L],
$$

The proofs of Theorems 3 and 4 follow from 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. The details are shown in the proofs of Theorem 5 and 6 in the full version of the paper in the supplementary material.

In the following, we analyze the trajectory of the discrete Res-Net from an $\varepsilon_1$-independent initialization. Recall Definition 2 that an $\varepsilon_1$-independent initialization induces a continuous Res-Net satisfying Assumption 3 which yields an unique neural feature flow $\Phi_\ast$ by Theorem 3. We show that scaled Gradient Descent from an $\varepsilon_1$-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 (Zinman (1991)). We compare the scaled Gradient Descent with an ideal discrete process determined by $\Phi_\ast$ as specified below:

- Actual process $(\tilde{v}_{[0,K]}, \tilde{\alpha}_{[0,K]}, \tilde{\beta}_{[0,K]})$ by executing Algorithm 1 in $K = \frac{T}{\eta}$ steps on the discrete Res-Net from $(\tilde{v}, \tilde{\alpha}, \tilde{\beta})$;
Ideal process \( \tilde{\beta}_1^\ell, \tilde{\alpha}_1^\ell, \tilde{\psi}_1^\ell, \tilde{\psi}_1^\ell, \tilde{\alpha}_1^{\ell, i, j}, \tilde{\psi}_1^{\ell, i, j} \) that evolves as neural feature flow:

\[
\begin{align*}
\tilde{\beta}_1^\ell, i, t & = \Phi_{\ell, t}^\beta (\tilde{\Theta}_i) (t), \quad \ell \in [L], \ i \in [m], \ t \in [0, T], \\
\tilde{\alpha}_1^\ell, i, t & = \Phi_{\ell, t}^\alpha (\tilde{\Theta}_i) (t), \quad \ell \in [2 : L], \ i \in [m], \ t \in [0, T], \\
\tilde{\psi}_1^\ell, i, t & = \Phi_{\ell, t}^\psi (\tilde{\Theta}_i) (t), \quad \ell \in [0, T], \\
\tilde{\psi}_1^{\ell, i, j} & = \Phi_{\ell, t}^{\psi_{i, j}} (\tilde{\Theta}_i, \tilde{\Theta}_j) (t), \quad \ell \in [2 : L], \ i \in [m], \ i \in [m], \ t \in [0, T], \\
\tilde{\psi}_1^{\ell, i, j} & = \Phi_{\ell, t}^{\psi_{i, j}} (\tilde{\Theta}_i, \tilde{\Theta}_j) (t), \quad \ell \in [0, T].
\end{align*}
\]

We also compare the discrete and the continuous losses denoted by \( \hat{L}_R^k := \frac{1}{n} \sum_{i=1}^{n} \hat{r}_R^k (\tilde{\beta}_{L+1,i}^k (n), y^n) \) and \( L_R^k := \frac{1}{n} \sum_{i=1}^{n} r_R (\beta_{L+1,i}^k (n), y^n) \). We have the following:

**Theorem 5.** Suppose \( \varepsilon_1 \leq O(1) \) and \( m \geq \tilde{O}(\varepsilon_1^{-2}) \), and test the parameters in assumptions and \( T \) as constants. Consider the actual process from an \( \varepsilon_1 \)-independent initialization in Definition 2, with step size \( \eta \leq \tilde{O}(\varepsilon_1) \). Then, the following holds with probability \( 1 - \delta \):

1. The two processes are close to each other:

\[
\sup_{k \in [0, K_1]} \left\{ \sup_{i \in [m]} \left\| \tilde{\psi}_1^k - \tilde{\psi}_1^{k, \eta} \right\|_\infty, \sup_{\ell \in [2 : L], i, j \in [m]} \left\| \tilde{\psi}_1^{\ell, i, j} - \tilde{\psi}_1^{\ell, i, j} \right\|_\infty \right\} \leq \tilde{O}(\varepsilon_1),
\]

\[
\sup_{k \in [0, K_1]} \left\{ \sup_{i \in [m]} \left\| \tilde{\alpha}_1^k - \tilde{\alpha}_1^{k, \eta} \right\|_\infty, \sup_{\ell \in [2 : L], i} \left\| \tilde{\alpha}_1^{\ell, i} - \tilde{\psi}_1^{\ell, i} \right\|_\infty, \sup_{\ell \in [L]} \left\| \tilde{\beta}_1^\ell - \tilde{\beta}_1^{\eta} \right\|_\infty \right\} \leq \tilde{O}(\varepsilon_1).
\]

2. The training losses are also close to each other:

\[
\sup_{k \in [0, K_1]} \left| \hat{L}_R^k - L_R^k \right| \leq O(\varepsilon_1).
\]

The proof of Theorem 5 is a Grönwall-type result by a similar argument to the “propagation of chaos”. The details are shown in the proof of Theorem 7 in the full version of this paper in the supplementary material.

**A.2.3 Verification of Algorithm 2**

In this subsection, we verify that Algorithm 2 produces an \( \varepsilon_1 \)-independent initial discrete Res-Net with \( \varepsilon_1 \leq \tilde{O}(1/\sqrt{m}) \). By Definition 2, 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 2. We specify \( p, \{v_\ell\}_{\ell=2}^{L+1} \), and the ideal discrete Res-Net below, and then we verify the properties in Lemma 2.

**Initial distribution.** We first define the distribution \( p \):

1. At the first layer, \( \beta_1 \sim p_1^\beta = \mathcal{N} (0^N, \sigma_1^2 K_0) \), where \( K_0 := \frac{1}{\beta} XX^\top \).
2. At the layer \( \ell \in [L - 1] \), let \( K_\ell^\beta := \int h_1 (\beta_i) h_\ell (\beta_\ell)^\top dp_\ell^\beta (\beta_\ell) \). We define the distribution of the residuals at layer \( \ell + 1 \) as

\[
p_{\ell+1}^\alpha = \mathcal{N} (0^N, \sigma_{\ell+1}^2 K_{\ell+1}^\beta).
\]

Defining the mapping \( \hat{f}_{\ell+1} (\beta_\ell, \alpha_{\ell+1}) := \beta_\ell + \hat{h}_2 (\alpha_{\ell+1}) \), the features at layer \( \ell + 1 \) is defined as the pushforward measure by \( \hat{f}_{\ell+1} \):

\[
p_{\ell+1}^\alpha = \hat{f}_{\ell+1} (p_{\ell}^\beta \times p_{\ell+1}^\alpha).
\]

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

\[
p (\nu, \alpha_1, \alpha_2, \ldots, \alpha_L) := p_1^\nu (\nu_1) \times p_2^\alpha (\alpha_2) \times p_3^\alpha (\alpha_3) \times \cdots \times p_L^\alpha (\alpha_L).
\]

**Weight functions.** Now we define the weight functions \( \{v_\ell\}_{\ell=2}^{L+1} \). Note that those gram matrices \( K^\beta_\ell \) are all positive definite under Assumption 2 (see Lemma 2) and thus are invertible. For \( \Theta = (\nu_1, \alpha_2, \ldots, \alpha_L), \Theta' = (\nu_1', \alpha_2', \ldots, \alpha_L'), \) we define the connecting weights between consecutive layers by

\[
v_{\ell+1} (\Theta, \Theta') = h_1 (\beta_{\ell-1} (\Theta))^\top \left[ K_{\ell-1}^\beta \right]^{-1} \alpha_{\ell-1}, \quad \ell \in [2 : L],
\]

and the definitions of \( K^\beta_\ell \). Therefore \( \{v_\ell\}_{\ell=2}^{L+1}, p \) constitutes a feasible continuous Res-Net.
Ideal discrete Res-Net. Finally we construct the initialization \((\nu, \alpha, \beta)\) of the ideal discrete Res-Net. Recall Algorithm 2 that the corresponding variables are initialized as \((\nu, \alpha, \beta)\). Let \(\nu_{i,1} : = \nu_{i,1}\) for \(i \in [m]\). For \(\ell \in [L - 1]\), define the empirical Gram matrix as
\[
K^\beta = \frac{1}{m} \sum_{i=1}^{m} h_1(\beta_{\ell,i}) h_1^\top(\beta_{\ell,i}).
\]
Let \(\alpha_{\ell+1,i,j} := (K^\beta)^{1/2} (K^\beta)^{-1/2} \alpha_{\ell+1,i,j}\) for all \(j \in [m]\) when \(K^\beta\) is invertible, and otherwise let \(\alpha_{\ell+1,i,j} \equiv p_{\ell+1}\). We use Definition 2 (2) for the values of \(\alpha\) in Assumption 3 are all satisfied.

We first consider the continuous Res-Net. By definition \(\epsilon\) for the values of \(\epsilon\), Under Assumption 2, Lemma 2. Let \(\alpha_{\ell+1,i,j} \sim p_{\ell+1}\) and they are conditionally independent Gaussian given \(\{\hat{\beta}_{\ell,i}\}_{i \in [m]}\). Furthermore, the conditional distribution of \(\alpha_{\ell+1,j}\) given \(\{\hat{\beta}_{\ell,i}\}_{i \in [m]}\) is \(N(0, \sigma^2 K^\beta) = p_{\ell+1}\). Therefore, marginally \(\alpha_{\ell+1,i,j} \sim p_{\ell+1}\) and they are independent of \(\nu_{i,1}\) and \(\alpha_{2,i}, \ldots, \alpha_{L,i}\) for \(i \in [m]\). So \(\{\hat{\Theta}_i\}_{i \in [m]} \sim p\) since \(p\) is a product distribution, all \(\nu_{i,1}, \alpha_{2,i}, \ldots, \alpha_{L,i}\) are all mutually independent.

Lastly we show the \(\tilde{O}(1/\sqrt{m})\)-closeness specified in Definition 2 (3). By Lemma 3 we have the following events with probability \(1 - \delta\):
\[
\|\alpha_{\ell+1,i}\|_2 \leq \tilde{O}(1),
\]
\[
\|K^\beta - K^\beta\|_2 \leq \epsilon_2,
\]
\[
\|\alpha_{\ell+1,i} - \alpha_{\ell+1,j}\|_2 \leq \epsilon_2 \|\hat{\Theta}\|_2,
\]
\[
\|\hat{\beta}_{\ell+1,i} - \hat{\beta}_{\ell+1,j}\|_2 \leq \epsilon_2 \|\hat{\Theta}\|_2,
\]
where \(\epsilon_2 \leq \tilde{O}(1/\sqrt{m})\). Under equation A.4, the matrix \(K^\beta\) is invertible, and it follows from Lemma 5 that
\[
\hat{\nu}_{\ell+1,i,j} = \hat{\alpha}_{\ell+1,j} \left[K^\beta\right]^{-1} h_1(\hat{\beta}_{\ell,i}), \quad \ell \in [L - 1], \ i, j \in [m].
\]

By the triangle inequality,
\[
\|\hat{\nu}_{\ell+1,i,j} - \nu_{\ell+1,i,j}\|_2 = \|\hat{\alpha}_{\ell+1,j} \left[K^\beta\right]^{-1} h_1(\hat{\beta}_{\ell,i}) - \alpha_{\ell+1,j} \left[K^\beta\right]^{-1} h_1(\beta_{\ell,i})\|_2
\]
\[
\leq \|\alpha_{\ell+1,j}\|_2 \left\|\left[K^\beta\right]^{-1} \right\|_2 \left\|h_1(\beta_{\ell,i}) - \hat{h}_1(\hat{\beta}_{\ell,i})\right\|_2 + \|\alpha_{\ell+1,j}\|_2 \left\|\left[K^\beta\right]^{-1} - \left[K^\beta\right]^{-1}\right\|_2 \left\|h_1(\beta_{\ell,i})\right\|_2
\]
\[
+ \|\alpha_{\ell+1,j} - \hat{\alpha}_{\ell+1,j}\|_2 \left\|\left[K^\beta\right]^{-1}\right\|_2 \left\|h_1(\hat{\beta}_{\ell,i})\right\|_2.
\]
We upper bound three terms separately. By the Lipschitz continuity of \(h_1\) and equation A.7, the first term is at most \(\tilde{O}(1/\sqrt{m})\|\Theta_j\|_2\); the second term is also at most \(\tilde{O}(1/\sqrt{m})\|\hat{\Theta}_j\|_2\) since \(\|\alpha_{\ell+1,j}\|_2 \leq \|\hat{\Theta}_j\|_2\).
\[
\left\|\left[K^\beta\right]^{-1} - \left[K^\beta\right]^{-1}\right\|_2 \leq \left\|\left[K^\beta\right]^{-1}\right\|_2 \left\|K^\beta - K^\beta\right\|_2 \left\|\left[K^\beta\right]^{-1}\right\|_2 \leq C \epsilon_2,
\]
and \(h_1\) is bounded; the third term is at most \(\tilde{O}(1/\sqrt{m})\|\hat{\Theta}_j\|_2\) by equation A.6.\]
With the above preparations, we are now ready to achieve Theorem 1.

Proof of Theorem 1. By Lemma 2, the discrete Res-Net and its corresponding continuous Res-Net satisfy our general initial conditions in Definition 2 and Assumptions 3 respectively. Then, Theorem 3 and Theorem 5 (2) indicate Theorem 1 (1) and (2), respectively.

A.2.4 PROOF OF ADDITIONAL LEMMAS

Lemma 3. \( \min_{\ell \in [L-1]} \lambda_{\min}(K_\ell^\beta) \geq C > 0. \)

Proof. Fix \( \ell \in [L-1] \). For \( (v_1, \alpha_2, \ldots, \alpha_L) \in \text{supp}(p) \), given \( \alpha_2, \ldots, \alpha_L \), we have \( \Theta = \Theta(v_1) \) and

\[
\left\| \beta_i(\Theta) - \frac{1}{d} X v_1 \right\|_{\infty} = \left\| \sum_{i=1}^L \hat{h}_2(\alpha_{\ell,i}) \right\|_{\infty} \leq LL_1. \quad (A.8)
\]

Note that \( v_1 \) is independent of \( \alpha_2, \ldots, \alpha_L \), and \( v_1 \sim \mathcal{N}(0, d \sigma^2 I) \) which is equivalent to the standard Gaussian distribution. By Assumption 2 with \( f_2(v_1) \equiv \sum_{i=1}^L h_2(\alpha_{\ell,i}) \) and the constant \( C_B = LL_1 \), we have

\[
E \left[ h_1(\beta_i(\Theta)) h_1^\top(\beta_i(\Theta)) \mid \alpha_2, \ldots, \alpha_L \right] \geq C I^N.
\]

Taking full expectation, we obtain Lemma 3.

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

Lemma 4. Let \( \varepsilon_2 \leq \tilde{O}(1/\sqrt{m}) \). With probability \( 1 - \delta \), for all \( \ell \in [L-1] \) and \( i \in [m] \),

\[
\left\| K_\ell^\beta - K_i^\beta \right\|_2 \leq \varepsilon_2, \quad \left\| \alpha_{\ell+1,i} - \hat{\alpha}_{\ell+1,i} \right\|_2 \leq \varepsilon_2 \left\| \hat{\Theta}_i \right\|_2, \quad \| \alpha_{\ell+1,i} \|_2 \leq \tilde{O}(1), \quad \left\| \beta_{\ell+1,i} - \hat{\beta}_{\ell+1,i} \right\|_2 \leq \varepsilon_2 \left\| \hat{\Theta}_i \right\|_2.
\]

Proof. In the proof of Lemma 2 we verified that \( v_{1,i}, \alpha_{2,i}, \ldots, \alpha_{L,i} \) for all \( i \in [m] \) are independent. Therefore, \( \hat{\beta}_{\ell,i} \sim p_{\ell,i}^\beta \) by the definitions of \( \hat{\beta}_{\ell,i} \) and \( p_{\ell,i}^\beta \). Consider auxiliary random matrices

\[
K_\ell^\beta := \frac{1}{m} \sum_{i=1}^m h_1(\beta_{\ell,i}) h_1^\top(\beta_{\ell,i}),
\]

Since \( h_1 \) is bounded, by the matrix Bernstein inequality [Tropp, 2015], with probability \( 1 - \frac{3}{4} \),

\[
\max_{\ell \in [L-1]} \left\| K_\ell^\beta - K_i^\beta \right\|_2 \leq \varepsilon_3 = \tilde{O}(1/\sqrt{m}). \quad (A.9)
\]

Due to the sub-gaussianess of \( p_i \), we have \( \| \alpha_{\ell+1,i} \|_2 \leq C \sqrt{\log(m/\delta)} = \tilde{O}(1) \) with probability \( 1 - \delta/3 \) [Vershynin, 2010]. We will also use the following upper bound that happen with probability \( 1 - \delta/3 \) by the sub-gaussianess of \( p_i \):

\[
\frac{1}{m} \sum_{i=1}^m \| \hat{\Theta}_i \|_2 \leq \beta_1 = \tilde{O}(1). \quad (A.10)
\]

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

\[
\left\| K_\ell^\beta - K_i^\beta \right\|_2 \leq (C \beta_1)^{\ell-1} C \varepsilon_3, \quad (A.11)
\]

\[
\left\| \alpha_{\ell+1,i} - \hat{\alpha}_{\ell+1,i} \right\|_2 \leq (C \beta_1)^{\ell-1} C \varepsilon_3 \left\| \hat{\Theta}_i \right\|_2, \quad (A.12)
\]

\[
\left\| \beta_{\ell+1,i} - \hat{\beta}_{\ell+1,i} \right\|_2 \leq (C \beta_1)^{\ell-1} C \varepsilon_3 \left\| \hat{\Theta}_i \right\|_2. \quad (A.13)
\]
For $\ell = 1$, by definition $K_1^\beta = K_1^\alpha$. The upper bound of $\left\| \left[ K_1^\alpha \right]^{1/2} - \left[ K_1^\beta \right]^{1/2} \right\|_2$ is achieved by matrix calculus [Bhatia 2013].

Section V.3. Since $\left\| K_1^\alpha - K_1^\beta \right\|_2 \leq \frac{\lambda_1}{2}$, then the eigenvalues of $K_1^\alpha$ are at least $\frac{\lambda_1}{2}$. Let $f(x) := \sqrt{x}$. Then $|f'(x)| \geq \frac{1}{\sqrt{2\lambda_1}}$ when $x$ is the eigenvalue of $K_1^\alpha$. Applying [Bhatia 2013] (V.20) yields that

$$\left\| \left[ K_1^\alpha \right]^{1/2} - \left[ K_1^\beta \right]^{1/2} \right\|_2 \leq \frac{N}{\sqrt{2\lambda_1}} \left\| K_1^\alpha - K_1^\beta \right\|_2 \leq C\varepsilon_3,$$

(A.14)

and

$$\left\| \hat{\alpha}_{2,i} - \hat{\alpha}_{2,i} \right\|_2 = \left\| \left( \left[ K_1^\alpha \right]^{1/2} \left[ K_1^\beta \right]^{1/2} - I \right) \hat{\alpha}_{2,i} \right\|_2 \leq C\varepsilon_3 \left\| \hat{\alpha}_{2,i} \right\|_2 \leq C\varepsilon_3 \left\| \tilde{\Theta}_i \right\|_2.$$  (A.15)

Then by the Lipschitz continuity of $h_2$, we have

$$\left\| \hat{\beta}_{2,i} - \hat{\beta}_{2,i} \right\|_2 = \left\| h_2(\hat{\alpha}_{2,i}) - h_2(\hat{\alpha}_{2,i}) \right\|_2 \leq C\varepsilon_3 \left\| \tilde{\Theta}_i \right\|_2.$$ (A.16)

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

$$\left\| \hat{\beta}_{\ell,i} - \hat{\beta}_{\ell,i} \right\|_2 \leq (C\beta_1)^{\ell-2} C\varepsilon_3 \left\| \tilde{\Theta}_i \right\|_2.$$ (A.17)

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

$$\left\| \hat{K}_\ell^\beta - \hat{K}_\ell^\beta \right\|_2 \leq \frac{C}{m} \sum_{i=1}^m \left\| h_1(\hat{\beta}_{\ell,i}) - h_1(\hat{\beta}_{\ell,i}) \right\|_2.$$ (A.18)

Applying the Lipschitz continuity of $h_1$ and equation A.17 yields that

$$\left\| \hat{K}_\ell^\beta - \hat{K}_\ell^\beta \right\|_2 \leq \frac{(C\beta_1)^{\ell-2} C\varepsilon_3}{m} \sum_{i=1}^m \left\| \tilde{\Theta}_i \right\|_2 \leq (C\beta_1)^{\ell-1} C\varepsilon_3.$$ (A.19)

where in the last inequality we used equation A.10. Then we obtain equation A.11 by triangle inequality from equation A.9 and equation A.18. The upper bound of $\left\| \hat{\beta}_{\ell+1,i} - \hat{\beta}_{\ell+1,i} \right\|_2$ for $\ell + 1$ follows from a similar argument of equation A.14 and equation A.15. Finally equation A.13 for $\ell + 1$ follows from equation A.12 and

$$\left\| \hat{\beta}_{\ell+1,i} - \hat{\beta}_{\ell+1,i} \right\|_2 = \left\| \sum_{j=2}^{\ell+1} \left[ h_2(\hat{\alpha}_{j,i}) - h_2(\hat{\alpha}_{j,i}) \right] \right\|_2 \leq C(C\beta_1)^{\ell-1} \varepsilon_3 \left\| \tilde{\Theta}_i \right\|_2.$$ (A.20)

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

**Lemma 5.** If $K_\ell^\beta$ is invertible, then

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

**Proof.** For a given layer $\ell$ and $j$, the $\ell_2$-regression problem in Algorithm 2 can be equivalently written as

$$\min_{\hat{v}} \quad \frac{1}{2} \| \hat{v} \|^2$$

s.t. \hspace{1cm} \frac{1}{m} \tilde{H} \hat{v} = \hat{\alpha}_{\ell+1,j},

(A.19)

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

$$\hat{v} = \tilde{H}^T \hat{z} + \hat{v'},$$

where $\hat{z} \in \mathbb{R}^m$ and $\tilde{H} \hat{v'} = 0$. Then equation A.19 is equivalent to

$$\min_{\hat{z}, \hat{v'}} \quad \frac{1}{2} \| \tilde{H}^T \hat{z} \|^2 + \frac{1}{2} \| \hat{v'} \|^2$$

s.t. \hspace{1cm} \frac{1}{m} \tilde{H} \tilde{H}^T \hat{z} = \hat{\alpha}_{\ell+1,j}.$$

Since $\frac{1}{m} \tilde{H} \tilde{H}^T = K_\ell^\beta$ is invertible, the optimal solution is $\hat{z} = \left[ K_\ell^\beta \right]^{-1} \hat{\alpha}_{\ell+1,j}$ and $\hat{v'} = 0^N$. \qed
A.3 Proof of Theorem \[2\]

Proof of Theorem \[2\] In the proof we use the following abbreviated notations: for \( t \in [0, \infty) \) and \( \Theta \in \text{supp}(p) \), let

\[
\begin{align*}
\Phi^\ell_1(\Theta) & = \Phi_{i,\ell}(\Theta)(t), \quad \ell \in [L] , \\
\Phi^\ell_{\Theta}(\Theta) & = \Phi_{\Theta,\ell}(\Theta)(t), \quad \ell \in [2: L] , \\
\Phi_{i,1}(\Theta) & = \Phi_{i,1}(\Theta)(t), \\
\Phi_{\Theta,1}(\Theta) & = \Phi_{\Theta,1}(\Theta)(t), \\
\Phi_{\Theta,L+1}(\Theta) & = \Phi_{\Theta,L+1}(\Theta)(t).
\end{align*}
\]

Then, since \( h_1 \) is bounded and Lipschitz continuous, we have

\[
\|\beta^\ell_{L+1} - \beta^\ell_{\Theta,L+1}\|_{\infty} \leq \bar{C}\varepsilon_2, \quad \ell \in [L].
\]

\[
\begin{align*}
\int |v^\ell_{L+1}(\Theta) - v^\ell_{\Theta,L+1}(\Theta)| p(\Theta) \leq \varepsilon_2, \quad \ell \in [L].
\end{align*}
\]

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

\[
\|\beta^\ell_{L}(\Theta) - \beta^\ell_{\Theta}(\Theta)\|_{\infty} \leq \varepsilon_2
\]

holds \( p \)-almost surely and

\[
\int |v^\ell_{L+1}(\Theta) - v^\ell_{\Theta,L+1}(\Theta)| p(\Theta) \leq \varepsilon_2, \quad \ell \in [L].
\]

Then, since \( h_1 \) is bounded and Lipschitz continuous, we have

\[
\|\beta^\ell_{L+1} - \beta^\ell_{\Theta,L+1}\|_{\infty} \leq \bar{C}\varepsilon_2, \quad \ell \in [L].
\]

The goal of the proof is to show that

\[
\|
\phi^\ell_{i}(\beta^\ell_{L+1})
\|_2 = 0.
\]

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

\[
f_{\varepsilon}(v_1) := \phi^\ell_{i}(\beta^\ell_{L+1})^T h_1 \left( \frac{1}{d} X v_1 + g_\varepsilon(v_1) \right),
\]

where the functions \( g_\varepsilon \) is uniformly bounded, such that \( |f_{\varepsilon}(v_1)| < \varepsilon \). Then it follows from equation \[A.24\] that

\[
\phi^\ell_{i}(\beta^\ell_{L+1}) = K^{-1} \int f_{\varepsilon}(v_1) h_1 \left( \frac{1}{d} X v_1 + g_\varepsilon(v_1) \right) d\bar{\nu}_1(v_1),
\]

where \( \bar{\nu}_1 = N(0^d, I^d) \) and \( K := \int h_1 \left( \frac{1}{d} X v_1 + g_\varepsilon(v_1) \right) h_1^T \left( \frac{1}{d} X v_1 + g_\varepsilon(v_1) \right) d\bar{\nu}_1(v_1) \) whose minimum eigenvalue is at least \( \lambda_1 > 0 \) by Assumption \[2\]. The boundedness of \( h_1 \) yields that

\[
\|\phi^\ell_{i}(\beta^\ell_{L+1})\|_2 \leq C\lambda_1^{-1}\varepsilon.
\]

Since \( \lambda_1 \) is independent of \( \varepsilon \), by letting \( \varepsilon \to 0 \), we obtain equation \[A.23\].

Next we construct \( g_\varepsilon \) and \( f \) in equation \[A.24\]. Let \( T \) be the time such that equation \[A.20\] and equation \[A.21\] hold with \( \varepsilon_2 < \varepsilon_2 \) for a constant \( c \) to be specified. Note that \( v_1^T \) is surjective by Lemma \[6\]. Let \( \bar{g} : \mathbb{R}^d \to \text{supp}(p) \) be the inverse function such that \( \nu^T(\bar{g}(v_1)) = v_1 \).

Define

\[
g_\varepsilon(v_1) = \sum_{\ell=2}^L h_2 \left( \alpha^\ell_1 \left( \bar{g}(v_1) \right) \right), \quad f_{\varepsilon}(v_1) = \phi^\ell_i(\beta^\ell_{L+1})^T h_1 \left( \beta^\ell_1 \left( \bar{g}(v_1) \right) \right),
\]

where \( g_\varepsilon \) is uniformly bounded by the boundedness of \( h_2 \). Suppose on the contrary that there exists \( \nu^T \) such that \( |f_{\varepsilon}(v_1^T)| > \varepsilon \). Let \( \Theta^T = \bar{g}(v_1) \). Since \( \Theta \mapsto \phi^\ell_i(\beta^\ell_{L+1})^T h_1 \left( \beta^\ell_1(\Theta) \right) \) is continuous by Theorem \[2\], there exists a ball around \( \Theta^T \), denoted by \( S \) such that \( p(S) > 0 \) and \( \phi^\ell_i(\beta^\ell_{L+1})^T h_1 \left( \beta^\ell_1(\Theta) \right) > \varepsilon/2 \) with the same sign for all \( \Theta \in S \). However, for \( t > T \),

\[
\begin{align*}
\int |v^\ell_{L+1}(\Theta) - v^\ell_{\Theta,L+1}(\Theta)| p(\Theta) & \geq \frac{1}{N} \int S \left( \left| \int_T^t \phi^\ell_i(\beta^\ell_{L+1})^T h_1 \left( \beta^\ell_1(\Theta) \right) dt \right| d\nu(\Theta) \right)
\geq \frac{1}{N} \int S \left( \left| \int_T^t \phi^\ell_i(\beta^\ell_{L+1})^T h_1 \left( \beta^\ell_1(\Theta) \right) dt - \int_T^t C\varepsilon_2 dt \right| d\nu(\Theta) \right).
\end{align*}
\]
where in the last step we used equation A.20, equation A.22, and the boundedness and Lipschitz continuity of $\phi_i$ and $h_1$. Let $c = \frac{1}{12}$. The lower bound in equation A.25 diverges with $t$, which contradicts equation A.21.

Finally from equation A.23 we show the convergence statement. Since $\phi$ is convex on the first argument, we obtain

$$\frac{c}{\alpha} \sum_{n=1}^{N} \phi(\beta_{n+1}^\infty(n), y^n) = \sum_{n=1}^{N} \min_{y'} \phi(y', y^n).$$

Since $\beta_{n+1}^1 \to \beta_{n+1}^\infty$ and $\phi$ is continuous, we obtain that

$$\lim_{t \to \infty} L^{t} = \sum_{n=1}^{N} \phi(\beta_{n+1}^\infty(n), y^n) = \sum_{n=1}^{N} \min_{y'} \phi(y', y^n),$$

which completes the proof.

**Lemma 6.** The function $t < \infty, v_1^t : \text{supp}(p) \to \mathbb{R}^d$ is a surjection.

**Proof.** Recall that at the initialization let $\Theta(v) = (v, 0^{D-d}) \in \mathbb{R}^D = \text{supp}(p)$. Given $t < \infty$, consider $f_t : \mathbb{R}^d \to \mathbb{R}^d$ as

$$f_t(v) = v_1^t(\Theta(v)).$$

It suffices to show that $f_t$ is surjective. Note that $f_t$ is continuous since $\Theta \mapsto v_1^t(\Theta)$ is continuous by Theorem 4. Furthermore, for any $v \in \mathbb{R}^d$, by Lemma 16 in the supplementary material, which states that the gradient of the weights $G_1 v$ is bounded, we have

$$\|f_t(v) - v\|_{\infty} = \int_0^1 \|G_1^v(\Theta(v); \Phi_*, s)\|_{\infty} ds \leq Ct.$$

For any $x \in \mathbb{R}^d$, consider $g(v) := x - f_t(v) - v$ which continuously maps $B_\infty(x, Ct)$ to itself. By the Brouwer’s fixed-point theorem (see, e.g. Granas & Dugundji (2013)), there exists $v_* \in B_\infty(x, Ct)$ such that $g(v_*) = v_*$; equivalently, we have $f_t(v_*) = x$. □

### A.4 PROOF OF LEMMA 1 AND PROPOSITION 1

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

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

where $a = (a_1, \ldots, a_N)$. Lemma 1 shows that, for $V' \sim p = \mathcal{N}(0^d, 1^d)$, the same result holds with a constant perturbation of the functions $h_1$; namely, by letting $g_i(V) = h_1(x_i \cdot v + C_i(V))$ where $\|C_i\|_{\infty} \leq C_B$,

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

where $\lambda'$ is uniform over all perturbations $\|C_i\|_{\infty} \leq C_B$. It suffices to prove equation A.26 for $V' \sim q = \text{Uniform}(R')$ where $R'$ is determined by $h$ and $C_B$, as the Radon–Nikodym derivative $\frac{dq}{dp}$ is bounded.

We first prove (1). Consider a compact region $R$ such that, for $V \sim \text{Uniform}(R)$ and any unit vector $a$,

$$\mathbb{E} \left\| \sum_{i=1}^{N} a_i h_3(x_i, V) \right\|_2^2 \geq \lambda_R > 0.$$

Then for any $\alpha > 0$, since $h_3$ is bounded and Lipschitz continuous, we have

$$\mathbb{E} \left\| \sum_{i=1}^{N} a_i h_3(x_i, V + \alpha C_1(V/a)) \right\|_2^2 \geq \lambda_R - C \alpha C_B \geq \frac{\lambda_R}{2},$$

when $\alpha \leq \frac{\lambda_R}{2 C_B C}$. Equivalently, $\mathbb{E}\|\sum_{i=1}^{N} a_i h_3(x_i, V/a + C_1(V/a))\|_2^2 \geq \frac{\lambda_R}{2}$. We achieve equation A.26 by letting $V' = V/a$. 18
For \( \text{(ii)} \), consider \( R = \{ v : 1/2 \leq \|v\|_2 \leq 1 \} \). Then, for \( V \sim \text{Uniform}(R) \) and any unit vector \( a \),

\[
E \left\| \sum_{i=1}^{N} a_i h_1(x_i \cdot V) \right\|_2^2 \geq \lambda_R > 0.
\]

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

\[
E \left\| \sum_{i=1}^{N} a_i h_1(x_i \cdot \beta V + C(\beta V)) \right\|_2^2 \geq \beta^{2\alpha} \lambda_R - C' \beta^{-2\alpha-1} C_B \geq \left( C' C_B \right)^{-2\alpha \frac{2}{\lambda_R}} .
\]

for a constant \( C' \) when \( \beta = \frac{2C'C_B}{\lambda_R} \). We achieve equation (A.26) by letting \( V' = \beta V \).

For \( \text{(iii)} \), we first show that there exists a compact set \( R \) such that, for all \( v \in R \) and \( x_i \),

\[
|x_i \cdot v| \geq c'.
\]

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

\[
P(\|v^T x\| < t \|x\|_2) = \frac{2\pi^{d-1} / \Gamma\left(\frac{d-1}{2}\right)}{2\pi^{\frac{d}{2}} / \Gamma\left(\frac{d}{2}\right)} \int_{-t}^{t} (1 - u^2)^{\frac{d-3}{2}} du < t^d.
\]

By a union bound, we have \( |x_i \cdot v| \geq \frac{\|x_i\|_2}{2N^\frac{d}{2}} \) with probability 0.5. Denote the set of \( v \in S^{d-1} \) by \( S' \). Since \( \min_i \|x_i\|_2 := C_x > 0 \), we obtain equation (A.27) with \( R = \{ t v : v \in S', \frac{2\sqrt{C_x N}}{C_x} \leq t \leq \frac{4C'C_B}{\sqrt{C_x}} \} \). Then, for \( V \sim \text{Uniform}(R) \) and any unit vector \( a \),

\[
E \left\| \sum_{i=1}^{N} a_i h_1(x_i \cdot V) \right\|_2^2 \geq \lambda_R > 0.
\]

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

\[
E \left\| \sum_{i=1}^{N} a_i h_1(x_i \cdot \beta V + C(\beta V)) \right\|_2^2 \geq \beta^{-2\alpha} \lambda_R - C' \beta^{-2\alpha-1} C_B \geq \left( C' C_B \right)^{-2\alpha \frac{2}{\lambda_R}} .
\]

for a constant \( C' \) when \( \beta = \frac{2C'C_B}{\lambda_R} \). We achieve equation (A.26) by letting \( V' = \beta V \). \( \square \)

**Proof of Proposition 7** Explicitly shown in [Nguyen & 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 7. \( \square \)