Mathematical Modeling and Convergence Analysis of Deep Neural Networks with Dense Layer Connectivities in Deep Learning

Jinshu Huang[1], Haibin Su[1], Xue-Cheng Tai[2], Chunlin Wu[1*]

[1]School of Mathematical Sciences, Nankai University, Tianjin, China.
[2]Norwegian Research Center (NORCE), Nygårdsgaten 112, 5008, Bergen, Norway.

*Corresponding author(s). E-mail(s): wucl@nankai.edu.cn;
Contributing authors: huangjsh@mail.nankai.edu.cn;
hbsu@mail.nankai.edu.cn; xtai@norceresearch.no;

Abstract

In deep learning, dense layer connectivity has become a key design principle in deep neural networks (DNNs), enabling efficient information flow and strong performance across a range of applications. In this work, we model densely connected DNNs mathematically and analyze their learning problems in the deep-layer limit. For a broad applicability, we present our analysis in a framework setting of DNNs with densely connected layers and general non-local feature transformations (with local feature transformations as special cases) within layers, which is called dense non-local (DNL) framework and includes standard DenseNets and variants as special examples. In this formulation, the densely connected networks are modeled as nonlinear integral equations, in contrast to the ordinary differential equation viewpoint commonly adopted in prior works. We study the associated training problems from an optimal control perspective and prove convergence results from the network learning problem to its continuous-time counterpart. In particular, we show the convergence of optimal values and the subsequence convergence of minimizers, using a piecewise linear extension and $\mathbf{\Gamma}$-convergence analysis. Our results provide a mathematical foundation for understanding densely connected DNNs and further suggest that such architectures can offer stability of training deep models.

1

1 Introduction

Deep learning technology has achieved significant breakthroughs in various fields [1]. A fundamental and key factor behind these achievements is the design of deep neural network (DNN) architectures. It determines how neurons are connected and how information flows through the neural network, which plays a crucial role in the expressive ability of DNNs and the efficiency of training algorithms.

The architecture of DNNs includes the layer connectivity methods and the computation schemes within layers. Regarding layer connections, the early proposed fully connected feedforward neural networks (FNNs) and convolutional neural networks (CNNs) [1] utilize a straightforward sequential connection structure. Such structures are compact and computationally simple, but they may become harder to train with an increasing number of layers and less effective at capturing complex dependencies. Later on, the residual networks (ResNets) [2] introduced skip connections and residual learning, which mitigate the vanishing gradient problem and enable the training of much deeper networks. After that, the dense convolutional network (DenseNet) [3] further enhanced information flow by connecting each layer to every other layer in a feedforward manner. This dense connectivity structure promotes feature reuse and thus can reduce the number of parameters, leading to more efficient models. These layer connectivity methods have led to the development of many effective deep neural networks. For instance, UNet [4] extends the typical CNN by incorporating an encoder-decoder structure and skip connections; [5–7] combined the dense layer connectivity methods with some attention mechanisms to get more effective DNN architectures. For more research in this direction, see, e.g., ResUNet [8], DenseUNet [9], PottsMGNet [10].

The computation scheme within each layer also plays an important role in the design of neural network architectures. There is a broad range of local and non-local feature transformations that can benefit the performances of DNNs. Initially, FNNs use affine transformations combined with nonlinear activations. Their fully connected layers are straightforward and capture global information within layers. After that, CNN layers introduce sparse convolution operations, applying filters to local regions of the state, and can effectively capture spatial hierarchies through multiple layers. Beyond these classical designs, more general feature transformation schemes have been explored, such as nonlinear operations beyond activation functions [11–14]. For example, the STD (Soft-Threshold-Dynamics) activation layer introduced in [15], which offered a mechanism to incorporate many well-known variational models into activation functions. The Transformers [11] employ self-attention mechanisms to model global dependencies. In each self-attention function, the output is calculated as a weighted sum of the values obtained by an affine transformation of inputs, where the weight assigned to each value is computed by the softmax$(\cdot)$ function [16]. Additionally, non-local neural networks (Non-local Nets) [12] expand the self-attention mechanism by computing interactions between all possible pairs of positions through a flexible non-local kernel, which can further capture global dependencies such as similarity.

Despite the great successes of these DNN architectures in applications, they were mainly handcrafted and lacked rigorous mathematical understanding. We note that

the connectivity between layers and the computation scheme within each layer corresponds exactly to a time variable and a space variable, respectively. Naturally, we can consider their dynamical system modeling. Such dynamical system modeling and analysis aim to study an associated continuous-time formulation in some detail, providing new perspectives and tools for theoretical research on DNN structures and network learning problems. Indeed, the works [17, 18] first established a connection between ResNets and ordinary differential equations (ODEs) by interpreting the forward propagation of the network as a time-discretization of an ODE, with each layer corresponding to a discrete time step. Lu et al. [19] further modeled some effective networks, such as RevNet [20], as different numerical discretizations of ODEs. Recently, Zhang and Schaeffer [21] represented the ResNets as an ordinary differential inclusion (ODI) to analyze its global stability property. Thorpe and Van Gennip [22] demonstrated some convergence results from the discrete-time to continuous-time learning problems for ResNets. For more continuous-time modeling and discussion of classic DNNs, one can see, e.g., [23–26]. We also mention that, for some deep unrolling/unfolding networks (see, e.g., [27] for a survey), it is natural to connect them as continuous-time systems and study the convergence of the associated learning problems in the deep-layer limit setting [28–30]. These dynamical system modeling and analysis methods not only provide a mathematical understanding but also aid in designing new network architectures through various numerical ordinary/partial differential equations or optimization schemes [10, 19, 31, 32].

DNNs with dense layer connectivities, such as DenseNet [3] and its variants [5–7, 9], have demonstrated excellent empirical performance across a wide range of deep learning applications. Despite their successes, a precise mathematical understanding of how such dense connections influence network behavior, particularly in the limit of infinite depth, remains unclear. In this work, we model and analyze densely connected DNNs through the dynamical system approach. Our analysis is presented within a general framework, namely DNL framework. Note that our goal is not to propose a new architecture, but rather to provide a unified theoretical perspective that applies across this class of DNNs. *Our central observation is that, with each network layer corresponding to a time step, dense layer connectivity yields a discrete accumulation across layers and naturally gives rise to a nonlinear integral equation in the deep-layer limit.* Our main contributions are as follows:

(i) We provide a dynamical system modeling for a class of densely connected DNNs. Specifically, within a general dense non-local framework, we formulate the DNNs in the deep-layer limit using nonlinear integral equations, which stands in contrast to prior works that typically model deep networks as ODEs or ODIs. We further cast the corresponding learning problems in both discrete- and continuous-time settings as optimal control problems with suitable regularization.

(ii) We establish convergence results from the learning problem of the discrete-time DNL framework to its continuous-time counterpart via Γ-convergence. In particular, we show convergence of the optimal values and subsequence convergence of the optimal solutions using a piecewise linear extension technique.

These dynamical system modeling and convergence analysis offer a mathematical understanding and evidence of DNN architectures with dense layer connectivities from the perspective of integral equations. Our analyses suggest that densely connected architectures may offer enhanced stability when training deep networks. Moreover, our results can be directly applied to, or analogously extended to, various DNNs with dense layer connections, such as DenseNet, DenseFormer [7], and Dense Residual Transformer [6].

The remainder of this paper is organized as follows. Section 2 introduces the dense non-local framework of DNNs. In Section 3, we present dynamical system modeling for the framework and our main theoretical result. Section 4 provides detailed proofs. In Section 5, we present some simple numerical experiments to validate the theoretical result. Finally, Section 6 concludes the paper.

2 A dense non-local (DNL) framework of deep neural networks

We denote by $\mathbb{N}$ the set of natural numbers and by $\mathbb{R}$ the field of real numbers. Scalars are denoted with italic letters, like $L, n \in \mathbb{N}$. Vectors are represented by lowercase letters and matrices by uppercase letters, all in upright font, for example, $\mathrm{a} \in \mathbb{R}^L$ and $\mathrm{U} \in \mathbb{R}^{n \times n}$. We also denote the l^2-norm for vectors in Euclidean space by $|\cdot|$ and the spectral norm for matrices by $\|\cdot\|_2$. Without confusion, we abbreviate $\|\cdot\|_2$ as $\|\cdot\|$. For a space $\mathbb{X}$, the notation $(\mathbb{X})^L$ represents the Cartesian product space of $\underbrace{\mathbb{X} \times \ldots \times \mathbb{X}}_{L}$. To facilitate the description of the network architecture, we define a general affine map.

Definition 1 Given a sequence of matrix-vector pairs $\{(\mathrm{M}^k, \mathrm{v}^k)\}_{k=0}^l \subset \mathbb{R}^{n \times n} \times \mathbb{R}^n$, we define an affine map $\mathbf{Aff}_{\{(\mathrm{M}^k, \mathrm{v}^k)\}_{k=0}^l} : \underbrace{\mathbb{R}^n \times \ldots \times \mathbb{R}^n}_{l+1} \to \mathbb{R}^n$ as

$$\mathbf{Aff}_{\{(\mathrm{M}^k, \mathrm{v}^k)\}_{k=0}^l}(\mathrm{x}^0, \ldots, \mathrm{x}^l) := \sum_{k=0}^{l} \left(\mathrm{M}^k \mathrm{x}^k + \mathrm{v}^k \right).$$

Such an affine map covers various types of linear operations commonly used in practice, including standard convolutions or matrix-vector multiplications in FNNs.

We now introduce the DNL framework and its associated learning problem. Since our formulation is not restricted to specific data modalities, we assume without loss of generality that both the input and output lie in $\mathbb{R}^n$. Given a layer number $L \in \mathbb{N}$, the DNL framework is defined as a class of densely connected neural networks with L layers. Each layer aggregates all preceding outputs through dense skip connections and applies a transformation that may be either local or non-local. Formally, the

4

architecture is given by

$$
\begin{aligned}
\mathrm{x}^0 &= \mathbf{Aff}_{\{(\mathrm{V}_L^0, \mathrm{b}_L^0)\}}\Big(\boldsymbol{\phi} \circ \mathbf{Aff}_{\{(\mathrm{U}_L^0, \mathrm{a}_L^0)\}}\big(\boldsymbol{\mathcal{A}_\kappa}(\mathrm{T}_L^0; \mathrm{d})\big)\Big) \\
&= \mathrm{V}_L^0 \boldsymbol{\phi} \circ (\mathrm{U}_L^0 \boldsymbol{\mathcal{A}_\kappa}(\mathrm{T}_L^0; \mathrm{d}) + \mathrm{a}_L^0) + \mathrm{b}_L^0, \\
\mathrm{x}^l &= \mathbf{Aff}_{\{(\mathrm{V}_L^l, \mathrm{b}_L^l)\}}\Big(\boldsymbol{\phi} \circ \mathbf{Aff}_{\{(\mathrm{U}_L^l, \mathrm{a}_L^l)\} \cup \{(\tau \mathrm{W}_L^{l,k+1}, \tau \mathrm{c}_L^{l,k+1})\}_{k=0}^{l-1}} \\
&\qquad\qquad \big(\boldsymbol{\mathcal{A}_\kappa}(\mathrm{T}_L^l; \mathrm{d}), \boldsymbol{\mathcal{A}_\kappa}(\mathrm{T}_L^l; \mathrm{x}^0), \dots, \boldsymbol{\mathcal{A}_\kappa}(\mathrm{T}_L^l; \mathrm{x}^{l-1})\big)\Big) \\
&= \mathrm{V}_L^l \boldsymbol{\phi} \circ \Big(\mathrm{U}_L^l \boldsymbol{\mathcal{A}_\kappa}(\mathrm{T}_L^l; \mathrm{d}) + \mathrm{a}_L^l + \tau \sum_{k=0}^{l-1}\big[\mathrm{W}_L^{l,k+1}\boldsymbol{\mathcal{A}_\kappa}(\mathrm{T}_L^l; \mathrm{x}^k) + \mathrm{c}_L^{l,k+1}\big]\Big) + \mathrm{b}_L^l,
\end{aligned}
\tag{1}
$$

for $1 \le l \le L$. Therein, the input $\mathrm{d} \in \mathbb{R}^n$; the affine parameters include $(\mathrm{U}_L^l, \mathrm{a}_L^l) \in$
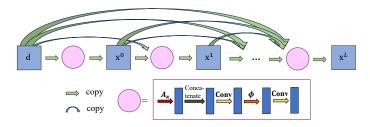

Fig. 1 Network architecture associated with (1).

$\mathbb{R}^{n \times n} \times \mathbb{R}^n$, $(\mathrm{V}_L^l, \mathrm{b}_L^l) \in \mathbb{R}^{n \times n} \times \mathbb{R}^n$ $(0 \le l \le L)$ and $(\mathrm{W}_L^l, \mathrm{c}_L^l) \in (\mathbb{R}^{n \times n})^l \times (\mathbb{R}^n)^l$ $(1 \le l \le L)$, where $\mathrm{W}_L^l = (\mathrm{W}_L^{l,1}, \dots, \mathrm{W}_L^{l,l}) \in (\mathbb{R}^{n \times n})^l$ and $\mathrm{c}_L^l = (\mathrm{c}_L^{l,1}, \dots, \mathrm{c}_L^{l,l}) \in (\mathbb{R}^n)^l$ due to the dense connectivity. Since $(\mathrm{W}_L^l, \mathrm{c}_L^l)$ is learnable, we extract a step size $\tau > 0$ and still denote it as $(\mathrm{W}_L^l, \mathrm{c}_L^l)$ without loss of generality. The activation function $\boldsymbol{\phi}$ is applied component-wise, and $\boldsymbol{\mathcal{A}_\kappa} : (\mathbb{R}^{n \times n})^3 \times \mathbb{R}^n \to \mathbb{R}^n$ is a parametrized general non-local transformation with kernel $\boldsymbol{\kappa}$ [12, 33]. For any vector $\mathrm{z} \in \mathbb{R}^n$,

$$
\boldsymbol{\mathcal{A}_\kappa}(\mathrm{T}_L^l; \mathrm{z})[i] = \frac{1}{\mathrm{Normalize}(\mathrm{T}_L^l; \mathrm{z})[i]} \sum_{j=1}^n \boldsymbol{\kappa}(((\mathrm{T}_1)_L^l \mathrm{z})[i], ((\mathrm{T}_2)_L^l \mathrm{z})[j]) \cdot ((\mathrm{T}_3)_L^l \mathrm{z})[j],
$$

for $1 \le i \le n$, where $\mathrm{T}_L^l = ((\mathrm{T}_1)_L^l, (\mathrm{T}_2)_L^l, (\mathrm{T}_3)_L^l) \in (\mathbb{R}^{n \times n})^3$ is the learnable parameter, the non-negative kernel $\boldsymbol{\kappa}(\cdot, \cdot)$ is a scalar value that quantifies the degree of relation or relevance between the features at positions i and j. The normalization factor $\mathrm{Normalize}(\cdot; \cdot) \ne 0$, which is commonly taken as $\mathrm{Normalize}(\mathrm{T}_L^l; \mathrm{z})[i] = \sum_{j=1}^n \boldsymbol{\kappa}(((\mathrm{T}_1)_L^l \mathrm{z})[i], ((\mathrm{T}_2)_L^l \mathrm{z})[j])$. Figure 1 illustrates the DNN architecture in (1).

Note that the DNL framework is quite general due to the flexibility of kernel $\boldsymbol{\kappa}$ and parameters $\mathrm{W}_L^l, \mathrm{c}_L^l$ $(1 \le l \le L)$. It characterizes a broad class of densely connected neural networks, encompassing several important examples and closely related variants as follows

- If $\boldsymbol{\kappa}(z[i], z[j]) = \boldsymbol{\delta}_{ij}$ with $\boldsymbol{\delta}_{i,j}$ being the Kronecker function, then $\boldsymbol{\mathcal{A}_\kappa}$ is a local operation, and the DNL framework reduces to a DenseNet [3].
- If $\boldsymbol{\kappa}(((T_1)^l_L z)[i], ((T_2)^l_L z)[j]) = \exp\left(((T_1)^l_L z)[i] \cdot ((T_2)^l_L z)[j]/\sqrt{n}\right)$ and $\text{Normalize}(T^l_L; z)[i] = \sum_{j=1}^n \boldsymbol{\kappa}(((T_1)^l_L z)[i], ((T_2)^l_L z)[j])$, the function $\boldsymbol{\mathcal{A}_\kappa}$ admits the self-attention operation in Transformer, and the DNL framework is a kind of Dense-Attention structure, which is closely similar to [5–7].
- If $(W^{l,k}_L, c^{l,k}_L) \equiv \mathbf{0}$, $1 \le k \le \lceil l/2 \rceil - 1, 2 \le l \le L$, then the DNL framework coincides with a partially densely connected neural network.

To facilitate the subsequent analysis, we package the learnable parameters together and introduce the notation Θ_L and the discrete learnable parameter set $\Omega_{\Theta;L}$ as follows

$$\Theta_L := (T_L, U_L, a_L, V_L, b_L, W_L, c_L) \in \Omega_{\Theta;L}, \tag{2}$$
$$\Omega_{\Theta;L} := ((\mathbb{R}^{n\times n})^3)^{L+1} \times (\mathbb{R}^{n\times n})^{L+1} \times (\mathbb{R}^n)^{L+1} \times (\mathbb{R}^{n\times n})^{L+1} \times (\mathbb{R}^n)^{L+1}$$
$$\times (\mathbb{R}^{n\times n})^{(1/2)L(L+1)} \times (\mathbb{R}^n)^{(1/2)L(L+1)},$$

where $T_L = ((T_1)_L, (T_2)_L, (T_3)_L), (T_i)_L = ((T_i)^0_L, \dots, (T_i)^L_L)$ $(i = 1, 2, 3)$, $U_L = (U^0_L, \dots, U^L_L)$, $a_L = (a^0_L, \dots, a^L_L)$, $V_L = (V^0_L, \dots, V^L_L)$, $b_L = (b^0_L, \dots, b^L_L)$, $W_L = (W^1_L, \dots, W^L_L)$ and $c_L = (c^1_L, \dots, c^L_L)$. It is natural to define a norm for $\Omega_{\Theta;L}$ as

$$\|\Theta_L\|_{\Omega_{\Theta;L}} := \max\{\|T_L\|_{T;L}, \|U_L\|_{U;L}, \|a_L\|_{a;L}, \|V_L\|_{V;L}, \|b_L\|_{b;L}, \|W_L\|_{W;L}, \|c_L\|_{c;L}\}, \tag{3}$$

where

$$\|T_L\|_{T;L} = \max_{i\in\{1,2,3\}} \max_{l\in\{0,1,\dots,L\}} \|(T_i)^l_L\|, \|U_L\|_{U;L} = \max_{l\in\{0,1,\dots,L\}} \|U^l_L\|,$$
$$\|a_L\|_{a;L} = \max_{l\in\{0,1,\dots,L\}} |a^l_L|, \|V_L\|_{V;L} = \max_{l\in\{0,1,\dots,L\}} \|V^l_L\|, \|b_L\|_{b;L} = \max_{l\in\{0,1,\dots,L\}} |b^l_L|,$$
$$\|W_L\|_{W;L} = \max_{l\in\{1,\dots,L\}} \max_{k\in\{1,\dots,l\}} \|W^{l,k}_L\|, \|c_L\|_{c;L} = \max_{l\in\{1,\dots,L\}} \max_{k\in\{1,\dots,l\}} |c^{l,k}_L|.$$

Without ambiguity, we will sometimes use $\|\Theta_L\|$ to denote $\|\Theta_L\|_{\Omega_{\Theta;L}}$ for simplicity.

For the DNL framework, we now introduce its learning problem. Denote $x^l(d; \Theta_L)$ as the output of l-th ($1 \le l \le L$) layer of (1) with input data $d \in \mathbb{R}^n$ and learnable parameter Θ_L. Let $\boldsymbol{\ell} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^+$ be the data loss function. Deep learning aims to minimize a loss function $\boldsymbol{\mathfrak{L}}_{\mathcal{S};L}(\Theta_L)$ on a training dataset $\mathcal{S} := \{(d_m, g_m)\}_{m=1}^M$, where d_m is the observed signal and g_m is the corresponding ground truth. Consequently, we have the following optimal control problem

$$(\mathcal{P}_L) : \begin{cases} \inf_{\Theta_L \in \Omega_{\Theta;L}} \left\{ \boldsymbol{\mathfrak{L}}_{\mathcal{S};L}(\Theta_L) = \frac{1}{M} \sum_{m=1}^M \boldsymbol{\ell}(x^L(d_m; \Theta_L), g_m) + \boldsymbol{\mathcal{R}}_L(\Theta_L) \right\} \\ \text{subject to:} \\ \{x^l(d_m; \Theta_L)\}_{l=0}^L \text{ satisfy Eq.(1) with input data } d_m \text{ and parameter } \Theta_L, \end{cases} \tag{4}$$

6

where $\boldsymbol{\mathcal{R}}_L : \Omega_{\Theta;L} \to [0, \infty)$ is a regularization function. In neural network training, it is common to employ regularization strategies [21, 22, 28, 34]. To clearly define the regularization term $\boldsymbol{\mathcal{R}}_L(\Theta_L)$, we introduce the following definitions.

Definition 2 Let $\mathbb{X}$ be a vector space. Given $\Xi_L \in \mathbb{Y}_L := \mathbb{X} \times (\mathbb{X})^2 \times \cdots \times (\mathbb{X})^L$, we define an operator $\mathbf{flip} : \mathbb{Y}_L \to (\mathbb{X})^{(L+1) \times (L+1)}$ by

flipping: $(\mathbf{flip}(\Xi_L))^{k,l} = (\mathbf{flip}(\Xi_L))^{l,k} = \Xi_L^{l,k}, 1 \le l \le L, 1 \le k \le l;$

with $(\mathbf{flip}(\Xi_L))_L^{0,0} = \Xi_L^{1,1}; (\mathbf{flip}(\Xi_L))^{l,0} = \Xi_L^{l,1}; (\mathbf{flip}(\Xi_L))^{0,k} = \Xi_L^{k,1}, 1 \le k, l \le L.$

We give an example in Figure 2 to help understand the $\mathbf{flip}(\cdot)$ operation.

Fig. 2 An illustration of the $\mathbf{flip}(\cdot)$ operation. The left and right pictures show the parameters $\mathbf{W}_3$ and $\mathbf{flip}(\mathbf{W}_3)$, respectively. The elements are arranged on the grid points according to their superscripts. The gray arrow represents the "copy" operation.

Let $\Theta_L = (\mathrm{T}_L, \mathrm{U}_L, \mathrm{a}_L, \mathrm{V}_L, \mathrm{b}_L, \mathrm{W}_L, \mathrm{c}_L) \in \Omega_{\Theta;L}$ and $\bar{\mathrm{W}}_L = \mathbf{flip}(\mathrm{W}_L), \bar{\mathrm{c}}_L = \mathbf{flip}(\mathrm{c}_L)$. We consider the following regularization term

$$\boldsymbol{\mathcal{R}}_L(\Theta_L) = \sum_{j=1}^{3} \boldsymbol{\mathcal{R}}_L^{(1)}((\mathrm{T}_j)_L) + \boldsymbol{\mathcal{R}}_L^{(1)}(\mathrm{U}_L) + \boldsymbol{\mathcal{R}}_L^{(2)}(\mathrm{a}_L) + \boldsymbol{\mathcal{R}}_L^{(1)}(\mathrm{V}_L) + \boldsymbol{\mathcal{R}}_L^{(2)}(\mathrm{b}_L) \\ + \boldsymbol{\mathcal{R}}_L^{(3)}(\bar{\mathrm{W}}_L) + \boldsymbol{\mathcal{R}}_L^{(4)}(\bar{\mathrm{c}}_L), \tag{5}$$

where

$$\mathcal{R}_L^{(1)}(\mathbf{U}_L) := \tau \sum_{l=1}^{L} \|\mathbf{U}_L^l\|^2 + \tau^{-1} \sum_{l=1}^{L} \|\mathbf{U}_L^l - \mathbf{U}_L^{l-1}\|^2,$$

$$\mathcal{R}_L^{(2)}(\mathbf{a}_L) := \tau \sum_{l=1}^{L} \|\mathbf{a}_L^l\|^2 + \tau^{-1} \sum_{l=1}^{L} \|\mathbf{a}_L^l - \mathbf{a}_L^{l-1}\|^2,$$

$$\mathcal{R}_L^{(3)}(\bar{\mathbf{W}}_L) := \tau^2 \sum_{l=1}^{L} \sum_{k=1}^{L} \|\bar{\mathbf{W}}_L^{l,k}\|^3$$
$$+ \tau^{-1} \Big(\sum_{l=1}^{L} \sum_{k=1}^{L} \|\bar{\mathbf{W}}_L^{l,k} - \bar{\mathbf{W}}_L^{l-1,k}\|^3 + \sum_{l=1}^{L} \sum_{k=1}^{L} \|\bar{\mathbf{W}}_L^{l,k} - \bar{\mathbf{W}}_L^{l,k-1}\|^3 \Big),$$

$$\mathcal{R}_L^{(4)}(\bar{\mathbf{c}}_L) := \tau^2 \sum_{l=1}^{L} \sum_{k=1}^{L} \|\bar{\mathbf{c}}_L^{l,k}\|^3$$
$$+ \tau^{-1} \Big(\sum_{l=1}^{L} \sum_{k=1}^{L} \|\bar{\mathbf{c}}_L^{l,k} - \bar{\mathbf{c}}_L^{l-1,k}\|^3 + \sum_{l=1}^{L} \sum_{k=1}^{L} \|\bar{\mathbf{c}}_L^{l,k} - \bar{\mathbf{c}}_L^{l,k-1}\|^3 \Big).$$

(6)

As can be seen, the definitions in (6) mimic some norms in some specific Sobolev spaces. We use the above regularization instead of standard l^2 or l^1 penalties to ensure compactness, as it controls both the magnitude and variation of parameters across layers, similar to the approach in [22, 35].

3 Deep-layer limit of the DNL framework: mathematical modeling and main convergence result

In this section, we consider the limit of the DNL framework (1) as the total layer number $L \to \infty$. We show dynamic system modeling for the DNL framework and present our main result on the convergence from the learning problem of the discrete-time DNL framework to that of the continuous-time DNL framework.

3.1 The continuous-time dynamical system modeling for DNL framework

We now model the forward propagation of the DNL framework (1) as a dynamical system in a continuous-time setting, and present its learning problem. For this, we give some notations in a continuous-time setting. The space of functions that are continuous on $\Omega \subset \mathbb{R}^d$ is denoted by $\mathcal{C}(\Omega)$. The Sobolev space of functions that are q-times weakly differentiable and each weak derivative is $\mathcal{L}^p$ integrable on Ω is denoted by $\mathcal{W}^{q,p}(\Omega)$. Specially, $\mathcal{H}^q(\Omega) := \mathcal{W}^{q,2}(\Omega)$. The functions are denoted with bold letters. For a vector-valued (resp., matrix-valued) function $\mathbf{a} : [0,1] \to \mathbb{R}^n$ (resp., $\mathbf{U} : [0,1] \to \mathbb{R}^{n \times n}$), we denote its supremum norm (if exists) as $\|\mathbf{a}\|_C$ (resp., $\|\mathbf{U}\|_C$) and denote its $\mathcal{L}^\infty$ norm

(if exists) as $\|\mathbf{a}\|_{\mathcal{L}^\infty}$ (resp., $\|\mathbf{U}\|_{\mathcal{L}^\infty}$). For a two-variable function $\mathbf{f}(t, s)$, we denote $\boldsymbol{\mathcal{D}}_t(\mathbf{f})$ and $\boldsymbol{\mathcal{D}}_{tt}(\mathbf{f})$ as its first-order and second-order weak partial derivative for variable t (if it exists). The notations $\boldsymbol{\mathcal{D}}_s(\mathbf{f})$ and $\boldsymbol{\mathcal{D}}_{ss}(\mathbf{f})$ are similarly defined. The identity operator is denoted by $\mathbf{Id}$.

Without loss of generality, we assume that the time interval for the continuous-time system is $[0, 1]$, and thus step size $\tau = 1/L$. We introduce the *ansatz* $\mathrm{x}^l \approx \mathbf{x}(l\tau)$, $\mathrm{T}^l \approx \mathbf{T}(l\tau)$, $(\mathrm{U}_L^l, \mathrm{a}_L^l) \approx (\mathbf{U}(l\tau), \mathbf{a}(l\tau))$, $(\mathrm{V}_L^l, \mathrm{b}_L^l) \approx (\mathbf{V}(l\tau), \mathbf{b}(l\tau))$ $(0 \le l \le L)$, and $(\mathrm{W}_L^{l,k}, \mathrm{c}_L^{l,k}) \approx (\mathbf{W}(l\tau, k\tau), \mathbf{c}(l\tau, k\tau))$ $(1 \le k \le l \le L)$, for some smooth curves $\mathbf{x}(t)$, $\mathbf{T}(t)$, $\mathbf{U}(t)$, $\mathbf{a}(t)$, $\mathbf{V}(t)$, $\mathbf{b}(t)$ defined for $0 \le t \le 1$ and $\mathbf{W}(t, s)$, $\mathbf{c}(t, s)$ defined for $0 \le t, s \le 1$. Let $t_L^l = l\tau$, $0 \le l \le L$. For small τ, we have $\mathbf{x}(t_L^l) \approx \mathrm{x}^l$, $\mathbf{T}(t_L^l) \approx \mathrm{T}^l$, $(\mathbf{U}(t_L^l), \mathbf{a}(t_L^l)) \approx (\mathrm{U}_L^l, \mathrm{a}_L^l)$, $(\mathbf{V}(t_L^l), \mathbf{b}(t_L^l)) \approx (\mathrm{V}_L^l, \mathrm{b}_L^l)$ and $(\mathbf{W}(t_L^l, t_L^k), \mathbf{c}(t_L^l, t_L^k)) \approx (\mathrm{W}_L^{l,k}, \mathrm{c}_L^{l,k})$. In this way, equation (1) implies that

$$\mathbf{x}(t_L^l) = \mathbf{V}(t_L^l)\boldsymbol{\phi}\circ\Big(\mathbf{U}(t_L^l)\boldsymbol{\mathcal{A}}_{\boldsymbol{\kappa}}(\mathbf{T}(t_L^l); \mathrm{d}) + \mathbf{a}(t_L^l)$$
$$+ \sum_{k=0}^{l-1}\int_{t_L^k}^{t_L^{k+1}}[\mathbf{W}(t_L^l, t_L^{k+1})\boldsymbol{\mathcal{A}}_{\boldsymbol{\kappa}}(\mathbf{T}(t_L^l); \mathbf{x}(t_L^k)) + \mathbf{c}(t_L^l, t_L^{k+1})]ds\Big) + \mathbf{b}(t_L^l).$$

This formulation can be seen as a discrete approximation of the following integral equation

$$\mathbf{x}(t) = \mathbf{V}(t)\boldsymbol{\phi}\circ\Big(\mathbf{U}(t)\boldsymbol{\mathcal{A}}_{\boldsymbol{\kappa}}(\mathbf{T}(t); \mathrm{d}) + \mathbf{a}(t) + \int_0^t\Big[\mathbf{W}(t, s)\boldsymbol{\mathcal{A}}_{\boldsymbol{\kappa}}(\mathbf{T}(t); \mathbf{x}(s)) + \mathbf{c}(t, s)\Big]ds\Big) + \mathbf{b}(t),$$
(7)

where $t \in [0, 1]$, $\mathrm{d} \in \mathbb{R}^n$ is the input data. It is noted that when $\boldsymbol{\phi} = \mathbf{Id}$, Eq.(7) is a Volterra integral equation of the second kind [36], which is often utilized to characterize the dynamic behavior of systems with memory properties or problems involving delay effects. We mention that this type of integral equation with memory properties is consistent with some original intention of DenseNets.

In the following, we refer to the integral equation (7) as continuous-time DNL framework. Similar to the discrete case, we package the continuous-time learnable parameters together and denote $\boldsymbol{\Theta} := (\mathbf{T}, \mathbf{U}, \mathbf{a}, \mathbf{V}, \mathbf{b}, \mathbf{W}, \mathbf{c})$, where $\mathbf{T} = (\mathbf{T}_1, \mathbf{T}_2, \mathbf{T}_3)$. For convenience, we let the parameter space

$$\begin{aligned}\mathcal{C}_{\boldsymbol{\Theta}} :=& \mathcal{C}([0, 1]; (\mathbb{R}^{n\times n})^3)\times\big(\mathcal{C}([0, 1]; \mathbb{R}^{n\times n})\times\mathcal{C}([0, 1]; \mathbb{R}^n)\big)^2 \\ &\times\{\mathbf{W}\in\mathcal{L}^\infty([0, 1]\times[0, 1]; \mathbb{R}^{n\times n}) : \mathbf{W}(t, s) = \mathbf{W}(s, t),\ 0 \le s, t \le 1\} \\ &\times\{\mathbf{c}\in\mathcal{L}^\infty([0, 1]\times[0, 1]; \mathbb{R}^n) : \mathbf{c}(t, s) = \mathbf{c}(s, t),\ 0 \le s, t \le 1\},\end{aligned}$$
(8)

and introduce a norm $\|\cdot\|_{\mathcal{C}_{\boldsymbol{\Theta}}}$ for $\boldsymbol{\Theta}\in\mathcal{C}_{\boldsymbol{\Theta}}$ as

$$\|\boldsymbol{\Theta}\|_{\mathcal{C}_{\boldsymbol{\Theta}}} := \max\{\|\mathbf{T}_1\|_C, \|\mathbf{T}_2\|_C, \|\mathbf{T}_3\|_C, \|\mathbf{U}\|_C, \|\mathbf{a}\|_C, \|\mathbf{V}\|_C, \|\mathbf{b}\|_C, \|\mathbf{W}\|_{\mathcal{L}^\infty}, \|\mathbf{c}\|_{\mathcal{L}^\infty}\}. \quad (9)$$

We sometimes abbreviate $\|\boldsymbol{\Theta}\|_{\mathcal{C}_{\boldsymbol{\Theta}}}$ as $\|\boldsymbol{\Theta}\|$. Here, to be consistent with the $\mathbf{flip}(\cdot)$ operation in discrete systems, we leverage symmetry to define the parameters $\mathbf{W}$ and $\mathbf{c}$ on

$[0,1] \times [0,1]$, thus avoiding some complex boundary condition discussions in subsequent analysis. We also introduce a parameter set

$$\Omega_{\boldsymbol{\Theta}} := \mathcal{C}_{\boldsymbol{\Theta}} \cap \Big(\mathcal{H}^1((0,1);(\mathbb{R}^{n \times n})^3) \times \big(\mathcal{H}^1((0,1);\mathbb{R}^{n \times n}) \times \mathcal{H}^1((0,1);\mathbb{R}^n) \big)^2$$
$$\times \mathcal{W}^{1,3}((0,1) \times (0,1);\mathbb{R}^{n \times n}) \times \mathcal{W}^{1,3}((0,1) \times (0,1);\mathbb{R}^n) \Big) \tag{10}$$

to define regularization terms for the continuous-time learning problem and enforce some compactness.

Let $\mathbf{x}(\cdot;\mathrm{d};\boldsymbol{\Theta})$ be the trajectory of Eq.(7) for parameter $\boldsymbol{\Theta}$ and input d. The learning problem of the continuous-time DNL framework can be given as follows

$$(\mathcal{P}): \begin{cases} \inf\limits_{\boldsymbol{\Theta} \in \Omega_{\boldsymbol{\Theta}}} \left\{ \mathfrak{L}_{\mathcal{S}}(\boldsymbol{\Theta}) = \frac{1}{M} \sum\limits_{m=1}^M \boldsymbol{\ell}(\mathbf{x}(1;\mathrm{d}_m;\boldsymbol{\Theta}),\mathrm{g}_m) + \mathcal{R}(\boldsymbol{\Theta}) \right\} \\ \text{subject to:} \\ \mathbf{x}(\cdot;\mathrm{d}_m;\boldsymbol{\Theta}) \text{ satisfies Eq.(7) with input data } \mathrm{d}_m \text{ and parameter } \boldsymbol{\Theta}, \end{cases} \tag{11}$$

where

$$\mathcal{R}(\boldsymbol{\Theta}) = \sum_{j=1}^3 \|\mathbf{T}_j\|^2_{\mathcal{H}^1((0,1);\mathbb{R}^{n \times n})} + \|\mathbf{U}\|^2_{\mathcal{H}^1((0,1);\mathbb{R}^{n \times n})} + \|\mathbf{a}\|^2_{\mathcal{H}^1((0,1);\mathbb{R}^n)} + \|\mathbf{V}\|^2_{\mathcal{H}^1((0,1);\mathbb{R}^{n \times n})}$$
$$+ \|\mathbf{b}\|^2_{\mathcal{H}^1((0,1);\mathbb{R}^n)} + \|\mathbf{W}\|^3_{\mathcal{W}^{1,3}((0,1) \times (0,1);\mathbb{R}^{n \times n})} + \|\mathbf{c}\|^3_{\mathcal{W}^{1,3}((0,1) \times (0,1);\mathbb{R}^n)}. \tag{12}$$

The regularization term $\mathcal{R}_L$ defined in Eq.(5) can be regarded as a discrete approximation of $\mathcal{R}$.

3.2 Main convergence result

In this subsection, we present the main result of this work on the convergence from the learning problem $(\mathcal{P}_L)$ of discrete-time DNL framework to the learning problem $(\mathcal{P})$ of continuous-time DNL framework, by using some extension operators defined as follows.

Definition 3 Let $\mathbb{X}$ be a vector space. Partition the time interval $[0,1]$ into L intervals $\{[(l-1)\tau, l\tau]\}_{l=1}^L$, where $\tau = 1/L$. We define the piecewise constant extension operator $\bar{\boldsymbol{\mathcal{I}}}_L : (\mathbb{X})^{L+1} \to \mathcal{M}_L([0,1];\mathbb{X})$ and the piecewise linear extension operator $\hat{\boldsymbol{\mathcal{I}}}_L : (\mathbb{X})^{L+1} \to \mathcal{C}([0,1];\mathbb{X})$ for $\Xi_L = (\xi_L^0, \xi_L^1, \ldots, \xi_L^L) \in (\mathbb{X})^{L+1}$ as

$$(\bar{\boldsymbol{\mathcal{I}}}_L \Xi_L)(t) = \begin{cases} \xi_L^l, & t \in ((l-1)\tau, l\tau], \ 1 \le l \le L, \\ \xi_L^0, & t = 0; \end{cases}$$

$$(\hat{\boldsymbol{\mathcal{I}}}_L \Xi_L)(t) = \xi_L^{l-1} + \Big(t - (l-1)\tau\Big) \frac{\xi_L^l - \xi_L^{l-1}}{\tau}, \ t \in [(l-1)\tau, l\tau], \ 1 \le l \le L,$$

where $\mathcal{M}_L([0,1];\mathbb{X})$ is the piecewise constant function space with L pieces.

Definition 4 Let $\mathbb{X}$ be a vector space. Divide the domain $[0,1] \times [0,1]$ equally into L^2 squares $\{[(l-1)\tau, l\tau] \times [(k-1)\tau, k\tau]\}_{l,k=1}^{L}$, where $\tau = 1/L$. Define the piecewise constant extension operator $\bar{\mathcal{BI}}_L : (\mathbb{X})^{(L+1)\times(L+1)} \to \mathcal{M}([0,1] \times [0,1]; \mathbb{X})$ and the piecewise bilinear extension operator $\hat{\mathcal{BI}}_L : (\mathbb{X})^{(L+1)\times(L+1)} \to \mathcal{C}([0,1] \times [0,1]; \mathbb{X})$ for $\Xi_L = (\xi_L^{l,k})_{0 \le l,k \le L} \in (\mathbb{X})^{(L+1)\times(L+1)}$ as

$$(\bar{\mathcal{BI}}_L \Xi_L)(t,s) = \begin{cases} \xi_L^{l,k}, & t \in ((l-1)\tau, l\tau], s \in ((k-1)\tau, k\tau], \ 1 \le l,k \le L, \\ \xi_L^{0,k}, & t = 0, s \in ((k-1)\tau, k\tau], \ 1 \le k \le L, \\ \xi_L^{l,0}, & s = 0, t \in ((l-1)\tau, l\tau], \ 1 \le l \le L, \\ \xi_L^{0,0}, & t = 0, s = 0. \end{cases}$$

$$(\hat{\mathcal{BI}}_L \Xi_L)(t,s) = \frac{k\tau - s}{\tau}\Big(\frac{l\tau - t}{\tau}\xi_L^{l-1,k-1} + \frac{t-(l-1)\tau}{\tau}\xi_L^{l,k-1}\Big) + \frac{s-(k-1)\tau}{\tau}\Big(\frac{l\tau - t}{\tau}\xi_L^{l-1,k}$$
$$+ \frac{t-(l-1)\tau}{\tau}\xi_L^{l,k}\Big), t \in [(l-1)\tau, l\tau], s \in [(k-1)\tau, k\tau], 1 \le k,l \le L,$$

where $\mathcal{M}([0,1] \times [0,1]; \mathbb{X})$ is the piecewise constant function space with $L \times L$ pieces.

These extension operators connect the learnable parameters of the discrete-time DNL framework (1) and continuous-time DNL framework (7). In particular, we denote $\hat{\mathcal{I}}_L \Theta_L = (\hat{\mathcal{I}}_L(\mathrm{T}_1)_L, \hat{\mathcal{I}}_L(\mathrm{T}_2)_L, \hat{\mathcal{I}}_L(\mathrm{T}_3)_L, \hat{\mathcal{I}}_L \mathrm{U}_L, \hat{\mathcal{I}}_L \mathrm{a}_L, \hat{\mathcal{I}}_L \mathrm{V}_L, \hat{\mathcal{I}}_L \mathrm{b}_L, \hat{\mathcal{BI}}_L(\mathbf{flip}(\mathrm{W}_L)),$ $\hat{\mathcal{BI}}_L(\mathbf{flip}(\mathrm{c}_L)))$ and $\bar{\mathcal{I}}_L \Theta_L = (\bar{\mathcal{I}}_L(\mathrm{T}_1)_L, \bar{\mathcal{I}}_L(\mathrm{T}_2)_L, \bar{\mathcal{I}}_L(\mathrm{T}_3)_L, \bar{\mathcal{I}}_L \mathrm{U}_L, \bar{\mathcal{I}}_L \mathrm{a}_L, \bar{\mathcal{I}}_L \mathrm{V}_L, \bar{\mathcal{I}}_L \mathrm{b}_L,$ $\bar{\mathcal{BI}}_L(\mathbf{flip}(\mathrm{W}_L)), \bar{\mathcal{BI}}_L(\mathbf{flip}(\mathrm{c}_L)))$.

We will use the following assumptions:

(A_1) The activation function $\boldsymbol{\phi} : \mathbb{R} \to \mathbb{R}$ is $L_{\boldsymbol{\phi}}$-Lipschitz, increasing, acts point-wise and takes 0 at 0.

(A_2) The transformation $\boldsymbol{\mathcal{A}}_{\boldsymbol{\kappa}} : (\mathbb{R}^{n\times n})^3 \times \mathbb{R}^n \to \mathbb{R}^n$ satisfies a local growth condition and a local Lipschitz condition in the following sense. There exist continuous functions $\boldsymbol{\mathcal{G}}_{\boldsymbol{\mathcal{A}}}(\cdot) : (\mathbb{R}^{n\times n})^3 \to \mathbb{R}^+$ and $\boldsymbol{\mathcal{L}}_{\boldsymbol{\mathcal{A}}}(\cdot,\cdot,\cdot) : (\mathbb{R}^{n\times n})^3 \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^+$ such that for all $\Xi = (\Xi_1, \Xi_2, \Xi_3), \Xi' = (\Xi_1', \Xi_2', \Xi_3') \in (\mathbb{R}^{n\times n})^3$ and $\mathrm{z}, \mathrm{z}' \in \mathbb{R}^n$

$$|\boldsymbol{\mathcal{A}}_{\boldsymbol{\kappa}}(\Xi, \mathrm{z})| \le \boldsymbol{\mathcal{G}}_{\boldsymbol{\mathcal{A}}}(\Xi) \cdot |\mathrm{z}|,$$

$$|\boldsymbol{\mathcal{A}}_{\boldsymbol{\kappa}}(\Xi, \mathrm{z}) - \boldsymbol{\mathcal{A}}_{\boldsymbol{\kappa}}(\Xi', \mathrm{z}')| \le \boldsymbol{\mathcal{L}}_{\boldsymbol{\mathcal{A}}}(\Xi, \mathrm{z}, \mathrm{z}') \cdot \Big(\sum_{i=1}^{3}\|\Xi_i - \Xi_i'\| + |\mathrm{z} - \mathrm{z}'|\Big).$$

(A_3) The loss function $\boldsymbol{\ell} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^+$ is continuous in its first argument.

Remark 1 These assumptions are mild and commonly used in deep learning theory. Many widely used activation functions, such as ReLU, parametric ReLU, and tanh [1], satisfy (A_1). Common loss functions such as mean squared error and cross-entropy satisfy (A_3). Moreover, a broad class of existing architectures satisfy (A_2). For example:

(i) Affine map in DenseNet: the kernel $\boldsymbol{\kappa}((\Xi_1 \mathrm{z})[i], (\Xi_2 \mathrm{z})[j]) = \boldsymbol{\delta}_{ij}$ and $\boldsymbol{\mathcal{A}}_{\boldsymbol{\kappa}}(\Xi, \mathrm{z}) = \mathrm{z}\Xi_3$, where $\boldsymbol{\delta}_{ij}$ is the Kronecker function. Then (A_2) is satisfied with

$$\boldsymbol{\mathcal{G}}_{\boldsymbol{\mathcal{A}}}(\Xi) = \|\Xi_3\|_2, \ \boldsymbol{\mathcal{L}}_{\boldsymbol{\mathcal{A}}}(\Xi, \mathrm{z}, \mathrm{z}') = \|\Xi_3\|_2 + |\mathrm{z}'|.$$

(ii) Self-attention in Transformers [11]: the kernel $\boldsymbol{\kappa}((\Xi_1 z)[i], (\Xi_2 z)[j]) = \exp((\Xi_1 z)[i] \cdot (\Xi_2 z)[j]/\sqrt{n})$, and $\boldsymbol{\mathcal{A}}_{\boldsymbol{\kappa}}(\Xi, z) = \operatorname{softmax}\left(\frac{\Xi_1 z \cdot (\Xi_2 z)^{\top}}{\sqrt{n}}\right) \Xi_3 z$. One can verify that assumption (A_2) holds with

$$\boldsymbol{\mathcal{G}}_{\boldsymbol{\mathcal{A}}}(\Xi) = \sqrt{n}\|\Xi_3\|_2,$$
$$\boldsymbol{\mathcal{L}}_{\boldsymbol{\mathcal{A}}}(\Xi, z, z') = \sqrt{n} \max\left\{\|\Xi\|^2 |z|^3, \|\Xi\|^3(|z|^2 + |z||z'|) + \|\Xi\|, \|\Xi\|^2 |z|^2 |z'|, |z'|\right\},$$

where $\|\Xi\| = \max\{\|\Xi_1\|_2, \|\Xi_2\|_2, \|\Xi_3\|_2\}$.

(iii) Gaussian non-local operator in Non-local Nets [12]: the kernel is given by $\boldsymbol{\kappa}((\Xi_1 z)[i], (\Xi_2 z)[j]) = \exp(z[i] \cdot z[j])$, and $\boldsymbol{\mathcal{A}}_{\boldsymbol{\kappa}}(\Xi, z) = \operatorname{softmax}\left(zz^{\top}\right) \Xi_3 z$. In this case, (A_2) is satisfied with

$$\boldsymbol{\mathcal{G}}_{\boldsymbol{\mathcal{A}}}(\Xi) = \sqrt{n}\|\Xi_3\|_2,$$
$$\boldsymbol{\mathcal{L}}_{\boldsymbol{\mathcal{A}}}(\Xi, z, z') = \max\left\{\|\Xi\||z|(|z| + |z'|) + \sqrt{n}\|\Xi\|, \sqrt{n}|z'|\right\}.$$

See the supplementary material for detailed derivations.

Our main result establishes both the convergence of optimal values and the subsequence convergence of optimal solutions for the discrete-time learning problem as the number of layers approaches infinity.

Theorem 1 *(Convergence from the discrete to continuous learning problem) Consider the problems* $(\mathcal{P}_L), (\mathcal{P})$ *defined in* (4) *and* (11), *respectively. Let the parameter sets* $\Omega_{\Theta;L}$ *and* Ω_{Θ} *be given by* (2) *and* (10). *If assumptions* $(A_1) - (A_3)$ *hold, then minimizers of* $(\mathcal{P}_L)$ *and* $(\mathcal{P})$ *exist. Moreover, for any sequence* $\{\Theta_L^*\}_L \subset \Omega_{\Theta;L}$, *where* $\Theta_L^* = (\mathrm{T}_L^*, \mathrm{U}_L^*, \mathrm{a}_L^*, \mathrm{V}_L^*, \mathrm{b}_L^*, \mathrm{W}_L^*, \mathrm{c}_L^*)$ *is a minimizer of* $(\mathcal{P}_L)$, *we have*

$$\min_{\Theta_L \in \Omega_{\Theta;L}} \boldsymbol{\mathfrak{L}}_{\mathcal{S};L}(\Theta_L) = \boldsymbol{\mathfrak{L}}_{\mathcal{S};L}\left(\Theta_L^*\right) \to \min_{\boldsymbol{\Theta} \in \Omega_{\boldsymbol{\Theta}}} \boldsymbol{\mathfrak{L}}_{\mathcal{S}}(\boldsymbol{\Theta}), \quad as\ L \to \infty.$$

Let $\boldsymbol{\Theta}_L^* := \hat{\boldsymbol{\mathcal{I}}}_L \Theta_L^*$, $L \geq 1$. *Then* $\{\boldsymbol{\Theta}_L^*\}_L$ *is relatively compact in* $\mathcal{C}_{\boldsymbol{\Theta}}$ *with respect to the norm* $\|\cdot\|_{\mathcal{C}_{\boldsymbol{\Theta}}}$, *and any limit point of* $\{\boldsymbol{\Theta}_L^*\}_L$ *in* $\Omega_{\boldsymbol{\Theta}}$ *is a minimiser of* $(\mathcal{P})$.

Theorem 1 establishes that the training process of the discrete-time DNL framework can be viewed as a consistent discretization of a well-posed continuous-time learning problem, which may provide a theoretical justification in designing dynamic system learning approaches for networks with densely connected layers. While our focus is on theoretical guarantees, these results imply that densely connected architectures can offer stability when training very deep models. The proof of the theorem is given in the subsequent sections.

4 Proof details

In this section, we provide a detailed proof for Theorem 1, primarily utilizing the properties of Γ-convergence. We begin by presenting some useful preliminary results

on the forward dynamics of the DNL framework. Recall that $\tau = 1/L$, $t_L^l := l \cdot \tau$, $1 \leq l \leq L$.

Lemma 2 *(Bound estimation for network states) Let assumptions $(A_1)(A_2)$ hold. Given a layer number $L \in \mathbb{N}$, a learnable parameter $\Theta_L \in \Omega_{\Theta;L}$ and an input $\mathrm{d} \in \mathbb{R}^n$. Then the state variables $\{\mathrm{x}^l(\mathrm{d}; \Theta_L)\}_{l=0}^L$ of DNL framework (1) exist uniquely and satisfy*

$$
\begin{aligned}
|\mathrm{x}^l(\mathrm{d}; \Theta_L)| \leq &L_{\boldsymbol{\phi}} \|\Theta_L\|^2 \bar{M}_{\boldsymbol{\mathcal{A}}} \big[L_{\boldsymbol{\phi}} \|\Theta_L\|^2 (\bar{M}_{\boldsymbol{\mathcal{A}}} |\mathrm{d}| + 2) + \|\Theta_L\| \big] \exp(L_{\boldsymbol{\phi}} \|\Theta_L\|^2 \bar{M}_{\boldsymbol{\mathcal{A}}}) \\
&+ L_{\boldsymbol{\phi}} \|\Theta_L\|^2 (\bar{M}_{\boldsymbol{\mathcal{A}}} |\mathrm{d}| + 2) + \|\Theta_L\|,
\end{aligned}
\tag{13}
$$

where $\bar{M}_{\boldsymbol{\mathcal{A}}} = \max_{0 \leq l \leq L} \{\boldsymbol{\mathcal{G}}_{\boldsymbol{\mathcal{A}}}(\mathrm{T}_L^l)\}$. Moreover, $\{\mathrm{x}^l(\mathrm{d}; \Theta_L)\}_{l=0}^L$ continuously depend on the learnable parameter Θ_L.

Proof The proof is standard and is supplied in supplementary materials. □

The next proposition shows that under the assumptions of (A_1) and (A_2), the continuous-time DNL framework (7) is well defined.

Proposition 3 *(Existence, uniqueness, and bound estimation for the continuous architecture) Let $\mathrm{d} \in \mathbb{R}^n$ be a given input data. Let assumptions (A_1) and (A_2) hold true. For any given parameter $\boldsymbol{\Theta} = (\mathbf{T}, \mathbf{U}, \mathbf{a}, \mathbf{V}, \mathbf{b}, \mathbf{W}, \mathbf{c}) \in \mathcal{C}_{\boldsymbol{\Theta}}$, the continuous-time DNL framework (7) has a unique continuous solution on $[0, 1]$ which satisfies*

$$
|\mathbf{x}(t)| \leq \Big[L_{\boldsymbol{\phi}} \|\mathbf{V}\|_C (\hat{M}_{\boldsymbol{\mathcal{A}}} \|\mathbf{U}\|_C |\mathrm{d}| + \|\mathbf{a}\|_C + \|\mathbf{b}\|_C) + \|\mathbf{c}\|_{\mathcal{L}^\infty} \Big] \cdot \exp(L_{\boldsymbol{\phi}} \hat{M}_{\boldsymbol{\mathcal{A}}} \|\mathbf{V}\|_C \|\mathbf{W}\|_{\mathcal{L}^\infty}),
$$

where $t \in [0, 1]$, $\hat{M}_{\boldsymbol{\mathcal{A}}} = \sup_{t \in [0,1]} \boldsymbol{\mathcal{G}}_{\boldsymbol{\mathcal{A}}}(\mathbf{T}(t)) < \infty$. Moreover, the solution $\mathbf{x}(\cdot)$ continuously depend on the learnable parameter $\boldsymbol{\Theta}$.

Proof The proof is given in supplementary materials. □

The next proposition shows that when the layer number $L \to \infty$, the state variables $\{\mathrm{x}^l(\mathrm{d}; \Theta_L)\}_l$ of the discrete-time DNL framework (1) converge to the trajectory of the continuous-time DNL framework (7). Such a result plays an important role in the proof of Theorem 1, as it infers the convergence of the data loss of $(\mathcal{P}_L)$.

Proposition 4 *(Forward convergence) Let assumptions $(A_1)(A_2)$ hold. Given data $\mathrm{d} \in \mathbb{R}^n$, parameters $\boldsymbol{\Theta} = (\mathbf{T}, \mathbf{U}, \mathbf{a}, \mathbf{V}, \mathbf{b}, \mathbf{W}, \mathbf{c}) \in \mathcal{C}_{\boldsymbol{\Theta}}$ and $\Theta_L = (\mathrm{T}_L, \mathrm{U}_L, \mathrm{a}_L, \mathrm{V}_L, \mathrm{b}_L, \mathrm{W}_L, \mathrm{c}_L) \in \Omega_{\Theta;L}$, $L \geq 1$. If*

$$
\bar{\boldsymbol{\mathcal{I}}}_L \Theta_L \xrightarrow{\mathcal{C}_{\boldsymbol{\Theta}}} \boldsymbol{\Theta},
\tag{14}
$$

as $L \to \infty$, then $\sup_{1 \leq l \leq L} \sup_{t \in [t_L^{l-1}, t_L^l]} |\mathrm{x}^l(\mathrm{d}; \Theta_L) - \mathbf{x}(t; \mathrm{d}; \boldsymbol{\Theta})| \to 0$, as $L \to \infty$.

13

Proof Since $\boldsymbol{\Theta} \in \Omega_{\boldsymbol{\Theta}}$ is given, $\{\Theta_L\}_L$ is bounded by (14). Hence, there exist constants

$$M_{\boldsymbol{\Theta}} := \max\{\sup_{L \geq 1} \|\Theta_L\|, \|\boldsymbol{\Theta}\|_{\Omega_{\boldsymbol{\Theta}}}\} < +\infty,$$

$$M_{\boldsymbol{\mathcal{A}}} := \max\Big\{ \sup_{L \geq 1} \max_{0 \leq l \leq L} \boldsymbol{\mathcal{G}}_{\boldsymbol{\mathcal{A}}}(\mathrm{T}_L^l), \ \sup_{0 \leq t \leq 1} \boldsymbol{\mathcal{G}}_{\boldsymbol{\mathcal{A}}}(\mathbf{T}(t))\Big\} < \infty,$$

by assumption (A_2). Additionally, there exist constants $B_{\mathrm{x}} > 0$ and $B_{\mathbf{x}} > 0$ such that

$$|\mathrm{x}^l(\mathrm{d};\Theta_L)| \leq B_{\mathrm{x}}, \ \forall 1 \leq l \leq L, \ L \geq 1; \ |\mathbf{x}(t;\mathrm{d};\boldsymbol{\Theta})| \leq B_{\mathbf{x}}, \ \forall t \in [0,1],$$

as a consequence of Lemma 2 and Proposition 3, respectively. Therefore, we can let

$$L_{\boldsymbol{\mathcal{A}}} := \max\Big\{ \sup_{\substack{L \geq 1, \\ |\mathrm{x}|, |\mathrm{x}'| \leq B_{\mathrm{x}}}} \max_{0 \leq l \leq L} \boldsymbol{\mathcal{L}}_{\boldsymbol{\mathcal{A}}}(\mathrm{T}_L^l, \mathrm{x}, \mathrm{x}'), \ \sup_{\substack{L \geq 1, \\ |\mathrm{x}|, |\mathrm{x}'| \leq B_{\mathrm{x}}}} \{\boldsymbol{\mathcal{L}}_{\boldsymbol{\mathcal{A}}}(\mathbf{T}(t), \mathrm{x}, \mathrm{x}')\}\Big\} < \infty.$$

For simplicity, we denote $\mathrm{x}_L^l = \mathrm{x}^l(\mathrm{d};\Theta_L)$, $\mathbf{x}(t_L^l) = \mathbf{x}(t_L^l;\mathrm{d};\boldsymbol{\Theta})$. Subtracting Eq.(1) from Eq.(7) and using the assumptions $(A_1)(A_2)$ give

$$\left|\mathrm{x}_L^l - \mathbf{x}(t_L^l)\right| \tag{15}$$

$$= \Big| \mathrm{V}_L^l \boldsymbol{\phi} \circ \big(\mathrm{U}_L^l \boldsymbol{\mathcal{A}}_{\boldsymbol{\kappa}}(\mathrm{T}_L^l;\mathrm{d}) + \mathrm{a}_L^l + \tau \sum_{k=0}^{l-1}[\mathrm{W}_L^{l,k+1} \boldsymbol{\mathcal{A}}_{\boldsymbol{\kappa}}(\mathrm{T}_L^l;\mathrm{x}_L^k) + \mathrm{c}_L^{l,k+1}]\big) + \mathrm{b}_L^l$$

$$- \mathbf{V}(t_L^l)\boldsymbol{\phi} \circ \Big(\mathrm{U}_L^l \boldsymbol{\mathcal{A}}_{\boldsymbol{\kappa}}(\mathrm{T}_L^l;\mathrm{d}) + \mathrm{a}_L^l + \tau \sum_{k=0}^{l-1}[\mathrm{W}_L^{l,k+1} \boldsymbol{\mathcal{A}}_{\boldsymbol{\kappa}}(\mathrm{T}_L^l;\mathrm{x}_L^k) + \mathrm{c}_L^{l,k+1}]\Big)$$

$$+ \mathbf{V}(t_L^l)\boldsymbol{\phi} \circ \Big(\mathrm{U}_L^l \boldsymbol{\mathcal{A}}_{\boldsymbol{\kappa}}(\mathrm{T}_L^l;\mathrm{d}) + \mathrm{a}_L^l + \tau \sum_{k=0}^{l-1}[\mathrm{W}_L^{l,k+1} \boldsymbol{\mathcal{A}}_{\boldsymbol{\kappa}}(\mathrm{T}_L^l;\mathrm{x}_L^k) + \mathrm{c}_L^{l,k+1}]\Big)$$

$$- \mathbf{V}(t_L^l)\boldsymbol{\phi} \circ \Big[\mathbf{U}(t_L^l)\boldsymbol{\mathcal{A}}_{\boldsymbol{\kappa}}(\mathbf{T}(t_L^l);\mathrm{d}) + \mathbf{a}(t_L^l)$$

$$+ \int_0^{t_L^l}(\mathbf{W}(t_L^l,s)\boldsymbol{\mathcal{A}}_{\boldsymbol{\kappa}}(\mathbf{T}(t_L^l);\mathbf{x}(s)) + \mathbf{c}(t_L^l,s))\mathrm{d}s\Big] - \mathbf{b}(t_L^l)\Big|$$

$$\leq L_{\boldsymbol{\phi}}\|\mathrm{V}_L^l - \mathbf{V}(t_L^l)\| \cdot \Big|\mathrm{U}_L^l \boldsymbol{\mathcal{A}}_{\boldsymbol{\kappa}}(\mathrm{T}_L^l;\mathrm{d}) + \mathrm{a}_L^l + \tau \sum_{k=0}^{l-1}[\mathrm{W}_L^{l,k+1} \boldsymbol{\mathcal{A}}_{\boldsymbol{\kappa}}(\mathrm{T}_L^l;\mathrm{x}_L^k) + \mathrm{c}_L^{l,k+1}]\Big|$$

$$+ |\mathrm{b}_L^l - \mathbf{b}(t_L^l)| + L_{\boldsymbol{\phi}}\|\mathbf{V}(t_L^l)\| \cdot |\mathrm{U}_L^l \boldsymbol{\mathcal{A}}_{\boldsymbol{\kappa}}(\mathrm{T}_L^l;\mathrm{d}) + \mathrm{a}_L^l - v\mathbf{U}(t_L^l)\boldsymbol{\mathcal{A}}_{\boldsymbol{\kappa}}(\mathbf{T}(t_L^l);\mathrm{d}) - \mathbf{a}(t_L^l)|$$

$$+ L_{\boldsymbol{\phi}}\|\mathbf{V}(t_L^l)\| \cdot \Big|\tau \sum_{k=0}^{l-1}\Big[\mathrm{W}_L^{l,k+1} \boldsymbol{\mathcal{A}}_{\boldsymbol{\kappa}}(\mathrm{T}_L^l;\mathrm{x}_L^k) + \mathrm{c}_L^{l,k+1}\Big]$$

$$- \int_0^{t_L^l}\Big[\mathbf{W}(t_L^l,s)\boldsymbol{\mathcal{A}}_{\boldsymbol{\kappa}}(\mathbf{T}(t_L^l);\mathbf{x}(s)) + \mathbf{c}(t_L^l,s)\Big]ds\Big|,$$

where $l \geq 1$. The first term on the right-hand side of Eq.(15) can be bounded as

$$L_{\boldsymbol{\phi}}\|\mathrm{V}_L^l - \mathbf{V}(t_L^l)\| \cdot \Big|\mathrm{U}_L^l \boldsymbol{\mathcal{A}}_{\boldsymbol{\kappa}}(\mathrm{T}_L^l;\mathrm{d}) + \mathrm{a}_L^l + \tau \sum_{k=0}^{l-1}[\mathrm{W}_L^{l,k+1} \boldsymbol{\mathcal{A}}_{\boldsymbol{\kappa}}(\mathrm{T}_L^l;\mathrm{x}_L^k) + \mathrm{c}_L^{l,k+1}]\Big|$$

$$\leq L_{\boldsymbol{\phi}}\|\bar{\boldsymbol{\mathcal{I}}}_L \mathrm{V}_L - \mathbf{V}\|_C\Big[\|\Theta_L\|(M_{\boldsymbol{\mathcal{A}}}|\mathrm{d}| + 1) + \tau \sum_{k=0}^{l-1}(\|\Theta_L\|M_{\boldsymbol{\mathcal{A}}}|\mathrm{x}_L^k| + \|\Theta_L\|)\Big]$$

$$\leq L_{\boldsymbol{\phi}}\|\bar{\boldsymbol{\mathcal{I}}}_L \mathrm{V}_L - \mathbf{V}\|_C \cdot M_{\boldsymbol{\Theta}}\big(M_{\boldsymbol{\mathcal{A}}}|\mathrm{d}| + M_{\boldsymbol{\mathcal{A}}}B_{\mathrm{x}} + 2\big),$$

14

where the first inequality is due to Definition (3) and assumption (A_2). In addition, we may bound the last two terms in the right-hand side of Eq.(15) by

$$L_{\boldsymbol{\phi}}\|\mathbf{V}(t_L^l)\|\big|\mathrm{U}_L^l\boldsymbol{\mathcal{A}_\kappa}(\mathrm{T}_L^l;\mathrm{d})+\mathrm{a}_L^l-\mathbf{U}(t_L^l)\boldsymbol{\mathcal{A}_\kappa}(\mathbf{T}(t_L^l);\mathrm{d})-\mathbf{a}(t_L^l)\big|$$

$$\leq L_{\boldsymbol{\phi}}M_\Theta\big(M_{\boldsymbol{\mathcal{A}}}|\mathrm{d}|\cdot\|\bar{\boldsymbol{\mathcal{I}}}_L\mathrm{U}_L-\mathbf{U}\|_C+M_\Theta L_{\boldsymbol{\mathcal{A}}}\|\bar{\boldsymbol{\mathcal{I}}}_L\mathrm{T}_L-\mathbf{T}\|_C+\|\bar{\boldsymbol{\mathcal{I}}}_L\mathrm{a}_L-\mathbf{a}\|_C\big),$$

and

$$L_{\boldsymbol{\phi}}\|\mathbf{V}(t_L^l)\|\cdot\Big|\tau\sum_{k=0}^{l-1}[\mathrm{W}_L^{l,k+1}\boldsymbol{\mathcal{A}_\kappa}(\mathrm{T}_L^l;\mathrm{x}_L^k)+\mathrm{c}_L^{l,k+1}]$$

$$-\int_0^{t_L^l}[\mathbf{W}(t_L^l,s)\boldsymbol{\mathcal{A}_\kappa}(\mathbf{T}(t_L^l);\mathbf{x}(s))+\mathbf{c}(t_L^l,s)]ds\Big|$$

$$\leq L_{\boldsymbol{\phi}}\|\boldsymbol{\Theta}\|\bigg\{\sum_{k=0}^{l-1}\int_{t_L^k}^{t_L^{k+1}}\|\mathrm{W}_L^{l,k+1}\|\big|\boldsymbol{\mathcal{A}_\kappa}(\mathrm{T}_L^l;\mathrm{x}_L^k)-\boldsymbol{\mathcal{A}_\kappa}(\mathbf{T}(t_L^l);\mathbf{x}(s))\big|$$

$$+\|\mathrm{W}_L^{l,k+1}-\mathbf{W}(t_L^l,s)\|\big|\boldsymbol{\mathcal{A}_\kappa}(\mathbf{T}(t_L^l);\mathbf{x}(s))\big|ds+\int_0^{t_L^l}\big|(\bar{\boldsymbol{\mathcal{BI}}}_L\bar{\mathrm{c}}_L)(t_L^l,s)-\mathbf{c}(t_L^l,s)\big|ds\bigg\}$$

$$\leq L_{\boldsymbol{\phi}}\|\boldsymbol{\Theta}\|\bigg\{\sum_{k=0}^{l-1}\|\Theta_L\|L_{\boldsymbol{\mathcal{A}}}\int_{t_L^k}^{t_L^{k+1}}\Big(|\mathrm{x}_L^k-\mathbf{x}(s)|+\|\mathrm{T}_L^l-\mathbf{T}(t_L^l)\|\Big)ds$$

$$+M_{\boldsymbol{\mathcal{A}}}B_\mathbf{x}\int_{t_L^k}^{t_L^{k+1}}\|\mathrm{W}_L^{l,k+1}-\mathbf{W}(t_L^l,s)\|ds+\int_0^{t_L^l}\big|(\bar{\boldsymbol{\mathcal{BI}}}_L\bar{\mathrm{c}}_L)(t_L^l,s)-\mathbf{c}(t_L^l,s)\big|ds\bigg\}$$

$$\leq L_{\boldsymbol{\phi}}\|\boldsymbol{\Theta}\|\|\Theta_L\|L_{\boldsymbol{\mathcal{A}}}\sum_{k=0}^{l-1}\int_{t_L^k}^{t_L^{k+1}}\Big(|\mathrm{x}_L^k-\mathbf{x}(t_L^k)|+|\mathbf{x}(t_L^k)-\mathbf{x}(s)|+\|\mathrm{T}_L^l-\mathbf{T}(t_L^l)\|\Big)ds$$

$$+L_{\boldsymbol{\phi}}\|\boldsymbol{\Theta}\|M_{\boldsymbol{\mathcal{A}}}B_\mathbf{x}\int_0^{t_L^l}\big\|(\bar{\boldsymbol{\mathcal{BI}}}_L\bar{\mathrm{W}}_L)(t_L^l,s)-\mathbf{W}(t_L^l,s)\big\|ds$$

$$+L_{\boldsymbol{\phi}}\|\boldsymbol{\Theta}\|\int_0^{t_L^l}\big|(\bar{\boldsymbol{\mathcal{BI}}}_L\bar{\mathrm{c}}_L)(t_L^l,s)-\mathbf{c}(t_L^l,s)\big|ds$$

$$\leq\tau L_{\boldsymbol{\phi}}L_{\boldsymbol{\mathcal{A}}}M_\Theta^2\sum_{k=0}^{l-1}|\mathrm{x}_L^k-\mathbf{x}(t_L^k)|+L_{\boldsymbol{\phi}}L_{\boldsymbol{\mathcal{A}}}M_\Theta^2\boldsymbol{\omega}_\mathbf{x}(\tau)+L_{\boldsymbol{\phi}}L_{\boldsymbol{\mathcal{A}}}M_\Theta^2\|\bar{\boldsymbol{\mathcal{I}}}_L\mathrm{T}_L-\mathbf{T}\|_C$$

$$+L_{\boldsymbol{\phi}}M_\Theta M_{\boldsymbol{\mathcal{A}}}B_\mathbf{x}\big\|\bar{\boldsymbol{\mathcal{BI}}}_L\bar{\mathrm{W}}_L-\mathbf{W}\big\|_{\mathcal{L}^\infty}+L_{\boldsymbol{\phi}}M_\Theta\big\|\bar{\boldsymbol{\mathcal{BI}}}_L\bar{\mathrm{c}}_L-\mathbf{c}\big\|_{\mathcal{L}^\infty},$$

where $\bar{\mathrm{W}}_L=\mathbf{flip}(\mathrm{W}_L),\bar{\mathrm{c}}_L=\mathbf{flip}(\mathrm{c}_L)$, $\boldsymbol{\omega}_\mathbf{x}(\cdot)$ is the modulus of continuity of $\mathbf{x}$. Combining the above four inequalities, we obtain for $1\leq l\leq L$ that

$$|\mathrm{x}_L^l-\mathbf{x}(t_L^l)|$$

$$\leq L_{\boldsymbol{\phi}}\|\bar{\boldsymbol{\mathcal{I}}}_L\mathrm{V}_L-\mathbf{V}\|_C\cdot M_\Theta\big(M_{\boldsymbol{\mathcal{A}}}|\mathrm{d}|+M_{\boldsymbol{\mathcal{A}}}B_\mathbf{x}+2\big)+\|\bar{\boldsymbol{\mathcal{I}}}_L\mathrm{b}_L-\mathbf{b}\|_C$$

$$+L_{\boldsymbol{\phi}}M_\Theta\big(M_{\boldsymbol{\mathcal{A}}}|\mathrm{d}|\cdot\|\bar{\boldsymbol{\mathcal{I}}}_L\mathrm{U}_L-\mathbf{U}\|_C+M_\Theta L_{\boldsymbol{\mathcal{A}}}\|\bar{\boldsymbol{\mathcal{I}}}_L\mathrm{T}_L-\mathbf{T}\|_C+\|\bar{\boldsymbol{\mathcal{I}}}_L\mathrm{a}_L-\mathbf{a}\|_C\big)$$

$$+\tau L_{\boldsymbol{\phi}}L_{\boldsymbol{\mathcal{A}}}M_\Theta^2\sum_{k=0}^{l-1}|\mathrm{x}_L^k-\mathbf{x}(t_L^k)|+L_{\boldsymbol{\phi}}L_{\boldsymbol{\mathcal{A}}}M_\Theta^2\boldsymbol{\omega}_\mathbf{x}(\tau)+L_{\boldsymbol{\phi}}L_{\boldsymbol{\mathcal{A}}}M_\Theta^2\|\bar{\boldsymbol{\mathcal{I}}}_L\mathrm{T}_L-\mathbf{T}\|_C \qquad(16)$$

$$+L_{\boldsymbol{\phi}}M_\Theta M_{\boldsymbol{\mathcal{A}}}B_\mathbf{x}\|\bar{\boldsymbol{\mathcal{BI}}}_L\bar{\mathrm{W}}_L-\mathbf{W}\|_{\mathcal{L}^\infty}+L_{\boldsymbol{\phi}}M_\Theta\|\bar{\boldsymbol{\mathcal{BI}}}_L\bar{\mathrm{c}}_L-\mathbf{c}\|_{\mathcal{L}^\infty}$$

$$=\tau L_{\boldsymbol{\phi}}L_{\boldsymbol{\mathcal{A}}}M_\Theta^2\sum_{k=0}^{l-1}|\mathrm{x}_L^k-\mathbf{x}(t_L^k)|+C_L,$$

15

with $C_L = L_{\boldsymbol{\phi}} L_{\boldsymbol{\mathcal{A}}} M_{\Theta}^2 \boldsymbol{\omega}_{\mathbf{x}}(\tau) + 2 L_{\boldsymbol{\phi}} L_{\boldsymbol{\mathcal{A}}} M_{\Theta}^2 \|\bar{\boldsymbol{\mathcal{I}}}_L \mathrm{T}_L - \mathbf{T}\|_C + L_{\boldsymbol{\phi}} M_{\Theta} M_{\boldsymbol{\mathcal{A}}} |\mathrm{d}| \|\bar{\boldsymbol{\mathcal{I}}}_L \mathrm{U}_L - \mathbf{U}\|_C + L_{\boldsymbol{\phi}} M_{\Theta} \|\bar{\boldsymbol{\mathcal{I}}}_L \mathrm{a}_L - \mathbf{a}\|_C + L_{\boldsymbol{\phi}} M_{\Theta} (M_{\boldsymbol{\mathcal{A}}} |\mathrm{d}| + M_{\boldsymbol{\mathcal{A}}} B_{\mathbf{x}} + 2) \|\bar{\boldsymbol{\mathcal{I}}}_L \mathrm{V}_L - \mathbf{V}\|_C + \|\bar{\boldsymbol{\mathcal{I}}}_L \mathrm{b}_L - \mathbf{b}\|_C + L_{\boldsymbol{\phi}} M_{\Theta} M_{\boldsymbol{\mathcal{A}}} B_{\mathbf{x}} \|\bar{\boldsymbol{\mathcal{B}}\boldsymbol{\mathcal{I}}}_L \bar{\mathrm{W}}_L - \mathbf{W}\|_{\mathcal{L}^\infty} + L_{\boldsymbol{\phi}} M_{\Theta} \|\bar{\boldsymbol{\mathcal{B}}\boldsymbol{\mathcal{I}}}_L \bar{\mathrm{c}}_L - \mathbf{c}\|_{\mathcal{L}^\infty}$.

Note that

$$|\mathrm{x}_L^0 - \mathbf{x}(0)| \leq L_{\boldsymbol{\phi}} \|\bar{\boldsymbol{\mathcal{I}}}_L \mathrm{V}_L - \mathbf{V}\|_C M_{\Theta} (M_{\boldsymbol{\mathcal{A}}} |\mathrm{d}| + 1) + \|\bar{\boldsymbol{\mathcal{I}}}_L \mathrm{b}_L - \mathbf{b}\|_C + M_{\Theta} L_{\boldsymbol{\phi}} (\|\bar{\boldsymbol{\mathcal{I}}}_L \mathrm{a}_L - \mathbf{a}\|_C$$
$$+ M_{\boldsymbol{\mathcal{A}}} |\mathrm{d}| \|\bar{\boldsymbol{\mathcal{I}}}_L \mathrm{U}_L - \mathbf{U}\|_C + M_{\Theta} L_{\boldsymbol{\mathcal{A}}} \|\bar{\boldsymbol{\mathcal{I}}}_L \mathrm{T}_L - \mathbf{T}\|_C) \leq C_L.$$

Therefore, applying discrete Gronwall's inequality [37, Lemma 100] to Eq.(16) and using $C_L \to 0$ as $L \to \infty$, we get

$$|\mathrm{x}_L^l - \mathbf{x}(t_L^l)| \leq C_L \left[L_{\boldsymbol{\phi}} L_{\boldsymbol{\mathcal{A}}} M_{\Theta}^2 \exp(L_{\boldsymbol{\phi}} L_{\boldsymbol{\mathcal{A}}} M_{\Theta}^2) + 1 \right] \to 0, \text{ as } L \to \infty.$$

Hence, $\sup_{t \in [t_L^{l-1}, t_L^l]} |\mathrm{x}_L^l - \mathbf{x}(t)| \leq |\mathrm{x}_L^l - \mathbf{x}(t_L^l)| + \boldsymbol{\omega}_{\mathbf{x}}(\tau) \to 0$ as $L \to \infty$, and then we complete the proof. $\qquad \square$

Corollary 5 *Under the assumptions of Proposition 4, suppose in addition that $\boldsymbol{\Theta} \in \Omega_{\boldsymbol{\Theta}}$ and that the discrete parameter $\Theta_L \in \Omega_{\Theta;L}$ is obtained by the following sampling strategy,*

$$(\mathrm{T}_L^l, \mathrm{U}_L^l, \mathrm{a}_L^l, \mathrm{V}_L^l, \mathrm{b}_L^l) = (\mathbf{T}(t_L^l), \mathbf{U}(t_L^l), \mathbf{a}(t_L^l), \mathbf{V}(t_L^l), \mathbf{b}(t_L^l)), \ 0 \leq l \leq L;$$

$$(\mathrm{W}_L^{l,k}, \mathrm{c}_L^{l,k}) = \left(\frac{1}{\tau^2} \int_{t_L^{l-1}}^{t_L^l} \int_{t_L^{k-1}}^{t_L^k} \mathbf{W}(t,s) dt ds, \frac{1}{\tau^2} \int_{t_L^{l-1}}^{t_L^l} \int_{t_L^{k-1}}^{t_L^k} \mathbf{c}(t,s) dt ds \right), \ 1 \leq k \leq l \leq L.$$

Then there exists a constant $C > 0$ such that

$$|\mathrm{x}^l(\mathrm{d}; \Theta_L) - \mathbf{x}(t_L^l; \mathrm{d}; \boldsymbol{\Theta})| \leq C \tau^{1/3}, \ 1 \leq l \leq L.$$

In particular, $\sup_{1 \leq l \leq L} \sup_{t \in [t_L^{l-1}, t_L^l]} |\mathrm{x}^l(\mathrm{d}; \Theta_L) - \mathbf{x}(t; \mathrm{d}; \boldsymbol{\Theta})|$ converges to 0 at rate $O(\tau^{1/3})$ as $L \to \infty$.

Proof The proof is a small modification of the above proposition and is provided in the supplementary materials. $\qquad \square$

Next, we study the learning problem of the DNL frameworks by using the results derived from the forward systems.

Proposition 6 *(Existence of solutions for learning problems) Under the assumptions of Theorem 1, the minimizers of $(\mathcal{P}_L)$ and $(\mathcal{P})$ exist in $\Omega_{\Theta;L}$ and $\Omega_{\boldsymbol{\Theta}}$, respectively.*

Proof The proof is straightforward by the compactness of the learnable parameters, the weak lower semicontinuity of the loss functions, and the continuous dependency of the DNL framework on learnable parameters in Lemma 2 and Proposition 3. Here, we omit the details. $\qquad \square$

We note that the functionals $\boldsymbol{\mathfrak{L}}_{\mathcal{S};L}$, $L \in \mathbb{N}$, are defined on different function spaces. To study the convergence of optimal solutions of $(\mathcal{P}_L)$ to those of $(\mathcal{P})$ via Γ-convergence (Definition 5 in Appendix), we need to expand their feasible sets to a common space

$\mathcal{C}_{\boldsymbol{\Theta}}$ (defined in (8)). Inspired by [28, 38], we define the discrete-to-continuum extension functional $\tilde{\mathfrak{L}}_{\mathcal{S};L} : \mathcal{C}_{\boldsymbol{\Theta}} \to [0, +\infty]$ as follows

$$\tilde{\mathfrak{L}}_{\mathcal{S};L}(\boldsymbol{\Theta}) = \begin{cases} \mathfrak{L}_{\mathcal{S};L}(\Theta_L), & \text{if } \boldsymbol{\Theta} = \hat{\boldsymbol{\mathcal{I}}}_L \Theta_L, \\ +\infty, & \text{otherwise}, \end{cases} \tag{17}$$

and the functional $\tilde{\mathfrak{L}} : \mathcal{C}_{\boldsymbol{\Theta}} \to [0, +\infty]$ as

$$\tilde{\mathfrak{L}}(\boldsymbol{\Theta}) = \begin{cases} \mathfrak{L}(\boldsymbol{\Theta}), & \text{if } \boldsymbol{\Theta} \in \Omega_{\boldsymbol{\Theta}}, \\ +\infty, & \text{otherwise}. \end{cases} \tag{18}$$

The following lemma shows the relationship between the optimal solution of $(\mathcal{P}_L)$ and the minimizer of $\tilde{\mathfrak{L}}_{\mathcal{S};L}$, as well as the relationship between $(\mathcal{P})$ and $\tilde{\mathfrak{L}}$. This enables us to study the Γ-convergence of $\tilde{\mathfrak{L}}_{\mathcal{S};L}$ to $\tilde{\mathfrak{L}}$ to establish the relationship between the optimal solutions of $(\mathcal{P}_L)$ and $(\mathcal{P})$.

Lemma 7 *Consider the problems $(\mathcal{P}_L), (\mathcal{P})$ defined in (4) and (11), respectively. Let $\tilde{\mathfrak{L}}_{\mathcal{S};L}$ and $\tilde{\mathfrak{L}}$ be given in (17) and (18). Let $\boldsymbol{\Theta}^* = (\mathbf{T}^*, \mathbf{U}^*, \mathbf{a}^*, \mathbf{V}^*, \mathbf{b}^*, \mathbf{W}^*, \mathbf{c}^*) \in \Omega_{\boldsymbol{\Theta}}$, $\Theta_L^* = (\mathrm{T}_L^*, \mathrm{U}_L^*, \mathrm{a}_L^*, \mathrm{V}_L^*, \mathrm{b}_L^*, \mathrm{W}_L^*, \mathrm{c}_L^*) \in \Omega_{\Theta;L}$ for $L \geq 1$. Then*

(i) Θ_L^* is an optimal solution of $(\mathcal{P}_L)$ if and only if $\hat{\boldsymbol{\mathcal{I}}}_L \Theta_L^*$ minimizes $\tilde{\mathfrak{L}}_{\mathcal{S};L}$;
(ii) $\boldsymbol{\Theta}^*$ is an optimal solution of $(\mathcal{P})$ if and only if $\boldsymbol{\Theta}^*$ minimizes $\tilde{\mathfrak{L}}$.

Proof The proof of this lemma is similar to [28, Lemma 3] and we omit it. $\qquad \square$

We then verify the compactness of parameter set $\Omega_{\Theta;L}$ after linear extension operation.

Lemma 8 *Given $\{\Theta_L \in \Omega_{\Theta;L} : L \in \mathbb{N}\}$ such that $\sup_{L \in \mathbb{N}}\{\boldsymbol{\mathcal{R}}_L(\Theta_L)\} < +\infty$, where $\boldsymbol{\mathcal{R}}_L(\Theta_L)$ is defined in (5). Then there exists a subsequence of $\{\hat{\boldsymbol{\mathcal{I}}}_L \Theta_L\}_{L \in \mathbb{N}}$ converging to a $\boldsymbol{\Theta} \in \Omega_{\boldsymbol{\Theta}}$ in $\mathcal{C}_{\boldsymbol{\Theta}}$.*

Proof We only need to prove the cases of $\{\hat{\boldsymbol{\mathcal{I}}}_L \mathrm{U}_L\}_L$ and $\{\hat{\boldsymbol{\mathcal{BI}}}_L \mathbf{flip}(\mathrm{W}_L)\}_L$. Since $\sup_{L \in \mathbb{N}}\{\boldsymbol{\mathcal{R}}_L(\Theta_L)\} < +\infty$, there exists a constant $M < +\infty$ such that, for all $L \geq 1$,

$$\max \Big\{ \boldsymbol{\mathcal{R}}_L^{(1)}((\mathrm{T}_1)_L), \boldsymbol{\mathcal{R}}_L^{(1)}((\mathrm{T}_2)_L), \boldsymbol{\mathcal{R}}_L^{(1)}((\mathrm{T}_3)_L), \boldsymbol{\mathcal{R}}_L^{(1)}(\mathrm{U}_L),$$
$$\boldsymbol{\mathcal{R}}_L^{(2)}(\mathrm{a}_L), \boldsymbol{\mathcal{R}}_L^{(1)}(\mathrm{V}_L), \boldsymbol{\mathcal{R}}_L^{(2)}(\mathrm{b}_L), \boldsymbol{\mathcal{R}}_L^{(3)}(\bar{\mathrm{W}}_L), \boldsymbol{\mathcal{R}}_L^{(4)}(\bar{\mathrm{c}}_L) \Big\} \leq M,$$

where $\bar{\mathrm{W}}_L := \mathbf{flip}(\mathrm{W}_L), \bar{\mathrm{c}}_L := \mathbf{flip}(\mathrm{c}_L)$. By the definition of $\{\hat{\boldsymbol{\mathcal{I}}}_L \mathrm{U}_L\}_L$, Jensen's inequality and $\tau \leq \frac{1}{\tau}$, we have

$$\|\hat{\boldsymbol{\mathcal{I}}}_L \mathrm{U}_L\|^2_{\mathcal{L}^2([0,1];\mathbb{R}^{n\times n})} \leq \sum_{l=1}^{L} \int_{t_L^{l-1}}^{t_L^l} \frac{t_L^l - t}{\tau} \|\mathrm{U}_L^{l-1}\|^2 + \frac{t - t_L^{l-1}}{\tau} \|\mathrm{U}_L^l\|^2 dt$$

$$\leq \tau \sum_{l=1}^{L} \|\mathrm{U}_L^l\|^2 + \tau \|\mathrm{U}_L^1\|^2 + \frac{1}{\tau} \|\mathrm{U}_L^1 - \mathrm{U}_L^0\|^2,$$

$$\|\boldsymbol{\mathcal{D}}_t(\hat{\boldsymbol{\mathcal{I}}}_L \mathrm{U}_L)\|^2_{\mathcal{L}^2([0,1];\mathbb{R}^{n\times n})} = \sum_{l=1}^{L} \int_{t_L^{l-1}}^{t_L^l} \left\| \frac{\mathrm{U}_L^l - \mathrm{U}_L^{l-1}}{\tau} \right\|^2 dt = \frac{1}{\tau} \sum_{l=1}^{L} \|\mathrm{U}_L^l - \mathrm{U}_L^{l-1}\|^2.$$

Adding the above two inequalities together, we obtain $\|\hat{\boldsymbol{\mathcal{I}}}_L \mathrm{U}_L\|^2_{\mathcal{H}^1((0,1);\mathbb{R}^{n\times n})} \leq 2M, \forall L \in \mathbb{N}$. Therefore, by Rellich-Kondrachov theorem [39, Theorem 6.3], there exists a subsequence of $\{\hat{\boldsymbol{\mathcal{I}}}_L \mathrm{U}_L\}_L$ converging to a $\mathbf{U} \in \mathcal{H}^1((0,1);\mathbb{R}^{n\times n})$ in $\mathcal{C}([0,1];\mathbb{R}^{n\times n})$.

As for the $\{\hat{\boldsymbol{\mathcal{BI}}}_L \mathbf{flip}(\mathrm{W}_L)\}_L$. By Jensen's inequality again, we have

$$\|\hat{\boldsymbol{\mathcal{BI}}}_L \bar{\mathrm{W}}_L\|^3_{\mathcal{L}^3([0,1]\times[0,1];\mathbb{R}^{n\times n})}$$

$$= \sum_{l=1}^{L} \sum_{k=1}^{L} \int_{t_L^{l-1}}^{t_L^l} \int_{t_L^{k-1}}^{t_L^k} \|(\hat{\boldsymbol{\mathcal{BI}}}_L \bar{\mathrm{W}}_L)(t,s)\|^3 dtds + \frac{s - t_L^{k-1}}{\tau} \left(\frac{t_L^l - t}{\tau} \bar{\mathrm{W}}_L^{l-1,k} + \frac{t - t_L^{l-1}}{\tau} \bar{\mathrm{W}}_L^{l,k} \right) \|^3 dtds$$

$$\leq \sum_{l=1}^{L} \sum_{k=1}^{L} \int_{t_L^{l-1}}^{t_L^l} \int_{t_L^{k-1}}^{t_L^k} \frac{t_L^k - s}{\tau} \left(\frac{t_L^l - t}{\tau} \|\bar{\mathrm{W}}_L^{l-1,k-1}\|^3 + \frac{t - t_L^{l-1}}{\tau} \|\bar{\mathrm{W}}_L^{l,k-1}\|^3 \right)$$

$$+ \frac{s - t_L^{k-1}}{\tau} \left(\frac{t_L^l - t}{\tau} \|\bar{\mathrm{W}}_L^{l-1,k}\|^3 + \frac{t - t_L^{l-1}}{\tau} \|\bar{\mathrm{W}}_L^{l,k}\|^3 \right) dtds$$

$$= \frac{\tau^2}{4} \sum_{l=1}^{L} \sum_{k=1}^{L} (\|\bar{\mathrm{W}}_L^{l-1,k-1}\|^3 + \|\bar{\mathrm{W}}_L^{l,k-1}\|^3 + \|\bar{\mathrm{W}}_L^{l-1,k}\|^3 + \|\bar{\mathrm{W}}_L^{l,k}\|^3).$$

Similarly, we can estimate the weak derivation of $\hat{\boldsymbol{\mathcal{BI}}}_L \bar{\mathrm{W}}_L$ as follows,

$$\|\boldsymbol{\mathcal{D}}_s(\hat{\boldsymbol{\mathcal{BI}}}_L \bar{\mathrm{W}}_L)\|^3_{\mathcal{L}^3([0,1]\times[0,1];\mathbb{R}^{n\times n})}$$

$$= \tau^{-3} \sum_{l=1}^{L} \sum_{k=1}^{L} \int_{t_L^{l-1}}^{t_L^l} \int_{t_L^{k-1}}^{t_L^k} \left\| \frac{t_L^l - t}{\tau} (\bar{\mathrm{W}}_L^{l-1,k} - \bar{\mathrm{W}}_L^{l-1,k-1}) + \frac{t - t_L^{l-1}}{\tau} (\bar{\mathrm{W}}_L^{l,k} - \bar{\mathrm{W}}_L^{l,k-1}) \right\|^3 dtds$$

$$\leq \frac{1}{2} \tau^{-1} \sum_{l=1}^{L} \sum_{k=1}^{L} (\|\bar{\mathrm{W}}_L^{l-1,k} - \bar{\mathrm{W}}_L^{l-1,k-1}\|^3 + \|\bar{\mathrm{W}}_L^{l,k} - \bar{\mathrm{W}}_L^{l,k-1}\|^3),$$

$$\|\boldsymbol{\mathcal{D}}_t(\hat{\boldsymbol{\mathcal{BI}}}_L \bar{\mathrm{W}}_L)\|^3_{\mathcal{L}^3([0,1]\times[0,1];\mathbb{R}^{n\times n})}$$

$$\leq \frac{1}{2} \tau^{-1} \sum_{l=1}^{L} \sum_{k=1}^{L} (\|\bar{\mathrm{W}}_L^{l,k-1} - \bar{\mathrm{W}}_L^{l-1,k-1}\|^3 + \|\bar{\mathrm{W}}_L^{l,k} - \bar{\mathrm{W}}_L^{l-1,k}\|^3).$$

Adding the above three inequalities together, we have $\|\hat{\boldsymbol{\mathcal{BI}}}_L \bar{\mathrm{W}}_L\|^3_{\mathcal{W}^{1,3}((0,1)\times(0,1);\mathbb{R}^{n\times n})} \leq 4M$ uniformly for $L \in \mathbb{N}$. Then there exists a subsequence of $\{\hat{\boldsymbol{\mathcal{BI}}}_L \mathbf{flip}(\mathrm{W}_L)\}_L$ converging to a $\mathbf{W} \in \mathcal{W}^{1,3}((0,1)\times(0,1);\mathbb{R}^{n\times n})$ in $\mathcal{L}^\infty([0,1]\times[0,1];\mathbb{R}^{n\times n})$ owing to Rellich-Kondrachov theorem again. $\qquad\square$

The next two lemmas will be used in proving the Γ-convergence of $\tilde{\boldsymbol{\mathfrak{L}}}_{\mathcal{S};L}$.

18

Lemma 9 *[22, Proposition 4.8] Let* $\mathbf{f}_L \in \mathcal{L}^2\left([0,1];\mathbb{R}^d\right), \mathbf{f} \in \mathcal{L}^2\left([0,1];\mathbb{R}^d\right)$ *and* $\varepsilon_L \to 0^+$ *as* $L \to \infty$. *Assume that* $\mathbf{f}_L \to \mathbf{f}$ *in* $\mathcal{L}^2\left([0,1];\mathbb{R}^d\right)$. *If*

$$\liminf_{L\to\infty} \frac{1}{\varepsilon_L^2} \int_{\varepsilon_L}^1 |\mathbf{f}_L(t) - \mathbf{f}_L(t-\varepsilon_L)|^2 dt < +\infty,$$

then $\mathbf{f} \in \mathcal{H}^1\left((0,1);\mathbb{R}^d\right)$ *and*

$$\liminf_{L\to\infty} \frac{1}{\varepsilon_L^2} \int_{\varepsilon_L}^1 |\mathbf{f}_L(t) - \mathbf{f}_L(t-\varepsilon_L)|^2 dt \geq \int_0^1 |\boldsymbol{\mathcal{D}}_t(\mathbf{f})(t)|^2 dt.$$

The following lemma generalizes Lemma 9 in a subtle way.

Lemma 10 *Let* $\mathbf{f}_L \in \mathcal{L}^3\left([0,1]\times[0,1];\mathbb{R}^d\right), \mathbf{f} \in \mathcal{L}^3\left([0,1]\times[0,1];\mathbb{R}^d\right)$ *and* $\varepsilon_L \to 0^+$ *as* $L \to \infty$. *Assume that* $\mathbf{f}_L \to \mathbf{f}$ *in* $\mathcal{L}^3\left([0,1]\times[0,1];\mathbb{R}^d\right)$. *If*

$$\liminf_{L\to\infty} \left\{ \frac{1}{\varepsilon_L^3} \int_{\varepsilon_L}^1 \int_0^1 |\mathbf{f}_L(t,s) - \mathbf{f}_L(t-\varepsilon_L,s)|^3 dt ds \right. \tag{19}$$
$$\left. + \frac{1}{\varepsilon_L^3} \int_0^1 \int_{\varepsilon_L}^1 |\mathbf{f}_L(t,s) - \mathbf{f}_L(t,s-\varepsilon_L)|^3 dt ds \right\} < +\infty,$$

then $\mathbf{f} \in \mathcal{W}^{1,3}\left((0,1)\times(0,1);\mathbb{R}^d\right)$ *and*

$$\liminf_{L\to\infty}\left\{ \frac{1}{\varepsilon_L^3} \int_{\varepsilon_L}^1 \int_0^1 |\mathbf{f}_L(t,s) - \mathbf{f}_L(t-\varepsilon_L,s)|^3 dt ds + \frac{1}{\varepsilon_L^3} \int_0^1 \int_{\varepsilon_L}^1 |\mathbf{f}_L(t,s) - \mathbf{f}_L(t,s-\varepsilon_L)|^3 dt ds \right\}$$
$$\geq \int_0^1 \int_0^1 |\boldsymbol{\mathcal{D}}_t(\mathbf{f})(t,s)|^3 + |\boldsymbol{\mathcal{D}}_s(\mathbf{f})(t,s)|^3 dt ds. \tag{20}$$

Proof We first show the following inequalities

$$\int_{\delta'}^{1-\delta'} \int_\delta^{1-\delta} |(J_\delta * \tilde{\mathbf{g}})(t,s) - (J_\delta * \tilde{\mathbf{g}})(t-\varepsilon_L,s)|^3 dt ds \leq \int_{\varepsilon_L}^1 \int_0^1 |\tilde{\mathbf{g}}(t,s) - \tilde{\mathbf{g}}(t-\varepsilon_L,s)|^3 dt ds, \tag{21}$$

$$\int_\delta^{1-\delta} \int_{\delta'}^{1-\delta'} |(J_\delta * \tilde{\mathbf{g}})(t,s) - (J_\delta * \tilde{\mathbf{g}})(t,s-\varepsilon_L)|^3 dt ds \leq \int_0^1 \int_{\varepsilon_L}^1 |\tilde{\mathbf{g}}(t,s) - \tilde{\mathbf{g}}(t,s-\varepsilon_L)|^3 dt ds, \tag{22}$$

for any $\tilde{\mathbf{g}} \in \mathcal{L}^3\left([0,1]\times[0,1];\mathbb{R}^d\right)$ and any $\delta, \delta' > 0$ that satisfy $\varepsilon_L + \delta < \delta'$, where J_δ is a standard 2D mollifier [39]; and

$$\int_{2\delta'}^{1-2\delta'} \int_{2\delta'}^{1-2\delta'} |(\boldsymbol{\mathcal{D}}_t \mathbf{g})(t,s)|^3 dt ds \leq \liminf_{L\to\infty} \frac{1}{\varepsilon_L^3} \int_{2\delta'}^{1-2\delta'} \int_{2\delta'}^{1-2\delta'} |\mathbf{g}_L(t,s) - \mathbf{g}_L(t-\varepsilon_L,s)|^3 dt ds \tag{23}$$

$$\int_{2\delta'}^{1-2\delta'} \int_{2\delta'}^{1-2\delta'} |\boldsymbol{\mathcal{D}}_s(\mathbf{g})(t,s)|^3 dt ds \leq \liminf_{L\to\infty} \frac{1}{\varepsilon_L^3} \int_{2\delta'}^{1-2\delta'} \int_{2\delta'}^{1-2\delta'} |\mathbf{g}_L(t,s) - \mathbf{g}_L(t,s-\varepsilon_L)|^3 dt ds, \tag{24}$$

for any $\mathbf{g}, \mathbf{g}_L \in C^\infty\left([\delta', 1-\delta'] \times [\delta', 1-\delta'] ; \mathbb{R}^d\right)$ with $\boldsymbol{\mathcal{D}}_t(\mathbf{g}_L) \to \boldsymbol{\mathcal{D}}_t(\mathbf{g})$, $\boldsymbol{\mathcal{D}}_s(\mathbf{g}_L) \to \boldsymbol{\mathcal{D}}_s(\mathbf{g})$ in $\mathcal{L}^\infty\left([\delta', 1-\delta'] \times [\delta', 1-\delta'] ; \mathbb{R}^d\right)$ and $\sup_L \|\boldsymbol{\mathcal{D}}_{tt}(\mathbf{g}_L)\|_{\mathcal{L}^\infty} < \infty$, $\sup_L \|\boldsymbol{\mathcal{D}}_{ss}(\mathbf{g}_L)\|_{\mathcal{L}^\infty} < \infty$.

We only need to prove (21) and (23). Note that $\varepsilon_L + \delta < \delta'$. To show (21), we have, by Minkowski's inequality for integrals [39, Theorem 2.9] and the property of mollifier, that

$$
\left(\int_{\delta'}^{1-\delta'} \int_{\delta}^{1-\delta} |(J_\delta * \tilde{\mathbf{g}})(t,s) - (J_\delta * \tilde{\mathbf{g}})(t - \varepsilon_L, s)|^3 dt ds\right)^{1/3}
$$

$$
\leq \left(\int_{\delta'}^{1-\delta'} \int_{\delta}^{1-\delta} \left| \iint_{B(0,\delta)} J_\delta(u,v) [\tilde{\mathbf{g}}(t-u, s-v) - \tilde{\mathbf{g}}(t - \varepsilon_L - u, s - v)] du dv\right|^3 dt ds\right)^{1/3}
$$

$$
\leq \iint_{B(0,\delta)} J_\delta(u,v) \left(\int_{\delta'}^{1-\delta'} \int_{\delta}^{1-\delta} |\tilde{\mathbf{g}}(t-u, s-v) - \tilde{\mathbf{g}}(t - \varepsilon_L - u, s - v)|^3 dt ds\right)^{1/3} du dv
$$

$$
\leq \iint_{B(0,\delta)} J_\delta(u,v) \left(\int_{\varepsilon_L}^{1} \int_{0}^{1} |\tilde{\mathbf{g}}(t,s) - \tilde{\mathbf{g}}(t - \varepsilon_L, s)|^3 dt ds\right)^{1/3} du dv
$$

$$
= \left(\int_{\varepsilon_L}^{1} \int_{0}^{1} |\tilde{\mathbf{g}}(t,s) - \tilde{\mathbf{g}}(t - \varepsilon_L, s)|^3 dt ds\right)^{1/3}.
$$

For inequality (23), the Taylor's theorem and the smoothness of $\mathbf{g}_L$ give

$$
\mathbf{g}_L(t,s) - \mathbf{g}_L(t - \varepsilon_L, s) = \varepsilon_L \boldsymbol{\mathcal{D}}_t(\mathbf{g}_L)(t,s) - \varepsilon_L^2 \boldsymbol{\mathcal{D}}_{tt}(\mathbf{g}_L)(r,s) \text{ for some } r \in [t - \varepsilon_L, t].
$$

Therefore, for $t \in [2\delta', 1 - 2\delta']$, $s \in [0,1]$ and $\varepsilon_L < \delta'$, we have

$$
\frac{|\mathbf{g}_L(t,s) - \mathbf{g}_L(t - \varepsilon_L, s)|}{\varepsilon_L} \geq |\boldsymbol{\mathcal{D}}_t(\mathbf{g}_L)(t,s)| - \varepsilon_L \|\boldsymbol{\mathcal{D}}_{tt}(\mathbf{g}_L)\|_{\mathcal{L}^\infty},
$$

due to the triangle inequality. For any $\eta > 0$, there exists $C_\eta > 0$ such that $|a + b|^3 \leq (1 + \eta)|a|^3 + C_\eta |b|^3$ for any $a, b \in \mathbb{R}$ by Young's inequality. Hence

$$
|\boldsymbol{\mathcal{D}}_t(\mathbf{g}_L)(t,s)|^3 \leq \left(\frac{|\mathbf{g}_L(t,s) - \mathbf{g}_L(t - \varepsilon_L, s)|}{\varepsilon_L} + \varepsilon_L \|\boldsymbol{\mathcal{D}}_{tt}(\mathbf{g}_L)\|_{\mathcal{L}^\infty}\right)^3
$$

$$
\leq (1 + \eta)\left|\frac{\mathbf{g}_L(t,s) - \mathbf{g}_L(t - \varepsilon_L, s)}{\varepsilon_L}\right|^3 + C_\eta \varepsilon_L^3 \|\boldsymbol{\mathcal{D}}_{tt}(\mathbf{g}_L)\|_{\mathcal{L}^\infty}^3.
$$

Due to $\sup_{L \in \mathbb{N}} \|\boldsymbol{\mathcal{D}}_{tt}(\mathbf{g}_L)\|_{\mathcal{L}^\infty} < \infty$ and Lebesgue's dominated convergence theorem, we see

$$
\int_{2\delta'}^{1-2\delta'} \int_{2\delta'}^{1-2\delta'} |\boldsymbol{\mathcal{D}}_t(\mathbf{g})(t,s)|^3 dt ds = \lim_{L \to \infty} \int_{2\delta'}^{1-2\delta'} \int_{2\delta'}^{1-2\delta'} |\boldsymbol{\mathcal{D}}_t(\mathbf{g}_L)(t,s)|^3 dt ds
$$

$$
\leq (1 + \eta) \liminf_{L \to \infty} \int_{2\delta'}^{1-2\delta'} \int_{2\delta'}^{1-2\delta'} \left|\frac{\mathbf{g}_L(t,s) - \mathbf{g}_L(t - \varepsilon_L, s)}{\varepsilon_L}\right|^3 dt ds.
$$

Taking $\eta \to 0$ yields (23).

We next use (21), (23), (22), and (24) to prove the existence of weak derivatives of function $\mathbf{f}$. By (21) and (19), there exists a constant M such that

$$
\liminf_{L \to \infty} \frac{1}{\varepsilon_L^3} \int_{\delta'}^{1-\delta'} \int_{\delta}^{1-\delta} |(J_\delta * \mathbf{f}_L)(t,s) - (J_\delta * \mathbf{f}_L)(t - \varepsilon_L, s)|^3 dt ds
$$

$$
\leq \liminf_{L \to \infty} \frac{1}{\varepsilon_L^3} \int_{\varepsilon_L}^{1} \int_{0}^{1} |\mathbf{f}_L(t,s) - \mathbf{f}_L(t - \varepsilon_L, s)|^3 dt ds \leq M.
$$

(25)

Furthermore, by the property of mollifiers and Hölder inequality, we have, for $(t,s) \in [\delta', 1 - \delta'] \times [\delta', 1 - \delta']$, that

$$|\boldsymbol{\mathcal{D}}_t(J_\delta * \mathbf{f}_L)(t,s) - \boldsymbol{\mathcal{D}}_t(J_\delta * \mathbf{f})(t,s)| = \left| \int_0^1 \int_0^1 \boldsymbol{\mathcal{D}}_t(J_\delta)(t-u, s-v)(\mathbf{f}_L(u,v) - \mathbf{f}(u,v))dudv \right|$$

$$\leq \|\boldsymbol{\mathcal{D}}_t(J_\delta)\|_{\mathcal{L}^{3/2}(\mathbb{R}^2)} \|\mathbf{f}_L - \mathbf{f}\|_{\mathcal{L}^3([0,1] \times [0,1])},$$

$$|\boldsymbol{\mathcal{D}}_{tt}(J_\delta * \mathbf{f}_L)| \leq \|\boldsymbol{\mathcal{D}}_{tt}(J_\delta)\|_{\mathcal{L}^{3/2}(\mathbb{R}^2)} \|\mathbf{f}_L\|_{\mathcal{L}^3([0,1] \times [0,1])}.$$

We can hence apply (23) with $\mathbf{g} = J_\delta * \mathbf{f}$ and $\mathbf{g}_L = J_\delta * \mathbf{f}_L$ and use (25) to get

$$\int_{2\delta'}^{1-2\delta'} \int_{2\delta'}^{1-2\delta'} |\boldsymbol{\mathcal{D}}_t(J_\delta * \mathbf{f})(t,s)|^3 dtds \leq M.$$

Therefore, there exists an $\mathbf{h} \in \mathcal{L}^3([2\delta', 1 - 2\delta'] \times [2\delta', 1 - 2\delta']; \mathbb{R}^d)$ and a subsequence of $\{\boldsymbol{\mathcal{D}}_t(J_\delta * \mathbf{f})\}_\delta$ (not relabeled) such that

$$\boldsymbol{\mathcal{D}}_t(J_\delta * \mathbf{f}) \rightharpoonup \mathbf{h} \text{ in } \mathcal{L}^3([2\delta', 1 - 2\delta'] \times [2\delta', 1 - 2\delta']; \mathbb{R}^d), \text{ as } \delta \to 0^+$$

by the reflexivity of $\mathcal{L}^3$. Hence, for any differentiable function $\boldsymbol{\psi}$ with compact support in $[2\delta', 1 - 2\delta'] \times [2\delta', 1 - 2\delta']$, we have

$$\int_{2\delta'}^{1-2\delta'} \int_{2\delta'}^{1-2\delta'} \boldsymbol{\psi}\mathbf{h} \leftarrow \int_{2\delta'}^{1-2\delta'} \int_{2\delta'}^{1-2\delta'} \boldsymbol{\psi}\boldsymbol{\mathcal{D}}_t(J_\delta * \mathbf{f}) = -\int_{2\delta'}^{1-2\delta'} \int_{2\delta'}^{1-2\delta'} \boldsymbol{\mathcal{D}}_t(\boldsymbol{\psi}) \cdot J_\delta * \mathbf{f}$$

$$\to -\int_{2\delta'}^{1-2\delta'} \int_{2\delta'}^{1-2\delta'} \boldsymbol{\mathcal{D}}_t(\boldsymbol{\psi}) \cdot \mathbf{f} = \int_{2\delta'}^{1-2\delta'} \int_{2\delta'}^{1-2\delta'} \boldsymbol{\psi}\boldsymbol{\mathcal{D}}_t(\mathbf{f}),$$

where we use the property of mollifiers in the second line. This shows $\boldsymbol{\mathcal{D}}_t(\mathbf{f}) = \mathbf{h}$ and in particular $\boldsymbol{\mathcal{D}}_t(\mathbf{f}) \in \mathcal{L}^3([2\delta', 1 - 2\delta'] \times [2\delta', 1 - 2\delta'])$. A similar discussion as above gives the existence of $\boldsymbol{\mathcal{D}}_s(\mathbf{f}) \in \mathcal{L}^3([2\delta', 1 - 2\delta'] \times [2\delta', 1 - 2\delta'])$.

Now, we prove (20). Applying (23) with $\mathbf{g} = J_\delta * \mathbf{f}$ and $\mathbf{g}_L = J_\delta * \mathbf{f}_L$, and by Eq.(21), we have

$$\int_{2\delta'}^{1-2\delta'} \int_{2\delta'}^{1-2\delta'} |\boldsymbol{\mathcal{D}}_t(J_\delta * \mathbf{f})|^3 dtds$$

$$\leq \liminf_{L \to \infty} \int_{2\delta'}^{1-2\delta'} \int_{2\delta'}^{1-2\delta'} \frac{|(J_\delta * \mathbf{f}_L)(t,s) - (J_\delta * \mathbf{f}_L)(t - \varepsilon_L, s)|^3}{\varepsilon_L^3} dtds$$

$$\leq \liminf_{L \to \infty} \int_{\varepsilon_L}^1 \int_0^1 \frac{|\mathbf{f}_L(t,s) - \mathbf{f}_L(t - \varepsilon_L, s)|^3}{\varepsilon_L^3} dtds.$$

Similarly, we can get

$$\int_{2\delta'}^{1-2\delta'} \int_{2\delta'}^{1-2\delta'} |\boldsymbol{\mathcal{D}}_s(J_\delta * \mathbf{f})|^3 dtds$$

$$\leq \liminf_{L \to \infty} \int_{2\delta'}^{1-2\delta'} \int_{2\delta'}^{1-2\delta'} \frac{|(J_\delta * \mathbf{f}_L)(t,s) - (J_\delta * \mathbf{f}_L)(t, s - \varepsilon_L)|^3}{\varepsilon_L^3} dtds$$

$$\leq \liminf_{L \to \infty} \int_0^1 \int_{\varepsilon_L}^1 \frac{|\mathbf{f}_L(t,s) - \mathbf{f}_L(t, s - \varepsilon_L)|^3}{\varepsilon_L^3} dtds.$$

By the property of mollifiers, $\boldsymbol{\mathcal{D}}_t(J_\delta * \mathbf{f})$ converges strongly to $\boldsymbol{\mathcal{D}}_t(\mathbf{f})$ in $\mathcal{L}^3([2\delta', 1 - 2\delta'] \times [2\delta', 1 - 2\delta']; \mathbb{R}^d)$, $\boldsymbol{\mathcal{D}}_s(J_\delta * \mathbf{f})$ converges strongly to $\boldsymbol{\mathcal{D}}_s(\mathbf{f})$ in $\mathcal{L}^3([2\delta', 1 - 2\delta'] \times [2\delta', 1 - 2\delta']; \mathbb{R}^d)$ as $\delta \to 0+$. Hence, for any $\delta' > 0$, we have

$$\int_{2\delta'}^{1-2\delta'} \int_{2\delta'}^{1-2\delta'} |\boldsymbol{\mathcal{D}}_t(\mathbf{f})|^3 dtds \leq \liminf_{L \to \infty} \int_{\varepsilon_L}^1 \int_0^1 \frac{|\mathbf{f}_L(t,s) - \mathbf{f}_L(t - \varepsilon_L, s)|^3}{\varepsilon_L^3} dtds,$$

$$\int_{2\delta'}^{1-2\delta'} \int_{2\delta'}^{1-2\delta'} |\boldsymbol{\mathcal{D}}_s(\mathbf{f})|^3 dtds \leq \liminf_{L \to \infty} \int_0^1 \int_{\varepsilon_L}^1 \frac{|\mathbf{f}_L(t,s) - \mathbf{f}_L(t, s - \varepsilon_L)|^3}{\varepsilon_L^3} dtds,$$

21

since the additional constraint imposed by $\delta' > \delta + \varepsilon_L$ vanishes when taking $\delta \to 0$ and $\varepsilon_L \to 0$. Adding the above two inequalities together and taking $\delta' \to 0^+$, we complete the proof. $\qquad\square$

Based on these preceding results, we can verify the $\liminf$ condition of Γ-convergence for $\{\tilde{\mathfrak{L}}_{\mathcal{S};L}\}_L$.

Lemma 11 *(Liminf condition) Let the assumptions in Theorem 1 hold. Then, for every learnable parameter function $\boldsymbol{\Theta} \in \mathcal{C}_{\boldsymbol{\Theta}}$ and every sequence $\{\boldsymbol{\Theta}_L\}_{L \in \mathbb{N}}$ converging to $\boldsymbol{\Theta}$ in $\mathcal{C}_{\boldsymbol{\Theta}}$, we have*

$$\liminf_{L \to \infty} \tilde{\mathfrak{L}}_{\mathcal{S};L}(\boldsymbol{\Theta}_L) \geq \tilde{\mathfrak{L}}(\boldsymbol{\Theta}),$$

where $\tilde{\mathfrak{L}}_{\mathcal{S};L}$ and $\tilde{\mathfrak{L}}$ are given in (17) and (18), respectively.

Proof If $\tilde{\mathfrak{L}}(\boldsymbol{\Theta}) = +\infty$, i.e., $\boldsymbol{\Theta} \notin \Omega_{\boldsymbol{\Theta}}$, we show $\liminf_{L \to \infty} \tilde{\mathfrak{L}}_{\mathcal{S};L}(\boldsymbol{\Theta}_L) = +\infty$ by contradiction. If $\liminf_{L \to \infty} \tilde{\mathfrak{L}}_{\mathcal{S};L}(\boldsymbol{\Theta}_L) < +\infty$, then there exists a subsequence $\{\boldsymbol{\Theta}_{L_k}\}_{L_k} \subset \{\boldsymbol{\Theta}_L\}_L$ such that $\lim_{L_k \to \infty} \tilde{\mathfrak{L}}_{\mathcal{S};L}(\boldsymbol{\Theta}_{L_k}) < +\infty$. Without loss of generality, we can assume $\boldsymbol{\Theta}_{L_k} = \hat{\boldsymbol{\mathcal{I}}}_{L_k} \Theta_{L_k}$ for some $\Theta_{L_k} \in \Omega_{\Theta;L}$ by the definition of $\tilde{\mathfrak{L}}_{\mathcal{S};L}$ in (17). Then, there exists a subsequence $\{\boldsymbol{\Theta}_{L_{k_r}}\}_{L_{k_r}}$ converging to some $\hat{\boldsymbol{\Theta}} \in \Omega_{\boldsymbol{\Theta}}$ in $\mathcal{C}_{\boldsymbol{\Theta}}$ due to Lemma 8. This contradicts $\boldsymbol{\Theta}_{L_k} \to \boldsymbol{\Theta} \notin \Omega_{\boldsymbol{\Theta}}$.

If $\tilde{\mathfrak{L}}(\boldsymbol{\Theta}) < +\infty$, we only need to consider the case when $\liminf_{L \to \infty} \tilde{\mathfrak{L}}_{\mathcal{S};L}(\boldsymbol{\Theta}_L) < +\infty$. Assume $\boldsymbol{\Theta}_L = \hat{\boldsymbol{\mathcal{I}}}_L \Theta_L$ for some $\Theta_L \in \Omega_{\Theta;L}$ by the definition of $\tilde{\mathfrak{L}}_{\mathcal{S};L}$. Due to the continuity of $\boldsymbol{\ell}$ (assumption (A_3)) and Proposition 4, we next prove

$$\frac{1}{M} \sum_{m=1}^{M} \boldsymbol{\ell}(\mathrm{x}^L(\mathrm{d}_m; \Theta_L), \mathrm{g}_m) \to \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{\ell}(\mathbf{x}(1; \mathrm{d}_m; \boldsymbol{\Theta}), \mathrm{g}_m) \text{ as } L \to \infty, \qquad (26)$$

by verifying the condition (14). Note that we only need to show

$$\bar{\boldsymbol{\mathcal{I}}}_L \mathrm{U}_L \to \mathbf{U} \text{ in } \mathcal{C}([0,1]; \mathbb{R}^{n \times n}); \ \boldsymbol{\mathcal{B}\bar{\mathcal{I}}}_L(\mathbf{flip}(\mathrm{W}_L)) \to \mathbf{W} \text{ in } \mathcal{L}^\infty([0,1] \times [0,1]; \mathbb{R}^{n \times n}). \qquad (27)$$

For any $\epsilon > 0$, there exists an $\check{L} \in \mathbb{N}$ such that $\|\hat{\boldsymbol{\mathcal{I}}}_L \mathrm{U}_L - \mathbf{U}\|_C < \epsilon/2$, $\forall L > \check{L}$, because $\{\boldsymbol{\Theta}_L\}_{L \in \mathbb{N}}$ converges to $\boldsymbol{\Theta}$ in $\mathcal{C}_{\boldsymbol{\Theta}}$. Besides, there exists an $\tilde{L} \in \mathbb{N}$ such that $\max_{|t_1 - t_2| \leq 1/L} \|\mathbf{U}(t_1) - \mathbf{U}(t_2)\| \leq \epsilon/2$ for $L \geq \tilde{L}$ due to the uniform continuity of $\mathbf{U}$ in $[0,1]$. Consequently, for all $L \geq \max\{\check{L}, \tilde{L}\}$, we have

$$\|(\bar{\boldsymbol{\mathcal{I}}}_L \mathrm{U}_L)(0) - \mathbf{U}(0)\| = \|(\hat{\boldsymbol{\mathcal{I}}}_L) \mathrm{U}_L(0) - \mathbf{U}(0)\| < \epsilon,$$

$$\|(\bar{\boldsymbol{\mathcal{I}}}_L \mathrm{U}_L)(t) - \mathbf{U}(t)\| \leq \|(\hat{\boldsymbol{\mathcal{I}}}_L \mathrm{U}_L)(t_L^l) - \mathbf{U}(t_L^l)\| + \|\mathbf{U}(t_L^l) - \mathbf{U}(t)\| < \epsilon, \ \ t \in (t_L^{l-1}, t_L^l],$$

where $1 \leq l \leq L$. Denote $\bar{\mathrm{W}}_L = \mathbf{flip}(\mathrm{W}_L)$. Since $\liminf_{L \to \infty} \tilde{\mathfrak{L}}_{\mathcal{S};L}(\boldsymbol{\Theta}_L) < +\infty$, there exists a constant M_d such that for all $L \in \mathbb{N}$,

$$\tau^{-1} \sum_{l=1}^{L} \|\mathrm{U}_L^l - \mathrm{U}_L^{l-1}\|^2 \leq M_d,$$

$$\tau^{-1} \Big(\sum_{l=1}^{L} \sum_{k=1}^{L} \|\bar{\mathrm{W}}_L^{l,k} - \bar{\mathrm{W}}_L^{l-1,k}\|^3 + \sum_{l=1}^{L} \sum_{k=1}^{L} \|\bar{\mathrm{W}}_L^{l,k} - \bar{\mathrm{W}}_L^{l,k-1}\|^3 \Big) \leq M_d, \qquad (28)$$

22

due to $\liminf_{L\to\infty} \tilde{\mathfrak{L}}_{\mathcal{S};L}(\mathbf{\Theta}_L) < +\infty$. This yields, for any $1 \leq k, l \leq L$, that

$$\sup_{L\in\mathbb{N}} \sup_{1\leq k,l\leq L} \max\{\|\bar{W}_L^{l,k} - \bar{W}_L^{l-1,k}\|, \|\bar{W}_L^{l,k} - \bar{W}_L^{l,k-1}\|\} \leq (M_d\tau)^{1/3}.$$

Hence, we derive

$$\|(\bar{\boldsymbol{\mathcal{B}}\boldsymbol{\mathcal{I}}}_L\bar{W}_L)(t,s) - \mathbf{W}(t,s)\|$$

$$\leq \|(\hat{\boldsymbol{\mathcal{B}}\boldsymbol{\mathcal{I}}}_L\bar{W}_L)(t_L^l,t_L^k) - (\hat{\boldsymbol{\mathcal{B}}\boldsymbol{\mathcal{I}}}_L\bar{W}_L)(t,s)\| + \|(\hat{\boldsymbol{\mathcal{B}}\boldsymbol{\mathcal{I}}}_L\bar{W}_L)(t,s) - \mathbf{W}(t,s)\|$$

$$\leq \left\|\bar{W}_L^{l,k} - \bar{W}_L^{l,k-1}\right\| + \left\|\bar{W}_L^{l,k-1} - \bar{W}_L^{l-1,k-1}\right\| + \left\|\bar{W}_L^{l,k} - \bar{W}_L^{l,k-1}\right\| + \left\|\bar{W}_L^{l,k} - \bar{W}_L^{l-1,k}\right\|$$

$$+ \|(\hat{\boldsymbol{\mathcal{B}}\boldsymbol{\mathcal{I}}}_L\bar{W}_L)(t,s) - \mathbf{W}(t,s)\|$$

$$\leq 4(M_d)^{1/3}\tau^{1/3} + \|(\hat{\boldsymbol{\mathcal{B}}\boldsymbol{\mathcal{I}}}_L\bar{W}_L)(t,s) - \mathbf{W}(t,s)\|,$$

where $(t,s) \in (t_L^{l-1}, t_L^l] \times (t_L^{k-1}, t_L^k]$ $(1 \leq l, k \leq L)$ is the Lebesgue points of $\mathbf{W}$, and we use $(\bar{\boldsymbol{\mathcal{B}}\boldsymbol{\mathcal{I}}}_L\bar{W}_L)(t,s) = (\hat{\boldsymbol{\mathcal{B}}\boldsymbol{\mathcal{I}}}_L\bar{W}_L)(t_L^l,t_L^k)$ for $(t,s) \in (t_L^{l-1}, t_L^l] \times (t_L^{k-1}, t_L^k]$ in the first inequality. Therefore, $\bar{\boldsymbol{\mathcal{B}}\boldsymbol{\mathcal{I}}}_L\bar{W}_L \to \mathbf{W}$ in $\mathcal{L}^\infty([0,1]\times[0,1];\mathbb{R}^{n\times n})$ due to $\hat{\boldsymbol{\mathcal{B}}\boldsymbol{\mathcal{I}}}_L\bar{W}_L \to \mathbf{W}$ in $\mathcal{L}^\infty([0,1]\times[0,1];\mathbb{R}^{n\times n})$.

To prove $\liminf_{L\to\infty} \tilde{\mathfrak{L}}_{\mathcal{S};L}(\mathbf{\Theta}_L) \geq \tilde{\mathfrak{L}}(\mathbf{\Theta})$, we now only need to show $\liminf_{L\to\infty}\boldsymbol{\mathcal{R}}_L(\Theta_L) \geq \boldsymbol{\mathcal{R}}(\mathbf{\Theta})$ since (26) holds. To achieve this goal, we show the following inequalities:

(i) $\liminf_{L\to\infty}\boldsymbol{\mathcal{R}}_L^{(1)}((\mathrm{T}_j)_L) \geq \|(\mathbf{T})_j\|_{\mathcal{H}^1((0,1);\mathbb{R}^{n\times n})}^2$, $j = 1, 2, 3$;

(ii) $\liminf_{L\to\infty}\boldsymbol{\mathcal{R}}_L^{(1)}(\mathrm{U}_L) \geq \|\mathbf{U}\|_{\mathcal{H}^1((0,1);\mathbb{R}^{n\times n})}^2$, $\liminf_{L\to\infty}\boldsymbol{\mathcal{R}}_L^{(2)}(\mathrm{a}_L) \geq \|\mathbf{a}\|_{\mathcal{H}^1((0,1);\mathbb{R}^n)}^2$;

(iii) $\liminf_{L\to\infty}\boldsymbol{\mathcal{R}}_L^{(1)}(\mathrm{V}_L) \geq \|\mathbf{V}\|_{\mathcal{H}^1((0,1);\mathbb{R}^{n\times n})}^2$, $\liminf_{L\to\infty}\boldsymbol{\mathcal{R}}_L^{(2)}(\mathrm{b}_L) \geq \|\mathbf{b}\|_{\mathcal{H}^1((0,1);\mathbb{R}^n)}^2$;

(iv) $\liminf_{L\to\infty}\boldsymbol{\mathcal{R}}_L^{(3)}(\mathbf{flip}(\mathrm{W}_L)) \geq \|\mathbf{W}\|_{\mathcal{W}^{1,3}((0,1)\times(0,1);\mathbb{R}^{n\times n})}^3$,

$\liminf_{L\to\infty}\boldsymbol{\mathcal{R}}_L^{(4)}(\mathbf{flip}(\mathrm{c}_L)) \geq \|\mathbf{c}\|_{\mathcal{W}^{1,3}((0,1)\times(0,1);\mathbb{R}^n)}^3$.

Note that we only need to show the $\liminf_{L\to\infty}\boldsymbol{\mathcal{R}}_L^{(1)}(\mathrm{U}_L) \geq \|\mathbf{U}\|_{\mathcal{H}^1((0,1);\mathbb{R}^{n\times n})}^2$ and $\liminf_{L\to\infty}\boldsymbol{\mathcal{R}}_L^{(3)}(\mathbf{flip}(\mathrm{W}_L)) \geq \|\mathbf{W}\|_{\mathcal{W}^{1,3}((0,1)\times(0,1);\mathbb{R}^{n\times n})}^3$, as the other cases are similar. By Definition 3, the convergence of $\{\bar{\boldsymbol{\mathcal{I}}}_L\mathrm{U}_L\}_L$ in Eq.(27), Eq.(28) and Lemma 9, we have

$$\liminf_{L\to\infty}\boldsymbol{\mathcal{R}}_L^{(1)}(\mathrm{U}_L)$$

$$\geq \liminf_{L\to\infty}\tau\sum_{l=1}^L\|\mathrm{U}_L^l\|^2 + \liminf_{L\to\infty}\frac{1}{\tau}\sum_{l=1}^L\|\mathrm{U}_L^l - \mathrm{U}_L^{l-1}\|^2$$

$$\geq \liminf_{L\to\infty}\int_0^1\|(\bar{\boldsymbol{\mathcal{I}}}_L\mathrm{U}_L)(t)\|^2 dt + \liminf_{L\to\infty}\frac{1}{\tau^2}\sum_{l=2}^L\int_{t_L^{l-1}}^{t_L^l}\|(\bar{\boldsymbol{\mathcal{I}}}_L\mathrm{U}_L)(t) - (\bar{\boldsymbol{\mathcal{I}}}_L\mathrm{U}_L)(t-\tau)\|^2 dt$$

$$\geq \int_0^1\|\mathbf{U}(t)\|^2 dt + \int_0^1\|\boldsymbol{\mathcal{D}}_t(\mathbf{U})(t)\|^2 dt = \|\mathbf{U}\|_{\mathcal{H}^1((0,1);\mathbb{R}^{n\times n})}^2.$$

23

Similarly, we can estimate $\liminf_{L\to\infty} \mathcal{R}_L^{(3)}(\bar{\mathrm{W}}_L)$ as

$$\liminf_{L\to\infty} \mathcal{R}_L^{(5)}(\bar{\mathrm{W}}_L)$$

$$\geq \liminf_{L\to\infty} \tau^2 \sum_{l=1}^{L}\sum_{k=1}^{L} \|\bar{\mathrm{W}}_L^{l,k}\|^3$$

$$+\liminf_{L\to\infty} \tau^{-1}\Big(\sum_{l=2}^{L}\sum_{k=1}^{L} \|\bar{\mathrm{W}}_L^{l,k}-\bar{\mathrm{W}}_L^{l-1,k}\|^3 + \sum_{l=1}^{L}\sum_{k=2}^{L} \|\bar{\mathrm{W}}_L^{l,k}-\bar{\mathrm{W}}_L^{l,k-1}\|^3 \Big)$$

$$=\liminf_{L\to\infty} \int_0^1\int_0^1 \|(\bar{\mathcal{B}\mathcal{I}}_L\bar{\mathrm{W}}_L)(t,s)\|^3 dtds$$

$$+\liminf_{L\to\infty}\Big\{ L^3 \int_{1/L}^1\int_0^1 \|(\bar{\mathcal{B}\mathcal{I}}_L\bar{\mathrm{W}}_L)(t,s)-(\bar{\mathcal{B}\mathcal{I}}_L\bar{\mathrm{W}}_L)\,(t-1/L,s)\,\|^3 dtds$$

$$+L^3 \int_0^1\int_{1/L}^1 \big\|(\bar{\mathcal{B}\mathcal{I}}_L\bar{\mathrm{W}}_L)(t,s)-(\bar{\mathcal{B}\mathcal{I}}_L\bar{\mathrm{W}}_L)\,(t,s-1/L)\big\|^3 dtds\Big\}$$

$$\geq \int_0^1\int_0^1 \|\mathbf{W}(t,s)\|^3 dtds + \int_0^1\int_0^1 \Big(\|\mathcal{D}_t(\mathbf{W}(t,s))\|^3 + \|\mathcal{D}_s(\mathbf{W}(t,s))\|^3 \Big) dtds$$

$$=\|\mathbf{W}\|^3_{\mathcal{W}^{1,3}((0,1)\times(0,1);\mathbb{R}^{n\times n})},$$

where we use the convergence of $\bar{\mathcal{B}\mathcal{I}}_L\bar{\mathrm{W}}_L$ in (27), Eq.(28) and Lemma 10 in the last inequality. This completes the proof. $\qquad\square$

Next, we verify the $\limsup$ condition of Γ-convergence for $\tilde{\mathfrak{L}}_{\mathcal{S};L}$. For any $\boldsymbol{\Theta}\in\Omega_{\boldsymbol{\Theta}}$ and $L\in\mathbb{N}$, we define the learnable parameter $\Theta_L\in\Omega_{\Theta;L}$ by

$$(\mathrm{T}_j)_L^0 = \frac{1}{\tau}\int_{t_L^0}^{t_L^1}(\mathbf{T}_j)(t)dt, (\mathrm{T}_j)_L^l = \frac{1}{2\tau}\int_{t_L^{l-1}}^{t_L^{l+1}}(\mathbf{T}_j)(t)dt, 1\leq l\leq L-1, (\mathrm{T}_j)_L^L = \frac{1}{\tau}\int_{t_L^{L-1}}^{t_L^L}(\mathbf{T}_j)(t)dt,$$

$$\mathrm{U}_L^0 = \frac{1}{\tau}\int_{t_L^0}^{t_L^1}\mathbf{U}(t)dt, \mathrm{U}_L^l = \frac{1}{2\tau}\int_{t_L^{l-1}}^{t_L^{l+1}}\mathbf{U}(t)dt, 1\leq l\leq L-1, \mathrm{U}_L^L = \frac{1}{\tau}\int_{t_L^{L-1}}^{t_L^L}\mathbf{U}(t)dt,$$

$$\mathrm{a}_L^0 = \frac{1}{\tau}\int_{t_L^0}^{t_L^1}\mathbf{a}(t)dt, \mathrm{a}_L^l = \frac{1}{2\tau}\int_{t_L^{l-1}}^{t_L^{l+1}}\mathbf{a}(t)dt, 1\leq l\leq L-1, \mathrm{a}_L^L = \frac{1}{\tau}\int_{t_L^{L-1}}^{t_L^L}\mathbf{a}(t)dt,$$

$$\mathrm{V}_L^0 = \frac{1}{\tau}\int_{t_L^0}^{t_L^1}\mathbf{V}(t)dt, \mathrm{V}_L^l = \frac{1}{2\tau}\int_{t_L^{l-1}}^{t_L^{l+1}}\mathbf{V}(t)dt, 1\leq l\leq L-1, \mathrm{V}_L^L = \frac{1}{\tau}\int_{t_L^{L-1}}^{t_L^L}\mathbf{V}(t)dt,$$

$$\mathrm{b}_L^0 = \frac{1}{\tau}\int_{t_L^0}^{t_L^1}\mathbf{b}(t)dt, \mathrm{b}_L^l = \frac{1}{2\tau}\int_{t_L^{l-1}}^{t_L^{l+1}}\mathbf{b}(t)dt, 1\leq l\leq L-1, \mathrm{b}_L^L = \frac{1}{\tau}\int_{t_L^{L-1}}^{t_L^L}\mathbf{b}(t)dt,$$

$$\mathrm{W}_L^{l,k} = \frac{1}{\tau^2}\int_{t_L^{l-1}}^{t_L^l}\int_{t_L^{k-1}}^{t_L^k}\mathbf{W}(t,s)dtds, 1\leq k\leq l\leq L, \tag{29}$$

$$\mathrm{c}_L^{l,k} = \frac{1}{\tau^2}\int_{t_L^{l-1}}^{t_L^l}\int_{t_L^{k-1}}^{t_L^k}\mathbf{c}(t,s)dtds, 1\leq k\leq l\leq L.$$

Lemma 12 *(Limsup condition) Let the assumptions in Theorem 1 hold. Then, for every learnable parameter function $\boldsymbol{\Theta} \in \mathcal{C}_{\boldsymbol{\Theta}}$, there exists a sequence $\{\boldsymbol{\Theta}_L\}_{L \in \mathbb{N}}$ converging to $\boldsymbol{\Theta}$ in $\mathcal{C}_{\boldsymbol{\Theta}}$ such that*

$$\limsup_{L \to \infty} \tilde{\mathfrak{L}}_{\mathcal{S};L}(\boldsymbol{\Theta}_L) \leq \tilde{\mathfrak{L}}(\boldsymbol{\Theta}),$$

where $\tilde{\mathfrak{L}}_{\mathcal{S};L}$ and $\tilde{\mathfrak{L}}$ are given in (17) and (18), respectively.

Proof We only need to consider the case when $\tilde{\mathfrak{L}}(\boldsymbol{\Theta}) < +\infty$, i.e., $\boldsymbol{\Theta} \in \Omega_{\boldsymbol{\Theta}}$. For every given $\boldsymbol{\Theta} \in \Omega_{\boldsymbol{\Theta}}$, let $\{\Theta_L\}_L$ be given by (29) associated with $\boldsymbol{\Theta}$, and let $\bar{\mathrm{W}}_L := \mathbf{flip}(\mathrm{W}_L), \bar{\mathrm{c}}_L := \mathbf{flip}(\mathbf{c}_L), \boldsymbol{\Theta}_L := \hat{\boldsymbol{\mathcal{I}}}_L \Theta_L$. We next prove that $\boldsymbol{\Theta}_L$ meets the requirements.

We show that the sequence $\{\boldsymbol{\Theta}_L\}_{L \in \mathbb{N}}$ converges to $\boldsymbol{\Theta}$ in $\mathcal{C}_{\boldsymbol{\Theta}}$ firstly. Note that we only need to show the convergence of $\{\hat{\boldsymbol{\mathcal{I}}}_L \mathrm{U}_L\}_L, \{\hat{\boldsymbol{\mathcal{BI}}}_L \bar{\mathrm{W}}_L\}_L$, as the other cases are similar. For $\tilde{t} \in [t_L^{l-1}, t_L^l], 2 \leq l \leq L-1$, we have

$$\|(\hat{\boldsymbol{\mathcal{I}}}_L \mathrm{U}_L)(\tilde{t}) - \mathbf{U}(\tilde{t})\| \leq \frac{t_L^l - \tilde{t}}{\tau} \|\mathrm{U}_L^{l-1} - \mathbf{U}(\tilde{t})\| + \frac{\tilde{t} - t_L^{l-1}}{\tau} \|\mathrm{U}_L^l - \mathbf{U}(\tilde{t})\|$$

$$\leq \Big\| \frac{1}{2\tau} \int_{t_L^{l-2}}^{t_L^l} \mathbf{U}(t)dt - \mathbf{U}(\tilde{t}) \Big\| + \Big\| \frac{1}{2\tau} \int_{t_L^{l-1}}^{t_L^{l+1}} \mathbf{U}(t)dt - \mathbf{U}(\tilde{t}) \Big\| \leq 2\boldsymbol{\omega}_{\mathbf{U}}(2\tau),$$

where $\boldsymbol{\omega}_{\mathbf{U}}$ is the modulus of continuity of $\mathbf{U}$. Similarly, by Eq.(29), we have for $\tilde{t} \in [t_L^0, t_L^1]$ and $\tilde{t} \in [t_L^{L-1}, t_L^L]$ that $\|(\hat{\boldsymbol{\mathcal{I}}}_L \mathrm{U}_L)(\tilde{t}) - \mathbf{U}(\tilde{t})\| \leq 2\boldsymbol{\omega}_{\mathbf{U}}(2\tau)$. Consequently, we obtain $\hat{\boldsymbol{\mathcal{I}}}_L \mathrm{U}_L \to \mathbf{U}$ in $\mathcal{C}([0,1], \mathbb{R}^{n \times n})$. In addition, by the definition of W_L in Eq.(29) and a similar proof of Morrey's inequality [40, Theorem 11.35], there exists constants $L_{\mathbf{W}}$ and $\tilde{L}_{\mathbf{W}}$ independent of τ, such that

$$\|(\hat{\boldsymbol{\mathcal{BI}}}_L \bar{\mathrm{W}}_L)(\tilde{t}, \tilde{s}) - \mathbf{W}(\tilde{t}, \tilde{s})\|$$

$$\leq \frac{t_L^k - \tilde{s}}{\tau} \frac{t_L^l - \tilde{t}}{\tau} \Big\| \frac{1}{\tau^2} \int_{t_L^{l-2}}^{t_L^{l-1}} \int_{t_L^{k-2}}^{t_L^{k-1}} \mathbf{W}(t,s)dtds - \mathbf{W}(\tilde{t}, \tilde{s}) \Big\| + \frac{t_L^k - \tilde{s}}{\tau} \frac{\tilde{t} - t_L^{l-1}}{\tau} \Big\| \frac{1}{\tau^2} \int_{t_L^{l-1}}^{t_L^l} \int_{t_L^{k-2}}^{t_L^{k-1}}$$

$$\mathbf{W}(t,s)dtds - \mathbf{W}(\tilde{t}, \tilde{s}) \Big\| + \frac{\tilde{s} - t_L^{k-1}}{\tau} \frac{t_L^l - \tilde{t}}{\tau} \Big\| \frac{1}{\tau^2} \int_{t_L^{l-2}}^{t_L^{l-1}} \int_{t_L^{k-1}}^{t_L^k} \mathbf{W}(t,s)dtds - \mathbf{W}(\tilde{t}, \tilde{s}) \Big\|$$

$$+ \frac{\tilde{s} - t_L^{k-1}}{\tau} \frac{\tilde{t} - t_L^{l-1}}{\tau} \Big\| \frac{1}{\tau^2} \int_{t_L^{l-1}}^{t_L^l} \int_{t_L^{k-1}}^{t_L^k} \mathbf{W}(t,s)dtds - \mathbf{W}(\tilde{t}, \tilde{s}) \Big\|$$

$$\leq \frac{1}{\tau^2} \int_{t_L^{l-2}}^{t_L^l} \int_{t_L^{k-2}}^{t_L^k} \|\mathbf{W}(t,s) - \mathbf{W}(\tilde{t}, \tilde{s})\| dtds$$

$$\leq \frac{1}{\tau^2} \cdot \int_{t_L^{l-2}}^{t_L^l} \int_{t_L^{k-2}}^{t_L^k} L_{\mathbf{W}} |(t,s) - (\tilde{t}, \tilde{s})|^{1/3} dtds$$

$$\leq \tilde{L}_{\mathbf{W}} \cdot \tau^{1/3}, \ \forall \tilde{t} \in [t_L^{l-1}, t_L^l], \tilde{s} \in [t_L^{k-1}, t_L^k], 2 \leq k \leq l \leq L.$$

Similarly, by Eq.(29) and a similar proof of Morrey's inequality [40, Theorem 11.35] again, we have for $\tilde{t} \in [t_L^{l-1}, t_L^l], \tilde{s} \in [t_L^0, t_L^1], 1 \leq l \leq L$, and $\tilde{t} \in [t_L^0, t_L^1], \tilde{s} \in [t_L^{k-1}, t_L^k], 1 \leq k \leq L$, that $\|(\hat{\boldsymbol{\mathcal{BI}}}_L \bar{\mathrm{W}}_L)(\tilde{t}, \tilde{s}) - \mathbf{W}(\tilde{t}, \tilde{s})\| \leq \tilde{L}_{\mathbf{W}} \tau^{1/3}$. Hence, $\hat{\boldsymbol{\mathcal{BI}}}_L(\mathbf{flip}(\mathrm{W}_L)) \to \mathbf{W}$ in $\mathcal{L}^\infty([0,1] \times [0,1], \mathbb{R}^{n \times n})$ owing to the symmetry of $\mathbf{W}$ and $\mathbf{flip}(\mathrm{W}_L)$.

We next show $\limsup_{L \to \infty} \mathcal{R}_L(\Theta_L) \leq \mathcal{R}(\boldsymbol{\Theta})$. Note that it is enough to show the following inequalities

25

(i) $\limsup_{L\to\infty} \mathcal{R}_L^{(1)}(\mathrm{T}_{j,L}) \leq \|\mathbf{T}_j\|_{\mathcal{H}^1((0,1);\mathbb{R}^{n\times n})}^2$, $j = 1, 2, 3$;

(ii) $\limsup_{L\to\infty} \mathcal{R}_L^{(1)}(\mathrm{U}_L) \leq \|\mathbf{U}\|_{\mathcal{H}^1((0,1);\mathbb{R}^{n\times n})}^2$, $\limsup_{L\to\infty} \mathcal{R}_L^{(2)}(\mathrm{a}_L) \leq \|\mathbf{a}\|_{\mathcal{H}^1((0,1);\mathbb{R}^n)}^2$;

(iii) $\limsup_{L\to\infty} \mathcal{R}_L^{(1)}(\mathrm{V}_L) \leq \|\mathbf{V}\|_{\mathcal{H}^1((0,1);\mathbb{R}^{n\times n})}^2$, $\limsup_{L\to\infty} \mathcal{R}_L^{(2)}(\mathrm{b}_L) \leq \|\mathbf{b}\|_{\mathcal{H}^1((0,1);\mathbb{R}^n)}^2$;

(iv) $\limsup_{L\to\infty} \mathcal{R}_L^{(3)}(\mathbf{flip}(\mathrm{W}_L)) \leq \|\mathbf{W}\|_{\mathcal{W}^{1,3}((0,1)\times(0,1);\mathbb{R}^{n\times n})}^3$,

$\limsup_{L\to\infty} \mathcal{R}_L^{(4)}(\mathbf{flip}(\mathrm{c}_L)) \leq \|\mathbf{c}\|_{\mathcal{W}^{1,3}((0,1)\times(0,1);\mathbb{R}^n)}^3$.

Since the proof of parts (i) - (iii) are analogous and the two inequalities in (iv) are analogous, we only need to show $\limsup_{L\to\infty} \mathcal{R}_L^{(1)}(\mathrm{U}_L) \leq \|\mathbf{U}\|_{\mathcal{H}^1((0,1);\mathbb{R}^{n\times n})}^2$ and $\limsup_{L\to\infty} \mathcal{R}_L^{(3)}(\mathbf{flip}(\mathrm{W}_L)) \leq \|\mathbf{W}\|_{\mathcal{W}^{1,3}((0,1)\times(0,1);\mathbb{R}^{n\times n})}^3$.

We now prove $\limsup_{L\to\infty} \mathcal{R}_L^{(1)}(\mathrm{U}_L) \leq \|\mathbf{U}\|_{\mathcal{H}^1((0,1);\mathbb{R}^{n\times n})}^2$. On one hand,

$$
\left| \tau \sum_{l=1}^{L} \|\mathrm{U}_L^l\|^2 - \int_0^1 \|\mathbf{U}(t)\|^2 dt \right| = \tau \sum_{l=1}^{L-1} \left(\|\mathrm{U}_L^l\| + \|\mathbf{U}(\xi_L^l)\| \right) \cdot \left\| \frac{1}{2\tau} \int_{t_L^{l-1}}^{t_L^{l+1}} \mathbf{U}(t) dt - \mathbf{U}(\xi_L^l) \right\|
$$

$$
+ \tau(\|\mathrm{U}_L^L\| + \|\mathbf{U}(\xi_L^L)\|) \cdot \left\| \frac{1}{\tau} \int_{t_L^{L-1}}^{t_L^L} \mathbf{U}(t) dt - \mathbf{U}(\xi_L^L) \right\|
$$

$$
\leq 2 \cdot \|\mathbf{U}\|_C \cdot \boldsymbol{\omega}_{\mathbf{U}}(2\tau) + 2\tau \|\mathbf{U}\|_C \cdot \boldsymbol{\omega}_{\mathbf{U}}(\tau),
$$

where $\xi_L^l \in [t_L^{l-1}, t_L^l]$, the first inequality is due to the mean value theorem for integrals. On the other hand, by Hölder's inequality and [40, Theorem 10.55], we get

$$
\frac{1}{\tau} \sum_{l=1}^{L} \|\mathrm{U}_L^l - \mathrm{U}_L^{l-1}\|^2 \tag{30}
$$

$$
\leq \frac{1}{4\tau^2} \int_{t_L^1}^{t_L^2} \|\mathbf{U}(t) - \mathbf{U}(t-\tau)\|^2 dt + \frac{1}{4\tau^2} \int_{t_L^{L-1}}^{t_L^L} \|\mathbf{U}(t) - \mathbf{U}(t-\tau)\|^2 dt
$$

$$
+ \frac{1}{2\tau^2} \sum_{l=2}^{L-1} \left[\int_{t_L^l}^{t_L^{l+1}} \|\mathbf{U}(t) - \mathbf{U}(t-\tau)\|^2 dt + \int_{t_L^{l-1}}^{t_L^l} \|\mathbf{U}(t) - \mathbf{U}(t-\tau)\|^2 dt \right]
$$

$$
= \frac{1}{\tau^2} \left[\frac{3}{4} \int_{t_L^1}^{t_L^2} \|\mathbf{U}(t) - \mathbf{U}(t-\tau)\|^2 dt + \sum_{l=2}^{L-2} \int_{t_L^l}^{t_L^{l+1}} \|\mathbf{U}(t) - \mathbf{U}(t-\tau)\|^2 dt \right.
$$

$$
\left. + \frac{3}{4} \int_{t_L^{L-1}}^{t_L^L} \|\mathbf{U}(t) - \mathbf{U}(t-\tau)\|^2 dt \right]
$$

$$
\leq L^2 \int_{1/L}^1 \|\mathbf{U}(t) - \mathbf{U}(t-\tfrac{1}{L})\|^2 dt \leq \int_0^1 \|\boldsymbol{\mathcal{D}}_t(\mathbf{U})(t)\|^2 dt.
$$

Combining the above two inequalities, we have

$$
\limsup_{L\to\infty} \mathcal{R}_L^{(1)}(\mathrm{U}_L) \leq \lim_{L\to\infty} \tau \sum_{l=1}^{L} \|\mathrm{U}_L^l\|^2 + \limsup_{L\to\infty} \frac{1}{\tau} \sum_{l=1}^{L} \|\mathrm{U}_L^l - \mathrm{U}_L^{l-1}\|^2
$$

$$
\leq \int_0^1 \|\mathbf{U}(t)\|^2 dt + \int_0^1 \|\boldsymbol{\mathcal{D}}_t(\mathbf{U})(t)\|^2 dt = \|\mathbf{U}\|_{\mathcal{H}^1((0,1);\mathbb{R}^{n\times n})}^2.
$$

We next show $\limsup_{L\to\infty} \mathcal{R}_L^{(3)}(\mathbf{flip}(W_L)) \leq \|\mathbf{W}\|_{\mathcal{W}^{1,3}((0,1)\times(0,1);\mathbb{R}^{n\times n})}^3$. A direct calculation yields

$$
\begin{aligned}
\mathcal{R}_L^{(3)}(\bar{W}_L) =& \tau^2 \sum_{l=1}^{L}\sum_{k=1}^{L} \|\bar{W}_L^{l,k}\|^3 + \tau^{-1}\Big(\sum_{l=1}^{L}\sum_{k=1}^{L}\|\bar{W}_L^{l,k}-\bar{W}_L^{l-1,k}\|^3 + \sum_{l=1}^{L}\sum_{k=1}^{L}\|\bar{W}_L^{l,k}-\bar{W}_L^{l,k-1}\|^3\Big) \\
=& \tau^2 \sum_{l=1}^{L}\sum_{k=1}^{L}\Big\|\frac{1}{\tau^2}\int_{t_L^{l-1}}^{t_L^l}\int_{t_L^{k-1}}^{t_L^k}\mathbf{W}(t,s)dtds\Big\|^3 + \tau^{-1}\Big(\sum_{l=2}^{L}\sum_{k=1}^{L}\Big\|\frac{1}{\tau^2}\int_{t_L^{l-1}}^{t_L^l}\int_{t_L^{k-1}}^{t_L^k}\mathbf{W}(t,s) \\
& -\mathbf{W}(t-\tau,s)dtds\Big\|^3 + \sum_{l=1}^{L}\sum_{k=2}^{L}\Big\|\frac{1}{\tau^2}\int_{t_L^{l-1}}^{t_L^l}\int_{t_L^{k-1}}^{t_L^k}\mathbf{W}(t,s)-\mathbf{W}(t,s-\tau)dtds\Big\|^3\Big) \\
\leq& \int_0^1\int_0^1 \|\mathbf{W}(t,s)\|^3 dtds + L^3\int_{1/L}^1\int_0^1\|\mathbf{W}(t,s)-\mathbf{W}(t-1/L,s)\|^3 dtds \\
& + L^3\int_0^1\int_{1/L}^1 \|\mathbf{W}(t,s)-\mathbf{W}(t,s-1/L)\|^3 dtds \\
\leq& \|\mathbf{W}\|_{\mathcal{W}^{1,3}((0,1)\times(0,1);\mathbb{R}^{n\times n})}^3,
\end{aligned}
$$

where the first inequality is owing to the Hölder's inequality and the second inequality is due to [40, Theorem 10.55]. Taking $L \to \infty$, we have $\limsup_{L\to\infty}\mathcal{R}_L^{(3)}(\bar{W}_L) \leq \|\mathbf{W}\|_{\mathcal{W}^{1,3}((0,1)\times(0,1);\mathbb{R}^{n\times n})}^3$.

The condition (14) is satisfied with the given $\{\Theta_L\}_L$ due to the convergence of $\{\mathbf{\Theta}_L\}_{L\in\mathbb{N}}$ and the proof of Lemma 11. Therefore, we get

$$
\frac{1}{M}\sum_{m=1}^{M}\boldsymbol{\ell}(x^L(d_m;\Theta_L),g_m) \to \frac{1}{M}\sum_{m=1}^{M}\boldsymbol{\ell}(\mathbf{x}(1;d_m;\mathbf{\Theta}),g_m) \text{ as } L\to\infty, \tag{31}
$$

by the continuity of $\boldsymbol{\ell}$ (assumption (A_3)) and Proposition 4. Thus, combining $\limsup_{L\to\infty}\mathcal{R}_L(\Theta_L) \leq \mathcal{R}(\mathbf{\Theta})$ and Eq.(31) completes the proof. $\square$

Now, we are ready to give a proof for Theorem 1.

Proof (**Proof of Theorem 1**) By Proposition 6, the optimal solutions of $(\mathcal{P})$ and $(\mathcal{P}_L)$ exist in $\Omega_{\mathbf{\Theta}}$ and $\Omega_{\Theta;L}$ respectively. Consequently, we obtain the existence of minimizers in $\mathcal{C}_{\mathbf{\Theta}}$ of $\tilde{\mathbf{\mathfrak{L}}}$ and $\tilde{\mathbf{\mathfrak{L}}}_{\mathcal{S};L}$, $L\in\mathbb{N}$, owing to Lemma 7. Moreover, $\tilde{\mathbf{\mathfrak{L}}}_{\mathcal{S};L}$ Γ-converges to $\tilde{\mathbf{\mathfrak{L}}}$ in $\mathcal{C}_{\mathbf{\Theta}}$ due to Lemma 11 and Lemma 12. Using the definition of $\tilde{\mathbf{\mathfrak{L}}}_{\mathcal{S};L}$, if $\mathbf{\Theta}_L^* \in \arg\min_{\mathbf{\Theta}\in\mathcal{C}_{\mathbf{\Theta}}}\tilde{\mathbf{\mathfrak{L}}}_{\mathcal{S};L}(\mathbf{\Theta})$, then there exists some $\Theta_L^* = (T_L^*, U_L^*, a_L^*, V_L^*, b_L^*, W_L^*, c_L^*) \in \Omega_{\Theta;L}$ such that $\mathbf{\Theta}_L^* = \hat{\mathcal{I}}_L\Theta_L^*$ for all $L\in\mathbb{N}$. It then follows from Lemma 8 that there exists a subsequence of $\{\mathbf{\Theta}_L^*\}_L$ converging to a $\tilde{\mathbf{\Theta}}\in\Omega_{\mathbf{\Theta}}$ in $\mathcal{C}_{\mathbf{\Theta}}$. Using the property of Γ-convergence [22, Theorem 3.2], we obtain

$$
\min_{\mathbf{\Theta}\in\mathcal{C}_{\mathbf{\Theta}}}\tilde{\mathbf{\mathfrak{L}}}_{\mathcal{S};L}(\mathbf{\Theta}) \to \min_{\mathbf{\Theta}\in\mathcal{C}_{\mathbf{\Theta}}}\tilde{\mathbf{\mathfrak{L}}}(\mathbf{\Theta}), \text{ as } L\to\infty.
$$

In addition, for any subsequence $\{\mathbf{\Theta}_{L_k}^*\}_{L_k\in\mathbb{N}}$ converging to some $\mathbf{\Theta}^*$ in $\mathcal{C}_{\mathbf{\Theta}}$, we have $\mathbf{\Theta}^* \in \arg\min_{\mathbf{\Theta}\in\mathcal{C}_{\mathbf{\Theta}}}\tilde{\mathbf{\mathfrak{L}}}(\mathbf{\Theta})$.

Therefore, for any sequence $\{\Theta_L^*\}_L$ with Θ_L^* being the minimiser of $(\mathcal{P}_L)$, we derive

$$
\min_{\Theta_L\in\Omega_{\Theta;L}}\mathbf{\mathfrak{L}}_{\mathcal{S};L}(\Theta_L) = \mathbf{\mathfrak{L}}_{\mathcal{S};L}(\Theta_L^*) = \tilde{\mathbf{\mathfrak{L}}}_{\mathcal{S};L}(\mathbf{\Theta}_L^*) = \min_{\mathbf{\Theta}\in\mathcal{C}_{\mathbf{\Theta}}}\tilde{\mathbf{\mathfrak{L}}}_{\mathcal{S};L}(\mathbf{\Theta})
$$

$$
\to \min_{\mathbf{\Theta}\in\mathcal{C}_{\mathbf{\Theta}}}\tilde{\mathbf{\mathfrak{L}}}(\mathbf{\Theta}) = \tilde{\mathbf{\mathfrak{L}}}(\mathbf{\Theta}^*) = \mathbf{\mathfrak{L}}(\mathbf{\Theta}^*) = \min_{\mathbf{\Theta}\in\Omega_{\mathbf{\Theta}}}\mathbf{\mathfrak{L}}(\mathbf{\Theta}), \text{ as } L\to\infty.
$$

Besides, $\hat{\mathcal{I}}_L\Theta_L^* \in \arg\min_{\mathbf{\Theta}\in\mathcal{C}_{\mathbf{\Theta}}}\tilde{\mathbf{\mathfrak{L}}}_{\mathcal{S};L}(\mathbf{\Theta})$ by Lemma 7. It then follows that any accumulation point of $\{\hat{\mathcal{I}}_L\Theta_L^*\}_{L\in\mathbb{N}}$ minimizes $\tilde{\mathbf{\mathfrak{L}}}$, and thus is an optimal solution of $(\mathcal{P})$. $\square$

5 Experiments

In this section, we conduct experiments to test the behavior of the training loss of the DNL framework to verify our convergence results in some sense. We investigate how it changes as the layer number L increases on image classification tasks on the SVHN and CIFAR-10 datasets. The SVHN dataset contains approximately 73,257 training images and 26,032 test images, while CIFAR-10 consists of 50,000 training images and 10,000 test images. Each image has 3 color channels and a resolution of 32×32 pixels, yielding 3,072 features per sample. We use the standard dataset splits without additional preprocessing.

We implement the standard DenseNet as an instance of the DNL framework to examine the deep-layer limit behavior established in our theoretical result. Each network consists of four DNL blocks, each of which has L layers with a growth rate of 6. For the loss function in (4), we take $\boldsymbol{\ell}$ as the cross entropy function, i.e.,

$$\boldsymbol{\ell}(\hat{g}, g) = -10^6 \times \sum_{i=1}^{n} g[i] \log(\hat{g}[i]), \ \hat{g}, g \in \mathbb{R}^n.$$

All models are trained for 50 epochs on the SVHN dataset and 150 epochs on CIFAR10 using SGD with an initial learning rate of 0.01 and momentum of 0.9. To reduce the influence of randomness, we repeat the training process five times for each hyperparameter setting and record the average training loss of the five trained models during testing. Since our theoretical results concern the training problem rather than generalization, we only report training losses. All experiments are implemented in PyTorch on an NVIDIA GeForce RTX 3090 GPU.

We test the DNL block with layer numbers $L = 6, 8, 10, 12, 14, 16$, yielding a step size $\tau_L = 1/L$. Figure 3 shows training losses versus epoch number for different L. On both SVHN and CIFAR10, the losses converge as L increases, i.e., larger L yields smaller training loss, and the improvement diminishes with further increases of L. This observation is consistent with the convergence result established in Theorem 1.

6 Conclusion

DNNs with dense layer connectivities form an important class of architectures in modern deep learning. In this work, we provided a dynamical system modeling and convergence analysis for such DNN architectures. Our study was presented within a general DNL framework, containing both standard dense layer connections and general (local/non-local) feature transformations within layers. In the deep-layer limit, we obtained a continuous-time formulation in the form of nonlinear integral equations and studied the associated learning problems from an optimal control perspective. By employing a piecewise linear extension technique and Γ-convergence tool, we established convergence properties between their learning problems, including the convergence of optimal values and the subsequence convergence of minimizers. Our theoretical result is applicable to a class of densely connected DNNs with various local/non-local feature transformations within layers. These findings provide a theoretical foundation for understanding densely connected networks and suggest directions
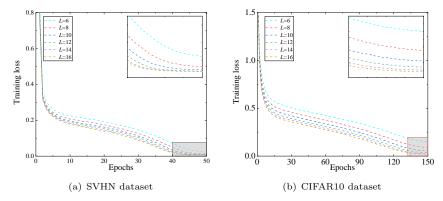
(a) SVHN dataset   (b) CIFAR10 dataset

Fig. 3 The plot of training loss v.s. epoch number for the DenseNet with different layer numbers L on the SVHN and CIFAR10 datasets. The training losses decrease with increasing L, closely matching the theoretical convergent result.

for future research, including the design of training algorithms and the study of loss landscapes based on continuous-time integral equations.

Supplementary information. The supplementary material contains the complete proofs of Remark 1, Lemma 2, Proposition 3, and Corollary 5 presented in the main text. Readers interested in the detailed derivations are referred to this file.

Appendix A   Γ-convergence

Definition 5 [38] Let $(\mathbb{U}, d_{\mathbb{U}})$ be a metric space, and $\boldsymbol{\mathcal{E}}_L : \mathbb{U} \to [0, \infty]$ be a sequence of functionals. Then the sequence $\{\boldsymbol{\mathcal{E}}_L\}_{L \in \mathbb{N}}$ Γ-converges with respect to metric $d_{\mathbb{U}}$ to the functional $\mathcal{E} : \mathbb{U} \to [0, \infty]$ as $L \to \infty$ if the following inequalities hold:

1. Liminf inequality: For every $u \in \mathbb{U}$ and every sequence $\{u_L\}_{L \in \mathbb{N}}$ converging to u,

$$\liminf_{L \to \infty} \boldsymbol{\mathcal{E}}_L (u_L) \geq \mathcal{E}(u);$$

2. Limsup inequality: For every $u \in \mathbb{U}$ there exists a sequence $\{u_L\}_{L \in \mathbb{N}}$ converging to u satisfying

$$\limsup_{L \to \infty} \boldsymbol{\mathcal{E}}_L (u_L) \leq \mathcal{E}(u).$$

We say that $\mathcal{E}$ is the Γ-limit of the sequence of functionals $\{\boldsymbol{\mathcal{E}}_L\}_{L \in \mathbb{N}}$ (with respect to the metric $d_{\mathbb{U}}$). A fundamental property of Γ-convergence is given in [22, Theorem 3.2].

Theorem 13 [22, Theorem 3.2] *Let $(\mathbb{U}, d_{\mathbb{U}})$ be a metric space and let $\boldsymbol{\mathcal{E}}_L : \mathbb{U} \to [0, \infty]$ be a sequence of functionals. Let u_L be a sequence of almost minimisers for $\boldsymbol{\mathcal{E}}_L$, i.e.*

$$\boldsymbol{\mathcal{E}}_L (u_L) \leq \max \left\{ \inf_{u \in \mathbb{U}} \boldsymbol{\mathcal{E}}_L (u_L) + \varepsilon_L, -\frac{1}{\varepsilon_L} \right\}$$

29

for some $\varepsilon_L \to 0^+$. Assume that $\boldsymbol{\mathcal{E}}$ is the Γ-limit of the sequence of functionals $\{\boldsymbol{\mathcal{E}}_L\}_{L \in \mathbb{N}}$ and $\{u_L\}_{L=1}^\infty$ are relatively compact. Then,

$$\inf_{u \in \mathbb{U}} \boldsymbol{\mathcal{E}}_L(u) \to \min_{u \in \mathbb{U}} \boldsymbol{\mathcal{E}}(u),$$

where the minimum of $\boldsymbol{\mathcal{E}}$ exists. Moreover, if a convergent subsequence $u_{L_k} \to u^*$, then u^* minimizes $\boldsymbol{\mathcal{E}}$.

References

[1] Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT press, Cambridge, MA (2016)

[2] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

[3] Huang, G., Liu, Z., Weinberger, K.Q.: Densely connected convolutional networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2261–2269 (2016)

[4] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, pp. 234–241 (2015). Springer

[5] Zhang, H., Qin, K., Zhang, Y., Li, Z., Xu, K.: Dense attention convolutional network for image classification. In: Journal of Physics: Conference Series, vol. 1651, p. 012184 (2020). IOP Publishing

[6] Yao, C., Jin, S., Liu, M., Ban, X.: Dense residual transformer for image denoising. Electronics **11**(3), 418 (2022)

[7] Ma, H., Li, X., Yuan, X., Zhao, C.: Denseformer: A dense transformer framework for person re-identification. IET Computer Vision **17**(5), 527–536 (2023)

[8] Diakogiannis, F.I., Waldner, F., Caccetta, P., Wu, C.: Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. ISPRS Journal of Photogrammetry and Remote Sensing **162**, 94–114 (2020)

[9] Cao, Y., Liu, S., Peng, Y., Li, J.: Denseunet: densely connected unet for electron microscopy image segmentation. IET Image Processing **14**(12), 2682–2689 (2020)

[10] Tai, X.-C., Liu, H., Chan, R.: Pottsmgnet: A mathematical explanation of encoder-decoder based neural networks. SIAM Journal on Imaging Sciences **17**(1), 540–594 (2024)

[11] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems **30** (2017)

[12] Wang, X., Girshick, R.B., Gupta, A.K., He, K.: Non-local neural networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7794–7803 (2017)

[13] Jia, F., Tai, X.-C., Liu, J.: Nonlocal regularized cnn for image segmentation. Inverse Problems & Imaging **14**(5), 891–911 (2020)

[14] Meng, J., Wang, F., Liu, J.: Learnable nonlocal self-similarity of deep features for image denoising. SIAM Journal on Imaging Sciences **17**(1), 441–475 (2024)

[15] Liu, J., Wang, X., Tai, X.-C.: Deep convolutional neural networks with spatial regularization, volume and star-shape priors for image segmentation. Journal of Mathematical Imaging and Vision **64**(6), 625–645 (2022)

[16] Gao, B., Pavel, L.: On the properties of the softmax function with application in game theory and reinforcement learning. arXiv preprint arXiv:1704.00805 (2017)

[17] E, W.: A proposal on machine learning via dynamical systems. Communications in Mathematics and Statistics **1**(5), 1–11 (2017)

[18] Haber, E., Ruthotto, L.: Stable architectures for deep neural networks. Inverse Problems **34**(1), 014004 (2017)

[19] Lu, Y., Zhong, A., Li, Q., Dong, B.: Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations. In: Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 3282–3291. PMLR, Stockholm, Sweden (2018)

[20] Gomez, A.N., Ren, M., Urtasun, R., Grosse, R.B.: The reversible residual network: Backpropagation without storing activations. Advances in Neural Information Processing Systems **30** (2017)

[21] Zhang, L., Schaeffer, H.: Forward stability of resnet and its variants. Journal of Mathematical Imaging and Vision **62**(3), 328–351 (2020)

[22] Thorpe, M., Gennip, Y.: Deep limits of residual neural networks. Research in the Mathematical Sciences **10**(1), 6 (2023)

[23] Chen, R.T.Q., Rubanova, Y., Bettencourt, J., Duvenaud, D.K.: Neural ordinary differential equations. In: Advances in Neural Information Processing Systems, vol. 31, pp. 6571–6583. Curran Associates, Inc., Red Hook, NY (2018)

[24] Lu, Y., Li, Z., He, D., Sun, Z., Dong, B., Qin, T., Wang, L., Liu, T.-Y.: Understanding and improving transformer from a multi-particle dynamic system point

of view. arXiv preprint arXiv:1906.02762 (2019)

[25] Sherstinsky, A.: Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. Physica D: Nonlinear Phenomena **404**, 132306 (2020)

[26] Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: International Conference on Learning Representations (2021)

[27] Monga, V., Li, Y., Eldar, Y.C.: Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. IEEE Signal Processing Magazine **38**(2), 18–44 (2021)

[28] Huang, J., Gao, Y., Wu, C.: On dynamical system modeling of learned primal-dual with a linear operator $\mathcal{K}$: stability and convergence properties. Inverse Problems **40**(7), 075006 (2024)

[29] Lin, X., Wu, C.: Deep layer limit and stability analysis of the basic forward-backward-splitting induced network (i): feed-forward systems. To appear in IMA Journal of Numerical Analysis (2025)

[30] Lin, X., Wu, C.: Deep layer limit and stability analysis of the basic forward-backward-splitting induced network (ii): learning problems. Submitted (2024)

[31] Haber, E., Lensink, K., Treister, E., Ruthotto, L.: Imexnet a forward stable deep neural network. In: International Conference on Machine Learning, pp. 2525–2534 (2019)

[32] Ruthotto, L., Haber, E.: Deep neural networks motivated by partial differential equations. Journal of Mathematical Imaging and Vision **62**(3), 352–364 (2020)

[33] Buades, A., Coll, B., Morel, J.-M.: A review of image denoising algorithms, with a new one. Multiscale modeling & simulation **4**(2), 490–530 (2005)

[34] Wei, C., Lee, J.D., Liu, Q., Ma, T.: Regularization matters: generalization and optimization of neural nets vs their induced kernel. Advances in Neural Information Processing Systems **32** (2019)

[35] Esteve, C., Geshkovski, B., Pighin, D., Zuazua, E.: Large-time asymptotics in deep learning. ArXiv **abs/2008.02491** (2020)

[36] Brunner, H.: Volterra Integral Equations: An Introduction to Theory and Applications. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, Cambridge (2017)

[37] Dragomir, S.S.: Some Gronwall Type Inequalities and Applications. Nova Science Publishers, Hauppauge, NY (2003)

[38] Braides, A.: Gamma-convergence for Beginners vol. 22. Oxford University Press, Oxford (2002)

[39] Adams, R.A., Fournier, J.J.F.: Sobolev Spaces vol. 140, 2nd edn. Elsevier Press, Amsterdam (2003)

[40] Leoni, G.: A First Course in Sobolev Spaces vol. 105. American Mathematical Society, Providence, RI (2009)