# THE EFFECTS OF NONLINEARITY ON APPROXIMATION CAPACITY OF RECURRENT NEURAL NETWORKS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

We study the effects of nonlinear recurrent activations on the approximation properties of recurrent neural networks (RNNs). Previous works indicate that in the linear setting, RNNs show good approximation performance when the target sequential relationship is smooth and has fast decaying memory. Otherwise, RNNs may suffer from the so-called "curse of memory", meaning that an exponentially large number of neurons is required for accurate approximation. A natural question is whether the recurrent nonlinearity has a substantial effect on RNNs' approximation capacity and approximation speed. In this paper, we present some negative results in this direction. We discover that, while the addition of nonlinearity does not shrink the hypothesis space, in the sense that nonlinear RNNs can still approximate linear functionals with the same approximation rates established for linear RNNs, it does not essentially alleviate the limitations of RNNs either. In particular, we prove that nonlinear RNNs fail to be universal approximators of arbitrary nonlinear functionals, and any linear functional that can be efficiently approximated must also possess an exponentially decaying memory.

## 1 INTRODUCTION

Recurrent neural networks (RNNs) (Rumelhart et al., 1986) are one of the most popular machine learning models to learn the relationship between sequential or temporal data. They have wide applications from time series prediction (Connor et al., 1994), text generation (Sutskever et al., 2011), speech recognition (Graves & Jaitly, 2014) to sentiment classification (Tang et al., 2015). However, when there are long-term dependencies in the data (Bengio et al., 1994; Hochreiter et al., 2001), empirical results show that RNN may encounter difficulties in learning.

In the previous works (Li et al., 2021; 2022a), a theoretical framework of functional approximation has been proposed to study the long-term dependencies, since the input-output relationship can be mathematically understood as a sequence of functionals. It is found that in the *linear* setting, the decaying memory, which is defined by the decay rates of linear target functionals, is a crucial property that enables efficient approximation. The key limitation of this analysis is that it only applies to RNNs with linear recurrent activations. We will hereafter call these models "linear RNNs".

In this paper, we study the effects of *nonlinear* recurrent activations on the approximation properties of recurrent neural networks. A natural question is whether nonlinear recurrent activations will improve RNNs' expressive power since the addition of nonlinearities to other neural network models can significantly improve their approximation capacity (Cybenko, 1989; Hornik et al., 1989; Hornik, 1991; 1993; Stinchcombe, 1999; Shen et al., 2019; 2022). However, it is not even an obvious result that RNNs with nonlinear recurrent activations ("nonlinear RNNs") can approximate arbitrary linear functionals. This is because the hypothesis space of nonlinear RNNs does not contain that of the linear RNNs. We discover that the addition of nonlinear recurrent activations does not reduce RNN's expressive power in the sense that nonlinear RNNs still can approximate the same class of linear functionals. The rates of the approximation by nonlinear RNNs are given and illustrated by numerical examples. These are analogous to the previous results on linear RNNs.

Since the nonlinear recurrent activations make the models nonlinear, we also investigate the approximation of nonlinear target functionals. However, unlike the fully-connected counterparts, the addition of nonlinear recurrent activations does not significantly increase RNNs' expressive capability. We give counter-examples showing that nonlinear RNNs are not universal approximators of

nonlinear functionals. In particular, we identify several classes of nonlinear functionals that cannot be universally approximated by nonlinear RNNs, including homogeneous functionals with degree greater than one, and disjoint-additive statistical functionals (Martin & Mizel, 1964). Furthermore, we prove an inverse approximation theorem that implies nonlinear RNNs do not outperform linear ones with respect to memory decay requirement, i.e. they also suffer from the "curse of memory" (Li et al., 2021). More precisely, we show that any linear functionals that can be efficiently approximated by nonlinear RNNs must also have an exponentially decaying memory - a result previously established for linear RNNs implying its inability to capture long-term memory.

To summarize, our main contributions are:

1. The hypothesis space of nonlinear RNNs is not smaller than the linear RNN hypothesis space (Theorem 4.1 and Theorem 4.2).
2. The hypothesis space of nonlinear RNNs is not much larger than the linear RNNs' hypothesis space.
    - There are several classes of nonlinear functionals that cannot be universally approximated (Proposition 4.1 and Proposition 4.2).
    - Linear functionals that can be efficiently approximated by nonlinear RNNs must have an exponentially decaying memory (Theorem 4.3).

**Organization.** In Section 2, we review the related work on approximation capacity and approximation rates of RNNs and effects of nonlinearity on other neural networks structures. The approximation problem formulation and relevant definitions are given in Section 3. The main theoretical results and corresponding numerical illustrations are presented in Section 4. All the proofs and numerical details are included in appendices.

## 2 Related work

In Li et al. (2021; 2022a), the universal approximation theorems and approximation rates results characterize the density and speed of linear RNNs applied to linear functionals. In particular, it has been proved that the linear functionals that can be efficiently approximated by linear RNNs must have an exponentially decaying memory. Li et al. (2022b) gives the universal approximation theorem and corresponding rates for recurrent encoder-decoder architectures (with linear RNNs as both encoders and decoders). These results establish a theoretical foundation for the empirical observation that recurrent architectures are ineffective in learning sequential relationships with long-term memories. However, the main limitation of the above results is the linear setting for both models and targets. In this work, we establish approximation results for nonlinear RNNs in the context of functional approximation.

A number of approximation results have been obtained for nonlinear recurrent neural networks (Funahashi & Nakamura, 1993; Doya, 1993; Chow & Li, 2000; Maass et al., 2007). However, most of these works focus on target sequential relationships generated by underlying dynamical systems in the form of difference or differential equations, and the time horizon therein is often limited to a compact set. By contrast, our work does not include such assumptions as the long-term memory is more reasonable to be investigated under the functional-approximation framework and infinite-time horizon. There are also several approximation results for randomized RNN-type networks, including reservoir computing (Grigoryeva & Ortega, 2019; Gonon et al., 2020) and echo state networks (Grigoryeva & Ortega, 2018; Gonon & Ortega, 2021; Li & Yang, 2022). The main difference is that the recurrent weights therein are randomly generated and then fixed without training. For example, (Grigoryeva & Ortega, 2018) shows that echo state networks can approximate arbitrary sequential relationships generated from an underlying dynamical system of the form $\frac{dh_t}{dt} = F(h_t, x_t), y_t = c^\top h_t$. We emphasize that the present work is concerned with general linear or nonlinear functionals, not necessarily possessing this representation, and it is shown that nonlinear RNNs (with or without trainable recurrent weights) are not universal approximators under this functional-approximation setting. Moreover, although nonlinear activations in fully-connected neural networks vastly improve the approximation power (Yarotsky, 2017; Petersen & Voigtlaender, 2018; Montanelli & Yang, 2020; Shen et al., 2019; 2021; 2022), we prove that nonlinear recurrent activations do not achieve this effect alone.

## 3 PROBLEM FORMULATION AND PRIOR RESULTS ON LINEAR RNNS

**Sequential approximation as a functional approximation problem.** The goal of sequential modeling is to learn a relationship between an input sequence $\mathbf{x} = \{\mathbf{x}_t\}$ and a corresponding output sequence $\{y_t\}$. The time index $t$ can be discrete or continuous. For ease of analysis, we adopt the continuous-time setting in (Li et al., 2021), where $t \in \mathbb{R}$. This is also a natural setting for irregularly sampled time series (Lechner & Hasani, 2020). We consider the $d$-dimensional input space $\mathcal{X} = C^1_{0,T,K}(\mathbb{R}, \mathbb{R}^d)$, which is continuously differentiable with respect to time and vanishes at infinity in the sense that $\mathbf{x}_t = \mathbf{x}_\infty$ for some $0 < T \leq t \leq \infty$, with the supremum norm $\|\mathbf{x}\|_{\mathcal{X}} := \sup_{t\in\mathbb{R}} \|\mathbf{x}_t\|_\infty$. Without loss of generality, we also assume $\mathbf{x}_t = 0$ for $t \leq 0$ and the derivatives of input is bounded: $\|\frac{d\mathbf{x}_t}{dt}\|_\infty \leq K, \forall t$.

We assume the input and output sequences are related by a family of underlying functionals

$$y_t = H_t(\{\mathbf{x}_t\}), \quad t \in \mathbb{R}, \tag{1}$$

where $H_t : \mathcal{X} \mapsto \mathbb{R}$ is a functional. Now, the sequential approximation problem can be formulated as the approximation of target functionals $\{H_t : t \in \mathbb{R}\}$ by model functionals $\{\widetilde{H}_t : t \in \mathbb{R}\}$, where $\widetilde{H}_t(\cdot)$ is contained in certain hypothesis space $\widetilde{\mathcal{H}}$, such as various classes of neural networks. We say a sequence of target functionals $\{H_t : t \in \mathbb{R}\}$ can be *uniformly approximated* by model functionals, if for any tolerance $\epsilon > 0$, there exists $\{\widetilde{H}_t : t \in \mathbb{R}\} \in \widetilde{\mathcal{H}}$ such that

$$\sup_{t\in\mathbb{R}} \sup_{\|\mathbf{x}\|_{\mathcal{X}}\leq 1} |H_t(\mathbf{x}) - \widetilde{H}_t(\mathbf{x})| \leq \epsilon. \tag{2}$$

In general, the overall hypothesis space $\widetilde{\mathcal{H}}$ can be divided into the union of hypothesis spaces with different complexities indexed by $m$:

$$\widetilde{\mathcal{H}} = \bigcup_{m=1}^{\infty} \widetilde{\mathcal{H}}_m, \tag{3}$$

where $m$ usually denotes the number of parameters. The *approximation rate* is characterized by the following relationship between the complexity $m$ and approximation error:

$$R(m) = \min_{\widetilde{H}_t \in \bigcup_{i=1}^m \widetilde{\mathcal{H}}_i} \sup_{t\in\mathbb{R}} \sup_{\|\mathbf{x}\|_{\mathcal{X}}\leq 1} |H_t(\mathbf{x}) - \widetilde{H}_t(\mathbf{x})|. \tag{4}$$

**The continuous-time RNN architecture.** First, we introduce the same continuous-time formulation of RNNs as (Li et al., 2021):

$$\frac{dh_t}{dt} = \sigma(Wh_t + U\mathbf{x}_t), \tag{5}$$

$$\hat{y}_t = c^\top h_t. \tag{6}$$

Here, $\hat{y}_t \in \mathbb{R}$ is the prediction, and $h_t \in \mathbb{R}^m$ denotes the hidden state. As a common practice, we set the initial hidden state $h_0 = 0$.[1] The hyper-parameter $m$ is also known as the hidden dimension of recurrent neural networks. For different hidden dimensions $m$, the RNN is parameterized by trainable weights $(c, W, U)$, where $c \in \mathbb{R}^m$ is the readout, $W \in \mathbb{R}^{m\times m}$ is the recurrent kernel and $U \in \mathbb{R}^{m\times d}$ is the input kernel. Obviously, the complexity of the RNN hypothesis space is characterized by the hidden dimension $m$. The nonlinearity arises from the activation $\sigma(\cdot)$, which is a scalar nonlinear function performed element-wisely, such as *tanh*, *hardtanh*, *sigmoid*, *ReLU* and so on. The hypothesis space of RNNs with different activations are denoted by $\mathcal{H}^{\sigma(\cdot)}$.

**Prior results on linear RNNs.** Before we present the nonlinear results, we first review the previous approximation theory established for linear RNNs. We begin with the definitions on functionals.

**Definition 3.1** *Let $\{H_t : t \in \mathbb{R}\}$ be a sequence of functionals.*

1. *(**Linear**) $H_t$ is linear if for any $\lambda, \lambda' \in \mathbb{R}$ and $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, $H_t(\lambda\mathbf{x} + \lambda'\mathbf{x}') = \lambda H_t(\mathbf{x}) + \lambda' H_t(\mathbf{x}')$.*

[1] This is consistent with practical applications such as TensorFlow and PyTorch, where the initial value of hidden state is set to be zero by default.

2. ***(Continuous)*** *$H_t$ is continuous if* $\sup_{\{\mathbf{x}\in\mathcal{X},\|\mathbf{x}\|_{\mathcal{X}}\leq 1\}} |H_t(\mathbf{x})| < \infty$.

3. ***(Time-homogeneous)*** *$\{H_t : t \in \mathbb{R}\}$ is time-homogeneous if the input-output relationship is invariant about time shift: $H_{t+\tau}(\mathbf{x}(\tau)) = H_t(\mathbf{x})$ for all $t, \tau \in \mathbb{R}$, where $\mathbf{x}(\tau)$ is defined by $\mathbf{x}(\tau)_s = \mathbf{x}_{s+\tau}$.*

4. ***(Causal)*** *$H_t$ is causal if it does not depend on future values of the input. That is, if $\mathbf{x}, \mathbf{x}'$ satisfy $\mathbf{x}_t = \mathbf{x}'_t$ for any $t \leq t_0$, then $H_t(\mathbf{x}) = H_t(\mathbf{x}')$ for any $t \leq t_0$.*

5. ***(Regular)*** *$H_t$ is regular if for any sequence $\{\mathbf{x}^{(n)} : n \in \mathbf{N}\}$ such that $\mathbf{x}_s^{(n)} \to 0$ for almost every $s \in \mathbb{R}$, then $\lim_{n\to\infty} H_t(\mathbf{x}^{(n)}) = 0$.*

Given the above definitions, we have the following representation for linear functionals.

**Theorem 3.1 (Riesz-Markov-Kakutani representation theorem)** *Assume $H : \mathcal{X} \mapsto \mathbb{R}$ is a linear and continuous functional, there exists a unique, vector-valued, regular, countably additive signed measure $\mu$ on $\mathbb{R}$ such that*

$$H(\mathbf{x}) = \int_{\mathbb{R}} \mathbf{x}_s^\top d\mu(s) = \sum_{i=1}^{d} \int_{\mathbb{R}} x_{s,i} d\mu_i(s). \tag{7}$$

*In addition, we have $\|H\| := \sup_{\|\mathbf{x}\|_{\mathcal{X}}\leq 1} |H(\mathbf{x})| = \|\mu\|_1(\mathbb{R}) := \sum_i |\mu_i|(\mathbb{R})$.*

Based on this representation theorem, one can further obtain that linear functionals $\{H_t : t \in \mathbb{R}\}$ with properties in Definition 3.1 can be represented by the following convolutional form

$$H_t(\mathbf{x}) = \int_{\mathbb{R}} \mathbf{x}_{t-s}^\top \boldsymbol{\rho}(s) ds, \quad t \in \mathbb{R}, \tag{8}$$

and the representation function $\boldsymbol{\rho} : [0, \infty) \to \mathbb{R}^d$ is a measurable and integrable function with $\|\boldsymbol{\rho}\|_{L^1([0,\infty))} = \sup_{t\in\mathbb{R}} \|H_t\|$. The details can be found in Appendix A.1 or Li et al. (2021).

It is straightforward to verify that the linear RNN hypothesis space also satisfies Definition 3.1. The key results of linear RNNs (Li et al., 2021; 2022a) are that: i) Any functionals satisfying Definition 3.1 can be universally approximated by linear RNNs, and the latter can be formulated as a convolution on input signals with exponential-sums: $\bar{H}_t(\mathbf{x}) = \int_{\mathbb{R}} c^\top e^{Ws} U \mathbf{x}_{t-s} ds$;[2] (ii) The approximation rate $R(m)$ can be characterized by the hidden dimension $m$ in target functionals' smoothness and memory decay. See Theorem A.1 and Theorem A.2 for precise statements. An important insight uncovered here is that the target memory pattern can be precisely defined as the decay rate of $\boldsymbol{\rho}(\cdot)$. Moreover, the inverse approximation theorem established in (Li et al., 2022a) states that the linear functionals that can be efficiently approximated by linear RNNs must have an exponentially decaying memory. These results together suggest the *curse of memory* phenomenon in the linear RNN setting.

We emphasize that all above results, including the Riesz representation for target functionals, exponential-sums representation for RNN functionals and the approximation theorems, rely on the linearity. However, general nonlinear functionals, and also nonlinear RNNs, do not possess a convolution form, hence it is impossible to derive approximation results via the above representation procedure directly. This leads to the main results in the next section.

## 4 Main results

In this section, we present the main approximation results for *nonlinear* RNNs when applied to *both* linear and nonlinear targets. We discuss in the following two aspects.

1. Nonlinearity is *not worse*. In Section 4.1, we prove that given a sequence of linear target functionals satisfying Definition 3.1, nonlinear RNNs can achieve the uniform approximation (see Equation (2)), and one can characterize the corresponding approximation rate (see Equation (4)). Both of them appear similar to those of linear RNNs.

[2]The functional approximation problem $\|\bar{H}_t - H_t\|$ is then reduced to the function approximation problem $\|c^\top e^{Ws} U - \rho(s)\|$.

2. Nonlinearity is *not better*. On the side of nonlinear targets (Section 4.2), we give several counter-examples of nonlinear functionals that satisfy properties 2-5 in Definition 3.1, cannot be uniformly approximated by nonlinear RNNs. These negative results indicate that the recurrent nonlinearity is not enough for RNNs to approximate arbitrary nonlinear functionals. On the side of linear targets (Section 4.3), we further prove an inverse approximation theorem of linear functionals satisfying Definition 3.1 for nonlinear RNNs with tanh/sigmoid activations. This is again a negative result, which suggests that even with nonlinear activations, the RNN still face the same problem of "curse of memory" as its linear counterpart.

### 4.1 Approximation of linear functionals by nonlinear RNNs

In this section, we first consider the approximation of functionals with properties defined in Definition 3.1 as they can be approximated universally by linear RNNs (Li et al., 2021; 2022a). The following result shows the approximation capacity of nonlinear RNNs is not smaller than linear RNNs.

**Theorem 4.1 (Universal approximation of linear functionals)** *Let $\{H_t : t \in \mathbb{R}\}$ be a family of linear, continuous, causal, regular and time-homogeneous functionals on $\mathcal{X}$. Assume the activation $\sigma(\cdot)$ is continuous and monotonically-increasing, and satisfies the following assumptions*

$$\sigma(0) = 0, \sigma'(0) = 1, |\sigma'(z)| \leq M, |\sigma'(z) - \sigma'(0)| \leq M|\sigma(z)|, |\sigma''(z)| \leq M|\sigma'(z)| \tag{9}$$

*for some constant $M > 0$. Then, for any $\epsilon > 0$, there exists a sequence of nonlinear RNNs $\{\widetilde{H}_t : t \in \mathbb{R}\} \in \widetilde{\mathcal{H}}^\sigma$ such that*

$$\sup_{t\in\mathbb{R}} \|H_t - \widetilde{H}_t\| \equiv \sup_{t\in\mathbb{R}} \sup_{\|\mathbf{x}\|_{\mathcal{X}}\leq 1} |H_t(\mathbf{x}) - \widetilde{H}_t(\mathbf{x})| \leq \epsilon. \tag{10}$$

One can verify that tanh, which is commonly-used in practice for RNNs, is an activation satisfying the conditions (9). The proof of Theorem 4.1 is given in Appendix A.3.

Based on the above universal approximation result, a natural question one may further have is, whether adding the nonlinear recurrent activations changes the approximation efficiency. We prove the same approximation rate as the linear setting (Li et al., 2021; 2022a). The key concepts are the memory and smoothness of target functionals. Define the maps $t \mapsto H_t(\mathbf{e}_i)$ for $i = 1, \ldots, d$, where $\mathbf{e}_i$ is the constant input signal defined by $\mathbf{e}_i = e_i \mathbf{1}_{\{t\geq 0\}}$ with $e_i$ as the standard basis vector in $\mathbb{R}^d$.

**Theorem 4.2 (Approximation rate of linear functionals)** *Assume the same conditions as in Theorem 4.1. Consider the output of constant signals*

$$y_i(t) = H_t(\mathbf{e}_i), \quad i = 1, \ldots, d. \tag{11}$$

*Suppose there exist constants $\alpha \in \mathbb{N}_+, \beta, \gamma > 0$, such that for $i = 1, \ldots, d$, $k = 1, \ldots, \alpha + 1$, $y_i(t) \in C^{(\alpha+1)}(\mathbb{R})$, and*

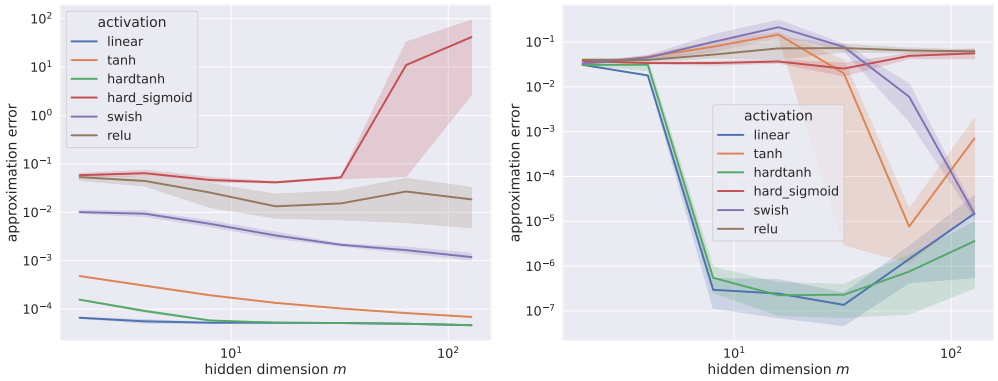
$$e^{\beta t} y_i^{(k)}(t) = o(1) \quad as\ t \to +\infty, \tag{12}$$

$$\sup_{t\geq 0} \frac{|e^{\beta t} y_i^{(k)}(t)|}{\beta^k} \leq \gamma. \tag{13}$$

*Then, there exists a universal constant $C(\alpha)$ only depending on $\alpha$, such that for any $m \in \mathbb{N}_+$, there exists a sequence of width-$m$ nonlinear RNN functionals $\{\widetilde{H}_t : t \in \mathbb{R}\} \in \widetilde{\mathcal{H}}^\sigma_m$ such that*

$$R(m) = \sup_{t\in\mathbb{R}} \|H_t - \widetilde{H}_t\| \equiv \sup_{t\in\mathbb{R}} \sup_{\|\mathbf{x}\|_{\mathcal{X}}\leq 1} |H_t(\mathbf{x}) - \widetilde{H}_t(\mathbf{x})| \leq \frac{C(\alpha)\gamma d}{\beta m^\alpha}. \tag{14}$$

The proof is included in Appendix A.4. Theorem 4.2 shows that, as long as the activation function $\sigma(\cdot)$ satisfies the conditions in Equation (9), such as the tanh activation, nonlinear RNNs can obtain the same approximation rate as linear RNNs.




(a) Exponential-sum target

(b) Airy function target

Figure 1: Numerical illustration for approximation capacity of RNNs with different activations. The target functionals are linear functionals with the exponential-sum memory pattern $\rho_1$ and Airy function memory pattern $\rho_2$. We see good performance of linear, tanh and hardtanh activations for both the exponential-sum and Airy example. Similar phenomena hold for the Airy function example. The increase of approximation errors at higher hidden dimensions might be related to optimization difficulties, since we select a fixed range of epochs for all hidden dimensions and activations. The details of numerical experiments can be found in Appendix B.

.

We now illustrate that RNNs with different activations can perform similarly with numerical examples (see Figure 1). Let the input dimension $d = 1$, we consider linear functionals in Equation (8) with two different memory patterns:

$$\rho_1(t) = \sum_{i=1}^{b} a_i e^{w_i t}, \qquad w_i \leq 0, \tag{15}$$

$$\rho_2(t) = \mathrm{Ai}\big(s(t - t_0)\big), \qquad \mathrm{Ai}(t) = \frac{1}{\pi} \lim_{\xi \to \infty} \int_0^{\xi} \cos(\frac{u^3}{3} + tu) du. \tag{16}$$

The linear functional with an exponential-sum memory pattern can be represented (exactly) by a linear RNN with a hidden dimension no less than $b$. The Airy function is an oscillatory function between the time $[0, t_0]$ and an exponential decaying function over $[t_0, \infty)$. The linear functional with Airy function memory pattern cannot be represented exactly but still can be approximated by RNNs with linear/tanh/hardtanh activations sufficiently well. In the numerical experiments, we estimate the approximation error between $H_t$ and $\widetilde{H}_t$ by

$$\sup_{t \in \mathbb{R}} \mathbb{E}_{\mathbf{x} \in \mathcal{X}} |H_t(\mathbf{x}) - \widetilde{H}_t(\mathbf{x})|. \tag{17}$$

The expectation is taken over samples of piece-wise constant signals with the values sampled from a uniform distribution $U[-0.1, 0.1]$. As the RNN is trained for a sufficiently long time, the estimated approximation error in Figure 1 reflects the similarity of approximation errors for different activations.

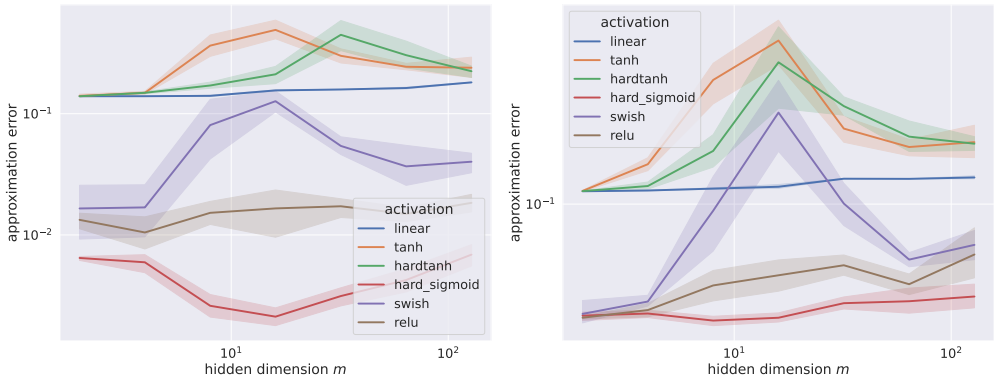
### 4.2 Approximation of nonlinear functionals by nonlinear RNNs

We now consider the nonlinear target functionals that satisfy properties 2-5 in Definition 3.1. The question is whether nonlinear RNNs can uniformly approximate *any* such nonlinear functionals. The answer is negative and we give some counter-examples.

#### 4.2.1 Homogeneous functionals cannot be approximated by tanh RNNs

Consider the **homogeneous** functionals with degree $p > 0$:

$$H_t(\kappa \mathbf{x}) = \kappa^p H_t(\mathbf{x}), \qquad \forall \kappa. \tag{18}$$




(a) Exponential-sum target

(b) Airy function target

Figure 2: Numerical illustrations for the difficulty of approximating homogeneous functionals by RNNs. The memory $\rho_1$ and $\rho_2$ are defined in Equation (15) and (16).

An example is the *monomial* functional $H_t(\mathbf{x}) = \int_0^\infty (\mathbf{x}_{t-s}^p)^\top \boldsymbol{\rho}_s ds$. This monomial functional is linear if and only if $p = 1$.

We show that the sequence of homogeneous functionals with degree $p > 1$ cannot be universally approximated by nonlinear RNNs with the tanh activation.

**Proposition 4.1** *There exists a sequence of homogeneous functionals $\{H_t : t \in \mathbb{R}\}$ with an homogeneity degree $p > 1$ and an $\epsilon > 0$, such that there is no tanh RNN $\{\widetilde{H}_t : t \in \mathbb{R}\}$ approximating $\{H_t : t \in \mathbb{R}\}$ uniformly with an approximation error smaller than $\epsilon$.*

The proof is given in Appendix A.5.

We illustrate this result with numerical examples (see Figure 2). Take $d = 1$ and $p = 2$, we construct the monomial functionals with the same memory patterns $\rho_1$ and $\rho_2$ (see Equation (15) and (16)) as the linear functionals. The approximation error is still estimated with the same approach as Figure 1. The results in Figure 2 reflect the difficulty of approximating homogeneous functional, as the estimated approximation error does not decrease as the hidden dimension increases.

### 4.2.2 Statistical functionals cannot be approximated by tanh RNNs

Besides the homogeneous functionals in the previous section, we further identify a class of nonlinear functionals - statistical functionals - that cannot be approximated by nonlinear RNNs. We begin with the definition of statistical functionals.

**Definition 4.1** *Two real-valued inputs $\mathbf{x}_1$ and $\mathbf{x}_2$ are said to be **equimeasurable**, if for every Borel set $B$ on $(-\infty, \infty)$, $\mathbf{x}_1^{-1}(B)$ and $\mathbf{x}_2^{-1}(B)$ are measurable and have equal measure.*

*As is defined in Martin & Mizel (1964), for equimeasurable input $\mathbf{x}_1$ and $\mathbf{x}_2$, **statistical** functionals satisfy the following property:*

$$H(\mathbf{x}_1) = H(\mathbf{x}_2). \tag{19}$$

According to the representation theorem of continuous disjoint-additive statistical functionals (Martin & Mizel, 1964), i.e. $H(\mathbf{x}) = \int_{\mathbb{R}} f(\mathbf{x}_s)^\top \boldsymbol{\rho}_s ds$, we further consider the sequence of disjoint-additive statistical functionals:

$$H_t(\mathbf{x}) = \int_{\mathbb{R}} f_t(\mathbf{x}_s, s) ds. \tag{20}$$

By Definition 3.1, the continuous, disjoint-additive, statistical, causal and time-homogeneous functionals can be represented by

$$H_t(\mathbf{x}) = \int_{-\infty}^{t} f(\mathbf{x}_s, t - s) ds. \tag{21}$$

Based on the homogeneity property stated in Section 4.2.1, we define the $p$-sub-homogeneity.

**Definition 4.2 ($p$-sub-homogeneity)** *For $p > 1$, we say a functional sequence $\{H_t : t \in \mathbb{R}\}$ has $p$-sub-homogeneity, if for any bounded input $\|x\|_{\mathcal{X}} < c_0$, and any $\kappa \in (0, 1)$,*

$$H_t(\kappa \boldsymbol{x}) \leq \kappa^p H_t(\boldsymbol{x}). \tag{22}$$

An example of nonlinear statistical functional is $H_t(\mathbf{x}) = \int_{-\infty}^{t} P(\mathbf{x})\rho(t-s)ds$, where $P(\mathbf{x}) = \sum_{p_i \in \mathcal{P}} q_i \mathbf{x}^{p_i}$ is a polynomial. If the smallest degree $\min \mathcal{P} > 1$, $H_t(\mathbf{x})$ has the $p$-sub-homogeneity.

We show that continuous, causal, disjoint-additive, statistical and time-homogeneous functionals with $p$-sub-homogeneity cannot be universally approximated by RNNs with tanh activations.

**Proposition 4.2** *There exists a sequence of continuous, causal, disjoint-additive, statistical and time-homogeneous functionals $\{H_t : t \in \mathbb{R}\}$ with $p$-sub-homogeneity ($p > 1$) and an $\epsilon > 0$, such that there is no tanh RNN $\{\widetilde{H}_t : t \in \mathbb{R}\}$ approximating $\{H_t : t \in \mathbb{R}\}$ uniformly with an approximation error smaller than $\epsilon$.*

The proof is given in Appendix A.6.

**Remark 4.1** *Gonon et al. (2020) investigates the approximation of nonlinear functionals using ReLU RNNs with randomly generated weights, and their approximation results are established in $L_2$-norm. By contrast, the negative results developed in this section are for tanh RNNs and in $L_\infty$-norm, which is quite different in the approximation setting.*

### 4.3 Inverse approximation of linear functionals by nonlinear RNNs

Previous results give some classes of functionals that cannot be approximated by nonlinear RNNs. Now, we show a different type of negative result, in the form of an inverse approximation theorem, which is also known as Bernstein-type approximation results (Bernstein, 1914).

In Li et al. (2022a), such a result was proved for linear RNNs: If a class of linear functionals can be efficiently approximated by linear RNNs, it must possess exponentially decaying memory pattern $\rho(t)$. That is, *only* functionals with exponentially decaying memory can be approximated efficiently by linear RNNs. An important question is whether the addition of nonlinearity alleviates this limitation and allows an efficient approximation of functionals. In this section, we show that the answer is negative, and a similar inverse approximation result holds for RNNs with tanh activations.

**Theorem 4.3 (Inverse approximation theorem)** *Assume that the activation function $\sigma(\cdot)$ is tanh, and $\{H_t : t \in \mathbb{R}\}$ is a family of linear, continuous, causal, regular and time-homogeneous functionals on $\mathcal{X}$. Consider the output of constant signals*

$$y_i(t) = H_t(\mathbf{e}_i) \in C^{(\alpha+1)}(\mathbb{R}), \quad i = 1, \ldots, d,\ \alpha \in \mathbb{N}_+. \tag{23}$$

*Suppose there exists a sequence of nonlinear RNNs $\{\widetilde{H}_t : t \in \mathbb{R}\} \in \widetilde{\mathcal{H}}^\sigma$ with an increasing hidden dimension $m$ approximating $\{H_t : t \in \mathbb{R}\}$ in the following sense:*

$$\lim_{m \to \infty} \sup_{t \geq 0} |\tilde{y}_{i,m}^{(k)}(t) - y_i^{(k)}(t)| = 0, \quad i = 1, \ldots, d,\ k = 1, \ldots, \alpha + 1, \tag{24}$$

*where*

$$\tilde{y}_{i,m}(t) = \widetilde{H}_t(\mathbf{e}_i), \quad i = 1, \ldots, d. \tag{25}$$

*Moreover, assume that the recurrent kernel matrix $W_m$ in the nonlinear RNN is a special Hurwitz matrix, that is, there exists a diagonal positive-definite matrix $D_m$ and a negative-definite matrix $N_m$ such that $W_m = D_m N_m$. Define $n_m = \max_{j \in [m]} \mathrm{Re}(\lambda_j)$, where $\{\lambda_j\}_{j=1}^m$ collects all the eigenvalues of $N_m$. Define $d_m = \min_{j \in [m]} (D_m)_{jj}$. Assume that the parameters in nonlinear RNNs are uniformly bounded and there exits constants $\beta, \theta > 0$ such that $\limsup_{m \to \infty} \frac{n_m}{\theta d_m} < -\beta$, then we have*

$$e^{\beta t} y_i^{(k)}(t) = o(1) \text{ as } t \to +\infty, \quad i = 1, \ldots, d,\ \ k = 1, \ldots, \alpha + 1. \tag{26}$$

Here, we only give the proof sketch to show the meaning and implications of Theorem 4.3. For the detailed proof, see Appendix A.7. Similar results can also be obtained for the *sigmoid* activation.

Define $v_t = \frac{dh_t}{dt}$, then $\tilde{y}^{(1)}_{i,m}(t) = c^\top v_{t,m}$. We know $\tilde{y}^{(1)}_{i,m}(t)$ decays exponentially if $v_{t,m}$ decays exponentially. For $|z|$ sufficiently small, the similarity of different activations $\sigma(z)$, i.e. tanh and linear, is quite clear (see the left panel of Figure 3). Therefore, the limiting behavior for different activations in this small region is the same (see the right panel of Figure 3), that is, $v_t$ decays exponentially if $v_t$ decays. The remaining problem is whether the nonlinear dynamics will enter this small region eventually. We show the validity and correctness via the Lyapunov function analysis.

We construct the Lyapunov function for the tanh RNN dynamics as follows:

$$V(v) = \sum_{i=1}^{m} \sum_{j=1}^{\infty} \frac{1}{2jD_{ii}} v_i^{2j}. \tag{27}$$

The derivative of this Lyapunov function shows that $v_t$ decays exponentially with the rate $\lim_{t\to\infty} e^{\beta t}\|v(t)\|_\infty = 0$, which implies $e^{\beta t}\tilde{y}^{(1)}_{i,m}(t) = o(1)$ for all $m \in \mathbb{N}_+$, and hence $e^{\beta t} y^{(1)}_i(t) = o(1)$. This is the essential reason for *tanh/sigmoid* recurrent activations do not fundamentally change the memory pattern of RNNs.

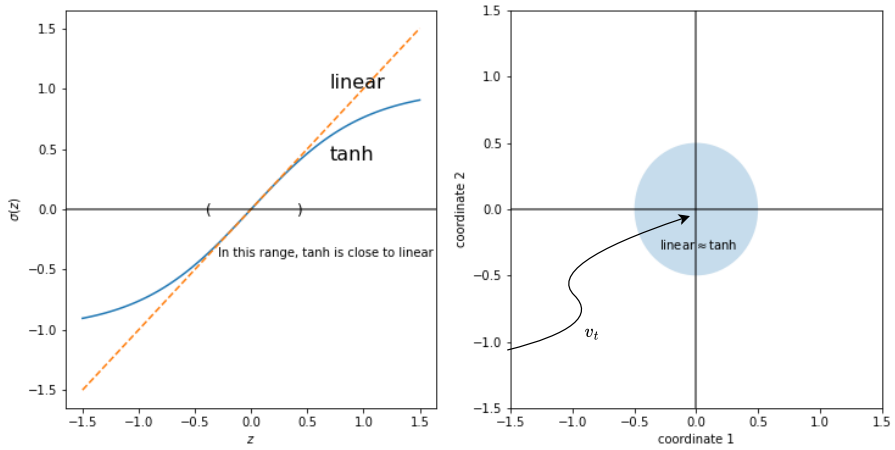



Figure 3: The left panel shows the closeness between tanh and linear activations when $|z|$ is small. If the dynamics enters this small region at any time, the behavior of nonlinear RNNs is similar to that of linear RNNs. This can be guaranteed by constructing a Lyapunov function $V(v)$ defined in Equation (27).

**Remark 4.2** *In general, the inverse approximation theorem is supposed to hold for a Hurwitz matrix $W$, which can be decomposed into the product of a symmetric positive-definite matrix $S$ and a negative-definite matrix $N$ (Duan & Patton, 1998). The Lyapunov function analysis here does not work for general Hurwitz matrices, but the numerical evidence in Appendix C.2 shows that $v_t$ still converges to 0 exponentially fast.*

## 5 Conclusion

In this paper, we analyze the approximation properties of RNNs with nonlinear recurrent activations. We show that nonlinear RNNs' performance on linear targets is not worse compared with linear RNNs. However, the additional recurrent nonlinearity does not give the same type of universal approximation of nonlinear targets, and also fail to enlarge the concept space of linear targets. Moreover, the approximation of linear targets by nonlinear RNNs is not better than linear RNNs in the sense that the "curse-of-memory" phenomenon still exists. In summary, adding nonlinear recurrent activations to RNNs does not significantly help in the approximation sense.

## REFERENCES


Yoshua. Bengio, Patrice. Simard, and Paolo. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994.

Serge Bernstein. Sur la meilleure approximation de |x| par des polynomes de degrés donnés. *Acta Mathematica*, 37(1):1–57, December 1914. ISSN 1871-2509. doi: 10.1007/BF02401828. URL https://doi.org/10.1007/BF02401828.

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

Jerome T Connor, R Douglas Martin, and Les E Atlas. Recurrent neural networks and robust time series prediction. *IEEE transactions on neural networks*, 5(2):240–254, 1994.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

Kenji Doya. Universality of fully connected recurrent neural networks. *Dept. of Biology, UCSD, Tech. Rep*, 1:1–6, 1993.

Guang-Ren Duan and Ron J Patton. A note on hurwitz stability of matrices. *Automatica*, 34(4): 509–511, 1998.

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

Lukas Gonon and Juan-Pablo Ortega. Fading memory echo state networks are universal. *Neural Networks*, 138:10–13, 2021.

Lukas Gonon, Lyudmila Grigoryeva, and Juan-Pablo Ortega. Approximation bounds for random neural networks and reservoir systems. *arXiv preprint arXiv:2002.05933*, 2020.

Alex Graves and Navdeep Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In *International conference on machine learning*, pp. 1764–1772. PMLR, 2014.

Lyudmila Grigoryeva and Juan-Pablo Ortega. Echo state networks are universal. *Neural Networks*, 108:495–508, 2018.

Lyudmila Grigoryeva and Juan-Pablo Ortega. Differentiable reservoir computing. *J. Mach. Learn. Res.*, 20(179):1–62, 2019.

Sepp Hochreiter, Yoshua Bengio, Paolo Frasconi, and Jürgen Schmidhuber. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. In S. C. Kremer and J. F. Kolen (eds.), *A Field Guide to Dynamical Recurrent Neural Networks*. IEEE Press, 2001.

Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4 (2):251–257, 1991.

Kurt Hornik. Some new results on neural network approximation. *Neural networks*, 6(8):1069–1072, 1993.

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

Mathias Lechner and Ramin Hasani. Learning long-term dependencies in irregularly-sampled time series. *arXiv preprint arXiv:2006.04418*, 2020.

Zhen Li and Yunfei Yang. Universality and approximation bounds for echo state networks with random weights. *arXiv preprint arXiv:2206.05669*, 2022.

Zhong Li, Jiequn Han, Weinan E, and Qianxiao Li. On the curse of memory in recurrent neural networks: Approximation and optimization analysis. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=8Sqhl-nF50.

Zhong Li, Jiequn Han, Weinan E, and Qianxiao Li. Approximation and optimization theory for linear continuous-time recurrent neural networks. *Journal of Machine Learning Research*, 23 (42):1–85, 2022a. URL http://jmlr.org/papers/v23/21-0368.html.

Zhong Li, Haotian Jiang, and Qianxiao Li. On the approximation properties of recurrent encoder-decoder architectures. In *International Conference on Learning Representations*, 2022b. URL https://openreview.net/forum?id=xDIvIqQ3DXD.

Wolfgang Maass, Prashant Joshi, and Eduardo D Sontag. Computational aspects of feedback in neural circuits. *PLoS computational biology*, 3(1):e165, 2007.

AD Martin and VJ Mizel. A representation theorem for certain nonlinear functionals. *Archive for Rational Mechanics and Analysis*, 15(5):353–367, 1964.

Hadrien Montanelli and Haizhao Yang. Error bounds for deep relu networks using the kolmogorov–arnold superposition theorem. *Neural Networks*, 129:1–6, 2020.

Philipp Petersen and Felix Voigtlaender. Optimal approximation of piecewise smooth functions using deep relu neural networks. *Neural Networks*, 108:296–330, 2018.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Nonlinear approximation via compositions. *Neural Networks*, 119:74–84, 2019.

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network with approximation error being reciprocal of width to power of square root of depth. *Neural Computation*, 33(4):1005–1036, 2021.

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Optimal approximation rate of relu networks in terms of width and depth. *Journal de Mathématiques Pures et Appliquées*, 157:101–135, 2022.

Maxwell B Stinchcombe. Neural network approximation of continuous functionals and continuous functions on compactifications. *Neural Networks*, 12(3):467–477, 1999.

Ilya Sutskever, James Martens, and Geoffrey E Hinton. Generating text with recurrent neural networks. In *ICML*, 2011.

Duyu Tang, Bing Qin, and Ting Liu. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 1422–1432, 2015.

Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural Networks*, 94: 103–114, 2017.

# A THEORETICAL RESULTS AND PROOFS

## A.1 UNIVERSAL APPROXIMATION AND APPROXIMATION RATE OF LINEAR FUNCTIONALS BY LINEAR RNNS

We include the statements of the universal approximation theorem and approximation rate for linear functionals by linear RNNs from Li et al. (2021).

**Theorem A.1 (Universal approximation for linear functionals by linear RNNs (Li et al., 2021))** *Let $\{H_t : t \in \mathbb{R}\}$ be a family of linear, continuous, causal, regular and time-homogeneous functionals on $\mathcal{X}$. Then, for any $\epsilon > 0$, there exists $\{\bar{H}_t : t \in \mathbb{R}\} \in \bar{\mathcal{H}}^{\text{linear}}$ such that*

$$\sup_{t\in\mathbb{R}} \|H_t - \bar{H}_t\| \equiv \sup_{t\in\mathbb{R}} \sup_{\|\mathbf{x}\|_{\mathcal{X}}\leq 1} |H_t(\mathbf{x}) - \bar{H}_t(\mathbf{x})| \leq \epsilon. \tag{28}$$

**Theorem A.2 (Approximation rates for linear functionals by linear RNNs (Li et al., 2021))** *Assume the same conditions as Theorem A.1. Consider the output of constant signals*

$$y_i(t) = H_t(\mathbf{e}_i), \quad i = 1, \ldots, d.$$

*Suppose there exists constants $\alpha \in \mathbb{N}_+, \beta, \gamma > 0$ such that for $i = 1, \ldots, d$, $k = 1, \ldots, \alpha + 1$, $y_i(t) \in C^{(\alpha+1)}(\mathbb{R})$ and*

$$e^{\beta t} y_i^{(k)}(t) = o(1) \quad as\ t \to +\infty,$$
$$\sup_{t\geq 0} \frac{|e^{\beta t} y_i^{(k)}(t)|}{\beta^k} \leq \gamma.$$

*Then, there exists a universal constant $C(\alpha)$ only depending on $\alpha$, such that for any $m \in \mathbb{N}_+$, there exists a sequence of width-$m$ RNN functionals $\{\bar{H}_t : t \in \mathbb{R}\} \in \bar{\mathcal{H}}_m^{\text{linear}}$ such that*

$$R(m) = \sup_{t\in\mathbb{R}} \|H_t - \bar{H}_t\| \equiv \sup_{t\in\mathbb{R}} \sup_{\|\mathbf{x}\|_{\mathcal{X}}\leq 1} |H_t(\mathbf{x}) - \bar{H}_t(\mathbf{x})| \leq \frac{C(\alpha)\gamma d}{\beta m^\alpha}. \tag{29}$$

**Remark A.1** *An important step to notice is the construction of $\bar{W}$ (the recurrent kernel matrix) in these approximation results is a diagonal negative definite matrix.*

## A.2 PROOFS OF PRELIMINARY PROPOSITIONS AND LEMMAS

**Proposition A.1 (Universal approximation for linear RNNs by nonlinear RNNs)** *For any linear RNN $\{\bar{H}_t : t \in \mathbb{R}\}$, there exists a sequence of nonlinear RNN $\{\widetilde{H}_t : t \in \mathbb{R}\}$ with activations satisfying Equation (9) that can approximate the linear RNN with the following error bound:*

$$\sup_{t\in\mathbb{R}} \sup_{\|\mathbf{x}\|_{\mathcal{X}}\leq 1} |\bar{H}_t(\mathbf{x}) - \widetilde{H}_t(\mathbf{x})| \leq \frac{\epsilon}{2}. \tag{30}$$

*Proof.* We first assume the linear RNN used in Theorem A.1 achieves an error $\frac{\epsilon}{2}$ is represented by the parameters $(\bar{c}, \bar{W}, \bar{U})$. After rescaling, this RNN is the same as the linear RNN represented by $(M\bar{c}, \bar{W}, \frac{1}{M}\bar{U})$ for any scalar $M \neq 0$. We only need to find $M_\epsilon$ such that the nonlinear RNN $\widetilde{H}_t$ represented by $(M_\epsilon\bar{c}, \bar{W}, \frac{1}{M_\epsilon}\bar{U})$ satisfies the following estimate:

$$\sup_{t\in\mathbb{R}} \sup_{\|\mathbf{x}\|_{\mathcal{X}}\leq 1} |\bar{H}_t(\mathbf{x}) - \widetilde{H}_t(\mathbf{x})| \leq \frac{\epsilon}{2},$$

where the linear RNN $\bar{H}_t$ is represented by $(M_\epsilon\bar{c}, \bar{W}, \frac{1}{M_\epsilon}\bar{U})$, and it is equivalent to find $M_\epsilon$ such that the nonlinear RNN $\widetilde{H}_t$ represented by $(M_\epsilon\bar{c}, \bar{W}, \bar{U})$ satisfies the following estimate:

$$\sup_{t\in\mathbb{R}} \sup_{\|\mathbf{x}\|_{\mathcal{X}}\leq \frac{1}{M_\epsilon}} |\bar{H}_t(\mathbf{x}) - \widetilde{H}_t(\mathbf{x})| \leq \frac{\epsilon}{2},$$

where the linear RNN $\bar{H}_t$ is constructed by $(M_\epsilon\bar{c}, \bar{W}, \bar{U})$.

Define $\tilde{v}_s = \frac{d\tilde{h}_s}{ds}, \bar{v}_s = \frac{d\bar{h}_s}{ds}$. With the conditions (9) on activations, one can write down the dynamics of $\tilde{v}_s$ and $\bar{v}_s$ by

$$\frac{d\tilde{v}_s}{ds} = \text{diag}\Big(\sigma'\big(\sigma^{-1}(\tilde{v}_s)\big)\Big)(\bar{W}\tilde{v}_s + \bar{U}\frac{d\mathbf{x}_s}{ds}),$$
$$\frac{d\bar{v}_s}{ds} = \bar{W}\bar{v}_s + \bar{U}\frac{d\mathbf{x}_s}{ds}.$$

Take the difference of $\frac{d\bar{v}_s}{ds}$ and $\frac{d\bar{v}_s}{ds}$, and notice that $\bar{W}$ is a diagonal matrix by construction, we have

$$\frac{d(\bar{v}_s - \tilde{v}_s)}{ds} = \bar{W}(\bar{v}_s - \tilde{v}_s) + \bigg(I - \text{diag}\Big(\sigma'\big(\sigma^{-1}(\tilde{v}_s)\big)\Big)\bigg)(\bar{W}\tilde{v}_s + \bar{U}\frac{d\mathbf{x}_s}{ds}),$$
$$\bar{v}_s - \tilde{v}_s = \int_0^s e^{\bar{W}(s-r)}\bigg(I - \text{diag}\Big(\sigma'\big(\sigma^{-1}(\tilde{v}_r)\big)\Big)\bigg)(\bar{W}\tilde{v}_r + \bar{U}\frac{d\mathbf{x}_r}{dr})dr.$$

Based on the above representation of $\bar{v}_s - \tilde{v}_s$, we can rewrite $\widetilde{H}_t(\mathbf{x}) - \bar{H}_t(\mathbf{x})$ in the following way:

$$\begin{aligned}
\bar{H}_t(\mathbf{x}) - \widetilde{H}_t(\mathbf{x}) =& M_\epsilon c^\top \int_0^t \tilde{v}_s - \bar{v}_s ds \\
=& M_\epsilon c^\top \int_0^t \int_0^s e^{\bar{W}(s-r)}(I - \text{diag}\Big(\sigma'\big(\sigma^{-1}(\tilde{v}_r)\big)\Big)\bigg)(\bar{W}\tilde{v}_r + \bar{U}\frac{d\mathbf{x}_r}{dr})drds \\
=& M_\epsilon c^\top \int_0^t \int_r^t e^{\bar{W}(s-r)}\bigg(I - \text{diag}\Big(\sigma'\big(\sigma^{-1}(\tilde{v}_r)\big)\Big)\bigg)(\bar{W}\tilde{v}_r + \bar{U}\frac{d\mathbf{x}_r}{dr})dsdr \\
=& M_\epsilon c^\top \int_0^t (-\bar{W})^{-1}(I - e^{(t-r)\bar{W}})\bigg(I - \text{diag}\Big(\sigma'\big(\sigma^{-1}(\tilde{v}_r)\big)\Big)\bigg)(\bar{W}\tilde{v}_r + \bar{U}\frac{d\mathbf{x}_r}{dr})dr \\
=& - M_\epsilon c^\top \int_0^t (I - e^{(t-r)\bar{W}})\bigg(I - \text{diag}\Big(\sigma'\big(\sigma^{-1}(\tilde{v}_r)\big)\Big)\bigg)\tilde{v}_r dr \\
&+ M_\epsilon c^\top \int_0^t (-\bar{W})^{-1}(I - e^{(t-r)\bar{W}})\bigg(I - \text{diag}\Big(\sigma'\big(\sigma^{-1}(\tilde{v}_r)\big)\Big)\bigg)\bar{U}\frac{d\mathbf{x}_r}{dr}dr.
\end{aligned}$$

According to Lemma A.1, we get $\|\tilde{v}_r\|_\infty < \frac{C}{M_\epsilon}$ for all $r$ if $M_\epsilon > M_1$, with $C$ as a constant related to $\bar{W}, \bar{U}, M$. Moreover, since $\frac{dx_r}{dr} = 0$ for $r > T$, we have $\|\tilde{v}_r\|_\infty < \frac{Ce^{-\beta_0(t-T)}}{M_\epsilon}$ for some constant $\beta_0 > 0$.

Also, since the derivative for inputs is bounded by $K$, the scaled input derivatives can be bounded by $\|\frac{d\mathbf{x}_r}{dr}\|_\infty \leq \frac{K}{M_\epsilon}$. According to Lemma A.2, there exists $M_2$ related to $\bar{c}, \bar{W}, \bar{U}, \epsilon, C, M, T, \beta_0$, such that for $M_\epsilon > M_2 > M_1$, we have

$$\|(I - e^{(t-r)\bar{W}})\bigg(I - \text{diag}\Big(\sigma'\big(\sigma^{-1}(\tilde{v}_r)\big)\Big)\bigg)\|_\infty < \frac{\epsilon}{4m\|c\|_\infty CT},$$
$$\|(-\bar{W})^{-1}(I - e^{(t-r)\bar{W}})\bigg(I - \text{diag}\Big(\sigma'\big(\sigma^{-1}(\tilde{v}_r)\big)\Big)\bigg)\bar{U}\|_\infty < \frac{\epsilon}{4m\|c\|_\infty KT},$$
$$\|(I - e^{(t-r)\bar{W}})\bigg(I - \text{diag}\Big(\sigma'\big(\sigma^{-1}(\tilde{v}_r)\big)\Big)\bigg)\|_\infty < \frac{\epsilon}{2m\|c\|_\infty C/\beta_0}.$$

(i) If $t < T$, we have

$$\begin{aligned}
|\bar{H}_t(x) - \widetilde{H}_t(x)| \leq& M_\epsilon m\|c\|_\infty \frac{\epsilon}{4m\|c\|_\infty CT} T \sup_{0<r<t}\|\tilde{v}_r\|_\infty \\
&+ M_\epsilon m\|c\|_\infty \frac{\epsilon}{4m\|c\|_\infty KT} T \sup_{0<r<t}\|\frac{dx_r}{dr}\|_\infty \\
<& \frac{\epsilon}{4} + \frac{\epsilon}{4} = \frac{\epsilon}{2}.
\end{aligned}$$

(ii) If $t > T$, we have

$$\begin{aligned}
|\bar{H}_t(x) - \widetilde{H}_t(x)| \leq & M_\epsilon m\|c\|_\infty \frac{\epsilon}{4m\|c\|_\infty CT} T \sup_{0<r<T} \|\tilde{v}_r\|_\infty \\
& + M_\epsilon m\|c\|_\infty \frac{\epsilon}{4m\|c\|_\infty KT} T \sup_{0<r<T} \|\frac{dx_r}{dr}\|_\infty \\
& + M_\epsilon m\|c\|_\infty \frac{\epsilon}{2m\|c\|_\infty C/\beta_0} \int_T^\infty \|v_r\|_\infty dr \\
\leq & \frac{\epsilon}{4} + \frac{\epsilon}{4} + M_\epsilon m\|c\|_\infty \cdot \frac{\epsilon}{2m\|c\|_\infty C/\beta_0} \frac{C/\beta_0}{M_\epsilon} \\
< & \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.
\end{aligned}$$

In summary, we proved that $|\bar{H}_t(x) - \widetilde{H}_t(x)| < \epsilon$ for all $x \in \mathcal{X}$. □

**Lemma A.1** *There exists $M_1$ depending on $\bar{W}, \bar{U}, M$, such that for all $M_\epsilon > M_1$, there exists a constant $C > 0$ such that for any $s$,*

$$\|\tilde{v}_s\|_\infty < \frac{C}{M_\epsilon}.$$

*Moreover, if $\frac{dx_s}{ds} = 0$ for $s > T$, we have $\|\tilde{v}_s\|_\infty < \frac{Ce^{-\beta_0(s-T)}}{M_\epsilon}, s > T$ for some $\beta_0 > 0$.*

*Proof.* By the general formula of $\tilde{v}_s$ and $x_0 = 0$, we get

$$\tilde{v}_s = \int_0^s e^{\int_r^s \text{diag}\left(\sigma'\left(\sigma^{-1}(\tilde{v}_q)\right)\right)\bar{W}dq} \text{diag}\Big(\sigma'\big(\sigma^{-1}(\tilde{v}_r)\big)\Big)\bar{U}\frac{d\mathbf{x}_r}{dr}dr. \tag{31}$$

By the integration by parts, we have

$$\begin{aligned}
\tilde{v}_s = & e^{\int_r^s \text{diag}\left(\sigma'\left(\sigma^{-1}(\tilde{v}_q)\right)\right)\bar{W}dq} \text{diag}\Big(\sigma'\big(\sigma^{-1}(\tilde{v}_r)\big)\Big)\bar{U}\mathbf{x}_r|_{r=0}^s \\
& - \int_0^s e^{\int_r^s \text{diag}\left(\sigma'\left(\sigma^{-1}(\tilde{v}_q)\right)\right)\bar{W}dq} \left[\frac{d}{dr}\text{diag}\Big(\sigma'\big(\sigma^{-1}(\tilde{v}_r)\big)\Big) - \text{diag}\Big(\sigma'\big(\sigma^{-1}(\tilde{v}_r)\big)\Big)^2\bar{W}\right]\bar{U}\mathbf{x}_r dr \\
= & \text{diag}\Big(\sigma'\big(\sigma^{-1}(\tilde{v}_r)\big)\Big)\bar{U}\mathbf{x}_s \\
& - \int_0^s e^{\int_r^s \text{diag}\left(\sigma'\left(\sigma^{-1}(\tilde{v}_q)\right)\right)\bar{W}dq} \left[\frac{d}{dr}\text{diag}\Big(\sigma'\big(\sigma^{-1}(\tilde{v}_r)\big)\Big) - \text{diag}\Big(\sigma'\big(\sigma^{-1}(\tilde{v}_r)\big)\Big)^2\bar{W}\right]\bar{U}\mathbf{x}_r dr \\
= & \text{diag}\Big(\sigma'\big(\sigma^{-1}(\tilde{v}_r)\big)\Big)\bar{U}\mathbf{x}_s \\
& - \int_0^s e^{\int_r^s \text{diag}\left(\sigma'\left(\sigma^{-1}(\tilde{v}_q)\right)\right)\bar{W}dq} \\
& \quad * \left[\text{diag}\Big(\sigma''\big(\sigma^{-1}(\tilde{v}_r)\big)\Big)\text{diag}\Big(\bar{W}\tilde{v}_R + U\frac{d\mathbf{x}_r}{dr}\Big) - \text{diag}\Big(\sigma'\big(\sigma^{-1}(\tilde{v}_r)\big)\Big)^2\bar{W}\right]\bar{U}\mathbf{x}_r dr.
\end{aligned}$$

First, we show that for sufficiently large $M_\epsilon$, $\|\tilde{v}_s\|_\infty < \frac{1}{2M}$ for all $s$. Notice the activation function satisfies $|\sigma'(z) - \sigma'(0)| \leq M\sigma(z)$, we have the following matrix inequality:

$$\frac{1}{2}I < I - M\text{diag}(\text{v}) < \text{diag}\big(\sigma'\big(\sigma^{-1}(\tilde{v}_q)\big)\big) < I + M\text{diag}(\text{v}) < \frac{3}{2}I.$$

Choose $M_1$ large enough, such that

$$M\|U\|_\infty \frac{1}{M_1} + \int_0^\infty e^{\frac{1}{2}w(s-r)}\Big(M(\|\bar{W}\|_\infty \frac{1}{2M} + \|U\|_\infty \frac{K}{M_1}) + M^2\|\bar{W}\|_\infty\Big)\|U\|_\infty \frac{1}{M_1}dr < \frac{1}{2M},$$

where $w$ is the eigenvalue of $\bar{W}$ with the largest real part. By the construction of $\bar{W}$, $w < 0$.

We further show that $T := \inf\{s \big| \|v_s\|_\infty = \frac{1}{2M}\} = \infty$. Otherwise, if $T < \infty$, by the conditions $|\sigma'(z)| \leq M, |\sigma'(z) - 1| \leq M|\sigma(z)|, |\sigma''(z)| \leq M|\sigma'(z)|$, we have

$$\begin{aligned}
\|\tilde{v}_T\|_\infty \leq & M\|U\|_\infty \|x_s\|_\infty \\
& + \int_0^s e^{\frac{1}{2}w(s-r)}(M(\|\bar{W}\|_\infty \frac{1}{2M} + \|U\|_\infty \frac{K}{M_1}) + M^2\|\bar{W}\|_\infty)\|U\|_\infty \|x_t\|_\infty dr \\
\leq & M\|U\|_\infty \|x_s\|_\infty \\
& + \int_0^\infty e^{\frac{1}{2}w(s-r)}(M(\|\bar{W}\|_\infty \frac{1}{2M} + \|U\|_\infty \frac{K}{M_1}) + M^2\|\bar{W}\|_\infty)\|U\|_\infty \|x_t\|_\infty dr \\
\leq & \frac{M\|U\|_\infty + \int_0^\infty e^{\frac{1}{2}w(s-r)}(M(\|\bar{W}\|_\infty \frac{1}{2M} + \|U\|_\infty \frac{K}{M_1}) + M^2\|\bar{W}\|_\infty)\|U\|_\infty dr}{M_\epsilon} \\
< & \frac{1}{2M}.
\end{aligned}$$

One can see that the integral representation of $\tilde{v}_s$ is bounded, therefore, we have for all $M_\epsilon > M_1$, $\|\tilde{v}_s\|_\infty < \frac{1}{2M}$ for any $s$. This results in the contradiction with $\|v_T\|_\infty = \frac{1}{2M}$.

Set $C = M\|U\|_\infty + \int_0^\infty e^{\frac{1}{2}w(s-r)}(M(\|\bar{W}\|_\infty \frac{1}{2M} + \|U\|_\infty \frac{K}{M_1}) + M^2\|\bar{W}\|_\infty)\|U\|_\infty dr$, repeating the above inequalities gives that $\|v_t\|_\infty < \frac{C}{M_\epsilon}$ for any $t$, if $M_\epsilon > M_1$.

If $\frac{dx_r}{dr} = 0$ for $r > T$, we get

$$\tilde{v}_r = e^{\int_T^r \sigma'(\sigma^{-1}(\tilde{v}_s))\bar{W}ds}\tilde{v}_T, \quad r > T.$$

Then we can obtain that for $\beta_0 = -\frac{1}{2}\max_{j\in[m]} \mathrm{Re}(\lambda_j)$, where $\{\lambda_j\}_{j=1}^m$ collects all the eigenvalues of $\bar{W}$,

$$\|\tilde{v}_r\|_\infty < \frac{Ce^{-\beta_0(t-T)}}{M_\epsilon}, \quad t > T.$$

□

**Lemma A.2** *There exists $M_2$ depending on $\bar{c}, \bar{W}, \bar{U}, \epsilon, C, M, T, \beta_0$, such that for all $M_\epsilon > M_2 > M_1$,*

$$\begin{aligned}
\|(I - e^{(t-r)\bar{W}})\Big(I - \mathit{diag}\Big(\sigma'\big(\sigma^{-1}(\tilde{v}_r)\big)\Big)\Big)\|_\infty &< \frac{\epsilon}{4m\|c\|_\infty CT}, \\
\|(-\bar{W})^{-1}(I - e^{(t-r)\bar{W}})\Big(I - \mathit{diag}\Big(\sigma'\big(\sigma^{-1}(\tilde{v}_r)\big)\Big)\Big)\bar{U}\|_\infty &< \frac{\epsilon}{4m\|c\|_\infty KT}, \\
\|(I - e^{(t-r)\bar{W}})\Big(I - \mathit{diag}\Big(\sigma'\big(\sigma^{-1}(\tilde{v}_r)\big)\Big)\Big)\|_\infty &< \frac{\epsilon}{2m\|c\|_\infty C/\beta_0}.
\end{aligned}$$

*Proof.* Based on the inequality of matrix norm and properties of activation functions, we get

$$\begin{aligned}
\left\|(I - e^{(t-r)\bar{W}})\Big(I - \mathrm{diag}\Big(\sigma'\big(\sigma^{-1}(\tilde{v}_r)\big)\Big)\Big)\right\|_\infty &< \left\|\Big(I - \mathrm{diag}\Big(\sigma'\big(\sigma^{-1}(\tilde{v}_r)\big)\Big)\Big)\right\|_\infty \\
&\leq M\|\mathrm{diag}(\tilde{v}_r)\|_\infty \\
&< \frac{MC}{M_\epsilon}.
\end{aligned}$$

$$\begin{aligned}
\left\|(-\bar{W})^{-1}(I - e^{(t-r)\bar{W}})\Big(I - \mathrm{diag}\Big(\sigma'\big(\sigma^{-1}(\tilde{v}_r)\big)\Big)\Big)\bar{U}\right\|_\infty &< \left\|(-\bar{W})^{-1}\|_\infty \|\Big(I - \mathrm{diag}\Big(\sigma'\big(\sigma^{-1}(\tilde{v}_r)\big)\Big)\Big)\right\|_\infty \|\bar{U}\|_\infty \\
&\leq M\|(-\bar{W})^{-1}\|_\infty \|\mathrm{diag}(\tilde{v}_r)\|_\infty \|\bar{U}\|_\infty \\
&< \|(-\bar{W})^{-1}\|_\infty \|\bar{U}\|_\infty \frac{MC}{M_\epsilon}.
\end{aligned}$$

In order to get the desired inequalities, it is sufficient to set

$$M_2 = \max\left\{M_1, \frac{4mMC^2\|\bar{c}\|_\infty T}{\epsilon}, \frac{4mMCK\|\bar{c}\|_\infty \|\bar{W}^{-1}\|_\infty \|\bar{U}\|_\infty T}{\epsilon}, \frac{2mMC^2\|\bar{c}\|_\infty/\beta_0}{\epsilon}\right\}.$$

□

### A.3 PROOF OF THEOREM 4.1

***Proof of Universal approximation for linear functionals by nonlinear RNNs***. By the universal approximation theorem for linear functionals by linear RNNs (Theorem A.1), and the universal approximation theorem for linear RNNs by tanh RNNs (Proposition A.1), we have

$$\sup_{t\in\mathbb{R}}\|H_t-\hat{H}_t\|\equiv\sup_{t\in\mathbb{R}}\sup_{\|\mathbf{x}\|_{\mathcal{X}}\leq 1}|H_t(\mathbf{x})-\hat{H}_t(\mathbf{x})|\leq\frac{\epsilon}{2},$$
$$\sup_{t\in\mathbb{R}}\|\hat{H}_t-\widetilde{H}_t\|\equiv\sup_{t\in\mathbb{R}}\sup_{\|\mathbf{x}\|_{\mathcal{X}}\leq 1}|\hat{H}_t(\mathbf{x})-\widetilde{H}_t(\mathbf{x})|\leq\frac{\epsilon}{2},$$

which yields the desired estimate

$$\sup_{t\in\mathbb{R}}\|H_t-\widetilde{H}_t\|\equiv\sup_{t\in\mathbb{R}}\sup_{\|\mathbf{x}\|_{\mathcal{X}}\leq 1}|H_t(\mathbf{x}-\widetilde{H}_t(\mathbf{x}))|\leq\epsilon.$$

□

The assumption for activations in Equation (9) does not hold for *hardtanh* but the approximation result still holds.

***Proof of universal approximation for linear functionals by hardtanh RNNs***. For any $\epsilon>0$, suppose that the linear RNN in Theorem A.1 with an error $\frac{\epsilon}{2}$ is represented by $(\bar{c},\bar{W},\bar{U})$. By the construction of linear RNNs, we know $\bar{W}$ is a diagonal negative matrix. With proper rescaling of $\bar{c}$ and $\hat{U}$, we obtain the following bound without changing the universal approximation estimate of linear RNNs:

$$\|\bar{h}_t\|_\infty=\left\|\int_{-\infty}^{t}e^{\bar{W}(t-s)}\bar{U}x_s ds\right\|_\infty\leq\frac{1}{2\|\bar{W}\|_\infty},$$
$$\|\bar{U}x_t\|_\infty\leq\frac{1}{2}.$$

Set $\hat{c}=\bar{c},\hat{W}=\bar{W},\hat{U}=\bar{U}$, we get

$$\sup_{t\in\mathbb{R}}\sup_{\|\mathbf{x}\|_{\mathcal{X}}\leq 1}|\bar{H}_t(\mathbf{x})-\hat{H}_t(\mathbf{x})|\equiv 0\leq\frac{\epsilon}{2}.$$

According to Theorem A.1, we have

$$\sup_{t\in\mathbb{R}}\|H_t-\bar{H}_t\|\equiv\sup_{t\in\mathbb{R}}\sup_{\|\mathbf{x}\|_{\mathcal{X}}\leq 1}|H_t(\mathbf{x})-\bar{H}_t(\mathbf{x})|\leq\frac{\epsilon}{2}.$$

Combining the above inequalities gives that

$$\begin{aligned}\sup_{t\in\mathbb{R}}\|H_t-\hat{H}_t\|&\equiv\sup_{t\in\mathbb{R}}\sup_{\|\mathbf{x}\|_{\mathcal{X}}\leq 1}|H_t(\mathbf{x})-\hat{H}_t(\mathbf{x})|\\&\leq\sup_{t\in\mathbb{R}}\sup_{\|\mathbf{x}\|_{\mathcal{X}}\leq 1}|H_t(\mathbf{x})-\bar{H}_t(\mathbf{x})|+\sup_{t\in\mathbb{R}}\sup_{\|\mathbf{x}\|_{\mathcal{X}}\leq 1}|\bar{H}_t(\mathbf{x})-\hat{H}_t(\mathbf{x})|\\&\leq\epsilon.\end{aligned}$$

□

### A.4 PROOF OF THEOREM 4.2

***Proof of approximation rates for linear functionals by tanh RNNs***. By the approximation rate theorem for linear functionals by linear RNNs (Theorem A.2), and proper rescaling from the universal approximation theorem for linear RNNs by tanh RNNs (Proposition A.1), we have

$$\sup_{t\in\mathbb{R}}\|H_t-\bar{H}_t\|\equiv\sup_{t\in\mathbb{R}}\sup_{\|\mathbf{x}\|_{\mathcal{X}}\leq 1}|H_t(\mathbf{x})-\bar{H}_t(\mathbf{x})|\leq\frac{C(\alpha)\gamma d}{\beta m^\alpha},$$
$$\sup_{t\in\mathbb{R}}\|\bar{H}_t-\widetilde{H}_t\|\equiv\sup_{t\in\mathbb{R}}\sup_{\|\mathbf{x}\|_{\mathcal{X}}\leq 1}|\bar{H}_t(\mathbf{x})-\widetilde{H}_t(\mathbf{x})|\leq\epsilon\frac{C(\alpha)\gamma d}{\beta m^\alpha},$$

which gives

$$\sup_{t\in\mathbb{R}}\|H_t-\widetilde{H}_t\|\equiv\sup_{t\in\mathbb{R}}\sup_{\|\mathbf{x}\|_{\mathcal{X}}\leq 1}|H_t(\mathbf{x}-\widetilde{H}_t(\mathbf{x}))|\leq(1+\epsilon)\frac{C(\alpha)\gamma d}{\beta m^{\alpha}}.$$

Let the quantity $M_\epsilon$ in Proposition A.1 go to infinity, we get $\epsilon\to 0$. This completes the proof. □

**Remark A.2** *The conditions (9) for activations do not hold for hardtanh, but the corresponding approximation rate can be similarly achieved. This result is natural based on the universal approximation of linear functionals by hardtanh RNNs.*

### A.5 Proof of Proposition 4.1

*Proof.* We will present the proof for $p=2$ and generalize it to any $p>1$ in the following remark.

Assume the universal approximation theorem holds for homogeneous functionals by tanh RNNs. For any $\epsilon>0$, there exists a tanh RNN represented by $(c,W,U)$, such that

$$\sup_{t}\sup_{\|x\|_{\mathcal{X}}\leq 1}|H_t(\mathbf{x})-\widetilde{H}_t(\mathbf{x})|\leq\epsilon.$$

Take $\mathbf{x}$ as an input such that $H_t(\mathbf{x})\neq 0$ for $t\geq 0$. Without loss of generality, assume $H_t(\mathbf{x})>0$ for $t\geq 0$. Moreover, we set $\epsilon=\frac{1}{2}|H_t(\kappa\mathbf{x})|=\frac{1}{2}|\kappa^2H_t(\mathbf{x})|$, where the constant $\kappa$ will be given later. By the universal approximation theorem, there exists $\widetilde{H}_t$ such that for any $t$,

$$\begin{aligned}|H_t(\kappa\mathbf{x})-\widetilde{H}_t(\kappa\mathbf{x})|&\leq\epsilon,\\ |H_t(\mathbf{x})-\widetilde{H}_t(\mathbf{x})|&\leq\epsilon.\end{aligned}$$

Construct a linear RNN $\bar{H}_t$ with the same parameters $(c,W,U)$. Since $\bar{H}_t(\mathbf{x})=\widetilde{H}_t(\mathbf{x})\equiv 0$ for $t\leq 0$, there always exists a $t>0$ such that for $\delta=\frac{1}{2}$,

$$\begin{aligned}|\bar{H}_t(\kappa\mathbf{x})-\widetilde{H}_t(\kappa\mathbf{x})|&\leq\delta|\widetilde{H}_t(\kappa\mathbf{x})|,\\ |\bar{H}_t(\mathbf{x})-\widetilde{H}_t(\mathbf{x})|&\leq\delta|\widetilde{H}_t(\mathbf{x})|.\end{aligned}$$

This is feasible due to $\bar{h}_0=\tilde{h}_0$, $\frac{d\bar{h}_0}{dt}=\frac{d\tilde{h}_0}{dt}$.

Now we take $\kappa=\frac{1}{10},\delta=\frac{1}{2}$, by continuity, we have $(1-\delta)(1-\frac{1}{2}\kappa^2)>(1+\delta)\frac{1}{\kappa}(\kappa^2+\frac{1}{2}\kappa^2)$, then

$$\begin{aligned}\bar{H}_t(\mathbf{x})&\geq(1-\delta)\widetilde{H}_t(\mathbf{x})\\ &\geq(1-\delta)(H_t(\mathbf{x})-\frac{1}{2}\kappa^2H_t(\mathbf{x}))\\ &>(1+\delta)\frac{1}{\kappa}(\kappa^2H_t(\mathbf{x})+\frac{1}{2}\kappa^2H_t(\mathbf{x}))\\ &=(1+\delta)\frac{1}{\kappa}(H_t(\kappa\mathbf{x})+\frac{1}{2}\kappa^2H_t(\mathbf{x}))\\ &\geq(1+\delta)\frac{1}{\kappa}\widetilde{H}_t(\kappa\mathbf{x})\\ &\geq\frac{1}{\kappa}\bar{H}_t(\kappa\mathbf{x}).\end{aligned}$$

However, we know that the linear RNN $\bar{H}_t$ satisfies $\bar{H}_t(\mathbf{x})=\frac{1}{\kappa}\bar{H}_t(\kappa\mathbf{x})$. Note that $\bar{H}_t(\mathbf{x})>(1-\delta)(1-\frac{1}{2}\kappa^3H_t(\mathbf{x}))H_t(\mathbf{x})>0$, we get the contradiction. □

**Remark A.3** *The above proof is given for $p=2$. For general $p\geq 1$, we just need to select a corresponding $\kappa\in(0,1)$, such that $(1-\delta)(1-\frac{1}{2}\kappa^p)>(1+\delta)\frac{1}{\kappa}(\kappa^p+\frac{1}{2}\kappa^p)$. This is equivalent to solve for a $\kappa\in(0,1)$, such that*

$$1>\frac{1}{2}\kappa^p+\frac{1+\delta}{1-\delta}\frac{1}{\kappa}(\kappa^p+\frac{1}{2}\kappa^p).$$

*This is feasible because* $\lim_{\kappa\to 0}\frac{1}{2}\kappa^p+\frac{1+\delta}{1-\delta}\frac{1}{\kappa}(\kappa^p+\frac{1}{2}\kappa^p)=0$.

### A.6 PROOF OF PROPOSITION 4.2

*Proof.* Based on the proof of homogeneous functionals, we can further utilize the property of sub-homogeneous, i.e. $\kappa^p H_t(\mathbf{x}) \geq H_t(\kappa\mathbf{x})$, to construct the contradiction in a similar approach for $p = 2$. For $\kappa = \frac{1}{10}, \delta = \frac{1}{2}$, by continuity we have $(1-\delta)(1-\frac{1}{2}\kappa^2) > (1+\delta)\frac{1}{\kappa}(\kappa^2 + \frac{1}{2}\kappa^2)$, then

$$\begin{aligned}
\bar{H}_t(\mathbf{x}) &\geq (1-\delta)\widetilde{H}_t(\mathbf{x}) \\
&\geq (1-\delta)(H_t(\mathbf{x}) - 0.5\kappa^2 H_t(\mathbf{x})) \\
&> (1+\delta)\frac{1}{\kappa}(\kappa^2 H_t(\mathbf{x}) + 0.5\kappa^2 H_t(\mathbf{x})) \\
&> (1+\delta)\frac{1}{\kappa}(H_t(\kappa\mathbf{x}) + 0.5\kappa^2 H_t(\mathbf{x})) \\
&\geq (1+\delta)\frac{1}{\kappa}\widetilde{H}_t(\kappa\mathbf{x}) \\
&\geq \frac{1}{\kappa}\bar{H}_t(\kappa\mathbf{x}).
\end{aligned}$$

Therefore, we get the contradiction when the statistical functionals with $p$-sub-homogeneity can be uniformly approximated by tanh RNNs. □

### A.7 PROOF OF THEOREM 4.3

***Proof of inverse approximation theorem for linear functionals by nonlinear tanh RNNs.***
Without loss of generality, we give the proof for $\alpha = 0$ and $k = 1$. The dynamics of hidden states of tanh RNNs are described by the following differential equations:

$$\begin{aligned}
\frac{dh_t}{dt} &= \sigma(Wh_t + U\mathbf{x}_t), \\
h_0 &= 0.
\end{aligned}$$

Define $v_t = \frac{dh_t}{dt}$, then we have $v_t = \sigma(Wh_t + U\mathbf{x}_t)$. Notice that $\sigma'(z) = 1 - \sigma(z)^2$ holds for the tanh activation, the dynamics of $v_t$ satisfies

$$\begin{aligned}
\frac{dv_t}{dt} &= \sigma'(Wh_t + U\mathbf{x}_t)(Wv_t + U\frac{d\mathbf{x}_t}{dt}) \\
&= (I - \text{diag}(\sigma(Wh_t + U\mathbf{x}_t))^2)(Wv_t + U\frac{d\mathbf{x}_t}{dt}), \\
&= (I - \text{diag}(v_t)^2)(Wv_t + U\frac{d\mathbf{x}_t}{dt}), \\
v_0 &= 0.
\end{aligned}$$

We also consider a special class of constant inputs $\boldsymbol{x}_t = x_0 1_{\{t\geq 0\}}, x_0 \in \mathbb{R}^d$. Then for any $t > 0$,

$$\begin{aligned}
\frac{dv_t}{dt} &= (I - \text{diag}(v_t)^2)Wv_t, \\
v_0 &= \sigma(Wx_0).
\end{aligned}$$

The proof is based on the construction of the Lyapunov function. Define

$$V(v) = \sum_{i=1}^{m}\sum_{j=1}^{\infty}\frac{1}{2jD_{ii}}v_i^{2j},$$

where $v_i$ is the $i$-th coordinate of $v$. It is straightforward to verify that $V(v) \geq 0$, and $V(v) = 0$ if and only if $v = 0$.

Since the special Hurwitz matrix $W$ can be decomposed into $W = DN$, where $D$ is a positive diagonal matrix and $N$ is a negative definite matrix. We can derive the global stability of $v_t$ as

follows:

$$\begin{aligned}\frac{dV(v_t)}{dt} &= \sum_{i=1}^{m}\sum_{j=1}^{\infty}\frac{1}{D_{ii}}v_{i,t}^{2j-1}\frac{dv_{i,t}}{dt}\\ &= v_t^\top D^{-1}\left(\sum_{j=1}^{\infty}\text{diag}(v_t)^{2j-2}\right)\frac{dv_t}{dt}\\ &= v_t^\top D^{-1}(I-\text{diag}(v_t)^2)^{-1}(I-\text{diag}(v_t)^2)Wv_t\\ &= v_t^\top D^{-1}Wv_t\\ &= v_t^\top Nv_t \le 0.\end{aligned}$$

For any bounded input, we get $\|v_0\| \le 1-\iota$ for some $\iota > 0$. Therefore, we further have

$$\frac{dV(v_{t,m})}{dt} \le -\alpha_m V(v_{t,m}),$$

where $m$ is the hidden dimension, $\alpha_m = \frac{n_m}{\theta d_m}$ with $\theta = \sum_{j=1}^{\infty}\frac{1}{2j}(1-\iota)^{2j-2}$ as a constant related to $\iota$. Define $\beta = \inf_{m\in\mathbb{N}_+}\alpha_m$. Based on the Lyapunov exponential stability theorem, we establish the exponential stability of $v_{t,m}$, i.e. $e^{\alpha_m t}v_{t,m} = o(1)$. Since $\tilde{y}_{i,m}^{(1)}(t) = c^\top v_{t,m}$, we get $e^{\beta t}\tilde{y}_{i,m}^{(1)}(t) = o(1)$ for all $m \in \mathbb{N}_+$, which gives $e^{\beta t}y_i^{(1)}(t) = o(1)$ for $i = 1, \cdots, d$. □

**Remark A.4** *The above proof is given for the tanh activation, but it can be generalized to the sigmoid scenario using the following Lyapunov analysis:*

$$\begin{aligned}\frac{dv_t}{dt} &= diag(v_t)\big(I - diag(v_t)\big)Wv_t,\\ V(v) &= \sum_{k=1}^{m}\frac{-\ln(1-v_k)}{D_{kk}},\\ \frac{dV(v_t)}{dt} &= v_t^\top Nv_t \le 0.\end{aligned}$$

## B Details of numerical illustrations

In all numerical experiments, the training epochs are set as 40000 and the batch size is 128. The optimizers and learning rates used are Adam (0.01) and SGD (0.1), respectively. The kernel weights ($W$ and $U$) are randomly initialized using the TensorFlow orthogonal initializer. As there is no significant difference in terms of loss curve trends, the results reported in the main text are all trained by Adam. Every training is repeated independently for 10 times.

In the exponential-sum example, we set the number of exponential components as $b = 8$ and the path length as 64. In the Airy function example, the constant $t_0$ is 3.0 while the scale $s_0$ is 2.5. The path length is also 64.

The RNNs used in numerical experiments to approximate target functionals are trained over a varied range of hidden dimensions, that is, $[2, 2^2, 2^3, \ldots, 2^7]$.

## C Details of Lyapunov analysis

In this section, we present the numerical evidence for recurrent matrices as special Hurwitz and general Hurwitz matrices. The special Hurwitz matrices form a subclass of Hurwitz matrices with the decomposition $W = DN$, where $D$ is a positive diagonal matrix and $N$ is a negative definite matrix. The general Hurwitz matrix can be decomposed into the product of a *symmetric* positive-definite matrix $S$ and a negative-definite matrix $N$ (Duan & Patton, 1998).

### C.1 Special Hurwitz: $W = DN$

In Figure 4 and Figure 5, we show the trajectories of hidden states and corresponding Lyapunov function values for the input dimension $d = 2$ and $d = 16$. The inputs here are constant signals with

random values $\boldsymbol{x}_t = x_0 1_{\{t \geq 0\}}$. The Lyapunov function is defined as

$$V(v) = \sum_{i=1}^{m} \sum_{j=1}^{\infty} \frac{1}{2jD_{ii}} v_i^{2j}.$$

The monotonically-decreasing function values show the stability of hidden dynamics.

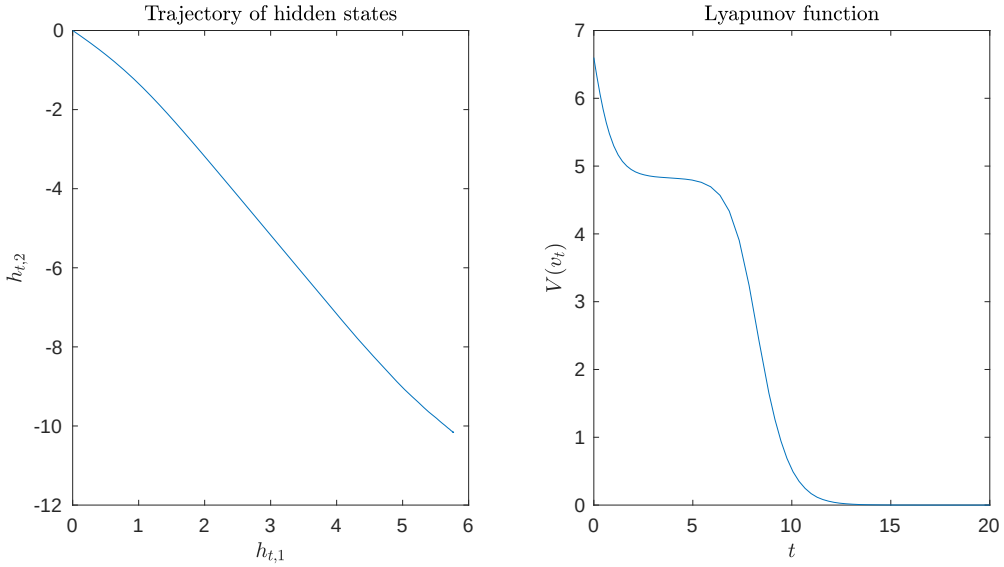



Figure 4: Trajectories of tanh RNNs with a special Hurwitz recurrent kernel $W = \begin{bmatrix} -2 & -1 \\ -0.5 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0.5 \end{bmatrix} \begin{bmatrix} -2 & -1 \\ -1 & -2 \end{bmatrix}$ and the corresponding Lyapunov function. The input dimension is $d = 2$.

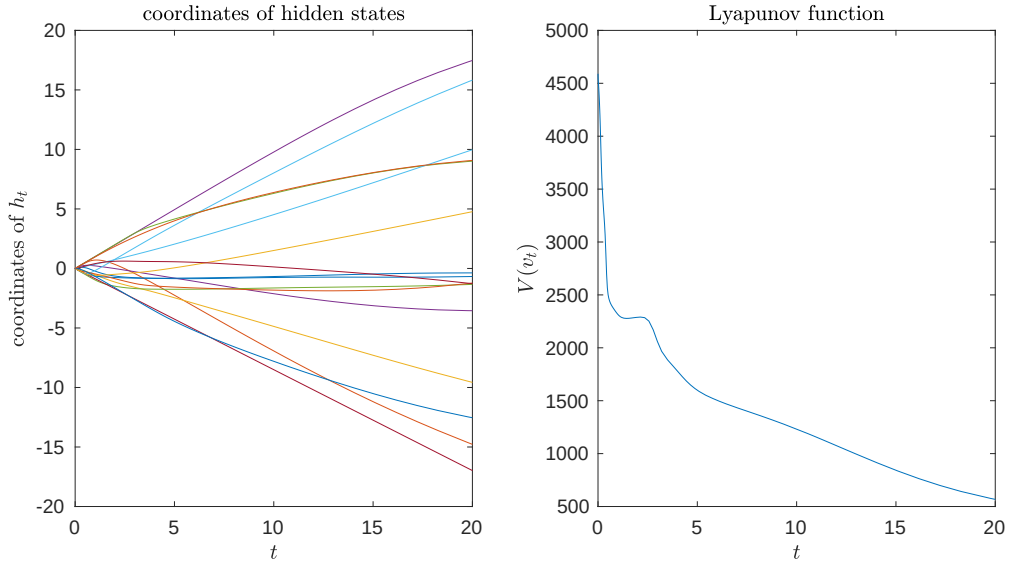



Figure 5: Trajectories of tanh RNNs with a randomly generated special Hurwitz recurrent kernel $W$ and the corresponding Lyapunov function. The input dimension is $d = 16$.

### C.2 GENERAL HURWITZ: $W = SN$

In Figure 6 and Figure 7, we show the trajectories of hidden states and corresponding quasi-Lyapunov function values for the input dimension $d = 2$ and $d = 16$. The inputs here are also constant signals with random values $\boldsymbol{x}_t = x_0 1_{\{t \geq 0\}}$. We construct the quasi-Lyapunov function by taking the diagonal components of $S$ as the matrix $D$. That is,

$$V^{\text{quasi}}(v) = \sum_{i=1}^{m} \sum_{j=1}^{\infty} \frac{1}{2jS_{ii}} v_i^{2j}.$$

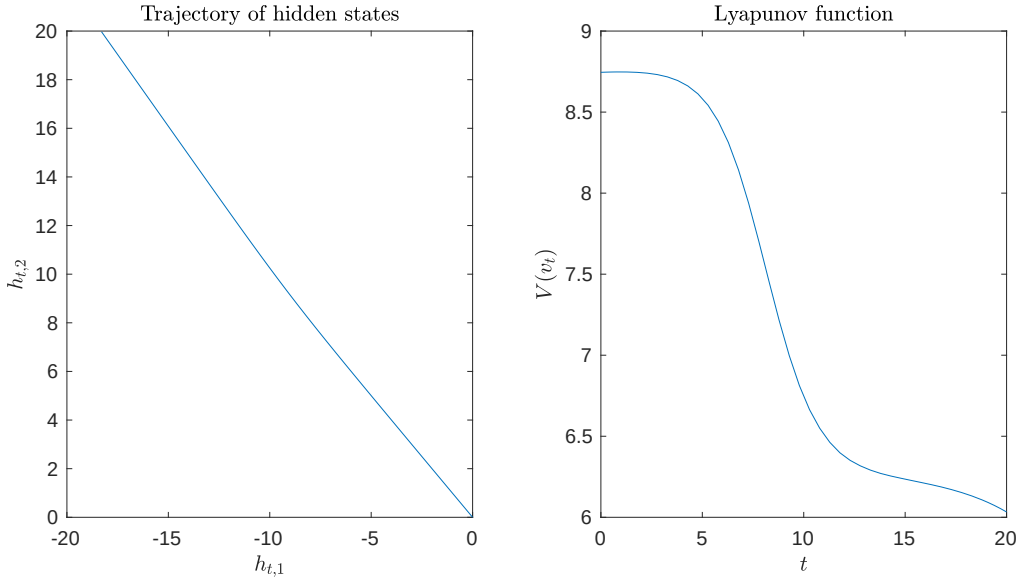



Figure 6: Trajectories of tanh RNNs with a Hurwitz recurrent kernel $W = \begin{bmatrix} -2.6 & -2.2 \\ -1.7 & -1.6 \end{bmatrix} = \begin{bmatrix} 1 & 0.6 \\ 0.6 & 0.5 \end{bmatrix} \begin{bmatrix} -2 & -1 \\ -1 & -2 \end{bmatrix}$ and the corresponding quasi-Lyapunov function. The input dimension is $d = 2$.

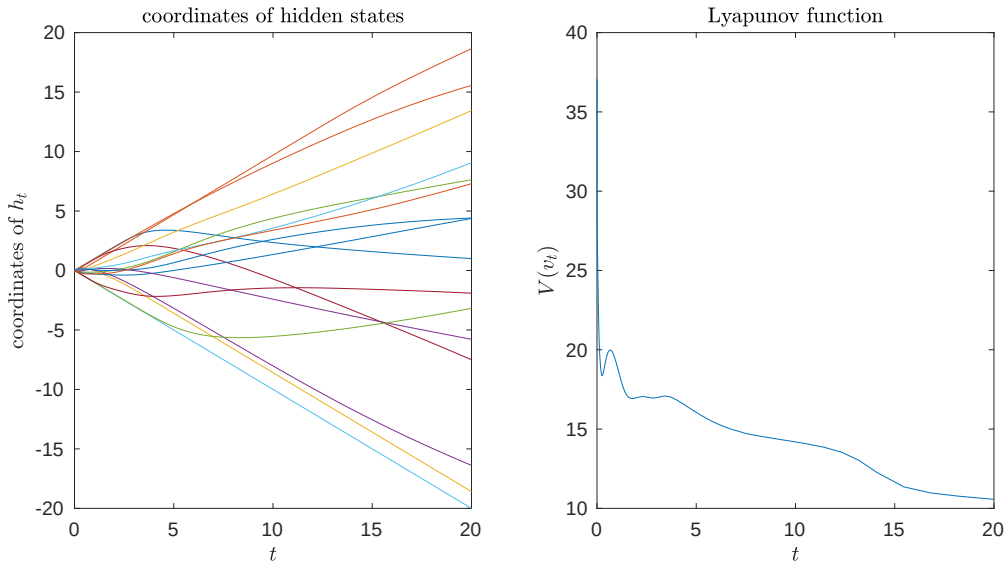



Figure 7: Trajectories of tanh RNNs with a randomly generated Hurwitz recurrent kernel $W$ and the corresponding quasi-Lyapunov function. The input dimension is $d = 16$.