Theoretical Exploration of Flexible Transmitter Model

Jin-Hui Wu, Shao-Qun Zhang, Yuan Jiang, and Zhi-Hua Zhou, *Fellow, IEEE*

***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 (Zhang and Zhou, 2021), has been proposed. It exhibits promising behaviors, particularly on temporal–spatial signals, even when simply embedded into the common feedforward network architecture. This article attempts to understand the properties of the FT network (FTNet) theoretically. Under mild assumptions, we show that: 1) FTNet is a universal approximator; 2) 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; and 3) any local minimum of FTNet is the global minimum, implying that it is possible to identify global minima by local search algorithms.**

***Index Terms*— Approximation complexity, flexible transmitter (FT) model, local minimum, neural networks.**

I. Introduction

DEEP neural networks have become mainstream in artificial intelligence and have exhibited excellent performance in many applications, such as disease detection [2], machine translation [3], emotion recognition [4], 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 [5] and the spiking neuron model [6], [7]. Recently, a new bio-plausible neuron model, flexible transmitter (FT) model [1], 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 [1] developed the FT network (FTNet), a feed-forward neural network (FNN) composed of FT neurons, which performs competitively with the 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. First, 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. Second, 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 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 (DODS) 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 article is organized as follows. Section II introduces related work. Section III provides basic notations, definitions, and the formulation of FTNet. Section IV proves the universal approximation of FTNet. Section V investigates the approximation complexity of FTNet. Section VI studies the property of the local minima of FTNet. Section VII concludes our work with prospect.

II. Related Works

A. 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], [9], [10]. Furthermore, Leshno et al. [11] point out that a nonpolynomial activation function is the necessary and sufficient

Manuscript received 14 December 2021; revised 4 May 2022; accepted 23 July 2022. Date of publication 26 July 2023; date of current version 1 March 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2020AAA0109401, in part by the NSFC under Grant 61921006 and Grant 62176117, and in part by the Collaborative Innovation Center of Novel Software Technology and Industrialization. *(Corresponding author: Zhi-Hua Zhou.)*

The authors are with the National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China (e-mail: wujh@lamda.nju.edu.cn; zhangsq@lamda.nju.edu.cn; jiangy@lamda.nju.edu.cn; zhouzh@lamda.nju.edu.cn).

Digital Object Identifier 10.1109/TNNLS.2022.3195909

condition for FNNs to achieve universal approximation. Later, some researchers extend the universal approximation theorems to other real-weighted neural networks with different architectures, such as RNN [12], [13], [14], [15], [16] and convolutional neural networks [17]. 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 [18], and that nonholomorphic and nonantiholomorphic activation functions are the necessary and sufficient condition of universal approximation [19]. 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.

B. 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 [20], [21], [22], [23]. Recent works prove separation results, i.e., one model cannot be expressed by another model with the same order of parameters [24], [25], [26], [27]. A notable work proves that one-hidden-layer FNN needs at least exponential parameters to express a given complex-reaction network (CRNet) [28]. 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.

C. 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 [29], [30], [31], [32]. These results are extended to multilayer FNN [33], [34] and other types of neural networks, such as deep ResNet [35], deep convolutional neural networks [36], deep linear networks [37], [38], and overparameterized deep neural networks [39]. 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 overparameterized neural networks [40], [41], [42], [43], [44], [45], [46], [47], [48], [49], [50]. 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.

III. Preliminary

We denote by $i = \sqrt{-1}$ 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-D outputs throughout this article. We focus on one-hidden-layer FTNet throughout this article. 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 [51]. Let $\boldsymbol{x}_t \in \mathbb{R}^I$ be the input vector at time t, and $\boldsymbol{x}_{1:T} = (\boldsymbol{x}_1; \boldsymbol{x}_2; \ldots; \boldsymbol{x}_T) \in \mathbb{R}^{IT}$ 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:

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

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, $\mathbf{W}_\times$, $\mathbf{V}_\times$, and $\boldsymbol{\alpha}_\times$ 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

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

and $\sigma_\times$ is an activation function applied componentwise. Notice that (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:

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

The zReLU activation function [52] is a promising choice of the activation function in FTNet, which extends the widely used real-valued activation function ReLU [53] to the complex-valued domain, and is defined as

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

Dynamical systems are of great interest when considering the universal approximation of neuron models in the recurrent architecture. We focus on the DODS 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 DODS is a mapping f_D defined by

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

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.

IV. Universal Approximation

We show the universal approximation of F-FTNet and R-FTNet in Sections IV-A and IV-B, respectively.

A. 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 $\sigma(x) = \mathrm{Re}[\sigma_\times(x + ci)]$ 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

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

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 [54], zReLU [52], and $\mathbb{C}$ReLU [55]. 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 [11], 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 [19]. Thus, FTNet successfully benefits from expressing real-valued functions instead of complex-valued ones. We begin our proof with the following lemma.

Lemma 1 [11, Th. 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

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

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

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

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

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

Thus, one has

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

where the first equality holds according to (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 (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. □

B. 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 (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

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

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 [16]. 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

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

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

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

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

Based on (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.

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

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

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

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

where the first equality holds from (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 (8). Thus, the conclusion holds for $t = \tau + 1$.

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

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

$$= \left[\mathbf{0}; \sigma_2(c\mathbf{1}, \mathbf{A}\boldsymbol{x}_t + \mathbf{B}\boldsymbol{q}_{t-1} - \boldsymbol{\zeta}); \mathbf{0}\right]$$
$$= \left[\mathbf{0}; \boldsymbol{p}_t; \mathbf{0}\right]$$

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 (10) has the same output as the additive FTNet defined by (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 (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. □

V. Approximation Complexity

We show the approximation advantage of FTNet over FNN and RNN in Section V-A and provide worst case guarantees in Section V-B. 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.

A. 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 $f_1 : \mathbb{R}^I \to \mathbb{R}$, such that

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 $\max\{3\ c_1^2 I^{15/2}/\varepsilon^2, 27I^2\}$ 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 $f_{\mathrm{D}} : \mathbb{R}^{IT} \to \mathbb{R}^T$, such that

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 [24], [25], [26]. A recent study shows exponential separation between one-hidden-layer CRNet and one-hidden-layer FNN [28]. 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 CRNet and an important lemma.

The CRNet is a recently proposed neural network with complex-valued operations [28]. 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 [28], 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:

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

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 [24, Th. 1] and [28, Th. 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.

First, we prove that F-FTNet can approximate the target function f_1 using polynomial parameters. Recalling the definition of CRNet in (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:

$$\mathbf{W}_\times = \begin{bmatrix}\mathbf{W}_{\mathrm{C},R} & -\mathbf{W}_{\mathrm{C},I} & \mathbf{0} & \boldsymbol{b}_{\mathrm{C},R}\\ \mathbf{W}_{\mathrm{C},I} & \mathbf{W}_{\mathrm{C},R} & \mathbf{0} & \boldsymbol{b}_{\mathrm{C},I}\\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0}\end{bmatrix}$$
$$\mathbf{V}_\times = \begin{bmatrix}\mathbf{W}_{\mathrm{C},I} & \mathbf{W}_{\mathrm{C},R} & \mathbf{0} & \boldsymbol{b}_{\mathrm{C},I}\\ \mathbf{W}_{\mathrm{C},R} & -\mathbf{W}_{\mathrm{C},I} & \mathbf{0} & \boldsymbol{b}_{\mathrm{C},R}\\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0}\end{bmatrix}$$
$$\boldsymbol{r}_0 = \mathbf{0}, \quad \boldsymbol{\alpha}_\times = \left[\boldsymbol{\alpha}_{\mathrm{C},R}; -\boldsymbol{\alpha}_{\mathrm{C},I}; \mathbf{0}\right]$$

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

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

It is observed that

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

Thus, one has

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

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

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

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 (16), indicates that

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

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

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

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

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

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

Combining the conclusions above completes the proof. □

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 (5).

First, 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

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

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 (13). Define the R-FTNet $f_{\times,\mathrm{R}}$ with $H_\times = 2H_\mathrm{C} + I + 1$ hidden neurons as follows:

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

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

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

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 (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

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

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

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

where the second equality holds from (17), (19), and the first inequality holds based on (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

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

Second, 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:

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

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:

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

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 (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

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

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 (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. □

B. 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 [24], [25], [26], [27], [28], 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 (22), such that

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

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

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

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

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

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

where the first equality holds based on (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 (4), one has

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

where the third equality holds because σ_{F} is the ReLU activation, and the fourth equality holds from (22). Finally, we prove that the constructed F-FTNet$_\times$ can $(\varepsilon, \mathcal{D})$-approximate the function f. According to (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 (20) satisfying the following inequality:

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

First, 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:

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

Second, we prove that the output of the constructed R-FTNet$_\times$ in (26) is the same as that of the RNN in (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 (26), the claim holds for $t = 0$.

TABLE I
APPROXIMATION COMPLEXITY OF FTNET AND FNN

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

TABLE II
APPROXIMATION COMPLEXITY OF FTNET AND RNN

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

2) *Induction:* Suppose that the claim holds for $t = \tau$ where $\tau \in \{0, 1, \ldots, T-1\}$. From (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

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

where the first equality holds from the construction in (26) and the hypothesis induction, the second equality holds according to (4), the third equality holds because σ_R is the ReLU activation function, and the fourth equality holds based on (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. From (25) and (27), one has

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{x}_{1:T}\sim\mathcal{D}}\big[(f_\mathrm{D}(\boldsymbol{x}_{1:T}) - f_{\times,\mathrm{R}}(\boldsymbol{x}_{1:T}))^2\big] \\ &\qquad = \mathbb{E}_{\boldsymbol{x}_{1:T}\sim\mathcal{D}}\big[(f_\mathrm{D}(\boldsymbol{x}_{1:T}) - f_\mathrm{R}(\boldsymbol{x}_{1:T}))^2\big] \leqslant \varepsilon\end{aligned}$$

which completes the proof. □

Tables I and II 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 [56] and open-environment machine learning [57], though the analysis is beyond the scope of this article.

VI. 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 ith sample, and $y_i \in \mathbb{R}$ represents the label of the ith sample. Consider the empirical loss of F-FTNet with the following form:

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

where $f_{\times,\mathrm{F}}$ is the mapping of F-FTNet defined in (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 [58, p. 2]: A function $g : \mathbb{C}^m \to \mathbb{C}$ is called holomorphic if for each point $\boldsymbol{w} = (w_1, w_2, \ldots, w_m) \in \mathbb{C}^m$, there exists an open set U, such that $\boldsymbol{w} \in U$, and the function g has a power series expansion

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

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

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: 1) l is analytic on $\mathbb{R}$; 2) $l(0) = 0$; and 3) 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 $l(x) = c^{-1}[\ln(e^{ax} + e^{-bx}) - \ln 2]$ 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 $\theta \in (0, 1)$ [59] 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:

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

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 [60], particular activation functions [37], [38], [61], [62], and over-parameterization together with special initialization [42], [45], [46], [47], [48]. 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 [63]. We begin our proof with two lemmas.

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

$$\big\|\Delta z^{(1)}\big\|_2^2 \leqslant \delta \quad \text{and} \quad \mathrm{Re}\big[g(z^{(0)} + \Delta z^{(1)})\big] > \mathrm{Re}\big[g(z^{(0)})\big]$$

$$\big\|\Delta z^{(2)}\big\|_2^2 \leqslant \delta \quad \text{and} \quad \mathrm{Re}\big[g(z^{(0)} + \Delta z^{(2)})\big] < \mathrm{Re}\big[g(z^{(0)})\big].$$

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 $\boldsymbol{z}^{(0)} = (z_1^{(0)}, z_2^{(0)}, \ldots, z_m^{(0)})$. Since the function g is holomorphic, by rearranging (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 $k \in [m]$, such that $\boldsymbol{z}^{(0)} \in U$, and for any $\boldsymbol{z} = (z_1, z_2, \ldots, z_m) \in U$, the following holds:

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

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:

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

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

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

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 $f_j^{(k_0)}(z_{k_0+1}, z_{k_0+2}, \ldots, z_m) \not\equiv 0$. Then the following j_0 is well-defined:

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

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 (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 (30), one has

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

meanwhile

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

where the first equality holds from (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, $\{\boldsymbol{x}^{(j)}\}_{j=1}^n \subset \mathbb{R}^m$ 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 [64, Th. 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

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

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 (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

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

where the equality holds from (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

$$\begin{aligned} M_l &= \max_{j \in [n]} \left|x_l^{(j)}\right| \\ m_{l,2} &= \begin{cases} \max_{j \in J_{l-1} \setminus J_l} \left|x_l^{(j)}\right|, & \text{if } J_{l-1} \neq J_l \\ 0, & \text{if } J_{l-1} = J_l \end{cases} \\ \mathcal{A}_l &= \left(\prod_{s=1}^{l-1} m_s^{q_s + r_s} \right) \left(\prod_{t=l+1}^{m} M_t^{q_t + r_t} \right). \end{aligned}$$

Thus, one has

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

where the second inequality holds because of $J_{l-1} \setminus J_l \subset J_{l-1} \subset [n]$. 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 $l \in [n] \setminus L$, 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

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

Based on the choice of r_l, one has

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

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 (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

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

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

Step 3: We prove that the necessary condition of constant g in (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 $j \in S_m \setminus \{j_0\}$, 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

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

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

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

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 $j \in J_m \setminus \{j_0\}$, there exists l_j such that $l_j \notin T_{j_0}$ and $l_j \in T_j$.

Thus, for the second term of (35), one has

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

where the second equality holds since $x_l^{(j)}<0$ and $\Pr[\sigma_{l_j}=0]=\Pr[\sigma_{l_j}=1]=1/2$. Substituting (36) and (37) into (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 (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 (3) and (28), one has

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

where $\boldsymbol{\kappa}_j$ and σ denote $\kappa(\boldsymbol{x}_j,H_\times)$ and $\sigma_\times$, respectively. Let $\mathbf{Z}=[z_1^\top;z_2^\top;\ldots;z_{H_\times}^\top]$ and $\boldsymbol{\alpha}=(\alpha_1;\alpha_2;\ldots;\alpha_{H_\times})$. 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 $\boldsymbol{\alpha}^\top\operatorname{Re}[\sigma(\mathbf{Z}\boldsymbol{\kappa}_{j_0})]\neq y_{j_0}$. Without loss of generality, we only consider the case of $\boldsymbol{\alpha}^\top\operatorname{Re}[\sigma(\mathbf{Z}\boldsymbol{\kappa}_{j_0})]>y_{j_0}$ 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 (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 $\Delta\boldsymbol{\alpha}=\mathbf{0}$, $z_k=\mathbf{0}$ for any $k\in[H_\times]\backslash\{k_0\}$, and $z_{k_0}=c\boldsymbol{v}$ where $c\in\mathbb{C}$ is a complex-valued variable. Then the change in loss becomes a function of c as follows:

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

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 $\alpha_{k_0}\sigma((z_{k_0}+c\boldsymbol{v})^\top\boldsymbol{\kappa}_{j_0})$ is holomorphic and not constant w.r.t. c. Then Lemma 3 implies that there exists $c\leqslant\delta/\|\boldsymbol{v}\|_2$, such that

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

and

$$\begin{aligned}&\operatorname{Re}\left[\alpha_{k_0}\sigma\left(\left(z_{k_0}+c\boldsymbol{v}\right)^\top\boldsymbol{\kappa}_{j_0}\right)\right]\\&\qquad\geqslant\operatorname{Re}\left[\alpha_{k_0}\sigma\left(z_{k_0}^\top\boldsymbol{\kappa}_j\right)\right]-\boldsymbol{\alpha}^\top\operatorname{Re}\left[\sigma\left(\mathbf{Z}\boldsymbol{\kappa}_{j_0}\right)\right]+y_{j_0}\end{aligned}\tag{38}$$

where (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 $\Delta z_k=\mathbf{0}$ and $\Delta\alpha_k=0$ for any $k\in[H_\times]\backslash\{1\}$. Then the change in loss becomes a function of Δz_1 and $\Delta\alpha_1$ as follows:

$$\Delta\hat{L}=\sum_{j=1}^{n}l\left(\Delta\alpha_1\operatorname{Re}\left[\sigma\left((z_1+\Delta z_1)^\top\boldsymbol{\kappa}_j\right)\right]-y_j\right)-l\left(-y_j\right).$$

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

$$\begin{aligned}\Delta\hat{L}&=\sum_{j=1}^{n}\sum_{p=1}^{\infty}\sum_{q=1}^{p}c_p\binom{p}{q}\left(-y_j\right)^{p-q}\\&\quad\times\left(\Delta\alpha_1\operatorname{Re}\left[\sigma\left((z_1+\Delta z_1)^\top\boldsymbol{\kappa}_j\right)\right]\right)^q\\&=\sum_{q=1}^{\infty}\sum_{j=1}^{n}\sum_{p=q}^{\infty}c_p\binom{p}{q}\left(-y_j\right)^{p-q}(\Delta\alpha_1)^q\\&\quad\times\left(\operatorname{Re}\left[\sigma\left((z_1+\Delta z_1)^\top\boldsymbol{\kappa}_j\right)\right]\right)^q\\&=\sum_{q=1}^{\infty}C_q(\Delta\alpha_1)^q\end{aligned}\tag{39}$$

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 Δz_1 defined by

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

with $R=\operatorname{Re}[\sigma((z_1+\Delta z_1)^\top\boldsymbol{\kappa}_j)]$.

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

$$\begin{aligned}C_1&=\sum_{j=1}^{n}\sum_{p=1}^{\infty}c_p p\operatorname{Re}\left[\sigma\left((z_1+\Delta z_1)^\top\boldsymbol{\kappa}_j\right)\right]\left(-y_j\right)^{p-1}\\&=\operatorname{Re}\left[\sum_{j=1}^{n}l'(-y_j)\sigma\left((z_1+\Delta z_1)^\top\boldsymbol{\kappa}_j\right)\right].\end{aligned}$$

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 $\sum_{j=1}^n l'(-y_j)\sigma((z_1+\Delta z_1)^\top\boldsymbol{\kappa}_j)$ is not a constant. Thus, there exists Δz_1, such that $\|\Delta z_1\|_2\leqslant\delta/2$ and $C_1\neq 0$.

Step 2.3: We give upper bounds for $\{C_q\}_{q=2}^\infty$. Provided Δz_1 in Step 2.2., we define $a=\max_{j\in[n]}|\operatorname{Re}[\sigma((z_1+\Delta z_1)^\top\boldsymbol{\kappa}_j)]|$. 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 $d>0$, such that $|c_p|\leqslant d/(4b)^p$ holds for any $p\geqslant 2$. Using these notations, coefficients $\{C_q\}_{q=2}^\infty$ can be bounded by

$$\begin{aligned}|C_q|&\leqslant\sum_{j=1}^{n}\sum_{p=q}^{\infty}|c_p|\binom{p}{q}\left|\operatorname{Re}\left[\sigma\left((z_1+\Delta z_1)^\top\boldsymbol{\kappa}_j\right)\right]\right|^q|y_j|^{p-q}\\&\leqslant\sum_{p=q}^{\infty}\frac{nd}{4^p}\binom{p}{q}\left(\frac{a}{b}\right)^q\end{aligned}\tag{41}$$

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 (39) can be rewritten as

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

where the first inequality holds according to the triangle inequality, the second inequality holds based on (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

$$\|\Delta\mathbf{Z}\|_F + \|\Delta\boldsymbol{\alpha}\|_2 = \|\Delta z_1\|_2 + |\Delta\alpha_1| \leqslant \delta/2 + \delta/2 = \delta.$$

Combining the results in all cases completes the proof. □

VII. Conclusion and Prospect

This work investigates the theoretical properties of FTNet via approximation and local minima. The main conclusions are three folds. First, we prove the universal approximation of F-FTNet and R-FTNet, which guarantees the possibility of expressing any continuous function and any DODS on any compact set arbitrarily well, respectively. Second, 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 toward 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.

Appendix A
Complete Proof of (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 (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 $K_1 \in \mathbb{R}^{H_D}$, such that $\boldsymbol{h}_t \in K_1$ holds for any $t \in \{0, 1, \ldots, T\}$. 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 $\boldsymbol{h} \in \mathbb{R}^{H_D}$. 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 $H_1, H_2 \in \mathbb{N}^+$, $\mathbf{A}_1 \in \mathbb{R}^{H_1\times I}$, $\mathbf{B}_1 \in \mathbb{R}^{H_1\times H_D}$, $\mathbf{C}_1 \in \mathbb{R}^{H_D\times H_1}$, $\boldsymbol{\theta}_1 \in \mathbb{R}^{H_1}$, $\mathbf{A}_2 \in \mathbb{R}^{H_2\times I}$, $\mathbf{B}_2 \in \mathbb{R}^{H_2\times H_D}$, $\mathbf{C}_2 \in \mathbb{R}^{H_D\times H_2}$, and $\boldsymbol{\theta}_2 \in \mathbb{R}^{H_2}$, where $\mathbf{C}_1$ and $\mathbf{C}_2$ are row independent, such that

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

Step 2: We prove that RNN using the same weight matrices as FNN in (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

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

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

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

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 (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

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

Step 3: Transformation is used to eliminate the matrices $\mathbf{C}_1$ and $\mathbf{C}_2$ in (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

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

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

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

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 $k \in \{0,1,\ldots,T-1\}$. 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 (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 (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

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

For $\boldsymbol{q}_t^{(3)}$, we claim that $\boldsymbol{q}_t^{(3)}=[\mathbf{x}_t;\boldsymbol{q}_t^{(2)}]$ holds for any $t\in[T]$, where $\mathbf{x}_t\in\mathbb{R}^{H_1\times 1}$ 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 (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 (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

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

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

$$\mathbf{C}_3=\begin{bmatrix}\mathbf{C}_1 & \mathbf{0}^{H_{\mathrm{D}}\times H_2}\end{bmatrix}.$$

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

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

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 (46), and the second inequality holds based on the definition of $u(a)$, $(\boldsymbol{x}_t,\boldsymbol{q}_{t-1}^{(1)})\in K_2$, (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

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

Step 5: The output function ψ in (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 $H_4\in\mathbb{N}^+$, $\mathbf{A}_4\in\mathbb{R}^{H_4\times H_{\mathrm{D}}}$, $\mathbf{B}_4\in\mathbb{R}^{O\times H_4}$, $\boldsymbol{\theta}_4\in\mathbb{R}^{H_4\times 1}$, such that

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

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 (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, such that $\boldsymbol{q}_t^{(2)}\in K_3$ 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 $\boldsymbol{\theta}_5\in\mathbb{R}^{H_5}$, such that

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

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 (50) and (51) imply that

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

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:

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

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

$$\boldsymbol{p}_1=\sigma_1(\mathbf{A}\boldsymbol{x}_1+\mathbf{B}\boldsymbol{q}_0-\boldsymbol{\zeta})$$

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

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

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

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

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

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

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 $u(a)$, (49) with $\boldsymbol{h} = \mathbf{C}_3\boldsymbol{p}_t^{(3)}$, and (53), and the third inequality holds in view of (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. □

Appendix 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 $\mathbf{U}_1 \in \mathbb{R}^{O\times H_1}$, such that

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

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

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

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

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

Thus, one has

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

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. □

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
$$\lim_{\varepsilon_1 \to 0_+} a_1 = \lim_{\varepsilon_1 \to 0_+} u(a_0) + \varepsilon_1 = 0 + 0 = 0.$$
Thus, the conclusion holds for $t = 1$.
2) If the conclusion holds for $t = k$ where $k \in [T-1]$, then
$$\begin{aligned}\lim_{\varepsilon_1 \to 0_+} a_{k+1} &= \lim_{\varepsilon_1 \to 0_+} u(a_k) + \varepsilon_1 \\ &= u\left(\lim_{\varepsilon_1 \to 0_+} a_k\right) + 0 \\ &= u(0) + 0 = 0.\end{aligned}$$
Thus, the conclusion holds for $t = k + 1$.

Then mathematical induction completes the proof. □

ACKNOWLEDGMENT

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 article.

REFERENCES

[1] S.-Q. Zhang and Z.-H. Zhou, "Flexible transmitter network," *Neural Comput.*, vol. 33, no. 11, pp. 2951–2970, 2021.
[2] J. Laguarta and B. Subirana, "Longitudinal speech biomarkers for automated Alzheimer's detection," *Frontiers Comput. Sci.*, vol. 3, Apr. 2021, Art. no. 624694.
[3] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Represent.*, 2015.
[4] W. Guo, G. Li, J. Lu, and J. Yang, "Singular learning of deep multilayer perceptrons for EEG-based emotion recognition," *Frontiers Comput. Sci.*, vol. 3, Dec. 2021, Art. no. 786964.
[5] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bull. Math. Biophys.*, vol. 5, no. 4, pp. 115–133, Dec. 1943.
[6] A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," *J. Physiol.*, vol. 117, no. 4, pp. 500–544, Aug. 1952.
[7] S.-Q. Zhang, Z.-Y. Zhang, and Z.-H. Zhou, "Bifurcation spiking neural network," *J. Mach. Learn. Res.*, vol. 22, no. 253, pp. 1–21, 2021.
[8] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Math. Control, Signals, Syst.*, vol. 2, no. 4, pp. 303–314, Dec. 1989.
[9] K.-I. Funahashi, "On the approximate realization of continuous mappings by neural networks," *Neural Netw.*, vol. 2, no. 3, pp. 183–192, Jan. 1989.
[10] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Netw.*, vol. 2, no. 5, pp. 359–366, Jan. 1989.
[11] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken, "Multilayer feedforward networks with a nonpolynomial activation function can approximate any function," *Neural Netw.*, vol. 6, no. 6, pp. 861–867, Jan. 1993.
[12] D. R. Seidl and R. D. Lorenz, "A structure by which a recurrent neural network can approximate a nonlinear dynamic system," in *Proc. IJCNN Seattle Int. Joint Conf. Neural Netw.*, 1991, pp. 709–714.
[13] K.-I. Funahashi and Y. Nakamura, "Approximation of dynamical systems by continuous time recurrent neural networks," *Neural Netw.*, vol. 6, no. 6, pp. 801–806, Jan. 1993.
[14] T. W. S. Chow and X.-D. Li, "Modeling of continuous time dynamical systems with input by recurrent neural networks," *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. 47, no. 4, pp. 575–578, Apr. 2000.
[15] X.-D. Li, J. K. L. Ho, and T. W. S. Chow, "Approximation of dynamical time-variant systems by continuous-time recurrent neural networks," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 52, no. 10, pp. 656–660, Oct. 2005.
[16] A. M. Schäfer and H. G. Zimmermann, "Recurrent neural networks are universal approximators," in *Proc. 30th Int. Conf. Artif. Neural Netw.*, 2006, pp. 632–640.
[17] D.-X. Zhou, "Universality of deep convolutional neural networks," *Appl. Comput. Harmon. Anal.*, vol. 48, no. 2, pp. 787–794, Mar. 2020.
[18] P. Arena, L. Fortuna, R. Re, and M. G. Xibilia, "On the capability of neural networks with complex neurons in complex valued functions approximation," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 1993, pp. 2168–2171.
[19] F. Voigtlaender, "The universal approximation theorem for complex-valued neural networks," 2020, *arXiv:2012.03351*.
[20] A. R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 930–945, May 1993.
[21] C. Debao, "Degree of approximation by superpositions of a sigmoidal function," *Approximation Theory Appl.*, vol. 9, no. 3, pp. 17–28, 1993.
[22] K. Hornik, M. Stinchcombe, H. White, and P. Auer, "Degree of approximation results for feedforward networks approximating unknown mappings and their derivatives," *Neural Comput.*, vol. 6, no. 6, pp. 1262–1275, Nov. 1994.
[23] H. N. Mhaskar and C. A. Micchelli, "Dimension-independent bounds on the degree of approximation by neural networks," *IBM J. Res. Develop.*, vol. 38, no. 3, pp. 277–284, May 1994.
[24] R. Eldan and O. Shamir, "The power of depth for feedforward neural networks," in *Proc. 29th Conf. Learn. Theory*, 2016, pp. 907–940.
[25] I. Safran and O. Shamir, "Depth-width tradeoffs in approximating natural functions with neural networks," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 2979–2987.
[26] M. Telgarsky, "Benefits of depth in neural networks," in *Proc. 29th Conf. Learn. Theory*, 2016, pp. 1517–1539.
[27] Z. Lu, H. Pu, F. Wang, Z. Hu, and L. Wang, "The expressive power of neural networks: A view from the width," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6232–6240.
[28] S.-Q. Zhang, W. Gao, and Z.-H. Zhou, "Towards understanding theoretical advantages of complex-reaction networks," *Neural Netw.*, vol. 151, pp. 80–93, Jul. 2022.
[29] T. Poston, C.-N. Lee, Y. Choie, and Y. Kwon, "Local minima and back propagation," in *Proc. IJCNN Seattle Int. Joint Conf. Neural Netw.*, 1991, pp. 173–176.
[30] X.-H. Yu, "Can backpropagation error surface not have local minima," *IEEE Trans. Neural Netw.*, vol. 3, no. 6, pp. 1019–1021, Jan. 1992.
[31] X.-H. Yu and G.-A. Chen, "On the local minima free condition of backpropagation learning," *IEEE Trans. Neural Netw.*, vol. 6, no. 5, pp. 1300–1303, Sep. 1995.
[32] D.-S. Huang, "The local minima-free condition of feedforward neural networks for outer-supervised learning," *IEEE Trans. Syst., Man Cybern. B, Cybern.*, vol. 28, no. 3, pp. 477–480, Jun. 1998.
[33] Q. Nguyen and M. Hein, "The loss surface of deep and wide neural networks," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 2603–2612.
[34] K. Kawaguchi and L. Kaelbling, "Elimination of all bad local minima in deep learning," in *Proc. 23rd Int. Conf. Artif. Intell. Statist.*, 2020, pp. 853–863.
[35] K. Kawaguchi and Y. Bengio, "Depth with nonlinearity creates no bad local minima in ResNets," *Neural Netw.*, vol. 118, pp. 167–174, 2019.
[36] Q. Nguyen and M. Hein, "Optimization landscape and expressivity of deep CNNs," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 3730–3739.
[37] K. Kawaguchi, "Deep learning without poor local minima," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 586–594.
[38] T. Laurent and J. Brecht, "Deep linear networks with arbitrary loss: All local minima are global," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 2902–2907.
[39] Q. Nguyen, M. C. Mukkamala, and M. Hein, "On the loss landscape of a class of deep neural networks with no bad local valleys," in *Proc. 6th Int. Conf. Learn. Represent.*, 2018.
[40] M. Gori and A. Tesi, "On the problem of local minima in backpropagation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 1, pp. 76–86, Jan. 1992.
[41] S. S. Du, X. Zhai, B. Poczos, and A. Singh, "Gradient descent provably optimizes over-parameterized neural networks," in *Proc. 6th Int. Conf. Learn. Represent.*, 2018.
[42] A. Jacot, F. Gabriel, and C. Hongler, "Neural tangent kernel: Convergence and generalization in neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 8580–8589.

[43] Y. Li and Y. Liang, "Learning overparameterized neural networks via stochastic gradient descent on structured data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 8168–8177.
[44] Z. Allen-Zhu, Y. Li, and Z. Song, "On the convergence rate of training recurrent neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 6676–6688.
[45] S. Du, J. Lee, H. Li, L. Wang, and X. Zhai, "Gradient descent finds global minima of deep neural networks," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 1675–1685.
[46] Z. Allen-Zhu, Y. Li, and Z. Song, "A convergence theory for deep learning via over-parameterization," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 242–252.
[47] D. Zou and Q. Gu, "An improved analysis of training over-parameterized deep neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 2055–2064.
[48] D. Zou, Y. Cao, D. Zhou, and Q. Gu, "Gradient descent optimizes over-parameterized deep ReLU networks," *Mach. Learn.*, vol. 109, no. 3, pp. 467–492, Mar. 2020.
[49] C. Liu, L. Zhu, and M. Belkin, "Loss landscapes and optimization in over-parameterized non-linear systems and neural networks," *Appl. Comput. Harmon. Anal.*, vol. 59, pp. 85–116, Jul. 2022.
[50] S.-Y. Zhao, Y.-P. Xie, and W.-J. Li, "On the convergence and improvement of stochastic normalized gradient descent," *Sci. China Inf. Sci.*, vol. 64, no. 3, pp. 1–13, Mar. 2021.
[51] Z.-H. Zhou, "Why over-parameterization of deep neural networks does not overfit?" *Sci. China Inf. Sci.*, vol. 64, no. 1, pp. 1–3, Jan. 2021.
[52] N. Guberman, "On complex valued convolutional neural networks," 2016, *arXiv:1602.09046*.
[53] K. Fukushima, "Visual feature extraction by a multilayered network of analog threshold elements," *IEEE Trans. Syst. Sci. Cybern.*, vol. SSC-5, no. 4, pp. 322–333, Jun. 1969.
[54] M. Arjovsky, A. Shah, and Y. Bengio, "Unitary evolution recurrent neural networks," in *Proc. 33rd Int. Conf. Mach. Learn.*, 2016, pp. 1120–1128.
[55] C. Trabelsi et al., "Deep complex networks," in *Proc. 5th Int. Conf. Learn. Represent.*, 2018.
[56] Z.-H. Zhou, "Rehearsal: Learning from prediction to decision," *Frontiers Comput. Sci.*, vol. 16, no. 4, pp. 1–3, Aug. 2022.
[57] Z.-H. Zhou, "Open-environment machine learning," *Nat. Sci. Rev.*, vol. 9, no. 8, Aug. 2022, Art. no. nwac123.
[58] R. C. Gunning and H. Rossi, *Analytic Functions Several Complex Variables*, vol. 368. Providence, RI, USA: American Mathematical Society, 2009.
[59] R. Koenker and G. Bassett Jr., "Regression quantiles," *Econometrica, J. Econ. Soc.*, vol. 46, no. 1, pp. 33–50, 1978.
[60] S. Liang, R. Sun, Y. Li, and R. Srikant, "Understanding the loss surface of neural networks for binary classification," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 2835–2843.
[61] M. Soltanolkotabi, "Learning ReLUs via gradient descent," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 2004–2014.
[62] S. Liang, R. Sun, J. D. Lee, and R. Srikant, "Adding one neuron can eliminate all bad local minima," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 4355–4365.
[63] T. Ding, D. Li, and R. Sun, "Suboptimal local minima exist for wide neural networks with smooth activations," *Math. Oper. Res.*, vol. 47, no. 4, pp. 2784–2814, Nov. 2022.
[64] E. M. Stein and R. Shakarchi, *Complex Analysis*, vol. 2. Princeton, NJ, USA: Princeton Univ. Press, 2010.

Jin-Hui Wu received the B.Sc. degree in information and computer science from Nanjing University, Nanjing, China, in 2019, where he is currently pursuing the Ph.D. degree with the School of Artificial Intelligence.

His research interests include machine learning and data mining.

Shao-Qun Zhang received the Ph.D. degree from the Department of Computer Science and Technology, Nanjing University, Nanjing, China.

He is currently an Assistant Professor with the School of Intelligence Science and Technology, Nanjing University. His research interests include neural computation, deep learning theory, and time series analysis.

Yuan Jiang received the Ph.D. degree in computer science from Nanjing University, Nanjing, China, in 2004.

In 2004, she has worked as a Faculty Member with the Department of Computer Science and Technology, Nanjing University, where she is currently a Professor. She was selected in the Program for New Century Excellent Talents in University, Ministry of Education, in 2009. She has published more than 50 papers in leading international/national journals and conferences. Her main research interests include the design of learning algorithms of machine learning and artificial intelligence.

Dr. Jiang has served as a program committee (PC) member for several international conferences.

Zhi-Hua Zhou (Fellow, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees (Hons.) in computer science from Nanjing University, Nanjing, China, in 1996, 1998, and 2000, respectively.

He joined the Department of Computer Science and Technology, Nanjing University, as an Assistant Professor in 2001, where he is currently a Professor, the Head of the Department of Computer Science and Technology, and the Dean of the School of Artificial Intelligence. He is also the Founding Director of the LAMDA Group, Nanjing University. He has authored the books *Ensemble Methods: Foundations and Algorithms*, *Evolutionary Learning: Advances in Theories and Algorithms*, and *Machine Learning*. He has published more than 200 papers in top-tier international journals or conference proceedings. He holds more than 30 patents. His main research interests include artificial intelligence, machine learning, and data mining.

Dr. Zhou is a Foreign Member of the Academia Europaea, and a fellow of ACM, AAAI, AAAS, IAPR, IET/IEE, CCF, and CAAI. He has received various awards/honors, including the National Natural Science Award of China, the IEEE Computer Society Edward J. McCluskey Technical Achievement Award, the Asian Conference on Machine Learning (ACML) Distinguished Contribution Award, the PAKDD Distinguished Contribution Award, the IEEE ICDM Outstanding Service Award, and the Microsoft Professorship Award. He founded ACML. He has served as an Advisory Committee Member for IJCAI from 2015 to 2016; a Steering Committee Member for ICDM, ACML, PAKDD, and PRICAI; and the Chair for various conferences, such as the Program Chair for AAAI 2019 and IJCAI 2021, the General Chair for ICDM 2016 and SDM 2022, and the Senior Area Chair for NeurIPS and ICML. He was the Chair of the CAAI Machine Learning Technical Committee from 2006 to 2015, the IEEE CIS Data Mining Technical Committee from 2015 to 2016, and CCF-AI from 2012 to 2019. He is a Series Editor of *Lecture Notes in Artificial Intelligence* (Springer), on the Advisory Board of *AI Magazine*. He serves as the Editor-in-Chief for *Frontiers of Computer Science*; the Associate Editor-in-Chief for *Science China Information Sciences*; and an Associate Editor for *Artificial Intelligence*, *Machine Learning*, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, and *ACM Transactions on Knowledge Discovery From Data*. He has served as the Associate Editor-in-Chief for *Chinese Science Bulletin* from 2008 to 2014, and an Associate Editor for IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING from 2008 to 2012, *ACM Transactions on Intelligent Systems and Technology* from 2009 to 2017, *Neural Networks* from 2014 to 2016, and IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS from 2014 to 2017.