Theoretical Exploration of Flexible Transmitter Model

Jin-Hui Wu, Shao-Qun Zhang, Yuan Jiang, Zhi-Hua Zhou[1]

National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210093, China
{wujh,zhangsq,jiangy,zhouzh}@lamda.nju.edu.cn

Abstract

Neural network models generally involve two important components, i.e., network architecture and neuron model. Although there are abundant studies about network architectures, only a few neuron models have been developed, such as the MP neuron model developed in 1943 and the spiking neuron model developed in the 1950s. Recently, a new bio-plausible neuron model, Flexible Transmitter (FT) model [58], has been proposed. It exhibits promising behaviors, particularly on temporal-spatial signals, even when simply embedded into the common feedforward network architecture. This paper attempts to understand the properties of the FT network (FTNet) theoretically. Under mild assumptions, we show that: i) FTNet is a universal approximator; ii) the approximation complexity of FTNet can be exponentially smaller than those of commonly-used real-valued neural networks with feedforward/recurrent architectures and is of the same order in the worst case; iii) any local minimum of FTNet is the global minimum, implying that it is possible to identify global minima by local search algorithms.

Key words: Neural Networks, Flexible Transmitter Model, Approximation Complexity, Local Minimum

1. Introduction

Deep neural networks have become mainstream in artificial intelligence and have exhibited excellent performance in many applications, such as disease detection [30], machine translation [5],

[1]Zhi-Hua Zhou is the corresponding author.

emotion recognition [20], etc. Typically, a neural network model is composed of a network architecture and a neuron model. The past decade has witnessed abundant studies about network architectures, whereas the modeling of neurons is relatively less considered. Typical neuron models include the MP neuron model [39] and the spiking neuron model [21, 57]. Recently, a new bio-plausible neuron model, *Flexible Transmitter* model [58], has been proposed. In contrast to the classical neuron models, the FT neuron model mimics neurotrophic potentiation and depression effects by a formulation of a two-variable function, exhibiting great potential for temporal-spatial data processing. Furthermore, Zhang and Zhou [58] developed the Flexible Transmitter Network, a feed-forward neural network composed of FT neurons, which performs competitively with state-of-the-art models when handling temporal-spatial data.

However, the theoretical properties of the FT model remain unknown. This work takes one step in this direction. We notice that the formulation of the FT model provides greater flexibility for the representation of neuron models, and its benefits are twofold. Firstly, the complex-valued implementation takes into account the magnitude and phase of variables and is thus good at processing data with norm-preserving and antisymmetric structures. Secondly, the modeling of neurotrophic potentiation and depression effects derives a local recurrent system, and FTNet intrinsically has temporal-spatial representation ability even in a feed-forward architecture. Inspired by these insights, we present the theoretical advantages over the Feedforward Neural Network (FNN) and Recurrent Neural Network (RNN) from the perspectives of approximation and local minima. Our main contributions can be summarized as follows:

1. FTNet is a universal approximator, i.e., a one-hidden-layer FTNet with admissible activation functions can approximate any continuous function and discrete-time open dynamical system on any compact set arbitrarily well, stated in Theorems 1 and 2, respectively.

2. We present the approximation-complexity advantages and the worst-case guarantees of FTNet over the FNN and RNN. Specifically, separation results exist between one-hidden-layer FTNet and one-hidden-layer FNN/RNN, as shown in Theorems 3 and 4, respectively. In addition, any function expressible by a one-hidden-layer FNN or RNN can be approximated by a one-hidden-layer FTNet with a similar number of hidden neurons, as shown in Theorems 5 and 6, respectively. These theorems imply that FTNet is capable of expressing functions more efficiently than FNN and RNN.

3. We show that FTNet in the feedforward architecture has no suboptimal local minimum using general activations and loss functions, as illustrated in Theorem 7. This implies that local search algorithms for FTNet have the potential to converge to the global minimum.

The rest of this paper is organized as follows. Section 2 introduces related work. Section 3 provides basic notations, definitions, and the formulation of FTNet. Section 4 proves the universal approximation of FTNet. Section 5 investigates the approximation complexity of FTNet. Section 6 studies the property of the local minima of FTNet. Section 7 concludes our work with prospect.

2. Related Works

Universal Approximation. The universal approximation confirms the powerful expressivity of neural networks. The earliest research is the universal approximation theorem of FNN, which proves that FNN with suitable activation functions can approximate any continuous function on any compact set arbitrarily well [8, 15, 22]. Furthermore, Leshno et al. point out that a non-polynomial activation function is the necessary and sufficient condition for FNNs to achieve universal approximation [32]. Later, some researchers extend the universal approximation theorems to other real-weighted neural networks with different architectures, such as RNN [47, 16, 7, 33, 46] and convolutional neural networks [60]. For complex-weighted neural networks, it has been proven that they can approximate any continuous complex-valued functions on any compact set using some activation function [3], and that non-holomorphic and non-antiholomorphic activation functions are the necessary and sufficient condition of universal approximation [52]. Our work investigates the universal approximation of FTNet, i.e., the capability of complex-weighted neural networks to approximate real-valued functions and dynamical systems.

Approximation Complexity. The universal approximation theorems only prove the possibility of approximating certain functions, but do not consider approximation complexity, i.e., the number of required hidden neurons for approximating particular functions. It is also important to consider approximation complexity, which reflects the efficiency of approximation. Early works focus on the degree of approximation of one-hidden-layer FNN, i.e., how approximation error depends on the input dimension and the number of hidden units [6, 9, 23, 40]. Recent works prove separation results, i.e., one model cannot be expressed by another model with the same order of parameters [13, 45, 50, 38]. A notable work proves that one-hidden-layer FNN needs at least exponential parameters to express a given complex-reaction network [56]. Our work not only provides the separation results between FTNet and FNN/RNN but also guarantees that any FNN/RNN can be expressed by FTNet with a similar hidden size.

Local Minima. Suboptimal local minima are undesirable points of the loss surface, without which it is tractable to train neural networks using local search algorithms. Early works show that one-hidden-layer FNN using the squared loss has no suboptimal local minimum under suitable conditions [44, 53, 54, 24]. These results are extended to multilayer FNN [41, 28] and other types of neural networks, such as deep ResNet [27], deep convolutional neural networks [42], deep linear networks [26, 31], and over-parameterized deep neural networks [43]. From another perspective of algorithms, some researchers prove that some commonly used gradient-based algorithms, e.g., GD and SGD, can converge to the global minimum or an almost optimal solution when optimizing over-parameterized neural networks [17, 12, 25, 34, 2, 11, 1, 64, 63, 37, 59]. Our work extends the classical results of FNN to FTNet in the feedforward architecture and generalizes the condition on the loss function from the squared loss to a large class of analytic functions.

3. Preliminary

We denote by $\mathrm{i} = \sqrt{-1}$ the imaginary unit. Let $\mathrm{Re}(z)$, $\mathrm{Im}(z)$, θ_z, and $\overline{z}$ be the real part, imaginary part, phase, and complex conjugate of the complex number z, respectively. Let $\mathbf{0}^{a\times b}$ denote the zero matrix with a rows and b columns.

This work considers FTNet with two typical architectures, that is, Recurrent FTNet (R-FTNet) and Feedforward FTNet (F-FTNet), and the time series regression task with 1-dimensional outputs throughout this paper. We focus on one-hidden-layer FTNet throughout this paper. For deep FTNet, it would be interesting to study feature space transformation, which might be a key to understanding the mysteries behind the success of deep neural networks [61]. Let $\boldsymbol{x}_t \in \mathbb{R}^I$ be the input vector at time t, and $\boldsymbol{x}_{1:T} = (\boldsymbol{x}_1; \boldsymbol{x}_2; \dots; \boldsymbol{x}_T) \in \mathbb{R}^{IT}$ denotes the concatenated input vector at time T. We employ the mapping $f_{\times,\mathrm{R}}$ to denote a one-hidden-layer R-FTNet with

$H_\times \geqslant I+1$ hidden neurons as follows

$$\begin{aligned} &f_{\times,\mathrm{R}} : \boldsymbol{x}_{1:T} \mapsto (y_{\times,1}, \dots, y_{\times,T}), \\ &\boldsymbol{s}_t + \boldsymbol{r}_t \mathrm{i} = \sigma_\times((\mathbf{W}_\times + \mathbf{V}_\times \mathrm{i})\,(\kappa(\boldsymbol{x}_t, H_\times) + \boldsymbol{r}_{t-1}\mathrm{i})), \\ &y_{\times,t} = \boldsymbol{\alpha}_\times^\top \boldsymbol{s}_t, \quad t \in [T], \end{aligned} \tag{1}$$

where $\boldsymbol{r}_t$, $\boldsymbol{s}_t \in \mathbb{R}^{H_\times}$, and $y_{\times,t} \in \mathbb{R}$ represent the receptor, stimulus, and output at time t, respectively, $\mathbf{W}_\times$, $\mathbf{V}_\times$, $\boldsymbol{\alpha}_\times$ denote real-valued weight parameters, $\kappa : \mathbb{R}^I \times \mathbb{N}^+ \to \mathbb{R}^{H_\times}$ stretches the input to a higher-dimensional space in which

$$\kappa(\boldsymbol{x}, H_\times) = (\boldsymbol{x}; 0; \dots; 0; 1) \in \mathbb{R}^{H_\times} \quad \text{with} \quad H_\times \geqslant I+1, \tag{2}$$

and $\sigma_\times$ is an activation function applied componentwise. Notice that Eq. (1) is a multiplicative (rather than additive) form of FTNet since multiplication is the last operation before applying the activation function. In addition, we also employ the mapping $f_{\times,\mathrm{F}}$ to denote a one-hidden-layer F-FTNet with $H_\times \geqslant I+1$ hidden neurons as follows

$$f_{\times,\mathrm{F}} : \boldsymbol{x} \mapsto \boldsymbol{\alpha}_\times^\top \mathrm{Re}\left[\sigma_\times\left((\mathbf{W}_\times + \mathbf{V}_\times \mathrm{i})\,\kappa(\boldsymbol{x}, H_\times)\right)\right]. \tag{3}$$

The zReLU activation function [18] is a promising choice of the activation function in FTNet, which extends the widely used real-valued activation function ReLU [14] to the complex-valued domain, and is defined as

$$\sigma(z) = \begin{cases} z, & \text{if} \quad \theta_z \in [0, \pi/2] \cup [\pi, 3\pi/2], \\ 0, & \text{otherwise}. \end{cases} \tag{4}$$

Dynamical systems are of great interest when considering the universal approximation of neuron models in the recurrent architecture. We focus on the discrete-time open dynamical system defined as follows.

Definition 1 *Given an initial hidden state $\boldsymbol{h}_0 \in \mathbb{R}^{H_\mathrm{D}}$ with $H_\mathrm{D} \in \mathbb{N}^+$, a Discrete-time Open Dynamical System (*DODS*) is a mapping f_D defined by*

$$\begin{aligned} &f_\mathrm{D} : \boldsymbol{x}_{1:T} \mapsto (y_1, \cdots, y_T), \\ &y_t = \psi(\boldsymbol{h}_t), \\ &\boldsymbol{h}_t = \varphi(\boldsymbol{x}_t, \boldsymbol{h}_{t-1}), \quad t \in [T], \end{aligned} \tag{5}$$

where $\boldsymbol{x}_t \in \mathbb{R}^I$, $\boldsymbol{h}_t \in \mathbb{R}^{H_\mathrm{D}}$, and $y_t \in \mathbb{R}$ represent the input, hidden state, and output at time t, respectively, $\varphi : \mathbb{R}^I \times \mathbb{R}^{H_\mathrm{D}} \to \mathbb{R}^{H_\mathrm{D}}$ and $\psi : \mathbb{R}^{H_\mathrm{D}} \to \mathbb{R}^O$ are continuous mappings.

4. Universal Approximation

We show the universal approximation of F-FTNet and R-FTNet in Subsections 4.1 and 4.2, respectively.

4.1. Universal Approximation of F-FTNet

Let $\|f\|_{L^\infty(\Omega)}$ denote the essential supremum of the function f on the domain Ω, i.e.,

$$\|f\|_{L^\infty(\Omega)} = \inf \{\lambda \mid \mu \{x : |f(x)| \geqslant \lambda\} = 0\},$$

where μ is the Lebesgue measure. We now present the universal approximation for F-FTNet as follows.

Theorem 1 *Let $K \subset \mathbb{R}^I$ be a compact set, g is a continuous function on K, and $\sigma_\times$ is the activation function of* F-FTNet*. Suppose there exists a constant $c \in \mathbb{R}$, such that the function $\sigma(x) = \mathrm{Re}\left[\sigma_\times(x + c\mathrm{i})\right]$ is continuous almost everywhere and not polynomial almost everywhere. Then for any $\varepsilon > 0$, there exists an* F-FTNet *$f_{\times,\mathrm{F}}$, such that*

$$\|f_{\times,\mathrm{F}} - g\|_{L^\infty(K)} \leqslant \varepsilon.$$

Theorem 1 indicates that F-FTNet with suitable activation functions can approximate any continuous function on any compact set arbitrarily well. The conditions in this theorem are satisfied by many commonly used activation functions, such as modReLU [4], zReLU [18], and $\mathbb{C}$ReLU [51]. Previous studies focus on the universal approximation of real-weighted networks with real-valued target functions or complex-weighted networks with complex-valued target functions. To our knowledge, Theorem 1 is the first result considering the approximation capability of complex-weighted networks with real-valued target functions. The condition about σ of FTNet is the same as that of FNN [32], but the activation $\sigma_\times$ of FTNet is more flexible since σ is just the restriction of $\sigma_\times$ on a particular direction. The requirement of not polynomial σ is weaker than non-holomorphic and non-antiholomorphic activation, which is the necessary requirement of complex-weighted networks with complex-valued target functions [52]. Thus, FTNet successfully benefits from expressing real-valued functions instead of complex-valued ones. We begin our proof with the following lemma.

Lemma 1 *[32, Theorem 1] Let $K \subset \mathbb{R}^I$ be a compact set, and g is a continuous function on K. Suppose the activation function σ_{F} is continuous almost everywhere and not polynomial almost everywhere. Then for any $\varepsilon > 0$, there exist $H_{\mathrm{F}} \in \mathbb{N}^+$, $\mathbf{W}_{\mathrm{F}} \in \mathbb{R}^{H_{\mathrm{F}} \times I}$, and $\boldsymbol{\theta}_{\mathrm{F}}, \boldsymbol{\alpha}_{\mathrm{F}} \in \mathbb{R}^{H_{\mathrm{F}}}$, such that*

$$\left\| \boldsymbol{\alpha}_{\mathrm{F}}^{\top} \sigma_{\mathrm{F}} \left(\mathbf{W}_{\mathrm{F}} \boldsymbol{x} - \boldsymbol{\theta}_{\mathrm{F}} \right) - g(\boldsymbol{x}) \right\|_{L^{\infty}(K)} \leqslant \varepsilon.$$

Lemma 1 shows that one-hidden-layer FNN can approximate any continuous function on any compact set arbitrarily well, using suitable activation functions.

Proof of Theorem 1. Based on Lemma 1, it suffices to construct an F-FTNet that has the same output as any given FNN. Since the function $\sigma(x)$ is continuous almost everywhere and not polynomial almost everywhere, it satisfies the conditions in Lemma 1. According to Lemma 1, there exist $H_{\mathrm{F}} \in \mathbb{N}^+$, $\mathbf{W}_{\mathrm{F}} \in \mathbb{R}^{H_{\mathrm{F}} \times I}$, and $\boldsymbol{\theta}_{\mathrm{F}}, \boldsymbol{\alpha}_{\mathrm{F}} \in \mathbb{R}^{H_{\mathrm{F}}}$, such that

$$\left\| \boldsymbol{\alpha}_{\mathrm{F}}^{\top} \sigma \left(\mathbf{W}_{\mathrm{F}} \boldsymbol{x} - \boldsymbol{\theta}_{\mathrm{F}} \right) - g(\boldsymbol{x}) \right\|_{L^{\infty}(K)} \leqslant \varepsilon. \tag{6}$$

We now construct an F-FTNet with $H_{\times} = \max\{I+1, H_{\mathrm{F}}\}$ hidden neurons as follows

$$\mathbf{W}_{\times} = \begin{bmatrix} \mathbf{W}_{\mathrm{F}} & \mathbf{0} & -\boldsymbol{\theta}_{\mathrm{F}} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad \mathbf{V}_{\times} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & c\mathbf{1} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad \boldsymbol{\alpha}_{\times} = \begin{bmatrix} \boldsymbol{\alpha}_{\mathrm{F}} \\ \mathbf{0} \end{bmatrix}.$$

Thus, one has

$$\begin{aligned} f_{\times,\mathrm{F}}(\boldsymbol{x}) &= \boldsymbol{\alpha}_{\times}^{\top} \mathrm{Re} \left[\sigma_{\times} \left(\left(\mathbf{W}_{\times} + \mathbf{V}_{\times} \mathrm{i} \right) \kappa(\boldsymbol{x}, H_{\times}) \right) \right] \\ &= \boldsymbol{\alpha}_{\times}^{\top} \mathrm{Re} \left[\sigma_{\times} \left(\left[\mathbf{W}_{\mathrm{F}} \boldsymbol{x} - \boldsymbol{\theta}_{\mathrm{F}} + c\mathbf{1}\mathrm{i}; \mathbf{0} \right] \right) \right] \\ &= \boldsymbol{\alpha}_{\mathrm{F}}^{\top} \mathrm{Re} \left[\sigma_{\times} \left(\mathbf{W}_{\mathrm{F}} \boldsymbol{x} - \boldsymbol{\theta}_{\mathrm{F}} + c\mathbf{1}\mathrm{i} \right) \right] \\ &= \boldsymbol{\alpha}_{\mathrm{F}}^{\top} \sigma \left(\mathbf{W}_{\mathrm{F}} \boldsymbol{x} - \boldsymbol{\theta}_{\mathrm{F}} \right), \end{aligned} \tag{7}$$

where the first equality holds according to Eq. (3), the second and third equalities hold from the construction of the F-FTNet, and the fourth equality holds because of the definition of the function σ. From Eqs. (6) and (7), we obtain

$$\left\| f_{\times,\mathrm{F}}(\boldsymbol{x}) - g(\boldsymbol{x}) \right\|_{L^{\infty}(K)} \leqslant \varepsilon,$$

which completes the proof. □

4.2. Universal Approximation of R-FTNet

We proceed to study the universal approximation for R-FTNet as follows.

Theorem 2 *Let $K \subset \mathbb{R}^I$ be a convex compact set, f_{D} is a* DODS *defined by Eq.* (5)*, and $\sigma_\times$ is the activation function of* R-FTNet *satisfying $\sigma_\times(0) = 0$. Suppose there exists a constant $c \in \mathbb{R}$, such that both $\sigma_1(x) = \mathrm{Re}[\sigma_\times(x + c\mathrm{i})]$ and $\sigma_2(x) = \mathrm{Im}[\sigma_\times(x + c\mathrm{i})]$ are continuous almost everywhere and not polynomials almost everywhere. Then for any $\varepsilon > 0$, there exists an* R-FTNet *$f_{\times,\mathrm{R}}$, such that*

$$\|f_{\times,\mathrm{R}} - f_{\mathrm{D}}\|_{L^\infty(K^T)} \leqslant \varepsilon.$$

Theorem 2 shows that R-FTNet is a universal approximator. The requirement of convex domain K is trivial since it is always possible to find a convex domain including a given compact domain. The conditions of the activation function are satisfied by many commonly used activation functions, such as modReLU, zReLU, and $\mathbb{C}$ReLU. Existing studies investigate the universal approximation of RNN, and the most general condition uses sigmoidal activation functions [46]. We extend the results to complex-weighted networks and generalize the requirement of activation functions.

Proof of Theorem 2. We start our proof with the universal approximation of an intermediate network, named additive FTNet. One-hidden-layer additive FTNet with H_+ hidden neurons can be viewed as a mapping $f_{+,\mathrm{R}}$, defined by

$$\begin{aligned}
&f_{+,\mathrm{R}} : \boldsymbol{x}_{1:T} \mapsto (y_{+,1}, \dots, y_{+,T}),\\
&\boldsymbol{p}_t = \sigma_1(\mathbf{A}\boldsymbol{x}_t + \mathbf{B}\boldsymbol{q}_{t-1} - \boldsymbol{\zeta}),\\
&\boldsymbol{q}_t = \sigma_2(\mathbf{A}\boldsymbol{x}_t + \mathbf{B}\boldsymbol{q}_{t-1} - \boldsymbol{\zeta}),\\
&y_{+,t} = \boldsymbol{\alpha}_+^\top \boldsymbol{p}_t, \quad t \in [T],
\end{aligned} \tag{8}$$

where $\mathbf{A} \in \mathbb{R}^{H_+ \times I}, \mathbf{B} \in \mathbb{R}^{H_+ \times H_+}, \boldsymbol{\alpha}_+, \boldsymbol{\zeta} \in \mathbb{R}^{H_+}$ indicate weight parameters, and $\boldsymbol{p}_t, \boldsymbol{q}_t \in \mathbb{R}^{H_+}$ denote hidden states. We claim that there exists an additive FTNet $f_{+,\mathrm{R}}$, such that

$$\|f_{+,\mathrm{R}} - f_{\mathrm{D}}\|_{L^\infty(K^T)} \leqslant \varepsilon. \tag{9}$$

This claim indicates the universal approximation of additive FTNet. The proof of Eq. (9) is similar to that of the universal approximation of RNN and is provided in Appendix.

Based on Eq. (9), it suffices to prove that any additive FTNet using induced activation functions σ_1 and σ_2 is equivalent to an R-FTNet using the activation function σ.

Firstly, provided an additive FTNet, the R-FTNet with $H_\times = I + H_+ + 1$ hidden neurons is constructed as follows

$$\mathbf{W}_\times = \begin{bmatrix} \mathbf{0}^{I\times I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} & -c\mathbf{1} \\ \mathbf{0} & \mathbf{0} & \mathbf{0}^{1\times 1} \end{bmatrix}, \boldsymbol{r}_0 = \begin{bmatrix} \mathbf{0} \\ \boldsymbol{q}_0 \\ \mathbf{0} \end{bmatrix}, \mathbf{V}_\times = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{A} & \mathbf{0} & -\boldsymbol{\zeta} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}, \boldsymbol{\alpha}_\times = \begin{bmatrix} \mathbf{0} \\ \boldsymbol{\alpha}_+ \\ \mathbf{0} \end{bmatrix}. \tag{10}$$

Secondly, we calculate the receptor, stimulus, and output of the above R-FTNet. We prove $\boldsymbol{r}_t = [\mathbf{0}^{I\times 1}; \boldsymbol{q}_t; \mathbf{0}^{1\times 1}]$ by mathematical induction as follows.

1. For $t = 0$, the conclusion holds according to Eq. (10).
2. Suppose that the conclusion holds for $t = \tau$ with $\tau \leqslant T - 1$. Thus, one has

$$\begin{aligned} \boldsymbol{r}_{\tau+1} &= \mathrm{Im}\left[\sigma_\times\left(\mathbf{0}; c\mathbf{1} + (\mathbf{A}\boldsymbol{x}_{\tau+1} + \mathbf{B}\boldsymbol{q}_\tau - \boldsymbol{\zeta})\,\mathrm{i}; \mathbf{0}\right)\right] \\ &= [\mathbf{0}; \sigma_2(\mathbf{A}\boldsymbol{x}_{\tau+1} + \mathbf{B}\boldsymbol{q}_\tau - \boldsymbol{\zeta}); \mathbf{0}] \\ &= [\mathbf{0}; \boldsymbol{q}_{\tau+1}; \mathbf{0}]\,, \end{aligned}$$

   where the first equality holds from Eqs. (1), (10) and the induction hypothesis, the second equality holds based on the definition of the activation function σ_2 in Theorem 2, and the third equality holds according to Eq. (8). Thus, the conclusion holds for $t = \tau + 1$.

For any $t \in [T]$, the stimulus satisfies

$$\begin{aligned} \boldsymbol{s}_t &= \mathrm{Re}\left[\sigma_\times\left(\mathbf{0}; c\mathbf{1} + (\mathbf{A}\boldsymbol{x}_t + \mathbf{B}\boldsymbol{q}_{t-1} - \boldsymbol{\zeta})\,\mathrm{i}; \mathbf{0}\right)\right] \\ &= [\mathbf{0}; \sigma_2(c\mathbf{1}, \mathbf{A}\boldsymbol{x}_t + \mathbf{B}\boldsymbol{q}_{t-1} - \boldsymbol{\zeta}); \mathbf{0}] \\ &= [\mathbf{0}; \boldsymbol{p}_t; \mathbf{0}]\,, \end{aligned}$$

which leads to $y_{\times,t} = \boldsymbol{\alpha}_\times^\top \boldsymbol{s}_t = \boldsymbol{\alpha}_+^\top \boldsymbol{p}_t = y_{+,t}$. Therefore, the R-FTNet defined by Eq. (10) has the same output as the additive FTNet defined by Eq. (8), i.e.,

$$f_{\times,\mathrm{R}}(\boldsymbol{x}_{1:T}) = f_{+,\mathrm{R}}(\boldsymbol{x}_{1:T}), \quad \forall\ \boldsymbol{x}_{1:T} \in K^T. \tag{11}$$

Finally, from Eqs. (9) and (11), one has

$$\|f_{\times,\mathrm{R}}(\boldsymbol{x}_{1:T}) - f_{\mathrm{D}}(\boldsymbol{x}_{1:T})\|_{L^\infty(K^T)} = \|f_{+,\mathrm{R}}(\boldsymbol{x}_{1:T}) - f_{\mathrm{D}}(\boldsymbol{x}_{1:T})\|_{L^\infty(K^T)} \leqslant \varepsilon,$$

which completes the proof. □

5. Approximation Complexity

We show the approximation advantage of FTNet over FNN and RNN in Subsection 5.1 and provide worst-case guarantees in Subsection 5.2. Let us introduce the $(\varepsilon, \mathcal{D})$-approximation, which is used throughout this section.

Definition 2 *Let g be a function from $\mathbb{R}^I$ to $\mathbb{R}$, $\mathcal{F}$ is a class of functions from $\mathbb{R}^I$ to $\mathbb{R}$, and $\mathcal{D}$ is a distribution over $\mathbb{R}^I$. The function g can be $(\varepsilon, \mathcal{D})$-approximated by function class $\mathcal{F}$ if there exists a function $f \in \mathcal{F}$, such that*

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[(g(\boldsymbol{x}) - f(\boldsymbol{x}))^2\right] \leqslant \varepsilon.$$

The $(\varepsilon, \mathcal{D})$-approximation means that the minimal expected squared difference between a function from the function class $\mathcal{F}$ and the target function g is small. Let $\mathcal{F}$ be the function space of a neural network, and g is the learning target. Then the $(\varepsilon, \mathcal{D})$-approximation indicates that it is possible to find a set of parameters for the neural network, such that the neural network suffers a negligible loss under the task of learning g.

5.1. *Approximation-Complexity Advantage of* FTNet

We now present two theorems showing the separation results between FTNet and FNN/RNN, respectively.

Theorem 3 *There exist constants $I_1 \in \mathbb{N}^+$, $\varepsilon_1 > 0$, and $c_1 > 0$, such that for any input dimension $I \geqslant I_1$, there exist a distribution $\mathcal{D}_1$ over $\mathbb{R}^I$ and a function $f_1 : \mathbb{R}^I \to \mathbb{R}$, s.t.*

1. *For any $\varepsilon > 0$, the target function f_1 can be $(\varepsilon, \mathcal{D}_1)$-approximated by one-hidden-layer* F-FTNet *with at most* $\max\{3c_1^2 I^{15/2}/\varepsilon^2, 27I^2\}$ *parameters using the* zReLU *activation function.*
2. *The target function f_1 cannot be $(\varepsilon_1, \mathcal{D}_1)$-approximated by one-hidden-layer* FNN *with at most $\varepsilon_1 \mathrm{e}^{\varepsilon_1 I}$ parameters using the* ReLU *activation function.*

Theorem 4 *There exist constants $I_2 \in \mathbb{N}^+$, $\varepsilon_2 > 0$, and $c_2 > 0$, such that for any input dimension $I \geqslant I_2$, there exist a distribution $\mathcal{D}_2$ over $\mathbb{R}^I$ and a* DODS *$f_{\mathrm{D}} : \mathbb{R}^{IT} \to \mathbb{R}^T$, s.t.*

1. *For any $\varepsilon > 0$, the* DODS f_{D} *can be* $(T\varepsilon, \mathcal{D}_2^T)$*-approximated by one-hidden-layer* R-FTNet *with at most* $3(c_2 I^{15/4}/\varepsilon + 3I)^2$ *parameters using the* zReLU *activation function.*

2. *The* DODS f_{D} *cannot be* $(T\varepsilon_2, \mathcal{D}_2^T)$*-approximated by one-hidden-layer* RNN *with at most* $\varepsilon_0 \mathrm{e}^{\varepsilon_2 I}/4$ *parameters using the* ReLU *activation function.*

Theorems 3 and 4 show the approximation-complexity advantage of FTNet over FNN and RNN, respectively, i.e., there exists a target function such that FTNet can express it with polynomial parameters, but FNN or RNN cannot approximate it unless exponential parameters are used. Previous studies usually demonstrate separation results between deep networks and shallow networks [13, 45, 50]. A recent study shows exponential separation between one-hidden-layer CRNet and one-hidden-layer FNN [56]. Our results consider both feedforward and recurrent architectures and demonstrate the advantage of FTNet by showing that it is sufficient for one-hidden-layer FTNet to possess exponential separation over FNN and RNN. We begin our proof by introducing the complex-reaction network (CRNet) and an important lemma.

The complex-reaction network (CRNet) is a recently proposed neural network with complex-valued operations [56]. The real-valued input vector $\boldsymbol{x} = (x_1; x_2; \ldots; x_I) \in \mathbb{R}^I$ is folded by a transformation mapping $\tau : \mathbb{R}^I \to \mathbb{C}^{I/2}$ to form a complex-valued vector, i.e.,

$$\tau : \boldsymbol{x} \mapsto \left(x_1; x_2; \ldots; x_{I/2}\right) + \left(x_{I/2+1}; x_{I/2+2}; \ldots; x_{I/2}\right) \mathrm{i},$$

where the input dimension I is assumed to be an even number without loss of generality. Recalling the formulation of CRNet [56], one-hidden-layer CRNet with H_{C} hidden neurons is a mapping $f_{\mathrm{C}} : \mathbb{R}^I \to \mathbb{R}$ of the following form

$$f_{\mathrm{C}} : \boldsymbol{x} \mapsto \mathrm{Re}\left[\boldsymbol{\alpha}_{\mathrm{C}}^\top \sigma_{\mathrm{C}}\left(\mathbf{W}_{\mathrm{C}} \tau(\boldsymbol{x}) + \boldsymbol{b}_{\mathrm{C}}\right)\right], \tag{12}$$

where $\mathbf{W}_{\mathrm{C}} \in \mathbb{C}^{H_{\mathrm{C}} \times d}$, $\boldsymbol{b}_{\mathrm{C}} \in \mathbb{C}^{H_{\mathrm{C}}}$, $\boldsymbol{\alpha}_{\mathrm{C}} \in \mathbb{C}^{H_{\mathrm{C}}}$ indicate weight parameters, and $\sigma_{\mathrm{C}} : \mathbb{C} \to \mathbb{C}$ is a complex-valued activation function applied componentwise.

Lemma 2 *[13, Theorem 1], [56, Theorem 2] There exist constants $I_0 \in \mathbb{N}^+$, $\varepsilon_0 > 0$, and $c_0 > 0$, such that for any input dimension $I \geqslant I_0$, there exist a distribution $\mathcal{D}_0$ over $\mathbb{R}^I$ and a function $f_0 : \mathbb{R}^I \to \mathbb{R}$, such that*

1. *For any* $\varepsilon > 0$, f_0 *can be* $(\varepsilon, \mathcal{D}_0)$*-approximated by one-hidden-layer* CRNet *with at most* $c_0 I^{19/4}/\varepsilon$ *parameters using the* zReLU *activation function.*

2. *The function* f_0 *cannot be* $(\varepsilon_0, \mathcal{D}_0)$*-approximated by one-hidden-layer* FNN *with at most* $\varepsilon_0 \mathrm{e}^{\varepsilon_0 I}$ *parameters using the* ReLU *activation function.*

Lemma 2 indicates the approximation-complexity advantage of CRNet over FNN, i.e., there exists a target function such that CRNet can express it with polynomial parameters, but FNN cannot express it unless exponential parameters are used.

Proof of Theorem 3. Let $I_1 = I_0$, $\varepsilon_1 = \varepsilon_0$, and $c_1 = c_0$, where I_0, ε_0, and c_0 are defined in Lemma 2. For any $I \geqslant I_1$, let $\mathcal{D}_1 = \mathcal{D}$ and $f_1 = f_0$. Without loss of generality, let the input dimension I be an even number.

Firstly, we prove that F-FTNet can approximate the target function f_1 using polynomial parameters. Recalling the definition of CRNet in Eq. (12), we define

$$\begin{aligned} &\mathbf{W}_{\mathrm{C}} = \mathbf{W}_{\mathrm{C},R} + \mathbf{W}_{\mathrm{C},I}\mathrm{i}, \\ &\boldsymbol{b}_{\mathrm{C}} = \boldsymbol{b}_{\mathrm{C},R} + \boldsymbol{b}_{\mathrm{C},I}\mathrm{i}, \\ &\boldsymbol{\alpha}_{\mathrm{C}} = \boldsymbol{\alpha}_{\mathrm{C},R} + \boldsymbol{\alpha}_{\mathrm{C},I}\mathrm{i}, \end{aligned} \tag{13}$$

where $\mathbf{W}_{\mathrm{C},R}, \mathbf{W}_{\mathrm{C},I} \in \mathbb{R}^{H_{\mathrm{C}} \times I/2}$, $\boldsymbol{b}_{\mathrm{C},R}, \boldsymbol{b}_{\mathrm{C},I} \in \mathbb{R}^{H_{\mathrm{C}}}$, and $\boldsymbol{\alpha}_{\mathrm{C},R}, \boldsymbol{\alpha}_{\mathrm{C},I} \in \mathbb{R}^{H_{\mathrm{C}}}$ are real-valued parameters. The rest of the proof is divided into several steps.

Step 1. We construct an F-FTNet with the same output as a given CRNet. The F-FTNet with $H_\times = \max\{2H_{\mathrm{C}}, I+1\}$ hidden neurons is constructed as follows

$$\mathbf{W}_\times = \begin{bmatrix} \mathbf{W}_{\mathrm{C},R} & -\mathbf{W}_{\mathrm{C},I} & \mathbf{0} & \boldsymbol{b}_{\mathrm{C},R} \\ \mathbf{W}_{\mathrm{C},I} & \mathbf{W}_{\mathrm{C},R} & \mathbf{0} & \boldsymbol{b}_{\mathrm{C},I} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad \mathbf{V}_\times = \begin{bmatrix} \mathbf{W}_{\mathrm{C},I} & \mathbf{W}_{\mathrm{C},R} & \mathbf{0} & \boldsymbol{b}_{\mathrm{C},I} \\ \mathbf{W}_{\mathrm{C},R} & -\mathbf{W}_{\mathrm{C},I} & \mathbf{0} & \boldsymbol{b}_{\mathrm{C},R} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix},$$

$$\boldsymbol{r}_0 = \mathbf{0}, \quad \boldsymbol{\alpha}_\times = \left[\boldsymbol{\alpha}_{\mathrm{C},R}; -\boldsymbol{\alpha}_{\mathrm{C},I}; \mathbf{0}\right],$$

and $\sigma_\times$ is the zReLU activation function. The output of the constructed F-FTNet above satisfies

$$\begin{aligned} f_{\times,\mathrm{F}}(\boldsymbol{x}) &= \boldsymbol{\alpha}_\times^\top \mathrm{Re}\left[\sigma_\times\left(\left(\mathbf{W}_\times + \mathbf{V}_\times \mathrm{i}\right)\left(\kappa(\boldsymbol{x}, H_\times) + \boldsymbol{r}_0\mathrm{i}\right)\right)\right] \\ &= \boldsymbol{\alpha}_\times^\top \mathrm{Re}\left[\sigma_{\mathrm{C}}\left(\left[\mathbf{W}_{\mathrm{C}}\tau(\boldsymbol{x}) + \boldsymbol{b}_{\mathrm{C}}; \overline{\mathbf{W}_{\mathrm{C}}\tau(\boldsymbol{x}) + \boldsymbol{b}_{\mathrm{C}}}\mathrm{i}; \mathbf{0}\right]\right)\right] \\ &= \boldsymbol{\alpha}_{\mathrm{C},R}^\top \mathrm{Re}\left[\sigma_{\mathrm{C}}\left(\mathbf{W}_{\mathrm{C}}\tau(\boldsymbol{x}) + \boldsymbol{b}_{\mathrm{C}}\right)\right] - \boldsymbol{\alpha}_{\mathrm{C},I}^\top \mathrm{Re}\left[\sigma_{\mathrm{C}}\left(\overline{\mathbf{W}_{\mathrm{C}}\tau(\boldsymbol{x}) + \boldsymbol{b}_{\mathrm{C}}}\mathrm{i}\right)\right], \end{aligned} \tag{14}$$

It is observed that

$$\mathrm{Re}\left[\sigma_{\mathrm{C}}(x+y\mathrm{i})\right] = \mathrm{Im}\left[\sigma_{\mathrm{C}}(\overline{x+y\mathrm{i}}\ \mathrm{i})\right] = \begin{cases} x, & \text{if } xy \geqslant 0, \\ 0, & \text{otherwise.} \end{cases}$$

Thus, one has

$$\begin{aligned} f_{\times,\mathrm{F}}(\boldsymbol{x}) &= \mathrm{Re}\left[(\boldsymbol{\alpha}_{\mathrm{C},R} + \boldsymbol{\alpha}_{\mathrm{C},I}\mathrm{i})^{\top} \sigma_{\mathrm{C}}\left(\mathbf{W}_{\mathrm{C}}\tau(\boldsymbol{x}) + \boldsymbol{b}_{\mathrm{C}}\right)\right] \\ &= \mathrm{Re}\left[\boldsymbol{\alpha}_{\mathrm{C}}^{\top}\sigma_{\mathrm{C}}\left(\mathbf{W}_{\mathrm{C}}\tau(\boldsymbol{x}) + \boldsymbol{b}_{\mathrm{C}}\right)\right] \\ &= f_{\mathrm{C}}(\boldsymbol{x}), \end{aligned} \tag{15}$$

which indicates that any CRNet with hidden size H_{C} can be expressed by an F-FTNet with hidden size $\max\{2H_{\mathrm{C}}, I+1\}$.

Step 2. We bound the number of required parameters in the constructed F-FTNet. From Lemma 2, for any $\varepsilon > 0$, there exists CRNet f_{C} with at most $(c_1 I^{19/4})/\varepsilon$ parameters using the zReLU activation function, such that

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\left(f_{\mathrm{C}}(\boldsymbol{x}) - f_1(\boldsymbol{x})\right)^2\right] \leqslant \varepsilon. \tag{16}$$

For CRNet with H_{C} hidden neurons, it has $2H_{\mathrm{C}}(I+2)$ parameters. Thus, the hidden size of CRNet satisfies

$$H_{\mathrm{C}} \leqslant \frac{c_1 I^{19/4}}{2(I+2)\varepsilon} \leqslant \frac{c_1 I^{15/4}}{2\varepsilon}.$$

According to Step 2, there exists an F-FTNet, with no more than $\max\{2H_{\mathrm{C}}, I+1\}$ hidden neurons, satisfying $f_{\times,\mathrm{F}}(\boldsymbol{x}) = f_{\mathrm{C}}(\boldsymbol{x})$. This property, together with Eq. (16), indicates that

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[\left(f_{\times,\mathrm{F}}(\boldsymbol{x}) - f_1(\boldsymbol{x})\right)^2\right] \leqslant \varepsilon,$$

and that the number of hidden neurons in the constructed F-FTNet $f_{\times,\mathrm{F}}$ is no more than

$$H_{\times} \leqslant \max\left\{2H_{\mathrm{C}}, I+1\right\} \leqslant \max\left\{c_1 I^{15/4}/\varepsilon, I+1\right\}.$$

For F-FTNet with $H_{\times}$ hidden neurons, it has $2H_{\times}^2 + H_{\times}$ parameters. Thus, the number of parameters in the constructed F-FTNet $f_{\times,\mathrm{F}}$ is no more than

$$2H_{\times}^2 + H_{\times} \leqslant 3H_{\times}^2 \leqslant \max\left\{3c_1^2 I^{15/2}/\varepsilon^2, 27I^2\right\}.$$

Secondly, Lemma 2 indicates that FNN needs at least exponential parameters to approximate the target function f_1.

Combining the conclusions above completes the proof. $\square$

Proof of Theorem 4. Let $I_2 = I_0$, $\varepsilon_2 = \varepsilon_0$, and $c_2 = c_0$, where I_0, ε_0, and c_0 are defined in Lemma 2. For any $I \geqslant I_2$, let $\mathcal{D}_2 = \mathcal{D}$. Without loss of generality, let the input dimension I be an even number. The DODS is constructed as follows. For any input $\boldsymbol{x} \in \mathbb{R}^I$ and hidden state $\boldsymbol{h} \in \mathbb{R}^{H_\mathrm{D}}$, let $\varphi(\boldsymbol{x}, \boldsymbol{h}) = \boldsymbol{x}$ and $\psi(\boldsymbol{h}) = f_0(\boldsymbol{h})$, where f_0 is the same function as that in Lemma 2. Thus, the output at time t is

$$y_t = \psi(\boldsymbol{h}_t) = \psi(\varphi(\boldsymbol{x}_t, \boldsymbol{h}_{t-1})) = f_0(\boldsymbol{x}_t), \tag{17}$$

which holds according to Eq. (5).

Firstly, we prove that R-FTNet can express f_D using polynomial parameters. The proof is divided into several steps.

Step 1. We construct an R-FTNet with the same output as a given CRNet. From the proof of Theorem 3, for any $\varepsilon > 0$, there exists a CRNet f_C with at most $H_\mathrm{C} = (c_1 I^{15/4})/(2\varepsilon)$ hidden neurons using the zReLU activation function, such that

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[(f_\mathrm{C}(\boldsymbol{x}) - f_0(\boldsymbol{x}))^2\right] \leqslant \varepsilon. \tag{18}$$

Let $\mathbf{W}_{\mathrm{C},R}$, $\mathbf{W}_{\mathrm{C},I}$, $\boldsymbol{b}_{\mathrm{C},R}$, $\boldsymbol{b}_{\mathrm{C},I}$, $\boldsymbol{\alpha}_{\mathrm{C},R}$, and $\boldsymbol{\alpha}_{\mathrm{C},I}$ be the real-valued weight matrices of the above CRNet, which are defined in the same way as those in Eq. (13). Define the R-FTNet $f_{\times,\mathrm{R}}$ with $H_\times = 2H_\mathrm{C} + I + 1$ hidden neurons as follows

$$\mathbf{W}_\times = \begin{bmatrix} \mathbf{0}^{I\times I/2} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{W}_{\mathrm{C},R} & -\mathbf{W}_{\mathrm{C},I} & \mathbf{0} & \boldsymbol{b}_{\mathrm{C},R} \\ \mathbf{W}_{\mathrm{C},I} & \mathbf{W}_{\mathrm{C},R} & \mathbf{0} & \boldsymbol{b}_{\mathrm{C},I} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad \mathbf{V}_\times = \begin{bmatrix} \mathbf{0}^{I\times I/2} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{W}_{\mathrm{C},I} & \mathbf{W}_{\mathrm{C},R} & \mathbf{0} & \boldsymbol{b}_{\mathrm{C},I} \\ \mathbf{W}_{\mathrm{C},R} & -\mathbf{W}_{\mathrm{C},I} & \mathbf{0} & \boldsymbol{b}_{\mathrm{C},R} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix},$$

$$\boldsymbol{r}_0 = \mathbf{0}, \quad \boldsymbol{\alpha}_\times = [\mathbf{0}; \boldsymbol{\alpha}_{\mathrm{C},R}; -\boldsymbol{\alpha}_{\mathrm{C},I}; \mathbf{0}],$$

and $\sigma_\times$ is the zReLU activation function. We then prove that the output of the above R-FTNet is the same as that of the CRNet in Eq. (18). Since the first I rows and the last row in $\mathbf{W}_\times$ and $\mathbf{V}_\times$ are all 0, one has $\boldsymbol{r}_t = [\mathbf{0}_{I\times 1}; \tilde{\boldsymbol{r}}_t; \mathbf{0}_{1\times 1}]$ for any $t \in [T]$, where $\tilde{\boldsymbol{r}}_t \in \mathbb{R}^{2H_\mathrm{C}}$ is an arbitrary vector. Then the output of the R-FTNet at time t is

$$\begin{aligned} y_{\times,t} =& \boldsymbol{\alpha}_\times^\top \mathrm{Re}\left[\sigma_\times\left((\mathbf{W}_\times + \mathbf{V}_\times \mathrm{i})\left(\kappa(\boldsymbol{x}_t, H_\times) + \boldsymbol{r}_{t-1}\mathrm{i}\right)\right)\right] \\ =& \boldsymbol{\alpha}_{\mathrm{C},R}^\top \mathrm{Re}\left[\sigma_\mathrm{C}\left(\mathbf{W}_\mathrm{C}\tau(\boldsymbol{x}_t) + \boldsymbol{b}_\mathrm{C}\right)\right] - \boldsymbol{\alpha}_{\mathrm{C},I}^\top \mathrm{Re}\left[\sigma_\mathrm{C}\left(\overline{\mathbf{W}_\mathrm{C}\tau(\boldsymbol{x}_t) + \boldsymbol{b}_\mathrm{C}}\mathrm{i}\right)\right]. \end{aligned}$$

The right-hand side of the above equation is the same as that of Eq. (14), except substituting $\boldsymbol{x}$ with $\boldsymbol{x}_t$. By similar derivation used in Eq. (15), one has

$$y_{\times,t} = f_{\mathrm{C}}(\boldsymbol{x}_t), \quad \forall\, t \in [T]. \tag{19}$$

Step 2. We prove that the R-FTNet constructed in Step 1 can approximate DODS f_{D} with a small expected squared loss and then bound the number of required parameters. The expected squared loss of the above R-FTNet is

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{x}_{1:T}\sim\mathcal{D}_2^T}\left[\|f_{\times,\mathrm{R}}(\boldsymbol{x}_{1:T}) - f_{\mathcal{D}}(\boldsymbol{x}_{1:T})\|^2\right] &= \mathbb{E}_{\boldsymbol{x}_{1:T}\sim\mathcal{D}_2^T}\left[\sum_{t=1}^{T}(y_{\times,t} - y_t)^2\right] \\
&= \mathbb{E}_{\boldsymbol{x}_{1:T}\sim\mathcal{D}_2^T}\left[\sum_{t=1}^{T}(f_{\mathrm{C}}(\boldsymbol{x}_t) - f_0(\boldsymbol{x}_t))^2\right] \\
&\leqslant T\varepsilon,
\end{aligned}$$

where the second equality holds from Eqs. (17), (19), and the first inequality holds based on Eq. (18). We then calculate the number of parameters in the above R-FTNet. Since FTNet with hidden size $H_\times$ has $2H_\times^2 + H_\times$ parameters, the number of parameters in the constructed R-FTNet is no more than

$$2H_\times^2 + H_\times \leqslant 3H_\times^2 \leqslant 3\left(c_1 I^{15/4}/\varepsilon + 3I\right)^2 .$$

Secondly, we prove that RNN needs at least exponential parameters to approximate the target DODS f_{D}. The proof is divided into several steps.

Step 1. We prove that if the total loss suffered by RNN is large, there exists a time point $t \in [T]$, such that RNN suffers a large loss at time t. For the unity of notations, we rewrite the one-hidden-layer RNN f_{R} with H_{R} hidden neurons as follows

$$\begin{aligned}
&f_{\mathrm{R}} : \boldsymbol{x}_{1:T} \mapsto (y_{\mathrm{R},1}, \dots, y_{\mathrm{R},T}), \\
&\boldsymbol{m}_t = \sigma_{\mathrm{R}}\left(\mathbf{W}_{\mathrm{R}}\boldsymbol{x}_t + \mathbf{V}_{\mathrm{R}}\boldsymbol{m}_{t-1} - \boldsymbol{\zeta}_{\mathrm{R}}\right), \\
&y_{\mathrm{R},t} = \boldsymbol{\alpha}_{\mathrm{R}}^\top \boldsymbol{m}_t, \quad \text{for} \quad t \in [T],
\end{aligned} \tag{20}$$

where $\boldsymbol{m}_t \in \mathbb{R}^{H_{\mathrm{R}}}$ and $y_{\mathrm{R},t} \in \mathbb{R}$ represent the memory and output at time t, respectively, $\mathbf{W}_{\mathrm{R}}$, $\mathbf{V}_{\mathrm{R}}$, $\boldsymbol{\zeta}_{\mathrm{R}}$, $\boldsymbol{\alpha}_{\mathrm{R}}$ denote weight parameters, and σ_{R} is the ReLU activation function applied componentwise. If the DODS can be $(T\varepsilon_0, \mathcal{D}_2^T)$-approximated by RNN, the following holds from Definition 2,

$$T\varepsilon_0 \geqslant \mathbb{E}_{\boldsymbol{x}_{1:T}\sim\mathcal{D}_2^T}\left[\|f_{\mathrm{R}}(\boldsymbol{x}_{1:T}) - f_{\mathrm{D}}(\boldsymbol{x}_{1:T})\|^2\right] = \mathbb{E}_{\boldsymbol{x}_{1:T}\sim\mathcal{D}_2^T}\left[\sum\nolimits_{t=1}^{T}(y_{\mathrm{R},t} - y_t)^2\right].$$

Since for any time $t \in [T]$, the squared term $(y_{\mathrm{R},t} - y_t)^2$ is always non-negative, there exists time $t_0 \in [T]$, such that $\mathbb{E}_{\boldsymbol{x}_{t_0} \sim \mathcal{D}_2}[(y_{\mathrm{R},t_0} - y_{t_0})^2] \leqslant \varepsilon_0$. According to the definitions of y_{R,t_0} and y_{t_0} in Eqs. (20) and (17), one has

$$\mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}_2}\left[\left(\boldsymbol{\alpha}_{\mathrm{R}}^{\top}\sigma_{\mathrm{R}}(\mathbf{W}_{\mathrm{R}}\boldsymbol{x} + \mathbf{V}_{\mathrm{R}}\boldsymbol{m}_{t_0-1} + \boldsymbol{\theta}_{\mathrm{R}}) - f_0(\boldsymbol{x})\right)^2\right] \leqslant \varepsilon_0. \tag{21}$$

Step 2. We use Lemma 2 to give a lower bound on the number of parameters of RNN with small loss. Before the proof, we rewrite the FNN in the mapping form for the unity of notation. One-hidden-layer FNN with H_{F} hidden neurons can be viewed as a mapping f_{F}, defined by

$$f_{\mathrm{F}} : \boldsymbol{x} \mapsto \boldsymbol{\alpha}_{\mathrm{R}}^{\top}\sigma_{\mathrm{F}}\left(\mathbf{W}_{\mathrm{F}}\boldsymbol{x} - \boldsymbol{\zeta}_{\mathrm{F}}\right), \tag{22}$$

where $\boldsymbol{x} \in \mathbb{R}^I$ represents input at time t, $\mathbf{W}_{\mathrm{F}} \in \mathbb{R}^{H_{\mathrm{F}} \times I}$, $\boldsymbol{\zeta}_{\mathrm{F}}$, $\boldsymbol{\alpha}_{\mathrm{F}} \in \mathbb{R}^{H_{\mathrm{F}}}$ denote weight parameters, and σ_{F} is the ReLU activation function applied componentwise. We now construct an FNN equivalent to the RNN at time t_0 as follows. Let $\boldsymbol{\alpha}_{\mathrm{F}} = \boldsymbol{\alpha}_{\mathrm{R}}$, $\mathbf{W}_{\mathrm{F}} = \mathbf{W}_{\mathrm{R}}$, and $\boldsymbol{b}_{\mathrm{F}} = \mathbf{V}_{\mathrm{R}}\boldsymbol{m}_{t_0-1} + \boldsymbol{\theta}_{\mathrm{R}}$. From Eq. (21), one has $\mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}_2}[(f_{\mathrm{F}}(\boldsymbol{x}) - f_0(\boldsymbol{x}))^2] \leqslant \varepsilon_0$. According to Lemma 2, the number of parameters of f_{F} is at least $\varepsilon_0 \mathrm{e}^{\varepsilon_0 I}$. For FNN with H_{F} hidden neurons, it has $2H_{\mathrm{F}}(I+1)$ parameters. Thus, the hidden size of the FNN satisfies

$$H_{\mathrm{F}} \geqslant \frac{\varepsilon_0 \mathrm{e}^{\varepsilon_0 I}}{2(I+1)} \geqslant \frac{\varepsilon_0 \mathrm{e}^{\varepsilon_0 I}}{4I}.$$

For RNN with H_{R} hidden neurons, it has $H_{\mathrm{R}}(I + H_{\mathrm{R}} + 2)$ parameters. Since the above FNN has the same hidden size as the RNN, i.e., $H_{\mathrm{R}} = H_{\mathrm{F}}$, one knows that the number of parameters of RNN satisfies

$$H_{\mathrm{R}}(I + H_{\mathrm{R}} + 2) > H_{\mathrm{R}} I = H_{\mathrm{F}} I \geqslant \frac{\varepsilon_0 \mathrm{e}^{\varepsilon_0 I}}{4}.$$

Combining the conclusions above completes the proof. □

5.2. Worst-Case Guarantee of FTNet

We proceed to provide the worst-case guarantees of approximation complexity for F-FTNet and R-FTNet.

Theorem 5 *Let f be a function from $\mathbb{R}^I$ to $\mathbb{R}$, and $\mathcal{D}$ is a distribution over $\mathbb{R}^I$. For any $\varepsilon > 0$, if f can be $(\varepsilon, \mathcal{D})$-approximated by one-hidden-layer* FNN *with hidden size H_{F} using the* ReLU *activation function, then f can be $(\varepsilon, \mathcal{D})$-approximated by one-hidden-layer* F-FTNet$_{\times}$ *with hidden size* $\max\{H_{\mathrm{F}}, I+1\}$ *using the* zReLU *activation function.*

Theorem 6 *Let f_{D} be a* DODS *from $\mathbb{R}^{IT}$ to $\mathbb{R}^{T}$, and $\mathcal{D}$ is a distribution over $\mathbb{R}^{TI}$. For any $\varepsilon > 0$, if f_{D} can be $(\varepsilon, \mathcal{D})$-approximated by one-hidden-layer* RNN *with hidden size H_{R} using the* ReLU *activation function, then f_{D} can be $(\varepsilon, \mathcal{D})$-approximated by one-hidden-layer* R-FTNet$_\times$ *with $2H_{\mathrm{R}} + I + 1$ hidden neurons using the* zReLU *activation function.*

Theorems 5 and 6 provide the worst-case guarantees for FTNet, saying that the disadvantages of FTNet over FNN and RNN are no more than constants. Previous studies only provide separation advantages of model A over model B when expressing particular functions [13, 45, 50, 38, 56], without considering the opposite problem, i.e., whether model B possesses separation advantages over model A when approximating other functions. To our knowledge, our work is the first one to realize the opposite problem and provide a negative answer.

Proof of Theorem 5. Since f can be $(\varepsilon, \mathcal{D})$-approximated by FNN, there exists an FNN defined by Eq. (22), such that

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\left[(f(\boldsymbol{x}) - f_{\mathrm{F}}(\boldsymbol{x}))^2\right] \leqslant \varepsilon. \tag{23}$$

Firstly, an F-FTNet$_\times$ $f_{\times,\mathrm{F}}$ with $H_\times = \max\{H_{\mathrm{F}}, I+1\}$ hidden neurons is constructed as follows

$$\mathbf{W}_\times = \begin{bmatrix} \mathbf{W}_{\mathrm{F}} & \mathbf{0} & \boldsymbol{b}_{\mathrm{F}} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad \mathbf{V}_\times = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{1} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad \boldsymbol{r}_0 = \mathbf{0}, \quad \boldsymbol{\alpha}_\times = \begin{bmatrix} \boldsymbol{\alpha}_{\mathrm{F}} \\ \mathbf{0} \end{bmatrix},$$

and $\sigma_\times$ is the zReLU activation function.

Secondly, we prove that the output of the above F-FTNet$_\times$ is the same as that of the FNN in Eq. (23). For any input $\boldsymbol{x} \in \mathbb{R}^I$, the output of the F-FTNet$_\times$ is

$$\begin{aligned} f_{\times,\mathrm{F}}(\boldsymbol{x}) &= \boldsymbol{\alpha}_\times^\top \mathrm{Re}\left[\sigma_\times\left((\mathbf{W}_\times + \mathbf{V}_\times \mathrm{i})\left(\kappa(\boldsymbol{x}, H_\times) + \boldsymbol{r}_0 \mathrm{i}\right)\right)\right] \\ &= \boldsymbol{\alpha}_\times^\top \mathrm{Re}\left[\sigma_\times\left([(\mathbf{W}_{\mathrm{F}}\boldsymbol{x} + \boldsymbol{b}_{\mathrm{F}}) + \mathbf{1}\mathrm{i}; \mathbf{0}]\right)\right] \\ &= \boldsymbol{\alpha}_{\mathrm{F}}^\top \mathrm{Re}\left[\sigma_\times\left(\mathbf{W}_{\mathrm{F}}\boldsymbol{x} + \boldsymbol{b}_{\mathrm{F}} + \mathbf{1}\mathrm{i}\right)\right], \end{aligned}$$

where the first equality holds based on Eq. (3), the second and third equalities hold from the construction of F-FTNet$_\times$ in Step 1. Recalling the definition of the zReLU activation function in Eq. (4), one has

$$
\begin{aligned}
f_{\times,\mathrm{F}}(\boldsymbol{x}) &= \boldsymbol{\alpha}_{\mathrm{F}}^{\top}\mathrm{Re}\left[(\mathbf{W}_{\mathrm{F}}\boldsymbol{x}+\boldsymbol{b}_{\mathrm{F}}+\mathbf{1}\mathrm{i})\circ\mathbb{I}\left(\mathbf{W}_{\mathrm{F}}\boldsymbol{x}+\boldsymbol{b}_{\mathrm{F}}\geqslant 0\right)\right]\\
&= \boldsymbol{\alpha}_{\mathrm{F}}^{\top}\left[(\mathbf{W}_{\mathrm{F}}\boldsymbol{x}+\boldsymbol{b}_{\mathrm{F}})\circ\mathbb{I}\left(\mathbf{W}_{\mathrm{F}}\boldsymbol{x}+\boldsymbol{b}_{\mathrm{F}}\geqslant 0\right)\right]\\
&= \boldsymbol{\alpha}_{\mathrm{F}}^{\top}\sigma_{\mathrm{F}}\left(\mathbf{W}_{\mathrm{F}}\boldsymbol{x}+\boldsymbol{b}_{\mathrm{F}}\right)\\
&= f_{\mathrm{F}}(\boldsymbol{x}),
\end{aligned}
\tag{24}
$$

where the third equality holds because σ_{F} is the ReLU activation, and the fourth equality holds from Eq. (22). Finally, we prove that the constructed F-FTNet$_{\times}$ can $(\varepsilon,\mathcal{D})$-approximate the function f. According to Eqs. (23) and (24), one has

$$
\mathbb{E}_{\boldsymbol{x}}\left[(f(\boldsymbol{x})-f_{\times,\mathrm{F}}(\boldsymbol{x}))^2\right]=\mathbb{E}_{\boldsymbol{x}}\left[(f(\boldsymbol{x})-f_{\mathrm{F}}(\boldsymbol{x}))^2\right]\leqslant\varepsilon,
$$

which completes the proof. □

Proof of Theorem 6. Since the target DODS f_{D} can be $(\varepsilon,\mathcal{D})$-approximated by RNN, there exists an RNN defined by Eq. (20) satisfying the following inequality

$$
\mathbb{E}_{\boldsymbol{x}_{1:T}\sim\mathcal{D}}\left[(f_{\mathrm{D}}(\boldsymbol{x}_{1:T})-f_{\mathrm{R}}(\boldsymbol{x}_{1:T}))^2\right]\leqslant\varepsilon. \tag{25}
$$

Firstly, we construct an R-FTNet$_{\times}$ $f_{\times,\mathrm{R}}$ with hidden size $H_{\times}=2H_{\mathrm{R}}+I+1$ using the zReLU activation as follows

$$
\begin{aligned}
&\mathbf{W}_{\times}=\begin{bmatrix}\mathbf{0}_{I\times I} & \mathbf{0} & \mathbf{0} & \mathbf{0}\\ \mathbf{W}_{\mathrm{R}} & \mathbf{0}_{H_{\mathrm{R}}\times H_{\mathrm{R}}} & \mathbf{0} & \boldsymbol{\theta}_{\mathrm{R}}\\ \mathbf{0} & \mathbf{0} & \mathbf{V}_{\mathrm{R}} & \mathbf{1}\\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0}_{1\times 1}\end{bmatrix},\quad \mathbf{V}_{\times}=\begin{bmatrix}\mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & -\mathbf{V}_{\mathrm{R}} & \mathbf{1}\\ \mathbf{W}_{\mathrm{R}} & \mathbf{0} & \mathbf{0} & \boldsymbol{\theta}_{\mathrm{R}}\\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0}\end{bmatrix},\\
&\boldsymbol{r}_0=\left[\mathbf{0}_{I\times 1};\mathbf{0}_{H_{\mathrm{R}}\times 1};\boldsymbol{m}_0;\mathbf{0}_{1\times 1}\right],\quad \boldsymbol{\alpha}_{\times}=\left[\mathbf{0}_{I\times 1};\boldsymbol{\alpha}_{\mathrm{R}};\mathbf{0}_{H_{\mathrm{R}}\times 1};\mathbf{0}_{1\times 1}\right].
\end{aligned}
\tag{26}
$$

Secondly, we prove that the output of the constructed R-FTNet$_{\times}$ in Eq. (26) is the same as that of the RNN in Eq. (25). Let $\boldsymbol{r}_t=[\boldsymbol{r}_{t,1};\boldsymbol{r}_{t,2};\boldsymbol{r}_{t,3};\boldsymbol{r}_{t,4}]$, where $\boldsymbol{r}_{t,1}\in\mathbb{R}^{I}$, $\boldsymbol{r}_{t,2}\in\mathbb{R}^{H_{\mathrm{R}}}$, $\boldsymbol{r}_{t,3}\in\mathbb{R}^{H_{\mathrm{R}}}$, and $\boldsymbol{r}_{t,4}\in\mathbb{R}$. We prove that $\boldsymbol{r}_{t,1}=\mathbf{0}_{I\times 1}$, $\boldsymbol{r}_{t,3}=\boldsymbol{m}_t$, and $\boldsymbol{r}_{t,4}=\mathbf{0}_{1\times 1}$ hold for any $t\leqslant T$ by mathematical induction.

1. Base case. From Eq. (26), the claim holds for $t=0$.

2. Induction. Suppose that the claim holds for $t = \tau$ where $\tau \in \{0, 1, \ldots, T-1\}$. From Eq. (26), it is observed that the first and fourth rows of the weight matrices $\mathbf{W}_\times$ and $\mathbf{V}_\times$ are all 0. Based on $\sigma_\times(0) = 0$, one knows that

$$\boldsymbol{r}_{\tau+1,1} = \sigma_\times(\mathbf{0}) = \mathbf{0} \quad \text{and} \quad \boldsymbol{r}_{\tau+1,4} = \sigma_\times(\mathbf{0}) = \mathbf{0}.$$

Furthermore, one has

$$\begin{aligned} \boldsymbol{r}_{\tau+1,3} &= \mathrm{Im}\left[\sigma_\times\left(\mathbf{1} + \left(\mathbf{W}_\mathrm{R}\boldsymbol{x}_{\tau+1} + \mathbf{V}_\mathrm{R}\boldsymbol{r}_{\tau,3} + \boldsymbol{\theta}_\mathrm{R}\right)\mathrm{i}\right)\right] \\ &= \left(\mathbf{W}_\mathrm{R}\boldsymbol{x}_{\tau+1} + \mathbf{V}_\mathrm{R}\boldsymbol{r}_{\tau,3} + \boldsymbol{\theta}_\mathrm{R}\right) \circ \mathbb{I}\left(\mathbf{W}_\mathrm{R}\boldsymbol{x}_{\tau+1} + \mathbf{V}_\mathrm{R}\boldsymbol{r}_{\tau,3} + \boldsymbol{\theta}_\mathrm{R} \geqslant 0\right) \\ &= \sigma_\mathrm{R}\left(\mathbf{W}_\mathrm{R}\boldsymbol{x}_{\tau+1} + \mathbf{V}_\mathrm{R}\boldsymbol{m}_\tau + \boldsymbol{\theta}_\mathrm{R}\right) \\ &= \boldsymbol{m}_{\tau+1}, \end{aligned}$$

where the first equality holds from the construction in Eq. (26) and the hypothesis induction, the second equality holds according to Eq. (4), the third equality holds because σ_R is the ReLU activation function, and the fourth equality holds based on Eq. (20). Thus, the claim holds for $t = \tau + 1$.

For any $t \in [T]$, let $\boldsymbol{s}_t = [\boldsymbol{s}_{t,1}; \boldsymbol{s}_{t,2}; \boldsymbol{s}_{t,3}; \boldsymbol{s}_{t,4}]$, where $\boldsymbol{s}_{t,1} \in \mathbb{R}^I$, $\boldsymbol{s}_{t,2} \in \mathbb{R}^{H_\mathrm{R}}$, $\boldsymbol{s}_{t,3} \in \mathbb{R}^{H_\mathrm{R}}$, and $\boldsymbol{s}_{t,4} \in \mathbb{R}$. Similar to the calculation of $\boldsymbol{r}_{t,3}$, one has $\boldsymbol{s}_{t,2} = \boldsymbol{m}_t$. Thus, the output of R-FTNet$_\times$ is the same as that of RNN, i.e.,

$$y_{\times,t} = \boldsymbol{\alpha}_\times^\top \boldsymbol{s}_t = \boldsymbol{\alpha}_\mathrm{R}^\top \boldsymbol{s}_{t,2} = \boldsymbol{\alpha}_\mathrm{R}^\top \boldsymbol{m}_t = y_{\mathrm{R},t}. \tag{27}$$

Finally, we prove that the constructed R-FTNet$_\times$ can $(\varepsilon, \mathcal{D})$-approximate f_D. According to Eqs. (25) and (27), one has

$$\mathbb{E}_{\boldsymbol{x}_{1:T}\sim\mathcal{D}}\left[\left(f_\mathrm{D}(\boldsymbol{x}_{1:T}) - f_{\times,\mathrm{R}}(\boldsymbol{x}_{1:T})\right)^2\right] = \mathbb{E}_{\boldsymbol{x}_{1:T}\sim\mathcal{D}}\left[\left(f_\mathrm{D}(\boldsymbol{x}_{1:T}) - f_\mathrm{R}(\boldsymbol{x}_{1:T})\right)^2\right] \leqslant \varepsilon,$$

which completes the proof. □

Table 1: Approximation Complexity of FTNet and FNN

Target	Width of FNN	Width of FTNet
In Theorem 3	$\Omega(\mathrm{e}^{\epsilon_1 I}/I)$	$O(I^{15/4})$
Any (Theorem 5)	H_F	$O(H_F)$

Table 2: Approximation Complexity of FTNet and RNN

Target	Width of RNN	Width of FTNet
In Theorem 4	$\Omega(\mathrm{e}^{\varepsilon_2 I})$	$O(I^{15/4})$
Any (Theorem 6)	H_R	$O(H_R)$

Tables 1 and 2 summarize the approximation complexity results of FTNet using asymptotic notations, where ϵ_1 and ϵ_2 are two constants irrelevant to the input dimension I. FTNet possesses exponential advantage when expressing particular functions, and requires hidden size of the same order in arbitrary cases. These results suggest that FTNet is able to exhibit dynamic reaction by the flexible formulation of the synapse, which would be demanded in decision making [55] and open-environment machine learning [62], though the analysis is beyond the scope of this paper.

6. Local Minima

This section investigates the empirical loss surface of F-FTNet. Let $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ be the training set, where $\boldsymbol{x}_i \in \mathbb{R}^I$ denotes the i-th sample, and $y_i \in \mathbb{R}$ represents the label of the i-th sample. Consider the empirical loss of F-FTNet with the following form

$$\hat{L} = \sum_{i=1}^{n} l\left(f_{\times,\mathrm{F}}(\boldsymbol{x}_i) - y_i\right), \tag{28}$$

where $f_{\times,\mathrm{F}}$ is the mapping of F-FTNet defined in Eq. (3), and $l : \mathbb{R} \to \mathbb{R}$ is a loss function. Let $\mathbf{Z} = \mathbf{W}_\times + \mathbf{V}_\times \mathrm{i}$ be the complex-valued weight matrix, and $\boldsymbol{\alpha}$ denotes $\boldsymbol{\alpha}_\times$ for simplicity. Then the empirical loss $\hat{L}$ is a function of $\mathbf{Z}$ and $\boldsymbol{\alpha}$, denoted by $\hat{L}(\mathbf{Z}, \boldsymbol{\alpha})$. Holomorphic activation functions are of interest in this section, and the definition of holomorphic functions is reviewed as follows.

Definition 3 *[19, Page 2] A function $g : \mathbb{C}^m \to \mathbb{C}$ is called holomorphic if for each point $\boldsymbol{w} = (w_1, w_2, \dots, w_m) \in \mathbb{C}^m$, there exists an open set U, such that $\boldsymbol{w} \in U$, and the function g has a power series expansion*

$$f(\boldsymbol{z}) = \sum_{(v_1,v_2,\dots,v_m)\in\mathbb{N}^m} a_{v_1,v_2,\dots,v_m} \prod_{j=1}^{m} (z_j - w_j)^{v_j}, \tag{29}$$

which converges for all $\boldsymbol{z} = (z_1, z_2, \dots, z_m) \in U$.

Let us define a class of loss functions called *well-posed regression loss functions*.

Definition 4 *A loss function $l : \mathbb{R} \to \mathbb{R}$ is called a well-posed regression loss function, if l satisfies the following conditions: i) l is analytic on $\mathbb{R}$; ii) $l(0) = 0$; iii) l is strictly decreasing on $(-\infty, 0)$ and strictly increasing on $(0, +\infty)$.*

The conditions in Definition 4 are satisfied by many commonly used loss functions for regression or their smooth variants, such as the squared loss $l(x) = x^2$, the parameterized cosh $l(x) = c^{-1}[\ln(\mathrm{e}^{ax} + \mathrm{e}^{-bx}) - \ln 2]$ with positive parameters a, b, and c, which can approximate the absolute loss $l(x) = |x|$ and the quantile loss $l(x) = (1-\theta)x\mathbb{I}_{x\geqslant 0} - \theta x \mathbb{I}_{x<0}$ with $\theta \in (0,1)$ [29] in the limit. The following theorem studies local minima of the empirical loss $\hat{L}$.

Theorem 7 *Suppose that all samples are linearly independent, the activation function $\sigma_\times$ is holomorphic and not polynomial, and that l is a well-posed regression loss function. If the loss $\hat{L}(\mathbf{Z}, \boldsymbol{\alpha})$ is positive, then for any $\delta > 0$, there exist $\Delta\mathbf{Z}$ and $\Delta\boldsymbol{\alpha}$ satisfying the following inequalities*

$$\hat{L}\left(\mathbf{Z} + \Delta\mathbf{Z}, \boldsymbol{\alpha} + \Delta\boldsymbol{\alpha}\right) < \hat{L}\left(\mathbf{Z}, \boldsymbol{\alpha}\right)$$

and

$$\|\Delta\mathbf{Z}\|_F + \|\Delta\boldsymbol{\alpha}\|_2 \leqslant \delta.$$

Theorem 7 shows that it is always possible to reduce the loss in the neighborhood as long as the loss is not 0. This indicates that any local minimum of $\hat{L}(\mathbf{Z}, \boldsymbol{\alpha})$ is the global minimum. The requirement of linearly independent samples holds with probability 1 when the sample size is no larger than the input dimension, and the samples are generated from a continuous distribution. Existing studies mostly investigate the local-minima-free condition of FNN using specific loss with strong conditions, such as linearly separable data [36], particular activation functions [26, 48, 35, 31], and over-parameterization together with special initialization [1, 11, 25, 64, 63]. Theorem 7 holds for a large class of activations and loss functions. The requirement of a large input dimension is reasonable since previous work proves that FNN has suboptimal local minima with low-dimensional input under general settings [10]. We begin our proof with two lemmas.

Lemma 3 *Let $g : \mathbb{C}^m \to \mathbb{C}$ be holomorphic and not constant. For any $\boldsymbol{z}^{(0)} \in \mathbb{C}^m$, $\delta \in (0,1)$, there exist $\Delta\boldsymbol{z}^{(1)}, \Delta\boldsymbol{z}^{(2)} \in \mathbb{C}^m$ satisfying the following inequalities*

$$\left\|\Delta\boldsymbol{z}^{(1)}\right\|_2^2 \leqslant \delta \quad \textit{and} \quad \mathrm{Re}\left[g(\boldsymbol{z}^{(0)} + \Delta\boldsymbol{z}^{(1)})\right] > \mathrm{Re}\left[g(\boldsymbol{z}^{(0)})\right],$$
$$\left\|\Delta\boldsymbol{z}^{(2)}\right\|_2^2 \leqslant \delta \quad \textit{and} \quad \mathrm{Re}\left[g(\boldsymbol{z}^{(0)} + \Delta\boldsymbol{z}^{(2)})\right] < \mathrm{Re}\left[g(\boldsymbol{z}^{(0)})\right].$$

Lemma 3 shows that the neighborhood of holomorphic functions possesses rich diversity, i.e., one can always find a point with a smaller real part and another point with a larger real part in the neighborhood.

Proof of Lemma 3. Let $\boldsymbol{z}^{(0)} = (z_1^{(0)}, z_2^{(0)}, \dots, z_m^{(0)})$. Since the function g is holomorphic, by rearranging Eq. (29), there exist an open set U and a series of holomorphic functions $f_j^{(k)} : \mathbb{C}^{m-k} \to \mathbb{C}$ with $j \in \mathbb{N}$ and $k \in [m]$, s.t. $\boldsymbol{z}^{(0)} \in U$, and for any $\boldsymbol{z} = (z_1, z_2, \dots, z_m) \in U$, the following holds

$$f_0^{(0)}(\boldsymbol{z}) := g(\boldsymbol{z}) = \sum_{j=0}^{\infty} f_j^{(1)}(z_2, z_3, \dots, z_m)\left(z_1 - z_1^{(0)}\right)^j,$$
$$f_0^{(1)}(z_2, z_3, \dots, z_m) = \sum_{j=0}^{\infty} f_j^{(2)}(z_3, z_4, \dots, z_m)\left(z_2 - z_2^{(0)}\right)^j,$$
$$f_0^{(2)}(z_3, z_4, \dots, z_m) = \sum_{j=0}^{\infty} f_j^{(3)}(z_4, z_5, \dots, z_m)\left(z_3 - z_3^{(0)}\right)^j,$$
$$\cdots\cdots$$
$$f_0^{(m-1)}(z_m) = \sum_{j=0}^{\infty} f_j^{(m)}\left(z_m - z_m^{(0)}\right)^j.$$

It is observed that $f_0^{(m)} = g(\boldsymbol{z}^{(0)})$ when $\boldsymbol{z} = \boldsymbol{z}^{(0)}$. Then the following k_0 is well-defined

$$k_0 = \min\left\{k \in [m] \;\middle|\; f_0^{(k)} \equiv g\left(\boldsymbol{z}^{(0)}\right)\right\}.$$

Let $\boldsymbol{z}^{(1)} = (z_1^{(0)}, \dots, z_{k_0-1}^{(0)}, z_{k_0}, \dots, z_m)$. Thus, one has

$$g\left(\boldsymbol{z}^{(1)}\right) = g\left(\boldsymbol{z}^{(0)}\right) + \sum_{j=1}^{\infty} f_j^{(k_0)}(z_{k_0+1}, \dots, z_m)\left(z_{k_0} - z_{k_0}^{(0)}\right)^j.$$

Since $f_0^{(k_0-1)} \not\equiv g(\boldsymbol{z}^{(0)})$ and $g(\boldsymbol{z}^{(1)}) = f_0^{(k_0-1)}$, one has $g(\boldsymbol{z}^{(1)}) \not\equiv g(\boldsymbol{z}^{(0)})$. Thus, there exists a positive integer $j \in \mathbb{N}^+$, such that $f_j^{(k_0)}(z_{k_0+1}, z_{k_0+2}, \dots, z_m) \not\equiv 0$. Then the following j_0 is well-defined

$$j_0 = \min\left\{ j \in \mathbb{N}^+ \;\middle|\; f_j^{(k_0)}(z_{k_0+1}, z_{k_0+2}, \ldots, z_m) \not\equiv 0 \right\}.$$

Therefore, there exist $z_{k_0+1}^{(1)}, z_{k_0+2}^{(1)}, \ldots, z_m^{(1)}$, such that

$$f_{j_0}^{(k_0)}\left(z_{k_0+1}^{(1)}, z_{k_0+2}^{(1)}, \ldots, z_m^{(1)}\right) \neq 0,$$

and

$$\sum_{k=k_0+1}^{m} \left(z_k^{(1)} - z_k^{(0)}\right)^2 \leqslant \frac{\delta}{2}. \tag{30}$$

Let $\boldsymbol{z}^{(2)} = (z_1^{(0)}, z_2^{(0)}, \ldots, z_{k_0-1}^{(0)}, z_{k_0}, z_{k_0+1}^{(1)}, z_{k_0+2}^{(1)}, \ldots, z_m^{(1)})$. Then the function value of g at $\boldsymbol{z}^{(2)}$ satisfies

$$g\left(\boldsymbol{z}^{(2)}\right) = g\left(\boldsymbol{z}^{(0)}\right) + \sum_{j=j_0}^{\infty} a_j \left(z_{k_0} - z_{k_0}^{(0)}\right)^j, \tag{31}$$

where $a_j = f_j^{(k_0)}(z_{k_0+1}^{(1)}, z_{k_0+2}^{(1)}, \ldots, z_m^{(1)})$ and $a_{j_0} \neq 0$. Since $\boldsymbol{z}^{(0)}$ is in the open set U, there exists $r > 0$, such that the ball $B(\boldsymbol{z}^{(0)}, r)$ is a subset of U. Then the radius of convergence of the series in Eq. (31) is at least r. Thus, one has $\limsup_{j\to\infty} |a_j|^{1/j} \leqslant 1/r$ from the Cauchy-Hadamard theorem. Since any series with finite limit superior is bounded, there exists $M \geqslant \max\{1, \sqrt{2\delta}/3\}$, such that $|a_j|^{1/j} \leqslant M$, i.e., $|a_j| \leqslant M^j$. Define the change of $\boldsymbol{z}^{(0)}$ as

$$\Delta\boldsymbol{z}^{(1)} = \left(0, \ldots, 0, \tilde{z}_{k_0}, z_{k_0+1}^{(1)} - z_{k_0+1}^{(0)}, \ldots, z_m^{(1)} - z_m^{(0)}\right),$$

where

$$\tilde{z}_{k_0} = \frac{\min\{1, |a_{j_0}|\}}{3M^{j_0+1}\sqrt{2/\delta}} \mathrm{e}^{-\mathrm{i}\theta_{a_{j_0}}/j_0}.$$

In view of $M \geqslant 1$ and Eq. (30), one has

$$\left\|\Delta\boldsymbol{z}^{(1)}\right\|_2^2 \leqslant |\tilde{z}_{k_0}|^2 + \sum_{k=k_0+1}^{m} \left(z_k^{(1)} - z_k^{(0)}\right)^2 \leqslant \delta,$$

meanwhile,

$$\begin{aligned}
\mathrm{Re}\left[g\left(\boldsymbol{z}^{(0)} + \Delta\boldsymbol{z}^{(1)}\right)\right] - \mathrm{Re}\left[g\left(\boldsymbol{z}^{(0)}\right)\right] &= \sum_{j=j_0}^{\infty} \mathrm{Re}\left[a_j \left(\tilde{z}_{k_0}\right)^j\right] \\
&\geqslant |a_{j_0}|\, |\tilde{z}_{k_0}|^{j_0} - \sum_{j=j_0+1}^{\infty} M^j\, |\tilde{z}_{k_0}|^j \\
&\geqslant \min\{1, |a_{j_0}|\}^{j_0+1} \frac{(\delta/2)^{j_0/2}}{3^{j_0} M^{j_0(j_0+1)}} \left[1 - \frac{2}{3\sqrt{2/\delta}}\right] \\
&> 0,
\end{aligned}$$

where the first equality holds from Eq. (31), the first inequality holds because of $\mathrm{Re}[z] \geqslant -|z|$ and $|a_j| \leqslant M^j$, the second inequality holds based on $M \geqslant \sqrt{2\delta}/3$, and the third inequality holds in view of $\delta < 1$. Let

$$\Delta \boldsymbol{z}^{(2)} = \left(0, \ldots, 0, \hat{z}_{k_0}, z_{k_0+1}^{(1)} - z_{k_0+1}^{(0)}, \ldots, z_m^{(1)} - z_m^{(0)}\right),$$

where

$$\hat{z}_{k_0} = \frac{\min\{1, |a_{j_0}|\}}{3M^{j_0+1}\sqrt{2/\delta}} \mathrm{e}^{-\mathrm{i}\left(\pi+\theta_{a_{j_0}}\right)/j_0}.$$

Then the conclusion about $\Delta \boldsymbol{z}^{(2)}$ can be proven similarly. □

Lemma 4 *Let $\sigma : \mathbb{C} \to \mathbb{C}$ be holomorphic and not polynomial, $\left\{\boldsymbol{x}^{(j)}\right\}_{j=1}^n \subset \mathbb{R}^m$ are n different vectors, $\{y_j\}_{j=1}^n \subset \mathbb{R}$ are not all zero, and $\boldsymbol{z} = (z_1, z_2, \ldots, z_{m+1})$ is a complex-valued vector. Then the function $g : \mathbb{C}^{m+1} \to \mathbb{C}$, defined by*

$$g(\boldsymbol{z}) = \sum_{j=1}^{n} y_j \sigma\left(x_1^{(j)} z_1 + x_2^{(j)} z_2 + \cdots + x_m^{(j)} z_m + z_{m+1}\right),$$

is not a constant function.

Lemma 4 provides a sufficient condition that the summation of the activation of weighted average is not a constant.

Proof of Lemma 4. The proof consists of several steps.

Step 1. We find the necessary condition that g is a constant. Since σ is holomorphic, and any holomorphic function coincides with its Taylor series in any open set within the domain of the function [49, Theorem 4.4], there exists $\{c_k\}_{k=0}^{\infty} \subset \mathbb{C}$, such that $\sigma(z) = \sum_{k=0}^{\infty} c_k z^k$ holds for any $z \in \mathbb{C}$. Since σ is not polynomial, there exists $\{n_k\}_{k=1}^{\infty} \subset \mathbb{N}^+$, such that $\sigma(z) = c_0 + \sum_{k=1}^{\infty} c_{n_k} z^{n_k}$, where $n_k < n_{k+1}$ and $c_{n_k} \neq 0$ hold for any $k \in \mathbb{N}^+$. Thus, the function g can be rewritten as

$$g(\boldsymbol{z}) = \sum_{j=1}^{n} y_j \left[c_0 + \sum_{k=1}^{\infty} c_{n_k} \left(z_{m+1} + \sum_{l=1}^{m} x_l^{(j)} z_l\right)^{n_k}\right] = h_0(\boldsymbol{z}) + \sum_{k=1}^{\infty} h_{n_k}(\boldsymbol{z}).$$

If g is a constant, then one has

$$\sum_{j=1}^{n} y_j \left(z_{m+1} + \sum_{l=1}^{m} x_l^{(j)} z_l\right)^{n_k} \equiv 0, \quad \forall\ k \in \mathbb{N}^+.$$

According to the multinomial theorem, one has

$$\sum_{j=1}^{n}\sum_{\boldsymbol{p}\in P_{n_k}} y_j c_{\boldsymbol{p}} z_{m+1}^{p_{m+1}} \prod_{l=1}^{m}\left(x_l^{(j)} z_l\right)^{p_l} \equiv 0, \quad \forall\, k\in\mathbb{N}^+,$$

where $\boldsymbol{p} = (p_1, p_2, \ldots, p_{m+1})$, $P_{n_k} = \{\boldsymbol{p} \mid \forall\, l \in [m+1], p_l \in \mathbb{N}, \|\boldsymbol{p}\|_1 = n_k\}$, and $c_{\boldsymbol{p}}$ is the multinomial coefficient. Since $z_1, z_2, \ldots, z_{m+1}$ are free variables, one has

$$\sum_{j=1}^{n} y_j \prod_{l=1}^{m}\left(x_l^{(j)}\right)^{p_l} = 0, \quad \forall\, k\in\mathbb{N}^+, \boldsymbol{p}\in P_{n_k}.$$

Since $n_k \to +\infty$ as $k \to +\infty$, we obtain the following necessary condition of constant g

$$\sum_{j=1}^{n} y_j \prod_{l=1}^{m}\left(x_l^{(j)}\right)^{p_l} = 0, \quad \forall\, p_1, p_2, \ldots, p_m \in\mathbb{N}. \tag{32}$$

Step 2. We restrict the summation domain of the necessary condition in Eq. (32) to obtain another necessary condition. Let $J_0 = \{j\in[n] \mid y_j \neq 0\}$. For any $l\in[m]$, we define

$$m_l = \max_{j\in J_{l-1}} \left|x_l^{(j)}\right| \quad \text{and} \quad J_l = \left\{ j\in J_{l-1} \;\middle|\; \left|x_l^{(j)}\right| = m_l \right\}.$$

Under the assumption that g is constant, we claim that

$$\sum_{j\in J_m} y_j \prod_{l=1}^{m}\left(x_l^{(j)}\right)^{p_l} = 0, \quad \forall\, p_1, p_2, \ldots, p_m \in\mathbb{N}. \tag{33}$$

Otherwise, there exist $q_1, q_2, \ldots, q_m \in \mathbb{N}$, such that

$$\sum_{j\in J_m} y_j \prod_{l=1}^{m}\left(x_l^{(j)}\right)^{q_l} = c_0 \neq 0.$$

Let $r_1, r_2, \ldots, r_m$ be even natural numbers. Thus, one has

$$0 = \left|\sum_{j\in J_0} y_j \prod_{l=1}^{m}\left(x_l^{(j)}\right)^{q_l+r_l}\right| \geqslant \left|\sum_{j\in J_m} y_j \prod_{l=1}^{m}\left(x_l^{(j)}\right)^{q_l+r_l}\right| - \sum_{l=1}^{m}\left|\sum_{j\in J_{l-1}\setminus J_l} y_j \prod_{l=1}^{m}\left(x_l^{(j)}\right)^{q_l+r_l}\right|,$$

where the equality holds from Eq. (32) and definition of J_0, and the inequality holds based on the triangle inequality. Let

$$y_M = \max_{j\in[n]} |y_j| \quad \text{and} \quad L = \left\{ l\in[m] \;\middle|\; |J_{l-1}\setminus J_l| \geqslant 1 \right\}.$$

For $l\in[m]$, define

$$M_l = \max_{j\in[n]} \left|x_l^{(j)}\right|,$$

$$m_{l,2} = \begin{cases} \max\limits_{j\in J_{l-1}\setminus J_l} \left|x_l^{(j)}\right|, & \text{if } J_{l-1}\neq J_l, \\ 0, & \text{if } J_{l-1} = J_l, \end{cases}$$

$$\mathcal{A}_l = \left(\prod_{s=1}^{l-1} m_s^{q_s+r_s}\right)\left(\prod_{t=l+1}^{m} M_t^{q_t+r_t}\right).$$

Thus, one has

$$\begin{aligned}0 &\geqslant |c_0| \prod_{l=1}^{m} m_l^{r_l} - \sum_{l=1}^{m} |J_{l-1} \backslash J_l|\, y_M \mathcal{A}_l m_{l,2}^{r_l} \\ &\geqslant |c_0| \prod_{l=1}^{m} m_l^{r_l} - \sum_{l \in L} |J_{l-1} \backslash J_l|\, y_M \mathcal{A}_l m_{l,2}^{r_l} \\ &\geqslant |c_0| \prod_{l=1}^{m} m_l^{r_l} - n y_M \sum_{l \in L} \mathcal{A}_l m_{l,2}^{r_l},\end{aligned}$$

where the second inequality holds because of $J_{l-1} \backslash J_l \subset J_{l-1} \subset [n]$. For $l \in L$, it is observed that $m_l > m_{l,2} \geqslant 0$. Thus, the second term in the above inequality will be much smaller than the first term when r_l is sufficiently large. More formally, we define r_l as follows.

Case 1. If $l \in [n] \backslash L$, define $r_l = 0$.

Case 2. If $l \in L$ and $m_{l',2} = 0$, define $r_l = 2$.

Case 3. Otherwise, define $\lceil x \rceil_{\mathrm{E}}$ as the smallest even integer no less than x. Let

$$r_{l'} = \left\lceil \frac{1}{\ln\left(\frac{m_{l',2}}{m_{l'}}\right)} \ln\left(\frac{|c_0| \prod_{t=l'+1}^{m} M_t^{q_t + r_t}}{2n^2 y_M \prod_{s=1}^{l'-1} m_s^{q_s} \prod_{t=l'+1}^{m} m_t^{r_t}}\right) \right\rceil_{\mathrm{E}}.$$

Based on the choice of r_l, one has

$$0 \geqslant |c_0| \prod_{l=1}^{m} m_l^{r_l} - n y_M \sum_{l \in L} \frac{|c_0| \prod_{l'=1}^{m} m_{l'}^{r_{l'}}}{2n^2 y_M} \geqslant \frac{|c_0|}{2} \prod_{l=1}^{m} m_l^{r_l},$$

where the second inequality holds because of $|L| \leqslant n$. When $m_l = 0$, one has $J_{l-1} = J_l$ from the definition of J_l. Thus, one has $l \in L$, which leads to $r_l = 0$. Since $c_0 \neq 0$, one has

$$0 \geqslant \frac{|c_0|}{2} \prod_{l=1}^{m} m_l^{r_l} > 0,$$

which is a contradiction. Thus, we have proven the claim in Eq. (33). It is observed that $|x_l^{(j)}| = m_l$ holds for any $j \in J_m$. Thus, the claim indicates that when g is a constant, one has

$$\sum_{j \in J_m} y_j \prod_{l=1}^{m} \mathrm{sign}\left(x_l^{(j)}\right)^{p_l} = 0, \quad \forall\ p_1, p_2, \ldots, p_m \in \mathbb{N}, \tag{34}$$

where $\mathrm{sign}(\cdot)$ denotes the sign function.

Step 3. We prove that the necessary condition of constant g in Eq. (34) does not hold by probabilistic methods. For any $j \in J_m$, let $N_j = \{l \in [m] \mid \boldsymbol{x}_l^{(j)} = -m_l\}$ denote the set of dimensions in which $\boldsymbol{x}^{(j)}$ is negative. Observe that $\{N_j\}_{j \in J_m}$ are different since $\{\boldsymbol{x}^{(j)}\}_{j=1}^{n}$ are different. Thus, there exists a minimal element among $\{N_j\}_{j \in J_m}$, i.e., there exists $j_0 \in J_m$, such that for any $j \in S_m \backslash \{j_0\}$, one has $T_j \not\subset T_{j_0}$. For any $l \in [m]$, we define a random variable σ_l as follows

$$\begin{cases} \Pr[\sigma_l = 0] = 1, & \text{if } l \in T_{j_0}, \\ \Pr[\sigma_l = 0] = \Pr[\sigma_l = 1] = 1/2, & \text{if } l \notin T_{j_0}. \end{cases}$$

Let $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_m)$. Thus, one has

$$\begin{aligned} 0 &= \mathbb{E}_{\boldsymbol{\sigma}} \left[\sum_{j \in J_m} y_j \prod_{l=1}^{m} \text{sign}\left(x_l^{(j)}\right)^{\sigma_l} \right] \\ &= \mathbb{E}_{\boldsymbol{\sigma}} \left[y_{j_0} \prod_{l=1}^{m} \text{sign}\left(x_l^{(j)}\right)^{\sigma_l} \right] + \mathbb{E}_{\boldsymbol{\sigma}} \left[\sum_{j \in J_m \backslash \{j_0\}} y_j \prod_{l=1}^{m} \text{sign}\left(x_l^{(j)}\right)^{\sigma_l} \right], \end{aligned} \tag{35}$$

where the first equality holds from Eq. (34). For the first term of Eq. (35), the following equation holds

$$\mathbb{E}_{\boldsymbol{\sigma}} \left[y_{j_0} \prod_{l=1}^{m} \text{sign}\left(x_l^{(j)}\right)^{\sigma_l} \right] = y_{j_0} \mathbb{E}_{\boldsymbol{\sigma}} \left[\prod_{l \in T_{j_0}} \text{sign}\left(x_l^{(j_0)}\right)^{\sigma_l} \prod_{l \notin T_{j_0}} \text{sign}\left(x_l^{(j_0)}\right)^{\sigma_l} \right] = y_{j_0}, \tag{36}$$

where the second equality holds because of $\sigma_l = 0$ for all $l \in T_{j_0}$ and $x_l^{(j_0)} > 0$ for all $l \notin T_{j_0}$. Since $T_j \not\subset T_{j_0}$ holds for any $j \in J_m \backslash \{j_0\}$, there exists l_j such that $l_j \notin T_{j_0}$ and $l_j \in T_j$. Thus, for the second term of Eq. (35), one has

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\sigma}} \left[\sum_{j \in J_m \backslash \{j_0\}} y_j \prod_{l=1}^{m} \text{sign}\left(x_l^{(j)}\right)^{\sigma_l} \right] &= \sum_{j \in J_m \backslash \{j_0\}} y_j \mathbb{E}_{\boldsymbol{\sigma}} \mathbb{E}_{\sigma_{l_j}} \left[\prod_{l=1}^{m} \text{sign}\left(x_l^{(j)}\right)^{\sigma_l} \right] \\ &= \sum_{j \in J_m \backslash \{j_0\}} y_j \mathbb{E}_{\boldsymbol{\sigma}} \left[\prod_{l=1, l \neq l_j}^{m} \text{sign}\left(x_l^{(j)}\right)^{\sigma_l} \cdot 0 \right] \\ &= 0, \end{aligned} \tag{37}$$

where the second equality holds since $x_l^{(j)} < 0$ and $\Pr[\sigma_{l_j} = 0] = \Pr[\sigma_{l_j} = 1] = 1/2$. Substituting Eqs. (36) and (37) into Eq. (35), one has $y_{j_0} = 0$, which contradicts the fact that $j_0 \in J_m \subset J_0$ and $y_j \neq 0$ for all $j \in J_0$. Thus, the necessary condition of constant g in Eq. (34) does not hold, which leads to the conclusion that g is not a constant. □

Proof of Theorem 7. Let $\Delta \hat{L}(\Delta \mathbf{Z}, \Delta \boldsymbol{\alpha}) = \hat{L}(\mathbf{Z} + \Delta \mathbf{Z}, \boldsymbol{\alpha} + \Delta \boldsymbol{\alpha}) - \hat{L}(\mathbf{Z}, \boldsymbol{\alpha})$ denote the change of empirical loss. Recalling Eqs. (3) and (28), one has

$$\Delta \hat{L}(\Delta \mathbf{Z}, \Delta \boldsymbol{\alpha}) = \sum_{j=1}^{n} -l \left(\boldsymbol{\alpha}^\top \text{Re}\left[\sigma\left(\mathbf{Z} \boldsymbol{\kappa}_j\right)\right] - y_j \right) + l \left((\boldsymbol{\alpha} + \Delta \boldsymbol{\alpha})^\top \text{Re}\left[\sigma\left(\left(\mathbf{Z} + \Delta \mathbf{Z}\right) \boldsymbol{\kappa}_j\right)\right] - y_j \right),$$

where $\boldsymbol{\kappa}_j$ and σ denote $\kappa(\boldsymbol{x}_j, H_\times)$ and $\sigma_\times$, respectively. Let $\mathbf{Z} = [\boldsymbol{z}_1^\top; \boldsymbol{z}_2^\top; \dots; \boldsymbol{z}_{H_\times}^\top]$ and $\boldsymbol{\alpha} = (\alpha_1; \alpha_2; \dots; \alpha_{H_\times})$. We prove the theorem by discussion.

Case 1. There exists $k_0 \in [H_\times]$, such that $\alpha_{k_0} \neq 0$. Since the loss $\hat{L}(\mathbf{Z}, \boldsymbol{\alpha})$ is positive and $l(0) = 0$, there exists $j_0 \in [n]$, such that $\boldsymbol{\alpha}^\top \mathrm{Re}[\sigma(\mathbf{Z}\boldsymbol{\kappa}_{j_0})] \neq y_{j_0}$. Without loss of generality, we only consider the case of $\boldsymbol{\alpha}^\top \mathrm{Re}[\sigma(\mathbf{Z}\boldsymbol{\kappa}_{j_0})] > y_{j_0}$ in this proof. The other case can be proven similarly. In view of the condition of all samples being linearly independent and the definition of κ in Eq. (2), one knows that $\{\boldsymbol{\kappa}_j\}_{j=1}^n$ are linearly independent. Thus, there exists a non-zero vector $\boldsymbol{v} \in \mathbb{C}^{H_\times}$, such that $\boldsymbol{v}^\top \boldsymbol{\kappa}_{j_0} \neq 0$ and $\boldsymbol{v}^\top \boldsymbol{\kappa}_j = 0$ hold for any $j \in [n] \backslash \{j_0\}$. Let $\Delta\boldsymbol{\alpha} = \mathbf{0}$, $\boldsymbol{z}_k = \mathbf{0}$ for any $k \in [H_\times] \backslash \{k_0\}$, and $\boldsymbol{z}_{k_0} = c\boldsymbol{v}$ where $c \in \mathbb{C}$ is a complex-valued variable. Then the change in loss becomes a function of c as follows

$$
\begin{aligned}
\Delta \hat{L}(c) = l\Big(&\boldsymbol{\alpha}^\top \mathrm{Re}\left[\sigma\left(\mathbf{Z}\boldsymbol{\kappa}_{j_0}\right)\right] + \alpha_{k_0} \mathrm{Re}\left[\sigma\left(\left(\boldsymbol{z}_{k_0} + c\boldsymbol{v}\right)^\top \boldsymbol{\kappa}_{j_0}\right)\right] \\
&- \alpha_{k_0} \mathrm{Re}\left[\sigma\left(\boldsymbol{z}_{k_0}^\top \boldsymbol{\kappa}_{j_0}\right)\right] - y_{j_0}\Big) - l\left(\boldsymbol{\alpha}^\top \mathrm{Re}\left[\sigma\left(\mathbf{Z}\boldsymbol{\kappa}_{j_0}\right)\right] - y_{j_0}\right),
\end{aligned}
$$

where the equality holds since the output of FTNet on $\boldsymbol{x}^{(j)}$ remains the same for any $j \neq j_0$. Since $\alpha_{k_0} \neq 0$, $\boldsymbol{v}^\top \boldsymbol{\kappa}_{j_0} \neq 0$, and σ is holomorphic and not constant, one knows that $\alpha_{k_0}\sigma((\boldsymbol{z}_{k_0} + c\boldsymbol{v})^\top \boldsymbol{\kappa}_{j_0})$ is holomorphic and not constant w.r.t. c. Then Lemma 3 implies that there exists $c \leqslant \delta / \|\boldsymbol{v}\|_2$, s.t.

$$
\mathrm{Re}\left[\alpha_{k_0}\sigma\left(\left(\boldsymbol{z}_{k_0} + c\boldsymbol{v}\right)^\top \boldsymbol{\kappa}_{j_0}\right)\right] < \mathrm{Re}\left[\alpha_{k_0}\sigma\left(\boldsymbol{z}_{k_0}^\top \boldsymbol{\kappa}_j\right)\right]
$$

and

$$
\mathrm{Re}\left[\alpha_{k_0}\sigma\left(\left(\boldsymbol{z}_{k_0} + c\boldsymbol{v}\right)^\top \boldsymbol{\kappa}_{j_0}\right)\right] \geqslant \mathrm{Re}\left[\alpha_{k_0}\sigma\left(\boldsymbol{z}_{k_0}^\top \boldsymbol{\kappa}_j\right)\right] - \boldsymbol{\alpha}^\top \mathrm{Re}\left[\sigma\left(\mathbf{Z}\boldsymbol{\kappa}_{j_0}\right)\right] + y_{j_0}, \tag{38}
$$

where Eq. (38) can be satisfied based on the continuity of holomorphic functions. Thus, one has $\|\Delta\mathbf{Z}\|_F + \|\Delta\boldsymbol{\alpha}\|_2 = \|c\boldsymbol{v}\|_2 = |c|\|\boldsymbol{v}\|_2 \leqslant \delta$ and $\Delta\hat{L}(c) < 0$ since the loss function l is strictly increasing on $(0, +\infty)$.

Case 2. For any $k \in [H_\times]$, $\alpha_k = 0$. Let $\Delta\boldsymbol{z}_k = \mathbf{0}$ and $\Delta\alpha_k = 0$ for any $k \in [H_\times] \backslash \{1\}$. Then the change in loss becomes a function of $\Delta\boldsymbol{z}_1$ and $\Delta\alpha_1$ as follows

$$
\Delta\hat{L} = \sum_{j=1}^n l\left(\Delta\alpha_1 \mathrm{Re}\left[\sigma\left(\left(\boldsymbol{z}_1 + \Delta\boldsymbol{z}_1\right)^\top \boldsymbol{\kappa}_j\right)\right] - y_j\right) - l\left(-y_j\right).
$$

The proof of this case is divided into several steps.

Step 2.1. We rewrite the change of loss in a power series form. Since the loss function l is analytic, there exist coefficients $\{c_p\}_{p=0}^{\infty}$, such that $l(y) = \sum_{p=0}^{\infty} c_p y^p$ holds for any $y \in \mathbb{R}$. Then the change of loss can be rewritten as

$$
\begin{aligned}
\Delta \hat{L} =& \sum_{j=1}^{n}\sum_{p=1}^{\infty}\sum_{q=1}^{p} c_p \binom{p}{q} (-y_j)^{p-q} \left(\Delta\alpha_1 \mathrm{Re}\left[\sigma\left((\boldsymbol{z}_1 + \Delta\boldsymbol{z}_1)^\top \boldsymbol{\kappa}_j\right)\right]\right)^q \\
=& \sum_{q=1}^{\infty}\sum_{j=1}^{n}\sum_{p=q}^{\infty} c_p \binom{p}{q} (-y_j)^{p-q} (\Delta\alpha_1)^q \left(\mathrm{Re}\left[\sigma\left((\boldsymbol{z}_1 + \Delta\boldsymbol{z}_1)^\top \boldsymbol{\kappa}_j\right)\right]\right)^q \\
=& \sum_{q=1}^{\infty} C_q (\Delta\alpha_1)^q ,
\end{aligned}
\tag{39}
$$

where the first equality holds from the binomial expansion, the second equality holds by changing the order of summation, and C_q is a function of $\Delta \boldsymbol{z}_1$ defined by

$$
C_q = \sum_{j=1}^{n}\sum_{p=q}^{\infty} c_p \binom{p}{q} R^q (-y_j)^{p-q} \tag{40}
$$

with $R = \mathrm{Re}[\sigma((\boldsymbol{z}_1 + \Delta\boldsymbol{z}_1)^\top \boldsymbol{\kappa}_j)]$.

Step 2.2. We prove that C_1 defined in Eq. (40) is not always zero. For $q = 1$, it is observed that

$$
C_1 = \sum_{j=1}^{n}\sum_{p=1}^{\infty} c_p p \mathrm{Re}\left[\sigma\left((\boldsymbol{z}_1 + \Delta\boldsymbol{z}_1)^\top \boldsymbol{\kappa}_j\right)\right] (-y_j)^{p-1} = \mathrm{Re}\left[\sum_{j=1}^{n} l'(-y_j)\sigma\left((\boldsymbol{z}_1 + \Delta\boldsymbol{z}_1)^\top \boldsymbol{\kappa}_j\right)\right].
$$

Since the loss function l is a well-posed regression loss function, one knows that the equation $l'(y) = 0$ has a unique solution $y = 0$. Since $\boldsymbol{\alpha} = 0$, all outputs of F-FTNet are 0. In view of positive loss, one knows that $\{y_j\}_{j=1}^n$ are not all 0, which indicates that $\{l'(-y_j)\}_{j=1}^n$ are not all 0. Since $\{\boldsymbol{\kappa}_j\}_{j=1}^n$ are linearly independent, they are different. Note that σ is holomorphic and not polynomial, Lemma 4 indicates that $\sum_{j=1}^n l'(-y_j)\sigma((\boldsymbol{z}_1 + \Delta\boldsymbol{z}_1)^\top \boldsymbol{\kappa}_j)$ is not a constant. Thus, there exists $\Delta\boldsymbol{z}_1$, such that $\|\Delta\boldsymbol{z}_1\|_2 \leqslant \delta/2$ and $C_1 \neq 0$.

Step 2.3. We give upper bounds for $\{C_q\}_{q=2}^{\infty}$. Provided $\Delta\boldsymbol{z}_1$ in Step 2.2., we define $a = \max_{j\in[n]} |\mathrm{Re}[\sigma((\boldsymbol{z}_1 + \Delta\boldsymbol{z}_1)^\top \boldsymbol{\kappa}_j)]|$. Let $b = \max_{j\in[n]} |y_j|$ for labels $\{y_j\}_{j=1}^{\infty}$. Since the loss function l is analytic on $\mathbb{R}$, the convergence radius of its Taylor series should be infinity. Thus, one has $\limsup_{p\to\infty} |c_p|^{1/p} = 0$ from the Cauchy-Hadamard theorem. Furthermore, there exists $d > 0$, s.t. $|c_p| \leqslant d/(4b)^p$ holds for any $p \geqslant 2$. Using these notations, coefficients $\{C_q\}_{q=2}^{\infty}$ can be bounded by

$$
|C_q| \leqslant \sum_{j=1}^{n}\sum_{p=q}^{\infty} |c_p| \binom{p}{q} \left|\mathrm{Re}\left[\sigma\left((\boldsymbol{z}_1 + \Delta\boldsymbol{z}_1)^\top \boldsymbol{\kappa}_j\right)\right]\right|^q |y_j|^{p-q} \quad \leqslant \sum_{p=q}^{\infty} \frac{nd}{4^p} \binom{p}{q} \left(\frac{a}{b}\right)^q , \tag{41}
$$

where the first inequality holds from the triangle inequality.

Step 2.4. We choose a proper $\Delta\alpha_1$ and give an upper bound for $\Delta\hat{L}$. Let $\Delta\alpha_1 = -\text{sign}(C_1)b/(ka)$, where $k \geqslant 1$ is a coefficient determined later. Thus, the change of loss in Eq. (39) can be rewritten as

$$\Delta\widehat{L} \leqslant C_1\Delta\alpha_1 + \sum_{q=2}^{\infty} |C_q|\,|\Delta\alpha_1|^q \leqslant -\frac{|C_1|b}{ka} + \sum_{p=2}^{\infty}\sum_{q=2}^{p} \frac{nd}{4^p}\binom{p}{q}\frac{1}{k^q} \leqslant -\frac{|C_1|b}{ka} + \frac{nd}{2k^2},$$

where the first inequality holds according to the triangle inequality, the second inequality holds based on Eq. (41), the choice of $\Delta\alpha_1$, and changing the order of summation, and the third inequality holds because of $k \geqslant 1$. We employ

$$k = \max\left\{1, \frac{nda}{|C_1|b}, \frac{2b}{a\delta}\right\},$$

and thus, one has $\Delta\hat{L} < 0$ and

$$\|\Delta\mathbf{Z}\|_F + \|\Delta\boldsymbol{\alpha}\|_2 = \|\Delta\boldsymbol{z}_1\|_2 + |\Delta\alpha_1| \leqslant \delta/2 + \delta/2 = \delta.$$

Combining the results in all cases completes the proof. □

7. Conclusion and Prospect

This work investigates the theoretical properties of FTNet via approximation and local minima. The main conclusions are three folds. Firstly, we prove the universal approximation of F-FTNet and R-FTNet, which guarantees the possibility of expressing any continuous function and any discrete-time open dynamical system on any compact set arbitrarily well, respectively. Secondly, we claim the approximation-complexity advantages and worst-case guarantees of one-hidden-layer F-FTNet/R-FTNet over FNN/RNN, i.e., F-FTNet and R-FTNet can express some functions with an exponentially fewer number of hidden neurons and can express a function with the same order of hidden neurons in the worst case, compared with FNN and RNN, respectively. Thirdly, we provide the feasibility of optimizing F-FTNet to the global minimum using local search algorithms, i.e., the loss surface of one-hidden-layer F-FTNet has no suboptimal local minimum using general activations and loss functions. Our theoretical results take one step towards the theoretical understanding of FTNet, which exhibits the possibility of ameliorating FTNet. In the future, it is important to investigate other properties or advantages of FTNet beyond classical neural networks, such as from the perspectives of optimization and generalization.

Acknowledgements

This research was supported by the National Key Research and Development Program of China (2020AAA0109401), NSFC (61921006, 62176117), and Collaborative Innovation Center of Novel Software Technology and Industrialization. The authors would like to thank Shoucheng Yu for helpful discussions about complex analysis and Shen-Huan Lyu, Peng Tan, and Zhi-Hao Tan for feedback on drafts of the paper.

A. Complete Proof of Eq. (9)

Proof. We only demonstrate the proof when σ_1 and σ_2 are continuous for simplicity. The case of almost everywhere continuous activation functions can be proven with a slight modification. The proof is divided into several steps.

Step 1. We prove that FNN with activation functions σ_1 and σ_2 can approximate the state transition function φ of DODS, defined in Eq. (5). Since the hidden state transition function φ is continuous, and the image of a continuous function defined on the compact set K is a compact set, there exists a convex compact set $K_1 \in \mathbb{R}^{H_\mathrm{D}}$, such that $\boldsymbol{h}_t \in K_1$ holds for any $t \in \{0, 1, \cdots, T\}$. Let $B_\infty(A, r) = \cup_{a\in A}\{b \mid \|b - a\|_\infty \leqslant r\}$ denote the neighborhood of the set A with radius r, and $K_2 = K \times B_\infty(K_1, 1)$ is the Cartesian product of K and $B_\infty(K_1, 1)$. It is easy to check that K_2 is convex and compact. Let $\boldsymbol{x} \in \mathbb{R}^I$ and $\boldsymbol{h} \in \mathbb{R}^{H_\mathrm{D}}$. Since both σ_1 and σ_2 are continuous almost everywhere and not polynomial almost everywhere, Lemma 5 indicates that for any $\varepsilon_1 > 0$, there exist $H_1, H_2 \in \mathbb{N}^+, \mathbf{A}_1 \in \mathbb{R}^{H_1\times I}, \mathbf{B}_1 \in \mathbb{R}^{H_1\times H_\mathrm{D}}, \mathbf{C}_1 \in \mathbb{R}^{H_\mathrm{D}\times H_1}, \boldsymbol{\theta}_1 \in \mathbb{R}^{H_1}, \mathbf{A}_2 \in \mathbb{R}^{H_2\times I}, \mathbf{B}_2 \in \mathbb{R}^{H_2\times H_\mathrm{D}}, \mathbf{C}_2 \in \mathbb{R}^{H_\mathrm{D}\times H_2}$, and $\boldsymbol{\theta}_2 \in \mathbb{R}^{H_2}$, where $\mathbf{C}_1$ and $\mathbf{C}_2$ are row independent, such that

$$\begin{aligned} &\sup_{(\boldsymbol{x},\boldsymbol{h})\in K_2} \|\varphi(\boldsymbol{x},\boldsymbol{h}) - \mathbf{C}_1\sigma_1\left(\mathbf{A}_1\boldsymbol{x} + \mathbf{B}_1\boldsymbol{h} - \boldsymbol{\theta}_1\right)\|_\infty \leqslant \varepsilon_1, \\ &\sup_{(\boldsymbol{x},\boldsymbol{h})\in K_2} \|\varphi(\boldsymbol{x},\boldsymbol{h}) - \mathbf{C}_2\sigma_2\left(\mathbf{A}_2\boldsymbol{x} + \mathbf{B}_2\boldsymbol{h} - \boldsymbol{\theta}_1\right)\|_\infty \leqslant \varepsilon_1. \end{aligned} \tag{42}$$

Step 2. We prove that RNN using the same weight matrices as FNN in Eq. (42) can approximate $\boldsymbol{h}_t$, the hidden state of DODS. Let $\boldsymbol{p}_0^{(1)} = \boldsymbol{q}_0^{(1)} = \boldsymbol{h}_0$. For any $t \in [T]$, define

$$\begin{aligned} \boldsymbol{p}_t^{(1)} &= \mathbf{C}_1\sigma_1\left(\mathbf{A}_1\boldsymbol{x}_t + \mathbf{B}_1\boldsymbol{p}_{t-1}^{(1)} - \boldsymbol{\theta}_1\right) \in \mathbb{R}^{H_\mathrm{D}}, \\ \boldsymbol{q}_t^{(1)} &= \mathbf{C}_2\sigma_2\left(\mathbf{A}_2\boldsymbol{x}_t + \mathbf{B}_2\boldsymbol{q}_{t-1}^{(1)} - \boldsymbol{\theta}_2\right) \in \mathbb{R}^{H_\mathrm{D}}. \end{aligned} \tag{43}$$

The above $\boldsymbol{p}_t^{(1)}$ and $\boldsymbol{q}_t^{(1)}$ are outputs of two different RNNs. We then prove that $\boldsymbol{p}_t^{(1)}$ and $\boldsymbol{q}_t^{(1)}$ can approximate $\boldsymbol{h}_t$. Let $u : [0, +\infty) \to \mathbb{R}$ be defined as

$$u(a) = \sup \left\{ \|\varphi(\boldsymbol{y}) - \varphi(\boldsymbol{z})\|_\infty \mid \boldsymbol{y}, \boldsymbol{z} \in K_2, \|\boldsymbol{y} - \boldsymbol{z}\|_\infty \leqslant a \right\}.$$

From Lemma 6, $u(a)$ is continuous. For any t, if $\|\boldsymbol{h}_{t-1} - \boldsymbol{p}_{t-1}^{(1)}\|_\infty \leqslant 1$, then $(\boldsymbol{x}_t, \boldsymbol{p}_{t-1}^{(1)}) \in K_2$, and one has

$$\begin{aligned}
\left\|\boldsymbol{h}_t - \boldsymbol{p}_t^{(1)}\right\|_\infty &= \left\|\varphi(\boldsymbol{x}_t, \boldsymbol{h}_{t-1}) - \mathbf{C}_1 \sigma_1 \left(\mathbf{A}_1 \boldsymbol{x}_t + \mathbf{B}_1 \boldsymbol{p}_{t-1}^{(1)} - \boldsymbol{\theta}_1\right)\right\|_\infty \\
&\leqslant \left\|\varphi(\boldsymbol{x}_t, \boldsymbol{h}_{t-1}) - \varphi\left(\boldsymbol{x}_1, \boldsymbol{p}_{t-1}^{(1)}\right)\right\|_\infty \\
&\quad + \left\|\varphi\left(\boldsymbol{x}_1, \boldsymbol{p}_{t-1}^{(1)}\right) - \mathbf{C}_1 \sigma_1 \left(\mathbf{A}_1 \boldsymbol{x}_t + \mathbf{B}_1 \boldsymbol{p}_{t-1}^{(1)} - \boldsymbol{\theta}_1\right)\right\|_\infty \\
&\leqslant u\left(\left\|\boldsymbol{h}_{t-1} - \boldsymbol{p}_{t-1}^{(1)}\right\|_\infty\right) + \varepsilon_1,
\end{aligned}$$

where the first equality holds from the definitions of DODS and $\boldsymbol{p}_t^{(1)}$, the first inequality holds because of the triangle inequality, and the second inequality holds based on the definition of $u(a)$, $(\boldsymbol{x}_t, \boldsymbol{p}_{t-1}^{(1)}) \in K_2$, and Eq. (42). Let $a_0 = 0$. For any $t \in [T]$, we define $a_t = u(a_{t-1}) + \varepsilon_1$. Then Lemma 7 indicates $\lim_{\varepsilon_1 \to 0_+} a_t = 0$ for any $t \in [T]$, i.e., for any $\varepsilon_2 \in (0, 1)$, there exists $\delta_1(\varepsilon_2) > 0$, such that for any $\varepsilon_1 \leqslant \delta_1(\varepsilon_2)$, $a_t \leqslant \varepsilon_2$ holds for any $t \in [T]$. When $\varepsilon_1 \leqslant \delta_1(\varepsilon_2)$, it is easy to see that $\|\boldsymbol{h}_t - \boldsymbol{p}_t^{(1)}\|_\infty \leqslant a_t \leqslant \varepsilon_2$ holds for any $t \in [T]$. The same conclusion can be proven for $\boldsymbol{q}_t^{(1)}$ in the same way. Thus, for any $\varepsilon_1 \leqslant \delta_1(\varepsilon_2)$, one has

$$\max_{t \in [T]} \left\|\boldsymbol{h}_t - \boldsymbol{p}_t^{(1)}\right\|_\infty \leqslant \varepsilon_2 \quad \text{and} \quad \max_{t \in [T]} \left\|\boldsymbol{h}_t - \boldsymbol{q}_t^{(1)}\right\|_\infty \leqslant \varepsilon_2. \tag{44}$$

Step 3. Transformation is used to eliminate the matrices $\mathbf{C}_1$ and $\mathbf{C}_2$ in Eq. (43), which is the preparation to approximate $\boldsymbol{h}_t$ using additive FTNet. Since $\mathbf{C}_1, \mathbf{C}_2$ are row independent, both $\mathbf{C}_1 \boldsymbol{x} = \boldsymbol{p}_0^{(1)}$ and $\mathbf{C}_2 \boldsymbol{x} = \boldsymbol{q}_0^{(1)}$ have solutions. Let $\boldsymbol{p}_0^{(2)}$ and $\boldsymbol{q}_0^{(2)}$ be the solutions of the above equations, respectively, i.e., $\mathbf{C}_1 \boldsymbol{p}_0^{(2)} = \boldsymbol{p}_0^{(1)}$ and $\mathbf{C}_2 \boldsymbol{q}_0^{(2)} = \boldsymbol{q}_0^{(1)}$. Define

$$\begin{aligned}
\boldsymbol{p}_t^{(2)} &= \sigma_1 \left(\mathbf{A}_1 \boldsymbol{x}_t + \mathbf{B}_1 \mathbf{C}_1 \boldsymbol{p}_{t-1}^{(2)} - \boldsymbol{\theta}_1\right) \in \mathbb{R}^{H_1}, \\
\boldsymbol{q}_t^{(2)} &= \sigma_2 \left(\mathbf{A}_2 \boldsymbol{x}_t + \mathbf{B}_2 \mathbf{C}_2 \boldsymbol{q}_{t-1}^{(2)} - \boldsymbol{\theta}_2\right) \in \mathbb{R}^{H_2}.
\end{aligned} \tag{45}$$

We claim that, for any $t \in \{0, 1, \cdots, T\}$,

$$\boldsymbol{p}_t^{(1)} = \mathbf{C}_1 \boldsymbol{p}_t^{(2)}, \quad \boldsymbol{q}_t^{(1)} = \mathbf{C}_2 \boldsymbol{q}_t^{(2)}. \tag{46}$$

Since the proof of $\boldsymbol{q}_t^{(2)}$ is similar to that of $\boldsymbol{p}_t^{(2)}$, we only give the proof of $\boldsymbol{p}_t^{(2)}$ using mathematical induction as follows.

1. For $t = 0$, the claim holds from the definition of $\boldsymbol{p}_0^{(2)}$.

2. Suppose that the claim holds for $t = k$, where $k \in \{0, 1, \cdots, T-1\}$. Thus, one has

$$\begin{aligned}
\boldsymbol{p}_{k+1}^{(1)} &= \mathbf{C}_1 \sigma_1 \left(\mathbf{A}_1 \boldsymbol{x}_{k+1} + \mathbf{B}_1 \boldsymbol{p}_k^{(1)} - \boldsymbol{\theta}_1 \right) \\
&= \mathbf{C}_1 \sigma_1 \left(\mathbf{A}_1 \boldsymbol{x}_{k+1} + \mathbf{B}_1 \mathbf{C}_1 \boldsymbol{p}_k^{(2)} - \boldsymbol{\theta}_1 \right) \\
&= \mathbf{C}_1 \boldsymbol{p}_{k+1}^{(2)},
\end{aligned}$$

where the first equality holds from the definition of $\boldsymbol{p}_t^{(1)}$ with $t = k + 1$ in Eq. (43), the second equality holds because of the induction hypothesis, and the third equality holds based on the definition of $\boldsymbol{p}_t^{(2)}$ with $t = k + 1$. Thus, the claim holds for $t = k + 1$.

Step 4. We prove that additive FTNet can approximate $\boldsymbol{h}_t$ by unifying the weight matrices in Eq. (45). Let $H_3 = H_1 + H_2$. For any $t \in [T]$, define

$$\begin{aligned}
\boldsymbol{p}_t^{(3)} &= \sigma_1 \left(\mathbf{A}_3 \boldsymbol{x}_t + \mathbf{B}_3 \boldsymbol{q}_{t-1}^{(3)} - \boldsymbol{\theta}_3 \right) \in \mathbb{R}^{H_3}, \\
\boldsymbol{q}_t^{(3)} &= \sigma_2 \left(\mathbf{A}_3 \boldsymbol{x}_t + \mathbf{B}_3 \boldsymbol{q}_{t-1}^{(3)} - \boldsymbol{\theta}_3 \right) \in \mathbb{R}^{H_3},
\end{aligned} \tag{47}$$

where

$$\boldsymbol{q}_0^{(3)} = \begin{bmatrix} \mathbf{0} \\ \boldsymbol{q}_0^{(2)} \end{bmatrix}, \quad \mathbf{A}_3 = \begin{bmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \end{bmatrix}, \quad \mathbf{B}_3 = \begin{bmatrix} \mathbf{0} & \mathbf{B}_1 \mathbf{C}_2 \\ \mathbf{0} & \mathbf{B}_2 \mathbf{C}_2 \end{bmatrix}, \quad \boldsymbol{\theta}_3 = \begin{bmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{bmatrix}.$$

For $\boldsymbol{q}_t^{(3)}$, we claim that $\boldsymbol{q}_t^{(3)} = [\times_t; \boldsymbol{q}_t^{(2)}]$ holds for any $t \in [T]$, where $\times_t \in \mathbb{R}^{H_1 \times 1}$ is a vector that we do not care, because it has no contribution to the iteration or output in the above additive FTNet. We prove the claim about $\boldsymbol{q}_t^{(3)}$ by mathematical induction as follows.

1. For $t = 1$, one has

$$\begin{aligned}
\boldsymbol{q}_1^{(3)} =& \sigma_2 \left(\mathbf{A}_3 \boldsymbol{x}_1 + \mathbf{B}_3 \boldsymbol{q}_0^{(3)} - \boldsymbol{\theta}_3 \right) \\
=& \sigma_2 \left(\begin{bmatrix} \mathbf{A}_1 \boldsymbol{x}_1 + \mathbf{B}_1 \mathbf{C}_2 \boldsymbol{q}_0^{(2)} - \boldsymbol{\theta}_1 \\ \mathbf{A}_2 \boldsymbol{x}_1 + \mathbf{B}_2 \mathbf{C}_2 \boldsymbol{q}_0^{(2)} - \boldsymbol{\theta}_2 \end{bmatrix} \right) \\
=& \left[\times_1; \boldsymbol{q}_1^{(2)} \right],
\end{aligned}$$

where the first equality holds according to the definition of $\boldsymbol{q}_t^{(3)}$ with $t = 1$ in Eq. (47), the second equality holds based on the definitions of $\mathbf{A}_3, \mathbf{B}_3, \boldsymbol{\theta}_3$, and the third equality holds from the definition of $\boldsymbol{q}_t^{(2)}$ with $t = 1$ in Eq. (45). Thus, the claim holds for $t = 1$.

2. Suppose that the claim holds for $t = k$ where $k \in [T-1]$. Thus, one has

$$\begin{aligned}\boldsymbol{q}_{k+1}^{(3)} =& \sigma_2\left(\mathbf{A}_3\boldsymbol{x}_{k+1} + \mathbf{B}_3\boldsymbol{q}_k^{(3)} - \boldsymbol{\theta}_3\right)\\ =& \sigma_2\left(\left[\begin{array}{c}\mathbf{A}_1\boldsymbol{x}_{k+1} + \mathbf{B}_1\mathbf{C}_2\boldsymbol{q}_k^{(2)} - \boldsymbol{\theta}_1\\ \mathbf{A}_2\boldsymbol{x}_{k+1} + \mathbf{B}_2\mathbf{C}_2\boldsymbol{q}_k^{(2)} - \boldsymbol{\theta}_2\end{array}\right]\right)\\ =& \left[\times_{k+1}; \boldsymbol{q}_{k+1}^{(2)}\right],\end{aligned}$$

where the first equality holds from the definition of $\boldsymbol{q}_t^{(3)}$ with $t = k+1$, the second equality holds because of the definitions of $\mathbf{A}_3, \mathbf{B}_3, \boldsymbol{\theta}_3$, and the third equality holds based on the definition of $\boldsymbol{q}_t^{(2)}$ with $t = k+1$. Thus, the claim holds for $t = k+1$.

We then study the property of $\boldsymbol{p}_t^{(3)}$. Let

$$\mathbf{C}_3 = \left[\begin{array}{cc}\mathbf{C}_1 & \mathbf{0}^{H_\mathrm{D}\times H_2}\end{array}\right].$$

If $\|\boldsymbol{h}_{t-1} - \boldsymbol{q}_{t-1}^{(1)}\|_\infty \leqslant 1$, then $(\boldsymbol{x}_t, \boldsymbol{q}_{t-1}^{(1)}) \in K_2$, and one has

$$\begin{aligned}\left\|\boldsymbol{h}_t - \mathbf{C}_3\boldsymbol{p}_t^{(3)}\right\|_\infty =& \left\|\varphi\left(\boldsymbol{x}_t, \boldsymbol{h}_{t-1}\right) - \mathbf{C}_1\sigma_1\left(\mathbf{A}_1\boldsymbol{x}_t + \mathbf{B}_1\mathbf{C}_2\boldsymbol{q}_{t-1}^{(2)} - \boldsymbol{\theta}_1\right)\right\|_\infty\\ \leqslant& \left\|\varphi\left(\boldsymbol{x}_t, \boldsymbol{h}_{t-1}\right) - \varphi\left(\boldsymbol{x}_t, \boldsymbol{q}_{t-1}^{(1)}\right)\right\|_\infty\\ &+ \left\|\varphi\left(\boldsymbol{x}_t, \boldsymbol{q}_{t-1}^{(1)}\right) - \mathbf{C}_1\sigma_1\left(\mathbf{A}_1\boldsymbol{x}_t + \mathbf{B}_1\boldsymbol{q}_{t-1}^{(1)} - \boldsymbol{\theta}_1\right)\right\|_\infty\\ \leqslant& u\left(\varepsilon_2\right) + \varepsilon_1,\end{aligned}$$

where the first equality holds because of the definitions of DODS, $\mathbf{C}_3$, and $\boldsymbol{p}_t^{(3)}$, the first inequality holds based on the triangle inequality and Eq. (46), and the second inequality holds based on the definition of $u(a)$, $(\boldsymbol{x}_t, \boldsymbol{q}_{t-1}^{(1)}) \in K_2$, Eqs. (42) and (44). Then according to the continuity of $u(a)$ and $u(0) = 0$, for any $\varepsilon_3 > 0$, there exist $\delta_2(\varepsilon_3)$ and $\delta_3(\varepsilon_3)$, such that if $\varepsilon_1 \leqslant \delta_2(\varepsilon_3)$ and $\varepsilon_2 \leqslant \delta_3(\varepsilon_3)$, one has

$$\max_{t\in[T]}\left\|\boldsymbol{h}_t - \mathbf{C}_3\boldsymbol{p}_t^{(3)}\right\|_\infty \leqslant \varepsilon_3. \tag{48}$$

Step 5. The output function ψ in Eq. (5) can be approximated by an FNN. Let τ_3 be any continuous non-polynomial function. According to Lemma 1, for any $\varepsilon_4 > 0$, there exist $H_4 \in \mathbb{N}^+, \mathbf{A}_4 \in \mathbb{R}^{H_4\times H_\mathrm{D}}, \mathbf{B}_4 \in \mathbb{R}^{O\times H_4}, \boldsymbol{\theta}_4 \in \mathbb{R}^{H_4\times 1}$, s.t.

$$\sup_{\boldsymbol{h}\in B(K_1,1)}\left\|\psi(\boldsymbol{h}) - \mathbf{B}_4\tau_3\left(\mathbf{A}_4\boldsymbol{h} - \boldsymbol{\theta}_4\right)\right\|_\infty \leqslant \varepsilon_4. \tag{49}$$

For any $t \in [T]$, define $\boldsymbol{y}_t^{(1)} = \mathbf{B}_4\tau_3(\mathbf{A}_4\mathbf{C}_3\boldsymbol{p}_t^{(3)} - \boldsymbol{\theta}_4)$. Substituting the definition of $\boldsymbol{p}_t^{(3)}$ in Eq. (47) into the above definition, one has, for any $t \in [T]$,

$$\boldsymbol{y}_t^{(1)} = \mathbf{B}_4\tau_3\left(\mathbf{A}_4\mathbf{C}_1\sigma_1\left(\mathbf{A}_1\boldsymbol{x}_t + \mathbf{B}_1\mathbf{C}_2\boldsymbol{q}_{t-1}^{(2)} - \boldsymbol{\theta}_3\right) - \boldsymbol{\theta}_4\right). \tag{50}$$

Since σ_2 is continuous, and $\boldsymbol{x}_t \in K$ holds for any $t \in [T]$, there exists a compact set K_3, s.t. $\boldsymbol{q}_t^{(2)} \in K_3$ holds for any $t \in [T]$. Since σ_1 is continuous and not polynomial, Lemma 1 implies that for any $\varepsilon_5 > 0$, there exist $H_5 \in \mathbb{N}^+, \mathbf{A}_5 \in \mathbb{R}^{H_5 \times I}, \mathbf{B}_5 \in \mathbb{R}^{H_5 \times H_2}, \mathbf{C}_5 \in \mathbb{R}^{O \times H_5}$, and $\boldsymbol{\theta}_5 \in \mathbb{R}^{H_5}$, s.t.

$$\sup_{(\boldsymbol{x},\boldsymbol{q}) \in K \times K_3} \left\| \mathbf{B}_4\tau_3\left(\mathbf{A}_4\mathbf{C}_1\sigma_1\left(\mathbf{A}_1\boldsymbol{x} + \mathbf{B}_1\mathbf{C}_2\boldsymbol{q} - \boldsymbol{\theta}_3\right) - \boldsymbol{\theta}_4\right) - \mathbf{C}_5\sigma_1\left(\mathbf{A}_5\boldsymbol{x} + \mathbf{B}_5\boldsymbol{q} - \boldsymbol{\theta}_5\right) \right\|_\infty \leqslant \varepsilon_5. \tag{51}$$

For any $t \in [T]$, define

$$\begin{aligned} \boldsymbol{p}_t^{(5)} &= \sigma_1\left(\mathbf{A}_5\boldsymbol{x}_t + \mathbf{B}_5\boldsymbol{q}_{t-1}^{(2)} - \boldsymbol{\theta}_5\right) \in \mathbb{R}^{H_5}, \\ \boldsymbol{q}_t^{(5)} &= \sigma_2\left(\mathbf{A}_5\boldsymbol{x}_t + \mathbf{B}_5\boldsymbol{q}_{t-1}^{(2)} - \boldsymbol{\theta}_5\right) \in \mathbb{R}^{H_5}. \end{aligned} \tag{52}$$

Then Eqs. (50) and (51) imply that

$$\max_{t \in [T]} \left\| \boldsymbol{y}_t^{(1)} - \mathbf{C}_5\boldsymbol{p}_t^{(5)} \right\|_\infty \leqslant \varepsilon_5. \tag{53}$$

Step 6. The final additive FTNet is constructed to approximate the target DODS. Let $H = H_3 + H_5$, and define the additive FTNet $f_{+,\mathrm{R}}$ as follows

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_3 \\ \mathbf{A}_5 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \mathbf{B}_3 & \mathbf{0} \\ \mathbf{B}_6 & \mathbf{0} \end{bmatrix}, \quad \boldsymbol{\zeta} = \begin{bmatrix} \boldsymbol{\theta}_3 \\ \boldsymbol{\theta}_5 \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} \mathbf{0} & \mathbf{C}_5 \end{bmatrix}, \quad \boldsymbol{q}_0 = \begin{bmatrix} \boldsymbol{q}_0^{(3)} \\ \boldsymbol{q}_0^{(5)} \end{bmatrix}, \tag{54}$$

where $\mathbf{B}_6 = [\mathbf{0}, \mathbf{B}_5]$ pads the matrix $\mathbf{B}_5$ with 0. We claim that $\boldsymbol{p}_t = [\boldsymbol{p}_t^{(3)}; \boldsymbol{p}_t^{(5)}]$ and $\boldsymbol{q}_t = [\boldsymbol{q}_t^{(3)}; \boldsymbol{q}_t^{(5)}]$ hold for any $t \in [T]$. The proof of $\boldsymbol{q}_t$ is similar to that of $\boldsymbol{p}_t$, and we only prove the claim of $\boldsymbol{p}_t$ using mathematical induction as follows.

1. For $t = 1$, one has

$$\begin{aligned} \boldsymbol{p}_1 =& \sigma_1\left(\mathbf{A}\boldsymbol{x}_1 + \mathbf{B}\boldsymbol{q}_0 - \boldsymbol{\zeta}\right) \\ =& \sigma_1\left(\begin{bmatrix} \mathbf{A}_3\boldsymbol{x}_1 + \mathbf{B}_3\boldsymbol{q}_0^{(3)} - \boldsymbol{\theta}_3 \\ \mathbf{A}_5\boldsymbol{x}_1 + \begin{bmatrix} \mathbf{0} & \mathbf{B}_5 \end{bmatrix} \boldsymbol{q}_0^{(3)} - \boldsymbol{\theta}_5 \end{bmatrix}\right) \\ =& \left[\boldsymbol{p}_1^{(3)}; \boldsymbol{p}_1^{(5)}\right], \end{aligned}$$

   where the first equality holds because of the definition of $\boldsymbol{p}_t$ with $t = 1$ in Eq. (8), the second equality holds according to Eq. (54), and the third equality holds based on the definition of $\boldsymbol{q}_0^{(3)}$ in Eq. (54), the definitions of $\boldsymbol{p}_t^{(3)}, \boldsymbol{p}_t^{(5)}$ with $t = 1$ in Eqs. (47) and (52).

2. Suppose that the claim holds for $t = k$ where $k \in [T-1]$. Thus, one has

$$\begin{aligned} \boldsymbol{p}_{k+1} =& \sigma_1\left(\mathbf{A}\boldsymbol{x}_{k+1} + \mathbf{B}\boldsymbol{q}_k - \boldsymbol{\zeta}\right) \\ =& \sigma_1\left(\left[\begin{array}{c} \mathbf{A}_3\boldsymbol{x}_{k+1} + \mathbf{B}_3\boldsymbol{q}_k^{(3)} - \boldsymbol{\theta}_3 \\ \mathbf{A}_5\boldsymbol{x}_{k+1} + \left[\begin{array}{cc} \mathbf{0} & \mathbf{B}_5 \end{array}\right]\boldsymbol{q}_k^{(3)} - \boldsymbol{\theta}_5 \end{array}\right]\right) \\ =& \left[\boldsymbol{p}_{k+1}^{(3)}; \boldsymbol{p}_{k+1}^{(5)}\right], \end{aligned}$$

where the first equality holds from the definition of $\boldsymbol{p}_t$ with $t = k+1$, the second equality holds because of the definitions of $\mathbf{A}, \mathbf{B}, \boldsymbol{\zeta}, \boldsymbol{q}_0$, and the third equality holds based on the definitions of $\boldsymbol{p}_t^{(3)}, \boldsymbol{p}_t^{(5)}$ with $t = k+1$, and the conclusion $\boldsymbol{q}_t^{(3)} = [\times_t; \boldsymbol{q}_t^{(2)}]$ with $t = k$.

For any $t \in [T]$, one has

$$\begin{aligned} \|y_t - y_{+,t}\|_\infty \leqslant& \left\|\psi(\boldsymbol{h}_t) - \psi\left(\mathbf{C}_3\boldsymbol{p}_t^{(3)}\right)\right\|_\infty + \left\|\psi\left(\mathbf{C}_3\boldsymbol{p}_t^{(3)}\right) - \boldsymbol{y}_t^{(1)}\right\|_\infty + \left\|\boldsymbol{y}_t^{(1)} - \mathbf{C}_5\boldsymbol{p}_t^{(5)}\right\|_\infty \\ \leqslant& u\left(\left\|\boldsymbol{h}_t - \mathbf{C}_3\boldsymbol{p}_t^{(3)}\right\|_\infty\right) + \varepsilon_4 + \varepsilon_5 \\ \leqslant& u(\varepsilon_3) + \varepsilon_4 + \varepsilon_5, \end{aligned}$$

where the first inequality holds because of the triangle inequality, the definitions of DODS and $\hat{\boldsymbol{y}}_t$, the second inequality holds based on the definition of $u(a)$, Eq. (49) with $\boldsymbol{h} = \mathbf{C}_3\boldsymbol{p}_t^{(3)}$, and Eq. (53), and the third inequality holds in view of Eq. (48). Since $u(a)$ is continuous, and $u(0) = 0$, for any $\varepsilon_6 > 0$, there exists $\delta_4(\varepsilon_6) > 0$, such that for any $\varepsilon_3 \leqslant \delta_4(\varepsilon_6)$, one has $u(\varepsilon_3) \leqslant \varepsilon_6$. Let $\varepsilon_5 = \varepsilon_4 = \varepsilon_6 = \varepsilon/3$, then one has $\max_{t\in[T]} \|y_t - y_{+,t}\|_\infty \leqslant \varepsilon$, i.e.,

$$\sup_{\boldsymbol{x}_{1:T}\in K^T} \|f_{\mathrm{D}}(\boldsymbol{x}_{1:T}) - f_{+,\mathrm{R}}(\boldsymbol{x}_{1:T})\|_\infty \leqslant \varepsilon,$$

which completes the proof. □

B. Useful Lemmas

Lemma 5 *Suppose that $\sigma : \mathbb{R} \to \mathbb{R}$ is continuous almost everywhere and not polynomial almost everywhere. Then for any $\varepsilon > 0$, any continuous function $f : \mathbb{R}^I \to \mathbb{R}^O$, and any compact set $K \subset \mathbb{R}^I$, there exist $H \in \mathbb{N}^+$, $\mathbf{W} \in \mathbb{R}^{H\times I}$, $\boldsymbol{\theta} \in \mathbb{R}^H$, and row independent $\mathbf{U} \in \mathbb{R}^{O\times H}$, such that*

$$\|f(\boldsymbol{x}) - \mathbf{U}\sigma(\mathbf{W}\boldsymbol{x} - \boldsymbol{\theta})\|_{L^\infty(K)} \leqslant \varepsilon.$$

Proof. For any $\varepsilon > 0$, continuous function $f : \mathbb{R}^I \to \mathbb{R}^O$, and compact set $K \subset \mathbb{R}^I$, Lemma 1 indicates that there exist $H_1 \in \mathbb{N}^+$, $\mathbf{W}_1 \in \mathbb{R}^{H_1 \times I}$, $\boldsymbol{\theta}_1 \in \mathbb{R}^{H_1}$, and $\mathbf{U}_1 \in \mathbb{R}^{O \times H_1}$, s.t.

$$\|f(\boldsymbol{x}) - \mathbf{U}_1\sigma\left(\mathbf{W}_1\boldsymbol{x} - \boldsymbol{\theta}_1\right)\|_{L^\infty(K)} \leqslant \varepsilon.$$

Define a new FNN with hidden size $H = H_1 + O$ as follows

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_1 \\ \mathbf{0} \end{bmatrix}, \quad \boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\theta}_1 \\ \mathbf{0} \end{bmatrix}, \quad \mathbf{U} = \begin{bmatrix} \mathbf{U}_1 & \mathbf{I}_O \end{bmatrix},$$

where $\mathbf{I}_O$ is the identity matrix of size $O \times O$. Then it is easy to see that $\mathbf{U}$ is row independent and

$$\|f(\boldsymbol{x}) - \mathbf{U}\sigma(\mathbf{W}\boldsymbol{x} - \boldsymbol{\theta})\|_{L^\infty(K)} = \|f(\boldsymbol{x}) - \mathbf{U}_1\sigma(\mathbf{W}_1\boldsymbol{x} - \boldsymbol{\theta}_1)\|_{L^\infty(K)} \leqslant \varepsilon,$$

which completes the proof. □

Lemma 6 *Let $\varphi : \mathbb{R}^n \to \mathbb{R}$ be continuous, and $K_2 \subset \mathbb{R}^n$ is a convex compact set. Then $u(a) = \sup\{\|\varphi(\boldsymbol{y}) - \varphi(\boldsymbol{z})\|_\infty \mid \boldsymbol{y}, \boldsymbol{z} \in K_2, \|\boldsymbol{y} - \boldsymbol{z}\|_\infty \leqslant a\}$ is continuous on $[0, +\infty)$.*

Proof. The proof is divided into several steps.

Step 1. We prove that u is well-defined and bounded. Since any continuous function is bounded on any compact set, there exists $U_\varphi \in \mathbb{R}$, such that $\|\varphi(\boldsymbol{y})\|_\infty \leqslant U_\varphi$ holds for any $\boldsymbol{y} \in K_2$. Then according to the triangle inequality, $\|\varphi(\boldsymbol{y}) - \varphi(\boldsymbol{z})\|_\infty \leqslant 2U_\varphi$ holds for any $\boldsymbol{y}, \boldsymbol{z} \in K_2$, i.e., $|u(a)| \leqslant 2U_\varphi$ holds for any $a \in [0, +\infty)$. Thus, $u(a)$ is well-defined and bounded on $[0, +\infty)$.

Step 2. It is obvious that $u(0) = 0$.

Step 3. We prove that u is monotonically increasing. Let $0 \leqslant a_1 < a_2$. For any $\varepsilon > 0$, according to the definition of supremum, there exist $\boldsymbol{y}_1, \boldsymbol{z}_1 \in K_2$, such that $\|\varphi(\boldsymbol{y}_1) - \varphi(\boldsymbol{z}_1)\|_\infty \geqslant u(a_1) - \varepsilon$ and $\|\boldsymbol{y}_1 - \boldsymbol{z}_1\|_\infty \leqslant a_1$. Since $a_1 < a_2$, one has $\|\boldsymbol{y}_1 - \boldsymbol{z}_1\|_\infty \leqslant a_2$. Thus, one has

$$u(a_2) \geqslant \|\varphi(\boldsymbol{y}_1) - \varphi(\boldsymbol{z}_1)\|_\infty \geqslant u(a_1) - \varepsilon.$$

According to the arbitrariness of ε, one has $u(a_2) \geqslant u(a_1)$. Therefore, $u(a)$ is a monotonically increasing function.

Step 4. We prove that u is right continuous. Let $b \in [0, +\infty)$ be an arbitrary non-negative real number. Since $u(a)$ is bounded and monotonically increasing on $[0, +\infty)$, the limit $\lim_{a \to b_+} u(a)$ exists. Let $u_+ = \lim_{a \to b_+} u(a)$ denote this limit. If $u_+ \neq u(b)$, then one has $u_+ > u(b)$ since $u(a)$ is monotonically increasing. Since any continuous function on a compact set is uniformly continuous, there exists $\delta > 0$, such that for any $\boldsymbol{y}, \boldsymbol{z} \in K_2$, $\|\boldsymbol{y} - \boldsymbol{z}\|_\infty \leqslant \delta$ indicates $\|\varphi(\boldsymbol{y}) - \varphi(\boldsymbol{z})\|_\infty \leqslant [u_+ - u(b)]/3$. Since $u(a)$ is monotonically increasing, one has $u(b+\delta) \geqslant u_+$. According to the definition of supremum, there exist $\boldsymbol{y}_2, \boldsymbol{z}_2 \in K_2$, such that $\|\boldsymbol{y}_2 - \boldsymbol{z}_2\|_\infty \leqslant b + \delta$ and

$$\|\varphi(\boldsymbol{y}_2) - \varphi(\boldsymbol{z}_2)\|_\infty \geqslant u_+ - [u_+ - u(b)]/3.$$

Let $\boldsymbol{\xi} = \lambda \boldsymbol{z}_2 + (1-\lambda)\boldsymbol{y}_2$, where $\lambda = b(b+\delta)^{-1} \in [0,1]$. Since $\boldsymbol{y}_2, \boldsymbol{z}_2 \in K_2$, and K_2 is convex, one has $\boldsymbol{\xi} \in K_2$. According to the homogeneity of norm, one has

$$\begin{aligned}
&\|\boldsymbol{\xi} - \boldsymbol{y}_2\|_\infty = \lambda \|\boldsymbol{z}_2 - \boldsymbol{y}_2\|_\infty \leqslant \lambda(b+\delta) = b, \\
&\|\boldsymbol{z}_2 - \boldsymbol{\xi}\|_\infty = (1-\lambda)\|\boldsymbol{z}_2 - \boldsymbol{y}_2\|_\infty \leqslant (1-\lambda)(b+\delta) = \delta.
\end{aligned}$$

Thus, one has

$$\begin{aligned}
u(b) &\geqslant \|\varphi(\boldsymbol{\xi}) - \varphi(\boldsymbol{y}_2)\|_\infty \\
&\geqslant \|\varphi(\boldsymbol{z}_2) - \varphi(\boldsymbol{y}_2)\|_\infty - \|\varphi(\boldsymbol{z}_2) - \varphi(\boldsymbol{\xi})\|_\infty \\
&\geqslant \ (u_+ - [u_+ - u(b)]/3) - [u_+ - u(b)]/3 \\
&= u(b) + [u_+ - u(b)]/3 \\
&> u(b),
\end{aligned}$$

where the first inequality holds from $\|\boldsymbol{\xi} - \boldsymbol{y}_2\|_\infty = b$, the second inequality holds based on the triangle inequality, and the third inequality holds because of the definitions of $\boldsymbol{y}_2, \boldsymbol{z}_2$, and $\|\boldsymbol{z}_2 - \boldsymbol{\xi}\|_\infty = b$. The above inequality is a contradiction, which means that $u_+ \neq u(b)$ does not hold. Therefore, one has $u_+ = u(b)$, which means that $u(a)$ is right continuous.

Step 5. Similarly, we can prove that $u(a)$ is left continuous. Therefore, $u(a)$ is continuous. $\square$

Lemma 7 *Let $a_0 = 0$. For any $t \in [T]$, let $a_t = u(a_{t-1}) + \varepsilon$, where $T \in \mathbb{N}^+$ is a positive integer, $u : \mathbb{R} \to \mathbb{R}$ is continuous, and $u(0) = 0$. Then $\lim_{\varepsilon \to 0+} a_t = 0$ holds for any $t \in [T]$.*

Proof. We prove this lemma by mathematical induction.

1. For $t = 1$, one has

$$\lim_{\varepsilon_1 \to 0_+} a_1 = \lim_{\varepsilon_1 \to 0_+} u(a_0) + \varepsilon_1 = 0 + 0 = 0,$$

   where the first equality holds from the definition of a_t with $t = 1$, and the second equality holds based on $a_0 = 0$ and $u(0) = 0$. Thus, the conclusion holds for $t = 1$.

2. If the conclusion holds for $t = k$ where $k \in [T-1]$, then

$$\lim_{\varepsilon_1 \to 0_+} a_{k+1} = \lim_{\varepsilon_1 \to 0_+} u(a_k) + \varepsilon_1 = u\left(\lim_{\varepsilon_1 \to 0_+} a_k\right) + 0 = u(0) + 0 = 0,$$

where the first equality holds from the definition of a_t with $t = k+1$, the second equality holds based on the continuity of $u(a)$, the third equality holds because of the induction hypothesis, and the fourth equality holds since $u(0) = 0$. Thus, the conclusion holds for $t = k+1$.

Then mathematical induction completes the proof. □

References

[1] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *Proceedings of the 36th International Conference on Machine Learning*, pages 242–252, 2019.

[2] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. On the convergence rate of training recurrent neural networks. In *Advances in Neural Information Processing Systems 32*, pages 6676–6688, 2019.

[3] Paolo Arena, Luigi Fortuna, R Re, and Maria Gabriella Xibilia. On the capability of neural networks with complex neurons in complex valued functions approximation. In *Proceedings of IEEE International Symposium on Circuits and Systems*, pages 2168–2171, 1993.

[4] Martin Arjovsky, Amar Shah, and Yoshua Bengio. Unitary evolution recurrent neural networks. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 1120–1128, 2016.

[5] Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

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

[7] Tommy WS Chow and Xiao-Dong Li. Modeling of continuous time dynamical systems with input by recurrent neural networks. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 47(4):575–578, 2000.

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

[9] Chen Debao. Degree of approximation by superpositions of a sigmoidal function. *Approximation Theory and its Applications*, 9(3):17–28, 1993.

[10] Tian Ding, Dawei Li, and Ruoyu Sun. Suboptimal local minima exist for wide neural networks with smooth activations. *Mathematics of Operations Research*, 2022.

[11] Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 1675–1685, 2019.

[12] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

[13] Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. In *Proceedings of the 29th Conference on Learning Theory*, pages 907–940, 2016.

[14] Kunihiko Fukushima. Visual feature extraction by a multilayered network of analog threshold elements. *IEEE Transactions on Systems Science and Cybernetics*, 5(4):322–333, 1969.

[15] Ken-Ichi Funahashi. On the approximate realization of continuous mappings by neural networks. *Neural Networks*, 2(3):183–192, 1989.

[16] Ken-ichi Funahashi and Yuichi Nakamura. Approximation of dynamical systems by continuous time recurrent neural networks. *Neural Networks*, 6(6):801–806, 1993.

[17] Marco Gori and Alberto Tesi. On the problem of local minima in backpropagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(1):76–86, 1992.

[18] Nitzan Guberman. On complex valued convolutional neural networks. *arXiv:1602.09046*, 2016.

[19] Robert Clifford Gunning and Hugo Rossi. *Analytic Functions of Several Complex Variables*, volume 368. American Mathematical Society, 2009.

[20] Weili Guo, Guangyu Li, Jianfeng Lu, and Jian Yang. Singular learning of deep multilayer perceptrons for eeg-based emotion recognition. *Frontiers in Computer Science*, 3:786964, 2021.

[21] Alan L Hodgkin and Andrew F Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology*, 117(4):500–544, 1952.

[22] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

[23] Kurt Hornik, Maxwell Stinchcombe, Halbert White, and Peter Auer. Degree of approximation results for feedforward networks approximating unknown mappings and their derivatives. *Neural Computation*, 6(6):1262–1275, 1994.

[24] De-Shuang Huang. The local minima-free condition of feedforward neural networks for outer-supervised learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 28(3):477–480, 1998.

[25] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems 31*, pages 8580–8589, 2018.

[26] Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems 29*, pages 586–594, 2016.

[27] Kenji Kawaguchi and Yoshua Bengio. Depth with nonlinearity creates no bad local minima in ResNets. *Neural Networks*, 118:167–174, 2019.

[28] Kenji Kawaguchi and Leslie Kaelbling. Elimination of all bad local minima in deep learning. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, pages 853–863, 2020.

[29] Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.

[30] Jordi Laguarta and Brian Subirana. Longitudinal speech biomarkers for automated alzheimer's detection. *Frontiers in Computer Science*, 3:624694, 2021.

[31] Thomas Laurent and James Brecht. Deep linear networks with arbitrary loss: All local minima are global. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2902–2907, 2018.

[32] Moshe Leshno, Vladimir Ya Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, 1993.

[33] Xiao-Dong Li, John KL Ho, and Tommy WS Chow. Approximation of dynamical time-variant systems by continuous-time recurrent neural networks. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 52(10):656–660, 2005.

[34] Y Li and Y Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems 31*, pages 8168–8177, 2018.

[35] Shiyu Liang, Ruoyu Sun, Jason D Lee, and R Srikant. Adding one neuron can eliminate all bad local minima. In *Advances in Neural Information Processing Systems 31*, pages 4355–4365, 2018.

[36] Shiyu Liang, Ruoyu Sun, Yixuan Li, and Rayadurgam Srikant. Understanding the loss surface of neural networks for binary classification. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2835–2843, 2018.

[37] Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:85–116, 2022.

[38] Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. The expressive power of neural networks: A view from the width. In *Advances in Neural Information Processing Systems 30*, pages 6232–6240, 2017.

[39] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4):115–133, 1943.

[40] Hrushikesh Narhar Mhaskar and Charles A Micchelli. Dimension-independent bounds on the degree of approximation by neural networks. *IBM Journal of Research and Development*, 38(3):277–284, 1994.

[41] Quynh Nguyen and Matthias Hein. The loss surface of deep and wide neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 2603–2612, 2017.

[42] Quynh Nguyen and Matthias Hein. Optimization landscape and expressivity of deep CNNs. In *Proceedings of the 35th International Conference on Machine Learning*, pages 3730–3739, 2018.

[43] Quynh Nguyen, Mahesh Chandra Mukkamala, and Matthias Hein. On the loss landscape of a class of deep neural networks with no bad local valleys. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

[44] Timothy Poston, C-N Lee, Y Choie, and Yonghoon Kwon. Local minima and back propagation. In *Proceedings of International Joint Conference on Neural Networks*, pages 173–176, 1991.

[45] Itay Safran and Ohad Shamir. Depth-width tradeoffs in approximating natural functions with neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 2979–2987, 2017.

[46] Anton Maximilian Schäfer and Hans Georg Zimmermann. Recurrent neural networks are universal approximators. In *Proceedings of the 30th International Conference on Artificial Neural Networks*, pages 632–640, 2006.

[47] David R Seidl and Robert D Lorenz. A structure by which a recurrent neural network can approximate a nonlinear dynamic system. In *Proceedings of International Joint Conference on Neural Networks*, pages 709–714, 1991.

[48] Mahdi Soltanolkotabi. Learning ReLUs via gradient descent. In *Advances in Neural Information Processing Systems 30*, pages 2004–2014, 2017.

[49] Elias M Stein and Rami Shakarchi. *Complex Analysis*, volume 2. Princeton University Press, 2010.

[50] Matus Telgarsky. Benefits of depth in neural networks. In *Proceedings of the 29th Conference on Learning Theory*, pages 1517–1539, 2016.

[51] Chiheb Trabelsi, Olexa Bilaniuk, Ying Zhang, Dmitriy Serdyuk, Sandeep Subramanian, Joao Felipe Santos, Soroush Mehri, Negar Rostamzadeh, Yoshua Bengio, and Christopher J Pal. Deep complex networks. In *Proceedings of the 5th International Conference on Learning Representation*, 2018.

[52] Felix Voigtlaender. The universal approximation theorem for complex-valued neural networks. *arXiv:2012.03351*, 2020.

[53] X-H Yu. Can backpropagation error surface not have local minima. *IEEE Transactions on Neural Networks*, 3(6):1019–1021, 1992.

[54] Xiao-Hu Yu and Guo-An Chen. On the local minima free condition of backpropagation learning. *IEEE Transactions on Neural Networks*, 6(5):1300–1303, 1995.

[55] Z.-H. Zhou. Rehearsal: learning from prediction to decision. *Frontiers of Computer Science*, 16(4):1–3, 2022.

[56] Shao-Qun Zhang, Wei Gao, and Zhi-Hua Zhou. Towards understanding theoretical advantages of complex-reaction networks. *Neural Networks*, 151:80–93, 2022.

[57] Shao-Qun Zhang, Zhao-Yu Zhang, and Zhi-Hua Zhou. Bifurcation spiking neural network. *Journal of Machine Learning Research*, 22(253):1–21, 2021.

[58] Shao-Qun Zhang and Zhi-Hua Zhou. Flexible transmitter network. *Neural Computation*, 33(11):2951–2970, 2021.

[59] Shen-Yi Zhao, Yin-Peng Xie, and Wu-Jun Li. On the convergence and improvement of stochastic normalized gradient descent. *Science China Information Sciences*, 64(3):1–13, 2021.

[60] Ding-Xuan Zhou. Universality of deep convolutional neural networks. *Applied and Computational Harmonic Analysis*, 48(2):787–794, 2020.

[61] Zhi-Hua Zhou. Why over-parameterization of deep neural networks does not overfit? *Science China Information Sciences*, 64(1):1–3, 2021.

[62] Zhi-Hua Zhou. Open-environment machine learning. *National Science Review*, 2022.

[63] Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep ReLU networks. *Machine Learning*, 109(3):467–492, 2020.

[64] Difan Zou and Quanquan Gu. An improved analysis of training over-parameterized deep neural networks. In *Advances in Neural Information Processing Systems 32*, pages 2055–2064, 2019.