On the approximation properties of recurrent encoder-decoder architectures

Anonymous authors
Paper under double-blind review

Abstract

Encoder-decoder architectures have recently gained popularity in sequence to sequence modelling, featuring in state-of-the-art models such as transformers. However, a mathematical understanding of their working principles remains limited. In this paper, we study the approximation properties of recurrent encoder-decoder architectures. Prior work established theoretical results for classical RNNs in the linear setting, where approximation capabilities can be related to smoothness and memory. Here, we find that the encoder and decoder together form a particular "temporal product structure" which determines the approximation efficiency. Moreover, the encoder-decoder architecture generalises RNNs with the capability to learn time-inhomogeneous relationships. Our results provide the theoretical understanding of approximation properties of the recurrent encoder-decoder architecture, which characterises, in the considered setting, the types of temporal relationships that can be efficiently learned.

1 Introduction

Encoder-decoder is an increasingly popular architecture for sequence to sequence modelling problems (Sutskever et al., 2014; Chiu et al., 2018; Venugopalan et al., 2015). The core of this architecture is to first encode the input sequence into a vector using the encoder, and then map the vector into the output sequence through the decoder. In particular, such architecture forms the main component in the transformer network model (Vaswani et al., 2017), which has become a powerful method for modelling sequence to sequence relationships (Parmar et al., 2018; Beltagy et al., 2020; Li et al., 2019).

The encoder-decoder family of structures differ significantly from direct application of classical recurrent neural networks (Rumelhart et al., 1986) and their generalisations (Hochreiter & Schmidhuber, 1997; Cho et al., 2014) for processing sequences. However, both architectures can be considered as modelling mappings between sequences, albeit with different underlying structures. Hence, a natural but unresolved question is: how are these approaches fundamentally different? Answering this question is not only of theoretical importance, but is also of practical interest. Currently, architectural selection for different time-series modelling tasks is predominantly empirical. Thus, it is desirable to develop a concrete mathematical framework to understand the key differences between separate architectures in order to guide practitioners in a principled way.

In this paper, we investigate the approximation properties of encoder-decoder architectures. Where approximation is one of the most basic problems for supervised learning, it considers to what extent can the model fit to a target. In particular, we prove a general approximation result in the linear setting, that characterises the types of temporal input-output relationships can be efficiently approximated by encoder-decoder architectures.

These results reveal that such architectures essentially generalise the RNN by lifting the requirement of time-homogeneity in the target relationships. Hence, it can be used to tackle a broader class of sequence to sequence problems. Furthermore, of particular interest is the identification of a "temporal product structure" — a precise property of the target temporal relationship that enables further reduction in required trainable parameters for

approximation. This highlights another intrinsic difference between these architectures and classical recurrent neural networks.

Our main contributions can be summarised as follows

1. We prove a universal approximation result for encoder-decoder architectures in the linear setting, including approximation rates.
2. We show that in the considered setting, encoder-decoder generalises RNN and can learn time-inhomogeneous relationships, which further adapt to additional product structures in the target relationship. This answers precisely how encoder-decoder are different from classical RNNs, at least in the considered setting.

Organisation. In section 2, we review the related work on encoder-decoder architectures and general approximation theory results of sequence modelling. Then, we formulate our approximation problem in section 3. Our main results, their consequences and numerical illustrations are presented in section 4. All the proofs are included in the appendix.

Notation. For consistency, we adhere to the following notation. Boldfaced letters are reserved for sequences or paths, which can be understood as functions of time. Lower case letters can mean vectors or scalars. Matrices are denoted by capital letters.

2 Related work

We will first review some previous works on sequence to sequence modelling. Encoder-decoder architecture first appeared in Kalchbrenner & Blunsom (2013), where they map the input sequence into a vector using convolutional neural networks, and then using a recurrent structure to map the vector to the output sequence. With the flexibility of manipulating the underlying structure of the encoder and decoder, numerous models based on this architecture have come out thereafter. For instance, Cho et al. (2014) use a gated RNN as encoder and decoder, while in their later work (Cho et al., 2015) they proposed a CNN based decoder. In Sutskever et al. (2014), they proposed a deep LSTM for both the encoder and decoder. Bahdanau et al. (2015) first introduces the attention mechanism, which was further developed in the well-known transformer networks (Vaswani et al., 2017). However, most of the researches on encoder-decoder architectures focus on applications. A theoretical understanding of the architecture is helpful for its further improvement and development.

From the theoretical point of view, Ye & Sung (2019) studied several theoretical properties of CNN encoder-decoders, including expressiveness, generalisation capability and optimisation landscape. Of particular relevance to the current work is expressiveness, which considers the relationships that can be generated from the model. However, this is not approximation properties. Yun et al. (2020) proved the universal approximation property of transformers for particular classes of functions, for example, permutation equivariant functions. But they did not consider the actual dynamical properties of the target relationship that affects approximation. Dynamical proprieties such as memory, smoothness and low rank structures are essential because they can precisely characterise different temporal relationships and affect the approximation capabilities of the models. Assuming the target is generated from a hidden dynamic system is one approach that is wildly applied (Maass et al., 2007; Schäfer & Zimmermann, 2007; Doya, 1993; Funahashi & Nakamura, 1993). In contrast, a functional analysis approach is introduced recently, where the targets are generated from functionals satisfying specific properties such as linearity, regularity, time-homogeneity (Li et al., 2021). In Li et al. (2021), the approximation properties of RNN models are studied, and the results show that the approximation efficiency is related to the memory structure. In Jiang et al. (2021), similar formulations are applied to study convolutional based architectures, where the results show that targets with certain spectrum regularity can be well approximated by dilated CNNs. We will follow the functional analysis formulation, where the results reveal that the encoder-decoders have a special temporal product structure which is intrinsically different from other architectures.

3 Problem formulation

In this section, mathematical formulations are given for the supervised learning problem of temporal modelling. That is, we precisely define the input space, output space, concept space and hypothesis space, respectively.

Functional formulation of temporal modelling. Firstly, we define the input and output space precisely.

A temporal sequence can be considered as a function of time t. The input space is defined by

$$\mathcal{X} = C_0((-\infty, 0], \mathbb{R}^d). \tag{1}$$

That is, the space of continuous functions from $(-\infty, 0]$ to $\mathbb{R}^d$ vanishing at infinity, where $d \in \mathbb{N}_+$ is the dimension. Denote the element in $\mathcal{X}$ by $\boldsymbol{x} := \{x_t \in \mathbb{R}^d : t \in (-\infty, 0]\}$, we equip $\mathcal{X}$ with the supremum norm $\|\boldsymbol{x}\|_{\mathcal{X}} := \sup_{t \leq 0} \|x_t\|_\infty$.

For the outputs, we take the space of bounded continuous functions from $[0, \infty)$ to $\mathbb{R}$:

$$\mathcal{Y} = C_b([0, \infty), \mathbb{R}). \tag{2}$$

We consider real value outputs since we can handle each dimension individually for vector-valued outputs.

The mapping between the two sequences can be formulated as a sequence of functionals where

$$y_t = H_t(\boldsymbol{x}), \quad t \geq 0. \tag{3}$$

The output y_t at time step t depends on the input sequence $\boldsymbol{x}$. The ground truth relation between input and output is formulated by the sequence of functionals $\boldsymbol{H} = \{H_t\}_{t \geq 0}$.

We provide an example to illustrate the above formulation. Given an input $\boldsymbol{x}$, the output $\boldsymbol{y}$ is the smoothed version of $\boldsymbol{x}$, resulting from convolve $\boldsymbol{x}$ with the Gaussian kernel $g(s) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{s^2}{2})$. This relation is mathematically formulated as

$$y_t = H_t(\boldsymbol{x}) = \int_0^\infty g(t+s) \boldsymbol{x}_{-s} ds. \tag{4}$$

For a supervised learning problem, our goal is to use a model to learn the target relationship $\boldsymbol{H}$. Next we will define the model.

The RNN encoder-decoder model. Among all different variants of the encoder-decoder architectures, the RNN encoder-decoder introduced in Cho et al. (2014) can be considered as the most simple and representative model, where the encoder and decoder are both traditional RNNs. We will study this particular model as we try to eliminate other factors and only focus on the encoder-decoder architecture itself.

Under our setting, the simplified model of Cho et al. (2014) with classical RNN as encoder and decoder is formulated as

$$\begin{aligned} h_t &= \sigma_E(W_E h_{t-1} + U_E x_t + b_E), \qquad v = h_\tau, \\ g_t &= \sigma_D(W_D h_{t-1} + U_D x_t + b_D), \quad g_0 = v, \\ o_t &= \sigma_O(W_O g_t + b_O), \end{aligned} \tag{5}$$

where h_t and g_t are *hidden states* of the encoder and decoder, respectively, c is the coding vector and $o_t \in \mathbb{R}$ denotes the model prediction. First, the encoder reads the entire input $\boldsymbol{x}$, and it encodes (summarises) the input sequence into a fixed size coding vector v, which is the last hidden state of the encoder. Next, the coding vector is passed into the decoder as its initial state, and the decoder will produce an output at each output time step.

Note that the encoder has a terminating time step, and the decoder has a starting time step. This is the reason why we take the input and output space to be semi-infinite sequences. We set 0 to be both the terminating and starting time step.

We will take a linear and continuous-time idealisation for our investigation, which is formulated as

$$\begin{aligned} \frac{d}{ds}h_s &= Wh_s + Ux_s, \qquad v = Qh_0, \quad s \le 0 \\ \frac{d}{dt}g_t &= Vg_t, \qquad\qquad g_0 = Pv, \\ o_t &= c^\top g_t, \qquad\qquad t \ge 0, \end{aligned} \tag{6}$$

where $W \in \mathbb{R}^{m_E \times m_E}$, $U \in \mathbb{R}^{m_E \times d}$, $Q \in \mathbb{R}^{N \times m_E}$, $V \in \mathbb{R}^{m_D \times m_D}$, $P \in \mathbb{R}^{m_D \times N}$ and $c \in \mathbb{R}^{m_D}$. Here m_E, m_D denotes the width of the encoder RNN and decoder RNN, respectively. N is the dimension of the coding vector v.

Since our goal is to investigate approximation problems over large time horizons, we are supposed to consider the stable encoder-decoder RNNs, where

$$W \in \mathcal{W}_{m_E} := \{W \in \mathbb{R}^{m_E \times m_E} : \text{eigenvalues of } W \text{ have negative real parts}\}, \tag{7}$$

$$V \in \mathcal{V}_{m_D} := \{V \in \mathbb{R}^{m_D \times m_D} : \text{eigenvalues of } V \text{ have negative real parts}\}. \tag{8}$$

The hypothesis space of RNN encoder-decoder models with arbitrary width and coding size is defined as

$$\begin{aligned} \widehat{\mathcal{H}} &:= \bigcup_{m_E, m_D, N \in \mathbb{N}_+} \widehat{\mathcal{H}}_{m_E, m_D, N}, \\ \widehat{\mathcal{H}}_{m_E, m_D, N} &:= \Big\{ \widehat{\boldsymbol{H}} : \widehat{H}_t(\boldsymbol{x}) = c^\top e^{Vt} P \int_0^\infty Q e^{Ws} U x_{-s} ds, \text{ with} \\ &\qquad (W, U, Q, V, P, c) \in \mathcal{W}_{m_E} \times \mathbb{R}^{m_E \times d} \times \mathbb{R}^{N \times m_E} \times \mathcal{V}_{m_D} \times \mathbb{R}^{m_D \times N} \times \mathbb{R}^{m_D} \Big\}. \end{aligned} \tag{9}$$

Here, the model widths m_E, m_D and the coding vector dimension N together control the capacity/complexity of the hypothesis space.

Due to the mathematical form (9), not all functionals can be represented by the RNN encoder-decoder. To achieve a good approximation, the target functionals must possess certain structures. We introduce the following definitions to clarify these structures of functionals.

Definition 3.1. *Let $\boldsymbol{H} = \{H_t\}_{t \ge 0}$ be a sequence of functionals.*

1. *For any $t \in \mathbb{R}$, the functional H_t is* linear *and* continuous *if for any $\lambda_1, \lambda_2 \in \mathbb{R}$ and $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$, we have $H_t(\lambda_1 \boldsymbol{x}_1 + \lambda_2 \boldsymbol{x}_2) = \lambda_1 H_t(\boldsymbol{x}_1) + \lambda_2 H_t(\boldsymbol{x}_2)$, and $\|H_t\| := \sup_{\boldsymbol{x} \in \mathcal{X}, \|\boldsymbol{x}\|_{\mathcal{X}} \le 1} |H_t(\boldsymbol{x})| < \infty$, where $\|H_t\|$ denotes the induced functional norm.*

2. *For any $t \in \mathbb{R}$, the functional H_t is* regular *if for any sequence $\{\boldsymbol{x}^{(n)}\}_{n=1}^\infty \subset \mathcal{X}$ such that $\lim_{n \to \infty} x_s^{(n)} = 0$ for almost every $s \le 0$ (Lebesgue measure), we have $\lim_{n \to \infty} H_t(\boldsymbol{x}^{(n)}) = 0$.*

For a sequence of functionals $\boldsymbol{H}$, we define its norm by $\|\boldsymbol{H}\| := \int_0^\infty \|H_t\| dt$.

Remark 3.1. *The definitions of linear and continuous functionals are standard. One can view regular functionals as those not determined by inputs on arbitrarily small time intervals, e.g. an infinitely thin spike (i.e. δ-functions).*

Given the above definitions, we immediately have the following observation. The proof is found in Appendix A.

Proposition 3.1. *Let $\widehat{\boldsymbol{H}} \in \widehat{\mathcal{H}}$ be a sequence of functionals in the RNN encoder-decoder hypothesis space (see (9)). Then for any $t \ge 0$, $\widehat{H}_t \in \widehat{\boldsymbol{H}}$ is a linear, continuous and regular functional. Furthermore, $\|\widehat{H}_t\|$ decays exponentially as a function of t.*

This proposition characterises properties the encoder-decoder hypothesis space possesses. In particular, it is different from the RNN hypothesis space discussed in Li et al. (2021) since the

encoder-decoder is not necessarily *time-homogeneous*. Where a sequence of functionals $\widehat{\boldsymbol{H}}$ is time-homogeneous if for any $t, \tau \geq 0$, $H_t(\boldsymbol{x}) = H_{t+\tau}(\boldsymbol{x}(\tau))$, with $x(\tau)_s = x_{s-\tau}$ for all $s \in \mathbb{R}$. Our primary concern is how RNN encoder-decoder approximate general target relationships without assuming time-homogeneity. We will talk about our main approximation results in this direction.

Relation with RNNs. Here, we emphasise the differences between encoder-decoder hypothesis space and the RNN hypothesis space discussed in Li et al. (2021), where $\widehat{H}_t^{(\mathrm{RNN})}(\boldsymbol{x}) = \int_0^\infty c^\top e^{W(t+s)} U x_{-s} ds$. A key difference is that the encoder-decoder has a structure involving two temporal parameters t and s, while the RNN has only one depending on $t+s$, due to time-homogeneity. Owing to this difference and the fact that $\widehat{\boldsymbol{H}}^{(\mathrm{RNN})} \in \widehat{\mathcal{H}}$, the encoder-decoder hypothesis space (9) contains the RNN hypothesis space and is more general, with the extra capabilities to learn time-inhomogeneous relationships. Furthermore, e^{Vt}, e^{Ws} adapts to a product structure, which is an intrinsic difference between encoder-decoder and other architectures. We will discuss this in detail when we present our main results.

4 Approximation results

One of the most fundamental problems for supervised learning is the approximation problem. That is, given the hypothesis space (model) and concept space (target), to what extent can the model fit the target? Or what is the capacity/complexity of the hypothesis space? In this section, we present our main approximation results for encoder-decoder. Particular attention is paid to discuss how does the encoder-encoder differ from classical RNNs in terms of approximation.

In general, there are two levels of approximation problems that can be discussed. The first is known as the universal approximation, or density results, which focuses on the existence of models to achieve (any) given approximation accuracy to some targets satisfying appropriate conditions. The second is the approximation rate, which aims to quantitatively characterise the approximation accuracy concerning the capacity/complexity of the hypothesis space (e.g. the number of trainable parameters). In this section, both of them are developed for RNN encoder-decoder. In addition, the estimate of approximation rates can be refined given some common structures of targets.

4.1 Universal approximation

We first present the most basic density result, which says that any linear, continuous, and regular temporal relationship can be approximated by RNN encoder-decoder up to arbitrary accuracy. The proof is found in Appendix B.

Theorem 4.1. *Let $\boldsymbol{H}$ be a sequence of linear, continuous, and regular functionals defined on $\mathcal{X}$, and satisfy $\|\boldsymbol{H}\| < \infty$. Then for any $\epsilon > 0$, there exists $\widehat{\boldsymbol{H}} \in \widehat{\mathcal{H}}$ such that*

$$\|\boldsymbol{H} - \widehat{\boldsymbol{H}}\| \equiv \int_0^\infty \|H_t - \widehat{H}_t\| dt < \epsilon. \tag{10}$$

Here, we highlight two important observations while deriving Theorem 4.1. First, one can show that each sequence of functionals $\boldsymbol{H} \in \mathcal{H}$ can be associated with a unique two-parameter "representation" $\rho(t,s)$, such that $H_t(\boldsymbol{x}) = \displaystyle\int_0^\infty x_{-s}^\top \rho(t,s) ds$. Recall the model in the hypothesis space has the form $\widehat{H}_t(\boldsymbol{x}) = \displaystyle\int_0^\infty x_{-s}^\top \hat{\rho}(t,s) ds$, where $\hat{\rho}(t,s) := [c^\top e^{Vt} P Q e^{Ws} U]^\top$ denotes the representation for the model. The functional approximation problems reduce to a function approximation problems on this representation, in the sense that $\|\boldsymbol{H} - \widehat{\boldsymbol{H}}\| \leq \|\rho - \hat{\rho}\|$. It turns out that the structure of ρ directly affects the rate of approximation and can give rise to intrinsic properties that further lead to parameter reduction, as we will discuss in section4.3.

Here, we again emphasise the difference between our work with Li et al. (2021). In their work the target relationships are assumed to be time-homogeneous and have the representation $H_t(\boldsymbol{x}) = \int_0^\infty \rho(t+s)x_{-s}ds$, which only depends on $t+s$. However, our setting does not assume time-homogeneity, thus have a more general representation where ρ depends on the two temporal directions t and s simultaneously.

4.2 General approximation rates

While the density result (Theorem 4.1) ensures the universal approximation property of the RNN encoder-decoder, it does not identify targets that can be efficiently approximated. To achieve this, we focus on approximation rates next. We characterise the temporal structure of a target relationship by looking at its responses of "constant" input signals.

Here we consider the approximation rates for a model with "large size" coding vector, where the dimension of the $N \geq \bar{m} := \min\{m_E, m_D\}$. This is the scenario where we fix the width but want to take an oversized coding vector. The proof is found in Appendix C.

Theorem 4.2. *Let $\boldsymbol{H}$ be a sequence of linear, continuous, and regular functionals defined on $\mathcal{X}$, and satisfy $\|\boldsymbol{H}\| < \infty$. Consider the output of piece-wise constant signals $y_i^{\mathrm{c}}(t,s) = H_t(e_i \mathbf{1}_{(-\infty,-s]})$, $t,s \geq 0$, $i = 1,2\ldots,d$, where $\{e_i\}_{i=1}^d$ denotes the standard basis of $\mathbb{R}^d$. Assume that there exist $\alpha \in \mathbb{N}_+$, $\beta > 0$, such that for any $i = 1,2\ldots,d$,*

$$y_i^{\mathrm{c}} \in C^{(\alpha+1)}([0,\infty)^2), \tag{11}$$

$$e^{\beta(t+s)} \frac{\partial^{k+l}}{\partial t^k \partial s^l} y_i^{\mathrm{c}}(t,s) = o(1) \text{ as } \|(t,s)\| \to \infty, \quad (k,l) \in \mathbb{N} \times \mathbb{N}_+,\ k+l \leq \alpha+1. \tag{12}$$

Then, for any $m_E, m_D, N \in \mathbb{N}_+$ with $N \geq \bar{m}$, there exists $\widehat{\boldsymbol{H}} \in \widehat{\mathcal{H}}_{m_E,m_D,N}$ such that

$$\|\boldsymbol{H} - \widehat{\boldsymbol{H}}\| \leq \frac{C(\alpha)\gamma d}{\beta^2}\left(\frac{1}{m_E^\alpha} + \frac{1}{m_D^\alpha}\right), \tag{13}$$

where $C(\alpha), \gamma > 0$ are both universal constants with dependence only on α and (α,β), respectively, and $\gamma := \max_{i\in\mathbb{N}_+,\, i\leq d}\ \max_{k,l\in\mathbb{N},\, k+l\leq\alpha+1}\ \sup_{t,s\geq 0} \beta^{-(k+l)} e^{\beta(t+s)} \left|\frac{\partial^{k+l}}{\partial t^k \partial s^l} y_i^{\mathrm{c}}(t,s)\right| < \infty$. Here, the number of trainable parameters is $dN(m_E+m_D)$, with $N \geq \bar{m} := \min\{m_E, m_D\}$.

First, note that the error bound does not depend on the coding vector size N, as long as $N \geq \bar{m}$. This is because further increasing N beyond $\bar{m}$ only increases number of trainable parameters but does not increase its model capacity (See Remark C.1). Only the width m_E, m_D affect the approximation capabilities.

Next, we focus on which class of target relationships can be well approximated. Here, α characterises the smoothness of $\boldsymbol{H}$, and β characterises the temporal decay rates of the output of a constant signal under $\boldsymbol{H}$. This is a notion of memory in the target relationship. By observing the error bound (13), we see that a target functional can be efficiently approximated by the encoder-decoder if it is smooth (large α), and has fast decaying memory (large β).

The smoothness and decay rate characterisation also appear in the RNN results (Li et al., 2021), where the upper bound is $\frac{C(\alpha)\gamma d}{\beta m^\alpha}$. However, our results for encoder-decoder show extra structures, where the bounds involve two (instead of one) temporal parameters together with smoothness and decay requirements in each. The two-parameter time dependence allows the encoder-decoder to approximate time-inhomogeneous relationships, thus generalising the RNN. This two-parameter structure further leads to adaptation to a specific low-rank type of target relationships, resulting in finer approximation rates as we discuss next.

4.3 Approximation rates with temporal product structure

Motivation of the temporal product structure. In contrast with Theorem 4.2, we next consider the models with $N < \bar{m} := \min\{m_E, m_D\}$. In this situation, the model has

fewer parameters, we characterise the target relationships by further exploiting the structure of the two-parameters representation $\rho(t,s)$, which leads to a finer approximation rate by considering m_E, m_D, N together.

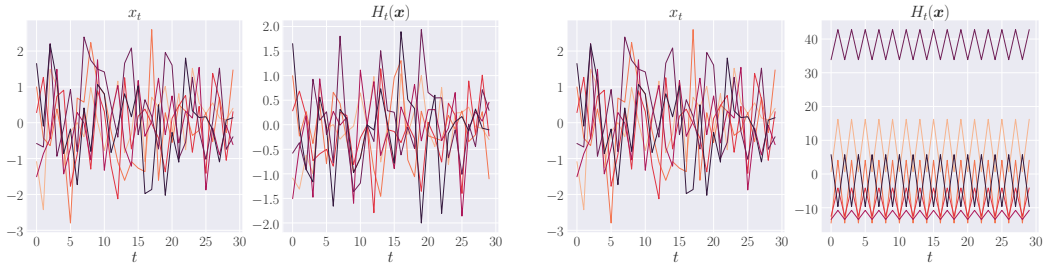
We first motivate how the "temporal product structure" arises, and how it relates to the approximation. Detailed discussions are found in the Appendix D. For illustration purposes, we take input dimension $d = 1$. Recall $Q \in \mathbb{R}^{N \times m_E}$, $P \in \mathbb{R}^{m_D \times N}$, and thus the representation $\hat{\rho}$ of the encoder-decoder functional can be written as

$$\begin{aligned}\hat{\rho}(t,s) = c^\top e^{Vt} P \cdot Q e^{Ws} u &= \sum_{n=1}^{N} \left(\sum_{i,j=1}^{m_D} c_i P_{jn} \left[e^{Vt}\right]_{ij} \right) \left(\sum_{i,j=1}^{m_D} u_i Q_{jn} \left[e^{Ws}\right]_{ij} \right) \\ &= \sum_{n=1}^{N} \hat{\varphi}_n(t) \hat{\phi}_n(s). \end{aligned} \tag{14}$$

This is a tensor product structure over the t, s time domain (determined by the encoder $\{\hat{\phi}_n\}$ and decoder $\{\hat{\varphi}_n\}$ successively). We call this the *temporal product structure*, and we will show that this structure significantly affects approximation rates. When $\{\hat{\phi}_n\}$ and $\{\hat{\varphi}_n\}$ are taken to be the "bases" along s, t direction, respectively, N is considered as the *rank* of the temporal product. We also define N as the rank of the model, which is understood as the maximum rank of temporal products the model can represent.

The rank concept of temporal relationships. Recall that the number of trainable parameters is $dN(m_E + m_D)$. Hence, a low rank model can achieve fewer trainable parameters. While considering what relationships can be well approximated by a low rank model, a natural conjecture would be "low rank" targets.

However, what does it mean for a temporal relationship to be "low rank"? We know that in linear algebra, a low rank operator means its range space is low dimensional. This idea also applies to temporal relationships. For a 'low rank" temporal relationship, the output sequence will be more regular, which means the output sequences (understood as functions) are in a low dimensional function space. We provide an intuitive numerical illustration for better understanding.



(a) high rank relationship

(b) low rank relationship

Figure 1: We construct a high rank target and a low rank target from the temporal product. For both (1a) and (1b), we plot the input x_t with its outputs $H_t(\boldsymbol{x})$. Detailed settings are found in Appendix E.1.

Figure 1 shows the outputs of a high rank (left) and a low rank (right) target relationship on the same set of random input sequences. Different colours refer to different instances of the inputs. In the first case (high rank), the temporal structure of the outputs is very complex and depends sensitively on the inputs. In the second case (low rank), the output sequence are much more regular, and only macroscopic structures (such as the scale or offset) depend on the input sequence.

Remark 4.1. *In the study of approximation theories for temporal sequences, prior work also related a notion of rank to the approximation property under the dilated convolutional structure (Jiang et al., 2021). We emphasise here that the notion of rank considered in the case of encoder-decoder is very different from that in Jiang et al. (2021), which mainly concerns*

the tensorisation of a one-parameter time sequence according to the width of convolution filters.

POD as an analogy of SVD Now, we need a mathematical way to characterise low rank and high rank temporal relationships. We will introduce the concepts informally, and rigorous definitions can be found in the Appendix D. For a matrix, we can assess its rank by looking at its singular value decomposition (SVD). This can be extended to the temporal relationships using proper orthogonal decomposition (POD). The idea is that we can decompose the function ρ into the following form

$$\rho(t,s) = \sum_{n=1}^{N_0} \sigma_n \varphi_n(t)\phi_n(s), \tag{15}$$

where $N_0 \leq \infty$, $\{\varphi_n\}$ and $\{\phi_n\}$ are orthonormal bases, and $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$ are the singular values. This procedure can be considered as applying SVD to an infinite dimensional space (when $N_0 = \infty$). An analogy of Eckart–Young theorem (Eckart & Young, 1936), which characterises best low-rank approximation errors, also exists for POD. It roughly states that

$$\inf_{\mathrm{rank}(\hat{\rho})=N} \|\rho - \hat{\rho}\|_{L^2}^2 = \sum_{n=N+1}^{\infty} \sigma_n^2. \tag{16}$$

That is, any target ρ has a rank N best approximation, and its error equals to the tail sum of its squared singular values. In other words, a target with fast decaying σ_n (low effective rank) has smaller errors. This forms the basis of our next result, which shows that if the target relationship possesses an effective low rank structure in terms of the decay of singular values, then we can achieve a more efficient approximation using encoder-decoder structures by limiting the size of the coding vector. Detailed definition for $\{\sigma_n\}$ and the proof are found in Appendix D.

Theorem 4.3. *Assume the same conditions as in Theorem 4.2. Then for any $m_E, m_D, N \in \mathbb{N}_+$ with $N \leq \bar{m}$, there exists $\widehat{\boldsymbol{H}} \in \widehat{\mathcal{H}}_{m_E,m_D,N}$ such that*

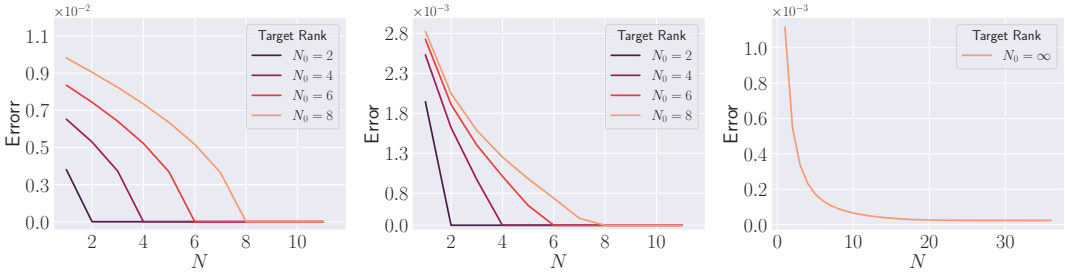
$$\|\boldsymbol{H} - \widehat{\boldsymbol{H}}\| \lesssim \frac{C(\alpha)\gamma d}{\beta^2} \Bigg\{ \left(1 + \sqrt{\bar{m} - N}\right) \cdot \left(\frac{1}{m_E^\alpha} + \frac{1}{m_D^\alpha}\right) + \left(\sum_{n=N+1}^{\bar{m}} \sigma_n^2\right)^{1/2} + \left(\sum_{n=N+1}^{\bar{m}} \sigma_n\right)^{1/2} \cdot \left(\frac{1}{m_E^{\alpha/2}} + \frac{1}{m_D^{\alpha/2}}\right) \Bigg\}, \tag{17}$$

where $\lesssim$ hides universal positive constants, and $\bar{m} = \min\{m_E, m_D\}$. Here, the number of trainable parameters is $dN(m_E + m_D)$ with $N \leq \bar{m}$.

This is a finer approximation rate compared to Theorem 4.2, where both the widths (m_E, m_D) and the coding vector size N affect the model capacity. Besides smoothness and memory decay, we have the additional rank structure of the target relationship, which is characterised by its singular values $\{\sigma_n\}$. We again look at the class of functionals that can be well approximated. Smoothness α and decay rate β is same as Theorem 4.2. We consider the rank structure $\{\sigma_n\}$, where we observe that the error bound is small if $\{\sigma_n\}$ have small tail sum $\sum_{n=N+1}^{\bar{m}} \sigma_n^2$. This means that a target with fast decaying $\{\sigma_n\}$ or low "effective rank" can be well approximated by the RNN encoder-decoder with fewer parameters. Due to the Eckart–Young-like low rank approximation form, we can choose a proper N based on the decay rate of singular values. This balances the number of parameters and the approximation error.

Here, we emphasise the temporal product is an *intrinsic* structure arising from the encoder-decoder architecture. Recall the key structure of encoder-decoder: it first encodes the input sequence into the coding vector and then decodes an entire output sequence from it. In this sense, the coding vector is the only interaction between the input and the output. Thus, the size of the coding vector N is an essential measure of model capacity concerning the dependence of outputs on the inputs. Here, we show that this concept can be formalised as a notion of rank, which can pinpoint the precise types of input-output relationships that encoder-decoder structures are well adapted to. This is one of the most unique properties of these structures.

Numerical illustrations. Here we use a numerical example to illustrate the above discussions. We look at how the singular value decay rate, rank N_0 of the target relationships, and model rank N affect the approximation error $\|\boldsymbol{H} - \widehat{\boldsymbol{H}}\|$.



(a) $\sigma_n = \begin{cases} n^{-\frac{1}{8}}, & n \leq N_0 \\ 0, & n > N_0 \end{cases}$ (b) $\sigma_n = \begin{cases} n^{-1}, & n \leq N_0 \\ 0, & n > N_0 \end{cases}$ (c) $\sigma_n = n^{-2}$

Figure 2: In (a), (b), (c) we consider target relationships with different singular values indicated in the respective caption. For (a), (b) we also consider targets with different rank, where $N_0 = 2, 4, 6, 8$. We use models with fixed width $m = m_E = m_D = 128$ with coding vector size N. Detail settings are found in Appendix E.2.

In Figure 2, we train linear encoder-decoder structures to learn three relationships of different ranks determined by the decay patterns of singular values, given in (a), (b) and (c). Different colours mean targets with different rank. From Figure 2 we have the following observations consistent with previous discussions. First, observe that increasing the model rank N makes approximation errors smaller, as expected. Moreover, note that when increasing N, the speeds of decrements of errors are different. If singular values decay fast, errors also decay fast. This implies that a target with fast decaying singular values can be approximated efficiently with fewer parameters (smaller N).

For each experiment, we are able to achieve low approximation error by choosing $N \ll m$. The error will remain unchanged or decreases much slower when further increasing N. This means that in practice, one can choose N such that it covers the major singular values of the target in order to improve the approximation efficiency while maintaining small errors. That is, the notion of model reduction.

5 Conclusion

We theoretically study the approximation properties of the RNN encoder-decoder in a linear setting. We prove a universal approximation result of linear functions by encoder-decoder structures and show that they generalise the RNN to the time-inhomogeneous setting. Moreover, we discover an important structure, which we call the temporal product structure, that characterises the types of input-output relationships that are especially suited for approximation using encoder-decoders. This elucidates the key differences between these novel architectures and classical methods for learning time-series, and forms a basic step towards understanding the intricacies of modern deep learning architectures for time-series analysis.

Reproducibility Statement. Detailed proofs for theoretical results, and complete settings of numerical examples are found in the appendix. The source code for numerical examples can be made available upon request.

Here is a quick reference:

Proposition 3.1	Properties of RNN encoder-decoder functionals	Appendix A
Theorem 4.1	Universal approximation theorem	Appendix B
Theorem 4.2	General approximation rates	Appendix C
Theorem 4.3	Approximation rates considering temporal product structure	Appendix D
Figure 1	Illustration of high rank, low rank temporal relationships	Appendix E.1
Figure 2	Numerical examples	Appendix E.2

References

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pp. 1–15, 2015.

Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

G Berkooz, P Holmes, and J L Lumley. The proper orthogonal decomposition in the analysis of turbulent flows. *Annual Review of Fluid Mechanics*, 25(1):539–575, 1993. doi: 10.1146/annurev.fl.25.010193.002543. URL `https://doi.org/10.1146/annurev.fl.25.010193.002543`.

Vladimir I Bogachev. *Measure theory*, volume 1. Springer Science & Business Media, 2007.

Ching-Hua Chang and Chung-Wei Ha. On eigenvalues of differentiable positive definite kernels. *Integral Equations and Operator Theory*, 33:1–7, 03 1999. doi: 10.1007/BF01203078.

Anindya Chatterjee. An introduction to the proper orthogonal decomposition. *Current Science*, 78(7):808–817, 2000. ISSN 00113891. URL `http://www.jstor.org/stable/24103957`.

Chung-Cheng Chiu, Tara N. Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J. Weiss, Kanishka Rao, Ekaterina Gonina, Navdeep Jaitly, Bo Li, Jan Chorowski, and Michiel Bacchiani. State-of-the-art speech recognition with sequence-to-sequence models, 2018.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pp. 1724–1734, 2014. doi: 10.3115/v1/d14-1179.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. pp. 103–111, 2015. doi: 10.3115/v1/w14-4012.

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

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Ken-ichi Funahashi and Yuichi Nakamura. Approximation of dynamical systems by continuous time recurrent neural networks. *Neural Networks*, 6(6):801 – 806, 1993. ISSN 0893-6080.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Haotian Jiang, Zhong Li, and Qianxiao Li. Approximation theory of convolutional architectures for time series modelling. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4961–4970. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/jiang21d.html`.

Nal Kalchbrenner and Phil Blunsom. Recurrent continuous translation models. *EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, (October):1700–1709, 2013.

Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. Neural speech synthesis with transformer network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 6706–6713, 2019.

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.

Y.C. Liang, H.P. Lee, S.P. Lim, W.Z. Lin, K.H. Lee, and C.G. Wu. Proper orthogonal decomposition and its applications—part i: Theory. *Journal of Sound and Vibration*, 252 (3):527–544, 2002. ISSN 0022-460X. doi: https://doi.org/10.1006/jsvi.2001.4041. URL https://www.sciencedirect.com/science/article/pii/S0022460X01940416.

G.G. Lorentz. *Approximation of Functions*. AMS Chelsea Publishing Series. Holt, Rinehart and Winston, 2005. ISBN 9780821840504. URL https://books.google.com.sg/books?id=8VMrOmTKSe0C.

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

Charles Bradfield Morrey. Multiple integrals in the calculus of variations. Springer-Verlag, 1966.

Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International Conference on Machine Learning*, pp. 4055–4064. PMLR, 2018.

W. Rudin, W.A. RUDIN, and Tata McGraw-Hill Publishing Company. *Real and Complex Analysis*. Higher Mathematics Series. McGraw-Hill Education, 1987. ISBN 9780070542341. URL https://books.google.com.sg/books?id=Z_fuAAAAMAAJ.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 1986. ISSN 00280836. doi: 10.1038/323533a0.

Anton Maximilian Schäfer and Hans-Georg Zimmermann. Recurrent neural networks are universal approximators. *International journal of neural systems*, 17(04):253–263, 2007.

Martin H. Schultz. L^∞-multivariate approximation theory. *SIAM Journal on Numerical Analysis*, 6(2):161–183, 1969. doi: 10.1137/0706017. URL https://doi.org/10.1137/0706017.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, 4(January):3104–3112, 2014. ISSN 10495258.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017-Decem(Nips):5999–6009, 2017. ISSN 10495258.

Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, pp. 4534–4542, 2015.

Jong Chul Ye and Woon Kyoung Sung. Understanding geometry of encoder-decoder CNNs. *36th International Conference on Machine Learning, ICML 2019*, 2019-June: 12245–12254, 2019.

Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J. Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions?, 2020.

A Properties of model functionals

In this section, we prove observations of the hypothesis space reported in Proposition 3.1.

Proof of Proposition 3.1. Recall that

$$\widehat{H}_t(\boldsymbol{x};\theta) = \int_0^\infty c^\top e^{Vt} M e^{Ws} U x_{-s} ds. \tag{18}$$

Fix any $\theta = (W, V, U, M, c)$. The linearity is obvious. Since both $W \in \mathcal{W}_{m_E}$ and $V \in \mathcal{V}_{m_D}$ have eigenvalues with negative real parts, there exist $c_1, c_2, c_1', c_2' > 0$, such that $\|e^{Vt}\|_\infty \leq c_1 e^{-c_2 t}$ and $\|e^{Ws}\|_\infty \leq c_1' e^{-c_2' s}$ for any $t, s \geq 0$, hence

$$\begin{aligned} |c^\top e^{Vt} M e^{Ws} U x_{-s}| &\leq \|c\|_1 \|e^{Vt} M e^{Ws} U x_{-s}\|_\infty \\ &\leq \|c\|_1 \|e^{Vt}\|_\infty \|M\|_\infty \|e^{Ws}\|_\infty \|U\|_\infty \|x_{-s}\|_\infty \\ &\lesssim e^{-c_2 t} e^{-c_2' s} \|\boldsymbol{x}\|_{\mathcal{X}}, \end{aligned} \tag{19}$$

where $\lesssim$ hides universal positive constants depending on parameters θ. Therefore

$$|\widehat{H}_t(\boldsymbol{x};\theta)| \lesssim e^{-c_2 t} \|\boldsymbol{x}\|_{\mathcal{X}} \Rightarrow \|\widehat{H}_t(\cdot;\theta)\| \lesssim e^{-c_2 t}. \tag{20}$$

That is, the functional $\widehat{H}_t(\cdot;\theta)$ is bounded (i.e. continuous) with an exponentially-decayed norm (as a function of t). Finally, by (19) and Lebesgue's dominated convergence theorem, the functional $\widehat{H}_t(\cdot;\theta)$ is also regular. The proof is completed. □

B Universal approximation

In this section, we provide the proof of Theorem 4.1, i.e. the universal approximation property of encoder-decoder RNNs. As is stated in the main text, the key step is to utilise the classical representation theorem below, which helps us to reduce the approximation problem of functionals to functions.

First, we list the background definitions and notations used in the following theorems. Let $(X, \mathcal{A})$ be a measure space (with $\mathcal{A}$ as the σ-algebra of subsets of X).

- The space X is called *locally compact* if for any $x \in X$, x has a compact neighbourhood.
- X is a *Hausdorff space* if all distinct points in X are pair-wisely separable by neighbourhoods. That is, for any $x, y \in X$, there exists a neighbourhood Δ_x of x and a neighbourhood Δ_y of y, such that $\Delta_x \cap \Delta_y = \varnothing$.
- The measure μ is called a *finite measure*, if it satisfies $\mu(X) < \infty$. The measure ν is called a *σ-finite measure*, if X can be covered with at most countably many measurable sets with finite measure. That is, there are measurable sets $\{A_n\}_{n=1}^\infty \subset \mathcal{A}$ with $\nu(A_n) < \infty$ for all $n \in \mathbb{N}_+$, such that $\bigcup_{n=1}^\infty A_n = X$. Obviously, a finite measure is also σ-finite.
- Let $X = (-\infty, 0]$. For a measure μ on the measure space $((-\infty, 0], \mathcal{A})$, μ is *absolutely continuous* with respect to the Lebesgue measure ν, if for every measurable set A, $\nu(A) = 0$ implies $\mu(A) = 0$, which is written as $\mu \ll \nu$.

Denote by $C_0(X)$ the linear space of continuous functions defined on X vanishing at infinity. We have the following classical representation theorem.

Theorem B.1 (Riesz-Markov-Kakutani representation theorem)**.** *Let X be a locally compact Hausdorff space. For any continuous linear functional ψ on $C_0(X)$, there is a unique, regular, countably additive and signed measure μ on X, such that*

$$\psi(f) = \int_X f(x) d\mu(x), \quad \forall f \in C_0(X), \tag{21}$$

with $\|\psi\| = |\mu|(X)$. Here, $|\mu|(X)$ denotes the total variation of (the signed measure) μ, which is defined as $|\mu|(X) := \sup_{\mathcal{P}} \sum_{i=1}^k |\mu(A_i)|$, where $\mathcal{P} : X = \bigcup_{i=1}^k A_i$ is a partition over X with $A_i \in \mathcal{A}$ for all $i = 1, 2, \cdots, k$.

Proof. Well-known, see e.g. Bogachev (2007) (CH 7.10.4). □

Remark B.1. *It is straightforward to verify that* $|\mu|(X) = \sup_{A\in\mathcal{A}}(|\mu(A)| + |\mu(A^c)|)$. *Furthermore, if* μ *has a density* $d\mu/d\nu$ *with respect to a countably additive, nonnegative measure* ν, *then we have* $|\mu|(X) = \|d\mu/d\nu\|_{L^1(\nu)}$.

To handle signed measures, the following Jordan decomposition theorem (Bogachev, 2007) is necessary.

Theorem B.2. *Let* μ *be a signed measure on the measure space* $(X, \mathcal{A})$. *Then, there are two mutually singular (non-negative) measures* μ^+ *and* μ^- *on* $(X, \mathcal{A})$, *such that* $\mu = \mu^+ - \mu^-$. *Moreover, such a pair* (μ^+, μ^-) *is unique.*

Based on this, we have the following proposition to characterise absolutely continuous signed measures.

Proposition B.1. *If* μ *and* ν *are signed measures, then we have* $\mu \ll \nu \Leftrightarrow \mu^+ \ll \nu$ *and* $\mu^- \ll \nu$.

We also need the following Radon-Nikodym theorem (Bogachev, 2007).

Theorem B.3. *Let* $(X, \mathcal{A}, \nu)$ *be a* σ*-finite measure space, and let* μ *be a* σ*-finite signed measure, such that* $\mu \ll \nu$. *Then there exists a unique measurable function* f, *such that* $\mu(A) = \int_A f d\nu$ *for every measurable set* A.

Lemma B.1. *Let* $\{H_t\}_{t\geq 0}$ *be a family of linear, continuous and regular functionals defined on* $\mathcal{X}$. *Then there exists a integrable function* $\rho : [0, \infty)^2 \to \mathbb{R}^d$, *i.e.*

$$\|\rho\|_{L^1([0,\infty)^2)} := \sum_{i=1}^{d} \|\rho_i\|_{L^1([0,\infty)^2)} < \infty, \tag{22}$$

such that

$$H_t(\boldsymbol{x}) = \int_0^\infty x_{-s}^\top \rho(t, s) ds, \quad \forall \boldsymbol{x} \in \mathcal{X}. \tag{23}$$

In particular, $\{H_t\}_{t\geq 0}$ *is bounded with* $\|\boldsymbol{H}\| = \int_0^\infty \|H_t\| dt = \|\rho\|_{L^1([0,\infty)^2)}$.

Proof. Obviously, $(-\infty, 0]$ is a locally compact Hausdorff space. For any $t \geq 0$, since H_t is linear continuous, according to the Riesz-Markov-Kakutani representation theorem (Theorem B.1), there exists a unique, regular, countably additive and signed measure μ_t, such that

$$H_t(\boldsymbol{x}) = \int_{-\infty}^0 x_s^\top d\mu_t(s), \quad \forall \boldsymbol{x} \in \mathcal{X}, \tag{24}$$

with $\sum_{i=1}^d |\mu_{t,i}|((-\infty, 0]) = \|H_t\|$. We show that for any $t \geq 0$, $i = 1, 2, \cdots, d$, $\mu_{t,i}$ is absolutely continuous with respect to ν (the Lebesgue measure), i.e. $\mu_{t,i} \ll \nu$. According to Theorem B.2 and Proposition B.1, one can assume $\mu_{t,i}$ to be non-negative without loss of generality. Take a measurable set $A \subset (-\infty, 0]$ with $\nu(A) = 0$, the aim is to show $\mu_{t,i}(A) = 0$. Let $A' = (-\infty, 0] \setminus A$. Since both A and A' are measurable, there exist $K_n \subset A$, $K'_n \subset A'$ with K_n, K'_n closed, such that $\mu_{t,i}(A \setminus K_n) \leq 1/n$, $\mu_{t,i}(A' \setminus K'_n) \leq 1/n$ and $\nu(A' \setminus K'_n) \leq 1/n$ for any $n \in \mathbb{N}_+$. Fix any $i \in \{1, 2, \ldots, d\}$, we construct the sequence of input signals $\{\boldsymbol{x}^{(n)}\}_{n=1}^\infty$ as

$$x_{s,j}^{(n)} = \begin{cases} 0, & s \leq 0,\ j \neq i, \\ 0, & s \in K'_n,\ j = i, \\ 1, & s \in K_n,\ j = i, \end{cases} \qquad j = 1, 2, \ldots, d, \tag{25}$$

which can then be continuously extended to $(-\infty, 0]$ by defining $x_{s,i}^{(n)} := \frac{d(s,K'_n)}{d(s,K_n)+d(s,K'_n)} \in [0, 1]$. [1]

[1] Here, $d(s, B) := \inf\{|s - a| : a \in B\}$ is the distance between a point s and a set B.

We deduce that $\lim_{n\to\infty} x_{s,i}^{(n)} = 0$ for ν-a.e. $s \le 0$. In fact, let $S := \{s \le 0 : \lim_{n\to\infty} x_{s,i}^{(n)} = 0\}$, we have $K_n' \subset S$ since for any $s \in K_n'$, $x_{s,i}^{(n)} = 0$. Hence, $(-\infty, 0] \setminus S \subset A \cup (A' \setminus K_n')$, which gives $\nu((-\infty,0]\setminus S) \le \nu(A) + \nu(A' \setminus K_n') \le 1/n \to 0$ as $n \to \infty$. Due to the regularity of H_t, we get $\lim_{n\to\infty} H_t(\boldsymbol{x}^{(n)}) = 0$. By (24) and (25), we have

$$\begin{aligned} H_t(\boldsymbol{x}^{(n)}) &= \sum_{j=1}^{d} \int_{-\infty}^{0} x_{s,j}^{(n)} d\mu_{t,j}(s) = \int_{-\infty}^{0} x_{s,i}^{(n)} d\mu_{t,i}(s) \\ &= \int_{K_n} + \int_{A\setminus K_n} + \int_{A'\setminus K_n'} x_{s,i}^{(n)} d\mu_{t,i}(s) = \mu_{t,i}(K_n) + I_{1,n} + I_{2,n}, \end{aligned} \tag{26}$$

where $\mu_{t,i}(K_n) = \mu_{t,i}(A) - \mu_{t,i}(A \setminus K_n) \in [\mu_{t,i}(A) - 1/n, \mu_{t,i}(A)]$, and $|I_{1,n}| + |I_{2,n}| \le \int_{A\setminus K_n} + \int_{A'\setminus K_n'} 1 d\mu_{t,i}(s) = \mu_{t,i}(A \setminus K_n) + \mu_{t,i}(A' \setminus K_n') \le 2/n$, which gives $\lim_{n\to\infty} H_t(\boldsymbol{x}^{(n)}) = \mu_{t,i}(A)$. Therefore, $\mu_{t,i}(A) = 0$.

Notice that $|\mu_{t,i}((-\infty,0])| \le |\mu_{t,i}|((-\infty,0]) \le \|H_t\| < \infty$ for a.e. $t \ge 0$ (since $\|\boldsymbol{H}\| = \int_0^\infty \|H_t\| dt < \infty$), we get that $((-\infty,0], \mathcal{A}, \mu_{t,i})$ is a finite measure space, and hence σ-finite. Obviously, $((-\infty,0], \mathcal{A}, \nu)$ is a σ-finite measure space. According to the Radon-Nikodym theorem (Theorem B.3), there exists a unique measurable function $\rho_{t,i} : (-\infty, 0] \to \mathbb{R}$, such that $\mu_{t,i}(A) = \int_A \rho_{t,i}(s) d\nu(s)$ for every measurable set A. Hence, we have

$$H_t(\boldsymbol{x}) = \int_{-\infty}^{0} x_s^\top \rho_t(s) ds = \int_0^\infty x_{-s}^\top \rho(t,s) ds, \quad \forall \boldsymbol{x} \in \mathcal{X}, \tag{27}$$

with $\rho(t,s) := \rho_t(-s)$. In addition, by Remark B.1, we have $|\mu_{t,i}|((-\infty,0]) = \int_{-\infty}^0 |\rho_{t,i}(s)| ds$, which gives

$$\|\boldsymbol{H}\| = \int_0^\infty \|H_t\| dt = \sum_{i=1}^{d} \int_0^\infty |\mu_{t,i}|((-\infty,0]) dt = \sum_{i=1}^{d} \|\rho_i\|_{L^1([0,\infty)^2)} = \|\rho\|_{L^1([0,\infty)^2)}. \tag{28}$$

The proof is completed. □

Based on this representation theorem, the problem of functional approximation is reduced as function approximation. That is,

$$\begin{aligned} \|\boldsymbol{H} - \widehat{\boldsymbol{H}}\| &= \int_0^\infty \|H_t - \widehat{H}_t\| dt = \int_0^\infty \sup_{\|\boldsymbol{x}\|_{\mathcal{X}} \le 1} \left| H_t(\boldsymbol{x}) - \widehat{H}_t(\boldsymbol{x};\theta) \right| dt \\ &= \int_0^\infty \sup_{\|\boldsymbol{x}\|_{\mathcal{X}} \le 1} \left| \int_0^\infty x_{-s}^\top (\rho(t,s) - \hat{\rho}(t,s)) ds \right| dt \\ &\le \int_0^\infty \sup_{\|\boldsymbol{x}\|_{\mathcal{X}} \le 1} \int_0^\infty \|x_{-s}\|_\infty \|\rho(t,s) - \hat{\rho}(t,s)\|_1 ds dt \\ &\le \sum_{i=1}^{d} \int_0^\infty \int_0^\infty |\rho_i(t,s) - \hat{\rho}_i(t,s)| ds dt, \end{aligned} \tag{29}$$

i.e.

$$\|\boldsymbol{H} - \widehat{\boldsymbol{H}}\| \le \|\rho - \hat{\rho}\|_{L^1([0,\infty)^2)} := \sum_{i=1}^{d} \|\rho_i - \hat{\rho}_i\|_{L^1([0,\infty)^2)}. \tag{30}$$

Lemma B.2. *Let $\rho(t,s) : [0,\infty)^2 \to \mathbb{R}$ with $\|\rho\|_{L^1([0,\infty)^2)} < \infty$. Then for any $\epsilon > 0$, there exists a polynomial $p(u,v) = \sum_{j=1,k=1}^{m} c_{jk} u^j v^k$, such that*

$$\int_0^\infty \int_0^\infty |\rho(t,s) - p(e^{-t}, e^{-s})| dt ds < \epsilon. \tag{31}$$

Proof. Fix any $\epsilon > 0$. Consider the following transformation

$$R(u,v) = \begin{cases} \frac{1}{uv}\rho(-\ln u, -\ln v), & u,v \in (0,1], \\ 0, & uv = 0. \end{cases} \tag{32}$$

This transformation preserves the norm with $\|\rho\|_{L^1([0,\infty)^2)} = \|R\|_{L^1([0,1]^2)}$.

First, according to the density of continuous functions in L^p space (Rudin et al., 1987, Theorem 3.14), there exists $\tilde{R} \in C([0,1]^2)$, such that $\|R - \tilde{R}\|_{L^1([0,1]^2)} < \epsilon/2$. Next, by the density of polynomials in the space of continuous functions (Lorentz, 2005, Theorem 6), there exists a polynomial $q(u,v) = \sum_{j,k=0}^{m} c_{jk} u^j v^k$, such that $\|\tilde{R} - q\|_{L^\infty([0,1]^2)} < \epsilon/2$. Finally, let $p(u,v) = uvq(u,v)$, we have

$$\begin{aligned} \int_0^\infty \int_0^\infty |\rho(t,s) - p(e^{-t}, e^{-s})| dt ds &= \int_0^1 \int_0^1 \left| R(u,v) - \frac{1}{uv} p(u,v) \right| du dv \\ &= \|R - q\|_{L^1([0,1]^2)} \\ &\le \|R - \tilde{R}\|_{L^1([0,1]^2)} + \|\tilde{R} - q\|_{L^\infty([0,1]^2)} \\ &< \epsilon/2 + \epsilon/2 = \epsilon, \end{aligned} \tag{33}$$

which completes the proof. □

Now we are ready to prove the universal approximation theorem.

Proof of Theorem 4.1. According to the representation theorem (Lemma B.1), we have

$$H_t(\boldsymbol{x}) = \int_0^\infty x_{-s}^\top \rho(t,s) ds, \quad \forall \boldsymbol{x} \in \mathcal{X}, \tag{34}$$

where $\|\rho\|_{L^1([0,\infty)^2)} = \|\boldsymbol{H}\| < \infty$. Therefore, by Lemma B.2, there exists $p_i(u,v) = \sum_{j,k=1}^{m} c_{jk}^{(i)} u^j v^k$, $i = 1, 2, \ldots, d$, where m is the maximal degree of $\{p_i\}_{i=1}^d$, such that

$$\sum_{i=1}^d \int_0^\infty \int_0^\infty |\rho_i(t,s) - p_i(e^{-t}, e^{-s})| dt ds < \epsilon. \tag{35}$$

Let

$$\begin{aligned} &c = u = \mathbf{1}_m,\ V = \tilde{W} = -\operatorname{diag}(1, 2, \ldots, m), \\ &W = \operatorname{diag}(\tilde{W}, \tilde{W}, \cdots, \tilde{W}) \in \mathbb{R}^{dm \times dm},\ U = \operatorname{diag}(u, u, \cdots, u) \in \mathbb{R}^{dm \times d}, \\ &M = PQ = (M_1, M_2, \cdots, M_d) \in \mathbb{R}^{m \times dm},\ [M_i]_{jk} = c_{jk}^{(i)}, \end{aligned} \tag{36}$$

we get

$$\begin{aligned} \hat{\rho}(t,s)^\top &= c^\top e^{Vt} PQ e^{Ws} U = c^\top e^{Vt} M e^{Ws} U \\ &= c^\top e^{Vt} \cdot (M_1, M_2, \cdots, M_d) \cdot \operatorname{diag}(e^{Ws}, e^{Ws}, \cdots, e^{Ws}) \cdot \operatorname{diag}(u, u, \cdots, u) \\ &= (c^\top e^{Vt} M_1, c^\top e^{Vt} M_2, \cdots, c^\top e^{Vt} M_d) \cdot \operatorname{diag}(e^{Ws} u, e^{Ws} u, \cdots, e^{Ws} u) \\ &= (c^\top e^{Vt} M_1 e^{Ws} u, c^\top e^{Vt} M_2 e^{Ws} u, \cdots, c^\top e^{Vt} M_d e^{Ws} u), \end{aligned} \tag{37}$$

with

$$\hat{\rho}_i(t,s) = c^\top e^{Vt} M_i e^{Ws} u = p_i(e^{-t}, e^{-s}), \quad i = 1, 2, \ldots, d. \tag{38}$$

Therefore, by (30) and (35), we have

$$\begin{aligned} \|\boldsymbol{H} - \widehat{\boldsymbol{H}}\| &\le \sum_{i=1}^d \|\rho_i - \hat{\rho}_i\|_{L^1([0,\infty)^2)} \\ &= \sum_{i=1}^d \int_0^\infty \int_0^\infty \left| \rho_i(t,s) - p_i(e^{-t}, e^{-s}) \right| ds dt < \epsilon, \end{aligned} \tag{39}$$

which completes the proof. □

C General approximation rates

In this section, the proof of Theorem 4.2 is given. Again, by (30), the aim now is to investigate the function approximation $\|\rho-\hat{\rho}\|$. Since one can handle each spatial dimension separately (similarly with (36) and (37)), we firstly derive the estimates by assuming $d=1$, and then extend the obtained results to the case of multi-dimensional inputs (general $d \in \mathbb{N}_+$).

Conditions on representation. To characterise the accuracy of using the model $c^\top e^{Vt} M e^{Ws} u$ (with $M := PQ$) to approximate the target $\rho(t,s)$, the first stuff is to translate the conditions on the output (of piece-wise constant signals) to the representation. Recall that $y^{\mathrm{c}}(t,s) = H_t(\mathbf{1}_{(-\infty,-s]})$, $t,s \geq 0$, we get $\rho(t,s) = -\frac{d}{ds} H_t(\mathbf{1}_{(-\infty,-s]})$. Hence, the assumptions on y^{c} in Theorem 4.2 is equivalent to the following smoothness and exponential decay conditions on ρ. That is, there exist $\alpha \in \mathbb{N}_+$, $\beta > 0$ such that

$$\rho \in C^\alpha([0,+\infty)^2), \tag{40}$$

$$e^{\beta(t+s)} \frac{\partial^{k+l}}{\partial t^k \partial s^l} \rho(t,s) = o(1) \text{ as } \|(t,s)\| \to \infty, \quad k,l \in \mathbb{N},\, k+l \leq \alpha. \tag{41}$$

Note that the last decay condition implies

$$\sup_{t,s\geq 0} \beta^{-(k+l)} e^{\beta(t+s)} \left| \frac{\partial^{k+l}}{\partial t^k \partial s^l} \rho(t,s) \right| \leq \gamma, \quad k,l \in \mathbb{N},\ k+l \leq \alpha \tag{42}$$

for some $\gamma > 0$.

Background theorems. Let $\Omega \in \mathbb{R}^d$ be a bounded set. Define the spaces

$$C^\alpha(\Omega) := \{f \in C(\bar{\Omega}) : D^{\boldsymbol{i}} f \in C(\bar{\Omega}) \text{ for all } |\boldsymbol{i}| \leq \alpha\}, \quad \alpha \in \mathbb{N}, \tag{43}$$

$$\begin{aligned} C^{\alpha,\mu}(\Omega) := \{f \in C^\alpha(\Omega) : |D^{\boldsymbol{i}} f(x) - D^{\boldsymbol{i}} f(y)| \leq K\|x-y\|_2^\mu \text{ for some } K > 0, \\ \text{for all } x,y \in \Omega \text{ and } |\boldsymbol{i}| = \alpha\}, \end{aligned} \tag{44}$$

and the "norm"

$$|f|_{\alpha,\mu,\Omega} := \sup_{|\boldsymbol{i}|=\alpha} \sup_{x,y\in\Omega} \frac{|D^{\boldsymbol{i}} f(x) - D^{\boldsymbol{i}} f(y)|}{\|x-y\|_2^\mu}, \quad \forall f \in C^{\alpha,\mu}(\Omega), \tag{45}$$

with the shorthand $\|\cdot\|_\Omega := |\cdot|_{0,0,\Omega}$.

Theorem C.1 (Multivariate Jackson's theorem (Schultz, 1969), Theorem 4.10)**.** *Let $\Omega \in \mathbb{R}^d$ be a regular,* [2] *bounded and open set, and $f \in C^{\alpha,\mu}(\bar{\Omega})$ for some $\alpha \in \mathbb{N}$, $\mu \in [0,1]$. Then for any $n \in \mathbb{N}_+$, we have*

$$\inf_{p \in \mathcal{P}_n^d} \|f-p\|_{\bar{\Omega}} \leq \frac{C(\alpha,\mu)}{n^{\alpha+\mu}} |f|_{\alpha,\mu,\bar{\Omega}}, \tag{46}$$

where $\mathcal{P}_n^d$ denotes the set of all polynomials with the degree of no more than n in each variable, $C(\alpha,\mu) > 0$ is a universal constant only depending on α, μ and Ω.

A commonly used case is when $\mu = 0$. That is, for $f \in C^\alpha(\bar{\Omega})$, we get

$$|f|_{\alpha,0,\bar{\Omega}} \leq 2 \max_{|\boldsymbol{i}|=\alpha} \|D^{\boldsymbol{i}} f\|_{L^\infty(\bar{\Omega})} < \infty. \tag{47}$$

For any $x, x_0 \in \bar{\Omega}$, let $\tilde{p}(x) := p(x) + f(x_0) - p(x_0)$, and $\tilde{e}(x) := f(x) - \tilde{p}(x)$. Then $\tilde{p} \in \mathcal{P}_n^d$, and

$$\begin{aligned} |f(x) - \tilde{p}(x)| &\leq |\tilde{e}(x_0)| + |\tilde{e}(x) - \tilde{e}(x_0)| \\ &\leq \sup_{x,y\in\bar{\Omega}} |\tilde{e}(x) - \tilde{e}(y)| = |\tilde{e}|_{0,0,\bar{\Omega}} = \|f-\tilde{p}\|_{\bar{\Omega}}, \quad \forall x \in \bar{\Omega}. \end{aligned} \tag{48}$$

This gives the following convenient corollary.

[2] It is proved that every bounded, open and convex set is regular. See Morrey (1966) (Lemma 3.4.1).

Corollary C.1. *Let $\Omega \in \mathbb{R}^d$ be a regular, bounded and open set, and $f \in C^\alpha(\bar{\Omega})$ for some $\alpha \in \mathbb{N}$. Then for any $n \in \mathbb{N}_+$, there exists $p \in \mathcal{P}_n^d$ such that*

$$\|f - p\|_{L^\infty(\bar{\Omega})} \leq \frac{C_\alpha}{n^\alpha} \max_{|\boldsymbol{i}|=\alpha} \|D^{\boldsymbol{i}} f\|_{L^\infty(\bar{\Omega})}, \tag{49}$$

where $\mathcal{P}_n^d$ denotes the set of all polynomials with the degree of no more than n in each variable, $C_\alpha > 0$ is a universal constant only depending on α and Ω.

Now we are ready to present the proof.

Proof of Theorem 4.2. Step 1: domain transform. Consider the transform from the infinite domain $[0, \infty)^2$ to the compact one $[0, 1]^2$:

$$R(u, v) = \begin{cases} \frac{1}{uv}\rho(-c_0 \ln u, -c_0 \ln v), & u, v \in (0, 1], \\ 0, & uv = 0, \end{cases} \tag{50}$$

where $c_0 := (\alpha + 1)/\beta > 0$ is a fixed constant. A straightforward computation by induction shows that, for any $k, l \in \mathbb{N}$, $k + l \leq \alpha$, and any $u, v \in (0, 1]$,

$$\frac{\partial^{k+l}}{\partial u^k \partial v^l} R(u, v) = \frac{(-1)^{k+l}}{u^{k+1} v^{l+1}} \sum_{i=0}^{k} \sum_{j=0}^{l} C(k, i) C'(l, j) c_0^{i+j} \frac{\partial^{i+j}}{\partial t^i \partial s^j} \rho(-c_0 \ln u, -c_0 \ln v), \tag{51}$$

where $C(k, i), C'(l, j)$ are some integer constants, and $(t, s) = (-c_0 \ln u, -c_0 \ln v)$ is a one-to-one mapping between $(0, 1]^2$ and $[0, \infty)^2$. By (42), we get

$$\begin{aligned}
\left| \frac{\partial^{k+l}}{\partial u^k \partial v^l} R(u, v) \right| &\leq \frac{1}{u^{k+1} v^{l+1}} \sum_{i=0}^{k} \sum_{j=0}^{l} |C(k, i)||C'(l, j)| c_0^{i+j} \left| \frac{\partial^{i+j}}{\partial t^i \partial s^j} \rho(-c_0 \ln u, -c_0 \ln v) \right|, \\
\left| \frac{\partial^{k+l}}{\partial u^k \partial v^l} R(e^{-\frac{t}{c_0}}, e^{-\frac{s}{c_0}}) \right| &\leq e^{\frac{(k+1)}{c_0} t} e^{\frac{(l+1)}{c_0} s} \sum_{i=0}^{k} \sum_{j=0}^{l} |C(k, i)||C'(l, j)| c_0^{i+j} \left| \frac{\partial^{i+j}}{\partial t^i \partial s^j} \rho(t, s) \right| \\
&\leq \sum_{i=0}^{k} \sum_{j=0}^{l} |C(k, i)||C'(l, j)| (\alpha + 1)^{i+j} \cdot \beta^{-(i+j)} e^{\beta(t+s)} \left| \frac{\partial^{i+j}}{\partial t^i \partial s^j} \rho(t, s) \right| \\
&\leq \sum_{i=0}^{k} \sum_{j=0}^{l} |C(k, i)||C'(l, j)| (\alpha + 1)^{i+j} \gamma \leq C(\alpha)\gamma,
\end{aligned} \tag{52}$$

where $C(\alpha) > 0$ is a universal constant only depending on α. According to the decay condition (41), we get

$$\begin{aligned}
\lim_{(u,v)\to(0,0)} \frac{\frac{\partial^{i+j}}{\partial t^i \partial s^j} \rho(-c_0 \ln u, -c_0 \ln v)}{u^{k+1} v^{l+1}} &= \lim_{(t,s)\to(+\infty,+\infty)} e^{\frac{(k+1)}{c_0} t} e^{\frac{(l+1)}{c_0} s} \frac{\partial^{i+j}}{\partial t^i \partial s^j} \rho(t, s) \\
&= \lim_{(t,s)\to(+\infty,+\infty)} e^{\frac{(k-\alpha)}{\alpha+1} \beta t} e^{\frac{(l-\alpha)}{\alpha+1} \beta s} \cdot e^{\beta(t+s)} \frac{\partial^{i+j}}{\partial t^i \partial s^j} \rho(t, s) \\
&= 0,
\end{aligned} \tag{53}$$

and similarly for $u_0, v_0 \in (0, 1]$,

$$\begin{aligned}
\lim_{(u,v)\to(u_0,0)} \frac{\frac{\partial^{i+j}}{\partial t^i \partial s^j} \rho(-c_0 \ln u, -c_0 \ln v)}{u^{k+1} v^{l+1}} &= \lim_{(t,s)\to(-c_0 \ln u_0,+\infty)} e^{\frac{(k+1)}{c_0} t} e^{\frac{(l+1)}{c_0} s} \frac{\partial^{i+j}}{\partial t^i \partial s^j} \rho(t, s) \\
&= \lim_{\|(t,s)\|\to+\infty} e^{\frac{(k-\alpha)}{\alpha+1} \beta t} e^{\frac{(l-\alpha)}{\alpha+1} \beta s} \cdot e^{\beta(t+s)} \frac{\partial^{i+j}}{\partial t^i \partial s^j} \rho(t, s) \\
&= 0,
\end{aligned} \tag{54}$$

$$\lim_{(u,v)\to(0,v_0)} \frac{\frac{\partial^{i+j}}{\partial t^i \partial s^j} \rho(-c_0 \ln u, -c_0 \ln v)}{u^{k+1} v^{l+1}} = 0. \tag{55}$$

This gives

$$\frac{\partial^{k+l}}{\partial u^k \partial v^l} R(u,v) = 0, \quad (u,v) \in [0,1] \times \{0\} \cup \{0\} \times [0,1], \tag{56}$$

and

$$M_0 := \max_{k,l \in \mathbb{N},\, k+l \le \alpha} \max_{(u,v) \in [0,1]^2} \left| \frac{\partial^{k+l}}{\partial u^k \partial v^l} R(u,v) \right| \le C(\alpha)\gamma \tag{57}$$

by (52) and (56). Hence, $R(u,v) \in C^\alpha([0,1]^2)$ with bounded derivatives.

Step 2: polynomial approximation. According to Corollary C.1 and (57), we obtain that there exists $\tilde{R}_n \in \mathcal{P}_n^2$, such that

$$\begin{aligned} \|R - \tilde{R}_n\|_{L^\infty([0,1]^2)} &\le \frac{C_\alpha}{n^\alpha} \max_{k,l \in \mathbb{N},\, k+l=\alpha} \left\| \frac{\partial^{k+l}}{\partial u^k \partial v^l} R(u,v) \right\|_{L^\infty([0,1]^2)} \\ &\le \frac{C(\alpha)\gamma}{n^\alpha}, \quad \forall n \in \mathbb{N}_+, \end{aligned} \tag{58}$$

where $C(\alpha) > 0$ is a universal constant only related to α. Furthermore, let $\hat{R}_n(u,v) := \tilde{R}_n(u,v) - \tilde{R}_n(u,0) - \tilde{R}_n(0,v) + \tilde{R}_n(0,0)$, we get $\hat{R}_n \in \mathcal{P}_n^2$ with $\hat{R}_n(u,0) = \hat{R}_n(0,v) = 0$ for any $u, v \in [0,1]$. By (56), we have $R(u,v) = 0$ for any $(u,v) \in [0,1] \times \{0\} \cup \{0\} \times [0,1]$, then

$$\begin{aligned} \|R - \hat{R}_n\|_{L^\infty([0,1]^2)} &\le \|R - \tilde{R}_n\|_{L^\infty([0,1]^2)} + \|\tilde{R}_n - \hat{R}_n\|_{L^\infty([0,1]^2)} \\ &\le \|R - \tilde{R}_n\|_{L^\infty([0,1]^2)} + \|\tilde{R}_n(u,0) - R(u,0)\|_{L^\infty([0,1]^2)} \\ &\quad + \|\tilde{R}_n(0,v) - R(0,v)\|_{L^\infty([0,1]^2)} + \|\tilde{R}_n(0,0) - R(0,0)\|_{L^\infty([0,1]^2)} \\ &\lesssim \|R - \tilde{R}_n\|_{L^\infty([0,1]^2)}, \end{aligned} \tag{59}$$

i.e. we can further require the approximator satisfying the zero half-boundary condition (i.e. vanishing on $(u,v) \in [0,1] \times \{0\} \cup \{0\} \times [0,1]$) without effecting the approximation accuracy. Let $\bar{m} := \min\{m_E, m_D\}$, we get

$$\|R - \tilde{R}\|_{L^\infty([0,1]^2)} \le \frac{C(\alpha)\gamma}{\bar{m}^\alpha} \le C(\alpha)\gamma \left(\frac{1}{m_E^\alpha} + \frac{1}{m_D^\alpha} \right), \tag{60}$$

where

$$\tilde{R} := \tilde{R}_{\bar{m}} \in \mathcal{P}_{\bar{m}}^2, \quad \tilde{R}(u,v) := \sum_{i=1}^{\bar{m}} \sum_{j=1}^{\bar{m}} \tilde{r}_{ij} u^i v^j. \tag{61}$$

Let

$$c = \mathbf{1}_{\bar{m}}, \quad u = \mathbf{1}_{\bar{m}}, \quad M = [\tilde{r}_{ij}] \in \mathbb{R}^{\bar{m} \times \bar{m}}, \tag{62}$$

$$V = -\mathrm{diag}(2, 3, \cdots, \bar{m}+1)/c_0, \quad W = -\mathrm{diag}(2, 3, \cdots, \bar{m}+1)/c_0, \tag{63}$$

then by (30) and (60), we have

$$\|\boldsymbol{H} - \widehat{\boldsymbol{H}}\| \le \|\rho - \hat{\rho}\|_{L^1} \tag{64}$$

$$\begin{aligned} &= \left\| \rho(t,s) - c^\top e^{Vt} M e^{Ws} u \right\|_{L^1([0,\infty)^2)} \\ &= \left\| \rho(t,s) - e^{-\frac{t}{c_0}} e^{-\frac{s}{c_0}} \tilde{R}(e^{-\frac{t}{c_0}}, e^{-\frac{s}{c_0}}) \right\|_{L^1([0,\infty)^2)} \\ &= c_0^2 \left\| R - \tilde{R} \right\|_{L^1([0,1]^2)} \end{aligned} \tag{65}$$

$$\le c_0^2 \left\| R - \tilde{R} \right\|_{L^\infty([0,1]^2)} \le \frac{C(\alpha)\gamma}{\beta^2 \bar{m}^\alpha} \le \frac{C(\alpha)\gamma}{\beta^2} \left(\frac{1}{m_E^\alpha} + \frac{1}{m_D^\alpha} \right). \tag{66}$$

That is, one can achieve an approximation accuracy scaling like $(1/\bar{m})^\alpha$ with $\bar{m}^2$ parameters. The proof is completed. □

Remark C.1. *Recall $M = PQ \in \mathbb{R}^{m_D \times m_E}$ with $P \in \mathbb{R}^{m_D \times N}$, $Q \in \mathbb{R}^{N \times m_E}$, we only need to investigate the case of $N \le \min\{m_E, m_D\} = \bar{m}$, since* $\mathrm{rank}(M) \le \bar{m}$.

D Approximation rates using temporal product structure

In this section, the proof of Theorem 4.3 is provided.

D.1 Proper Orthogonal Decomposition

Proper orthogonal decomposition (POD) (Liang et al., 2002; Berkooz et al., 1993; Chatterjee, 2000) is a method for model reduction, which is commonly applied to numerical PDEs and fluids mechanics. It can be viewed as an extension of singular value decomposition (SVD) and principal component analysis (PCA) to infinite-dimensional spaces.

Fix any $R \in L^\infty([0,1]^2)$. [3] Define the POD operator

$$\mathcal{K}: \ \phi(v) \mapsto \int_0^1 \int_0^1 R(u,v)\phi(v)dv \cdot R(u,v)du, \quad \phi(v) \in L^2[0,1]. \tag{67}$$

Proposition D.1. *The operator $\mathcal{K}$ is linear, bounded, compact, self-adjoint and non-negative.*

Proof. (i) The linearity is obvious.

(ii) Let

$$\mathcal{R}: \ \phi(v) \mapsto \int_0^1 R(u,v)\phi(v)dv, \quad \phi(v) \in L^2[0,1], \tag{68}$$

then

$$(\mathcal{K}\phi)(v) = \int_0^1 R(u,v)(\mathcal{R}\phi)(u)du, \quad \phi(v) \in L^2[0,1]. \tag{69}$$

By the Cauchy-Schwartz inequality, we get

$$\|\mathcal{R}\phi\|_{L^2[0,1]} \le \|R\|_{L^2([0,1]^2)}\|\phi\|_{L^2[0,1]}, \tag{70}$$

which gives

$$\|\mathcal{K}\phi\|_{L^2[0,1]} \le \|R\|_{L^2([0,1]^2)}\|\mathcal{R}\phi\|_{L^2[0,1]} \le \|R\|^2_{L^2([0,1]^2)}\|\phi\|_{L^2[0,1]}. \tag{71}$$

Note that $R(u,v) \in L^\infty([0,1]^2) \subset L^2([0,1]^2)$, hence both $\mathcal{K}$ and $\mathcal{R}$ are bounded operator from $L^2[0,1]$ to itself.

(iii) It is well-known that the Hilbert–Schmidt integral operator

$$(\mathcal{C}\psi)(v) = \int_0^1 R(u,v)\psi(u)du, \quad \psi(u) \in L^2[0,1] \tag{72}$$

is a compact operator from $L^2[0,1]$ to itself. Therefore

$$\mathcal{K}\phi = \mathcal{C}\mathcal{R}\phi, \quad \phi \in L^2[0,1], \tag{73}$$

which gives $\mathcal{K} = \mathcal{C}\mathcal{R}$. Since $\mathcal{R}$ is bounded and $\mathcal{C}$ is compact, we get $\mathcal{K}$ is also compact.

(iv) By Fubini's theorem, it is straightforward to verify that

$$\begin{aligned}
\langle \mathcal{K}\phi, \psi\rangle_{L^2[0,1]} &= \int_0^1 \int_0^1 R(u,w)\left(\int_0^1 R(u,v)\phi(v)dv\right)du \cdot \psi(w)dw \\
&= \int_0^1 \int_0^1 \int_0^1 R(u,w)R(u,v)\phi(v)\psi(w)dvdudw \\
&= \int_0^1 \int_0^1 R(u,v)\left(\int_0^1 R(u,w)\psi(w)dw\right)du \cdot \phi(v)dv \\
&= \langle \phi, \mathcal{K}\psi\rangle_{L^2[0,1]}, \quad \phi, \psi \in L^2[0,1].
\end{aligned} \tag{74}$$

[3] Here, we use the same notation as (50), since the function defined there is also bounded.

(v) By Fubini's theorem, it is straightforward to verify that

$$\begin{aligned}\langle \mathcal{K}\phi, \phi\rangle_{L^2[0,1]} &= \int_0^1 \int_0^1 (\mathcal{R}\phi)(u) R(u,w) du \cdot \phi(w) dw \\ &= \int_0^1 \int_0^1 (\mathcal{R}\phi)(u) R(u,w) \phi(w) du dw \\ &= \int_0^1 (\mathcal{R}\phi)(u) \left(\int_0^1 R(u,w)\phi(w) dw \right) du \\ &= \int_0^1 (\mathcal{R}\phi)^2(u) du \geq 0, \quad \phi \in L^2[0,1]. \end{aligned} \tag{75}$$

The proof is completed. □

Combining (i)—(iv) and applying Hilbert–Schmidt's expansion theorem, we obtain that $L^2[0,1]$ has an orthonormal basis $\{\phi_n\}_{n\in\mathcal{N}} \bigcup \{\psi_\xi\}_{\xi\in\Xi}$, such that

- $\mathcal{K}\phi_n = \lambda_n \phi_n$, $\lambda_n \neq 0$ for $n \in \mathcal{N}$, and $\mathcal{K}\psi_\xi = 0$ for $\xi \in \Xi$, where $\mathcal{N}$ is a finite or countable set. If $\mathcal{N}$ is not finite, we have $\lim_{n\to\infty} \lambda_n = 0$;
- For any $\psi \in L^2[0,1]$, we have

$$\psi = \sum_{n\in\mathcal{N}} \langle \psi, \phi_n\rangle_{L^2[0,1]} \phi_n + \sum_{\xi\in\Xi} \langle \psi, \psi_\xi\rangle_{L^2[0,1]} \psi_\xi, \tag{76}$$

  where the second summation has at most countable non-zero terms, and

$$\mathcal{K}\psi = \sum_{n\in\mathcal{N}} \lambda_n \langle \psi, \phi_n\rangle_{L^2[0,1]} \phi_n. \tag{77}$$

Here, all the series converge under the norm $\|\cdot\|_{L^2[0,1]}$. Without loss of generality, $\mathcal{N} = \{1, 2, \cdots, N_0\}$ for $N_0 \in \mathbb{N}_+$ or $N_0 = +\infty$ (i.e. $\mathcal{N} = \mathbb{N}_+$). By (75), we get

$$0 \leq \langle \mathcal{K}\phi_n, \phi_n\rangle_{L^2[0,1]} = \lambda_n \langle \phi_n, \phi_n\rangle^2_{L^2[0,1]} = \lambda_n, \quad \forall n \in \mathcal{N}, \tag{78}$$

i.e. all the eigenvalues of $\mathcal{K}$ are non-negative, and $\lambda_n \neq 0$ for $n \in \mathcal{N}$ implies $\lambda_n > 0$, $\forall n \in \mathcal{N}$. In addition, $\lim_{n\to\infty} \lambda_n = 0$ implies that one can index all the eigenvalues in a non-increasing sequence: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq \cdots \geq 0$.

Then we can present and derive the POD estimate.

Theorem D.1. *For any $R \in L^\infty([0,1]^2)$, we have*

$$\int_0^1 \left\| R(u,v) - \sum_{n=1}^{N} \langle R(u,v), \phi_n(v)\rangle_{L^2[0,1]} \phi_n(v) \right\|^2_{L^2[0,1]} du = \sum_{n=N+1}^{N_0} \lambda_n, \quad \forall N \in \mathbb{N}. \tag{79}$$

Proof. Combining (75) and (78) gives

$$\lambda_n = \langle \mathcal{K}\phi_n, \phi_n\rangle_{L^2[0,1]} = \int_0^1 (\mathcal{R}\phi_n)^2(u) du, \quad \forall n \in \mathcal{N}. \tag{80}$$

Similarly,

$$\int_0^1 (\mathcal{R}\psi_\xi)^2(u) du = \langle \mathcal{K}\psi_\xi, \psi_\xi\rangle_{L^2[0,1]} = \sum_{n\in\mathcal{N}} \lambda_n \langle \psi_\xi, \phi_n\rangle^2_{L^2[0,1]} = 0, \quad \forall \xi \in \Xi, \tag{81}$$

which gives

$$(\mathcal{R}\psi_\xi)(u) = \int_0^1 R(u,v)\psi_\xi(v) dv = 0, \quad a.e.\ u \in [0,1]. \tag{82}$$

Notice that $R_u(v) := R(u,v) \in C^\alpha[0,1] \subset L^2[0,1]$ for any $u \in [0,1]$. By (76) and (82), we get

$$\begin{aligned}\|R_u\|^2_{L^2[0,1]} &= \Bigg\langle \sum_{n\in\mathcal{N}} \langle R_u, \phi_n\rangle_{L^2[0,1]}\phi_n + \sum_{\xi\in\Xi}\langle R_u, \psi_\xi\rangle_{L^2[0,1]}\psi_\xi, \\ &\qquad \sum_{n\in\mathcal{N}} \langle R_u, \phi_n\rangle_{L^2[0,1]}\phi_n + \sum_{\xi\in\Xi}\langle R_u, \psi_\xi\rangle_{L^2[0,1]}\psi_\xi \Bigg\rangle_{L^2[0,1]} \\ &= \sum_{n\in\mathcal{N}} \langle R_u, \phi_n\rangle^2_{L^2[0,1]} + \sum_{\xi\in\Xi}\langle R_u, \psi_\xi\rangle^2_{L^2[0,1]} \\ &= \sum_{n\in\mathcal{N}} (\mathcal{R}\phi_n)^2(u), \quad a.e.\ u \in [0,1]. \end{aligned} \tag{83}$$

Hence for any $N \in \mathbb{N}$, we have

$$\left\| R_u - \sum_{n=1}^N \langle R_u, \phi_n\rangle_{L^2[0,1]}\phi_n \right\|^2_{L^2[0,1]} = \|R_u\|^2_{L^2[0,1]} - \sum_{n=1}^N \langle R_u, \phi_n\rangle^2_{L^2[0,1]} \tag{84}$$

$$= \sum_{n\in\mathcal{N}} (\mathcal{R}\phi_n)^2(u) - \sum_{n=1}^N (\mathcal{R}\phi_n)^2(u)$$

$$= \sum_{n=N+1}^{N_0} (\mathcal{R}\phi_n)^2(u), \tag{85}$$

where the summation is zero by convention if the subscript is larger than the superscript. This by (80) implies

$$\int_0^1 \left\| R(u,v) - \sum_{n=1}^N \langle R(u,v), \phi_n(v)\rangle_{L^2[0,1]}\phi_n(v) \right\|^2_{L^2[0,1]} du = \sum_{n=N+1}^{N_0} \lambda_n. \tag{86}$$

Here, the equality holds as a consequence of Beppo Levi's monotone convergence lemma and Lebesgue's dominated convergence theorem, and one has $\sum_{n\in\mathcal{N}} \lambda_n < +\infty$. In fact, for $N_0 = +\infty$, let $S_n = \sum_{k=1}^n \lambda_k$, we get S_n increasing (since $\lambda_k \geq 0$ for all $k \in \mathcal{N}$). By (80) and (84), we have

$$S_n = \sum_{k=1}^n \int_0^1 (\mathcal{R}\phi_k)^2(u)du = \int_0^1 \sum_{k=1}^n \langle R_u, \phi_k\rangle^2_{L^2[0,1]} du \leq \int_0^1 \|R_u\|^2_{L^2[0,1]} du = \|R\|^2_{L^2([0,1]^2)}, \tag{87}$$

which gives that S_n converges as $n \to \infty$. The proof is completed. □

Remark D.1. *Let $\varphi_n(u) := \langle R(u,v), \phi_n(v)\rangle_{L^2[0,1]}$, then we have the POD estimate $R(u,v) \approx \sum_n \varphi_n(u)\phi_n(v)$, where the error is characterised by the tail sum of eigenvalues of the POD operator.*

Recall the POD operator defined in (67). We similarly define

$$\tilde{\mathcal{K}}: \ \phi(v) \mapsto \int_0^1 \int_0^1 \tilde{R}(u,v)\phi(v)dv \cdot \tilde{R}(u,v)du, \quad \phi(v) \in L^2[0,1], \tag{88}$$

where $\tilde{R}$ is defined as (61), i.e. the approximator constructed in the general approximation theorem before. Obviously, as a polynomial, $\tilde{R} \in L^\infty([0,1]^2)$. Hence, by Proposition D.1, $\tilde{\mathcal{K}}$ is also linear, bounded, compact, self-adjoint and non-negative. In addition, according to Theorem D.1, we have the following POD estimate

$$\int_0^1 \left\| \tilde{R}(u,v) - \sum_{n=1}^N \left\langle \tilde{R}(u,v), \tilde{\phi}_n(v)\right\rangle_{L^2[0,1]} \tilde{\phi}_n(v) \right\|^2_{L^2[0,1]} du = \sum_{n=N+1}^{\tilde{N}_0} \tilde{\lambda}_n, \tag{89}$$

where $\{\tilde{\lambda}_n\}_{n=1}^{\tilde{N}_0}$ are eigenvalues of $\tilde{\mathcal{K}}$ satisfying $\tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq \cdots \geq \tilde{\lambda}_n \geq \cdots \geq 0$, and $\{\tilde{\phi}_n\}_{n=1}^{\tilde{N}_0} \subset L^2[0,1]$ are the corresponding orthonormal eigenfunctions, i.e. $\tilde{\mathcal{K}}\tilde{\phi}_n = \tilde{\lambda}_n\tilde{\phi}_n$, $\tilde{\lambda}_n > 0$ for $n \in \{1, 2, \cdots, \tilde{N}_0\}$.

Lemma D.1. *$\tilde{\mathcal{K}}$ is a finite-rank operator. That is, $\tilde{N}_0 \leq \bar{m} = \min\{m_E, m_D\} < \infty$.*

Proof. Let $\tilde{\varphi}_n(u) := \left\langle \tilde{R}(u,v), \tilde{\phi}_n(v) \right\rangle_{L^2[0,1]}$. We first show that both $\tilde{\varphi}_n(u)$ and $\tilde{\phi}_n(v)$ are polynomials. In fact, since $\tilde{R}(u,v) = \sum_{i=1}^{\bar{m}} \sum_{j=1}^{\bar{m}} \tilde{r}_{ij} u^i v^j$, we have

$$\begin{aligned} \left| \frac{\partial^k}{\partial u^k} \tilde{R}(u,v) \right| &= \left| \sum_{i=k}^{\bar{m}} \sum_{j=1}^{\bar{m}} \tilde{r}_{ij} \frac{i!}{(i-k)!} u^{i-k} v^j \right| \\ &\leq \sum_{i=k}^{\bar{m}} \sum_{j=1}^{\bar{m}} |\tilde{r}_{ij}| \frac{i!}{(i-k)!} \triangleq C_1(k, \bar{m}), \quad k = 1, 2, \cdots, \end{aligned} \tag{90}$$

with the convention that the summation is zero if the subscript is larger than the superscript, i.e. $C_1(k, \bar{m}) = 0$ for any $k > \bar{m}$. Let $C_1(\bar{m}) := \max\{\|\tilde{R}\|_{L^\infty([0,1]^2)}, \max_{1 \leq k \leq \bar{m}} C_1(k, \bar{m})\}$, then we have

$$\left| \frac{\partial^k}{\partial u^k} \tilde{R}(u,v) \tilde{\phi}_n(v) \right| \leq C_1(\bar{m}) |\tilde{\phi}_n(v)| \in L^2[0,1] \subset L^1[0,1], \quad k = 0, 1, \cdots. \tag{91}$$

According to Lebesgue's dominated convergence theorem, we get by induction that

$$\frac{d^k}{du^k} \tilde{\varphi}_n(u) = \frac{d^k}{du^k} \int_0^1 \tilde{R}(u,v) \tilde{\phi}_n(v) dv = \int_0^1 \frac{\partial^k}{\partial u^k} \tilde{R}(u,v) \tilde{\phi}_n(v) dv, \quad k = 0, 1, \cdots. \tag{92}$$

Similarly, we get

$$\begin{aligned} \left| \frac{\partial^k}{\partial v^k} \tilde{R}(u,v) \right| &= \left| \sum_{i=1}^{\bar{m}} \sum_{j=k}^{\bar{m}} \tilde{r}_{ij} u^i \frac{j!}{(j-k)!} v^{j-k} \right| \\ &\leq \sum_{i=1}^{\bar{m}} \sum_{j=k}^{\bar{m}} |\tilde{r}_{ij}| \frac{j!}{(j-k)!} \triangleq C_2(k, \bar{m}), \quad k = 1, 2, \cdots, \end{aligned} \tag{93}$$

and $C_2(k, \bar{m}) = 0$ for any $k > \bar{m}$. Let $C_2(\bar{m}) := \max\{\|\tilde{R}\|_{L^\infty([0,1]^2)}, \max_{1 \leq k \leq \bar{m}} C_2(k, \bar{m})\}$, then we have

$$\begin{aligned} \left| \frac{\partial^k}{\partial v^k} \tilde{R}(u,v) \int_0^1 \tilde{R}(u,v) \tilde{\phi}_n(v) dv \right| &\leq C_2(\bar{m}) \|\tilde{R}\|_{L^\infty([0,1]^2)} \|\tilde{\phi}_n\|_{L^1[0,1]} \\ &\leq C_2^2(\bar{m}) \|\tilde{\phi}_n\|_{L^2[0,1]} = C_2^2(\bar{m}) \subset L^1[0,1], \quad k = 0, 1, \cdots. \end{aligned} \tag{94}$$

According to Lebesgue's dominated convergence theorem, we get by induction that

$$\frac{d^k}{dv^k} \tilde{\phi}_n(v) = \frac{1}{\tilde{\lambda}_n} \int_0^1 \frac{\partial^k}{\partial v^k} \tilde{R}(u,v) \int_0^1 \tilde{R}(u,v) \tilde{\phi}_n(v) dv du, \quad k = 0, 1, \cdots. \tag{95}$$

That is, $\tilde{\varphi}_n, \tilde{\phi}_n \in C^\infty[0,1]$ for any $n = 1, 2, \cdots, \tilde{N}_0$. Since $\tilde{R}(u,0) = \tilde{R}(0,v) = 0$ for $u, v \in [0,1]$, we get $\tilde{\varphi}_n(0) = \tilde{\phi}_n(0) = 0$. Furthermore, we have $\frac{d^k}{du^k} \tilde{\varphi}_n(u) = \frac{d^k}{dv^k} \tilde{\phi}_n(v) = 0$ for $k > \bar{m}$, hence $\tilde{\varphi}_n, \tilde{\phi}_n \in \mathcal{P}_{\bar{m}}$. Since $\{\tilde{\phi}_n\}_{n=1}^{\tilde{N}_0} \subset L^2[0,1]$ are orthonormal, we must have $\tilde{N}_0 \leq \bar{m} < \infty$. [4] The proof is completed. □

[4] In fact, for any $p \in \mathcal{P}_{\bar{m}}$ with $p(0) = 0$, we have $p \in \text{span}\{v, v^2, \cdots, v^{\bar{m}}\}$. Through a standard Schmidt-orthogonalization, we can get $e_k(v) \in \mathcal{P}_k$, $k = 1, 2, \cdots, \bar{m}$, such that $\langle e_i, e_j \rangle_{L^2[0,1]} = \delta_{ij}$, and $p \in \text{span}\{e_1(v), e_2(v), \cdots, e_{\bar{m}}(v)\}$. That is, if $\langle p, q \rangle_{L^2[0,1]} = 0$ for some p, $q \in \mathcal{P}_{\bar{m}}$ with $p(0) = 0$, $q(0) = 0$, then their coordinates under the basis $\{e_k\}_{k=1}^{\bar{m}}$ are orthogonal. Hence, the $\tilde{N}_0$ orthogonal $\bar{m}$-dimensional coordinates here leads to at most $\bar{m}$ non-zeros.

D.2 Approximation rates

Perturbation of Eigenvalues. First, we need to bound the gap between the eigenvalues $\{\tilde{\lambda}_n\}_{n=1}^{\tilde{N}_0}$ and $\{\lambda_n\}_{n=1}^{N_0}$ (corresponding to the function R defined in (50)). The following theorem is necessary.

Theorem D.2 (Courant–Fischer–Weyl min-max principle)**.** *Let $\mathcal{B}$ be a compact, self-adjoint operator on a Hilbert space $\mathcal{H}$, whose positive eigenvalues are listed in a decreasing order $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_k \geq \cdots > 0$. Then*

$$\max_{\mathcal{S}_k} \min_{x \in \mathcal{S}_k,\, \|x\|_{\mathcal{H}}=1} \langle \mathcal{B}x, x\rangle_{\mathcal{H}} = \mu_k, \tag{96}$$

where $\mathcal{S}_k \subset \mathcal{H}$ is any k-dimensional linear subspace.

Based on it, we have the following lemma to characterise the perturbation of singular values.

Lemma D.2. *For any $R_1, R_2 \in L^\infty([0,1]^2)$, we have the estimate*

$$\left|\sqrt{\lambda_k^{R_1}} - \sqrt{\lambda_k^{R_2}}\right| \leq \|R_1 - R_2\|_{L^2([0,1]^2)}. \tag{97}$$

Proof. According to Theorem D.2 and by (75), we have

$$\lambda_k^R = \max_{\mathcal{S}_k} \min_{\phi \in \mathcal{S}_k,\, \|\phi\|_{L^2[0,1]}=1} \langle \mathcal{K}_R\phi, \phi\rangle_{L^2[0,1]} = \max_{\mathcal{S}_k} \min_{\phi \in \mathcal{S}_k,\, \|\phi\|_{L^2[0,1]}=1} \|\mathcal{R}_R\phi\|^2_{L^2[0,1]}. \tag{98}$$

Note that $\mathcal{R}_{R_1} - \mathcal{R}_{R_2} = \mathcal{R}_{R_1 - R_2}$ and by (70), we have

$$\begin{aligned}
\sqrt{\lambda_k^{R_1}} &= \max_{\mathcal{S}_k} \min_{\phi \in \mathcal{S}_k,\, \|\phi\|_{L^2[0,1]}=1} \|\mathcal{R}_{R_1}\phi\|_{L^2[0,1]} \\
&\leq \max_{\mathcal{S}_k} \min_{\phi \in \mathcal{S}_k,\, \|\phi\|_{L^2[0,1]}=1} \big(\|(\mathcal{R}_{R_1} - \mathcal{R}_{R_2})\phi\|_{L^2[0,1]} + \|\mathcal{R}_{R_2}\phi\|_{L^2[0,1]}\big) \\
&\leq \max_{\mathcal{S}_k} \min_{\phi \in \mathcal{S}_k,\, \|\phi\|_{L^2[0,1]}=1} \big(\|\mathcal{R}_{R_1 - R_2}\| + \|\mathcal{R}_{R_2}\phi\|_{L^2[0,1]}\big) \\
&\leq \max_{\mathcal{S}_k} \min_{\phi \in \mathcal{S}_k,\, \|\phi\|_{L^2[0,1]}=1} \|\mathcal{R}_{R_2}\phi\|_{L^2[0,1]} + \|R_1 - R_2\|_{L^2([0,1]^2)} \\
&= \sqrt{\lambda_k^{R_2}} + \|R_1 - R_2\|_{L^2([0,1]^2)},
\end{aligned} \tag{99}$$

and similarly,

$$\begin{aligned}
\sqrt{\lambda_k^{R_2}} &= \max_{\mathcal{S}_k} \min_{\phi \in \mathcal{S}_k,\, \|\phi\|_{L^2[0,1]}=1} \|\mathcal{R}_{R_2}\phi\|_{L^2[0,1]} \\
&\leq \max_{\mathcal{S}_k} \min_{\phi \in \mathcal{S}_k,\, \|\phi\|_{L^2[0,1]}=1} \big(\|(\mathcal{R}_{R_2} - \mathcal{R}_{R_1})\phi\|_{L^2[0,1]} + \|\mathcal{R}_{R_1}\phi\|_{L^2[0,1]}\big) \\
&\leq \max_{\mathcal{S}_k} \min_{\phi \in \mathcal{S}_k,\, \|\phi\|_{L^2[0,1]}=1} \big(\|\mathcal{R}_{R_2 - R_1}\| + \|\mathcal{R}_{R_1}\phi\|_{L^2[0,1]}\big) \\
&\leq \max_{\mathcal{S}_k} \min_{\phi \in \mathcal{S}_k,\, \|\phi\|_{L^2[0,1]}=1} \|\mathcal{R}_{R_1}\phi\|_{L^2[0,1]} + \|R_2 - R_1\|_{L^2([0,1]^2)} \\
&= \sqrt{\lambda_k^{R_1}} + \|R_1 - R_2\|_{L^2([0,1]^2)},
\end{aligned} \tag{100}$$

which completes the proof. □

Now we are ready to derive the final estimate.

Proof of Theorem 4.3. By Lemma D.2 and (60), we get

$$\left|\sqrt{\lambda_k} - \sqrt{\tilde{\lambda}_k}\right| \leq \|R - \tilde{R}\|_{L^2([0,1]^2)} \leq \|R - \tilde{R}\|_{L^\infty([0,1]^2)} \leq \frac{C(\alpha)\gamma}{\bar{m}^\alpha}. \tag{101}$$

Combining (60), (89) and (101) gives that

$$
\begin{aligned}
&\frac{1}{c_0^2}\left\|\rho(t,s)-\sum_{n=1}^{N} e^{-\frac{t}{c_0}}\tilde{\varphi}_n(e^{-\frac{t}{c_0}})\cdot e^{-\frac{s}{c_0}}\tilde{\phi}_n(e^{-\frac{s}{c_0}})\right\|_{L^1([0,\infty)^2)}\\
&=\left\|R(u,v)-\sum_{n=1}^{N}\tilde{\varphi}_n(u)\tilde{\phi}_n(v)\right\|_{L^1([0,1]^2)}\\
&\le \left\|R(u,v)-\tilde{R}(u,v)\right\|_{L^1([0,1]^2)}+\left\|\tilde{R}(u,v)-\sum_{n=1}^{N}\tilde{\varphi}_n(u)\tilde{\phi}_n(v)\right\|_{L^1([0,1]^2)}\\
&\le \left\|R(u,v)-\tilde{R}(u,v)\right\|_{L^\infty([0,1]^2)}+\left\|\tilde{R}(u,v)-\sum_{n=1}^{N}\tilde{\varphi}_n(u)\tilde{\phi}_n(v)\right\|_{L^2([0,1]^2)}\\
&\le \frac{C(\alpha)\gamma}{\bar{m}^\alpha}+\sqrt{\sum_{n=N+1}^{\tilde{N}_0}\tilde{\lambda}_n}\le \frac{C(\alpha)\gamma}{\bar{m}^\alpha}+\sqrt{\sum_{n=N+1}^{\tilde{N}_0}|\tilde{\lambda}_n-\lambda_n|+\sum_{n=N+1}^{\tilde{N}_0}\lambda_n}\\
&\le \frac{C(\alpha)\gamma}{\bar{m}^\alpha}+\sqrt{\sum_{n=N+1}^{\tilde{N}_0}\lambda_n}+\sqrt{\sum_{n=N+1}^{\tilde{N}_0}\left|\sqrt{\tilde{\lambda}_n}-\sqrt{\lambda_n}\right|\left|\sqrt{\tilde{\lambda}_n}+\sqrt{\lambda_n}\right|}\\
&\le \frac{C(\alpha)\gamma}{\bar{m}^\alpha}+\sqrt{\sum_{n=N+1}^{\tilde{N}_0}\lambda_n}+\sqrt{\sum_{n=N+1}^{\tilde{N}_0}\left|\sqrt{\tilde{\lambda}_n}-\sqrt{\lambda_n}\right|^2}+\sqrt{2}\sqrt{\sum_{n=N+1}^{\tilde{N}_0}\sqrt{\lambda_n}\left|\sqrt{\tilde{\lambda}_n}-\sqrt{\lambda_n}\right|}\\
&\lesssim C(\alpha)\gamma\left\{\left(1+\sqrt{\tilde{N}_0-N}\right)\cdot\frac{1}{\bar{m}^\alpha}+\sqrt{\sum_{n=N+1}^{\tilde{N}_0}\lambda_n}+\sqrt{\sum_{n=N+1}^{\tilde{N}_0}\sqrt{\lambda_n}}\cdot\frac{1}{\bar{m}^{\alpha/2}}\right\}. \qquad (102)
\end{aligned}
$$

For the corresponding parameters, recall that $\tilde{\varphi}_n, \tilde{\phi}_n \in \mathcal{P}_{\bar{m}}$ with $\tilde{\varphi}_n(0)=\tilde{\phi}_n(0)=0$, we can write

$$
\tilde{\varphi}_n(u)=\sum_{i=1}^{\bar{m}} P_{in}u^i, \quad \tilde{\phi}_n(v)=\sum_{j=1}^{\bar{m}} Q_{nj}v^j \qquad (103)
$$

for any $n=1,2,\cdots,\tilde{N}_0$. Let

$$
c=\mathbf{1}_{\bar{m}}, \quad u=\mathbf{1}_{\bar{m}}, \qquad (104)
$$

$$
V=-\mathrm{diag}(2,3,\cdots,\bar{m}+1)/c_0, \quad W=-\mathrm{diag}(2,3,\cdots,\bar{m}+1)/c_0, \qquad (105)
$$

$$
M=PQ \text{ with } P=[P_{in}]\in\mathbb{R}^{\bar{m}\times N},\ Q=[Q_{nj}]\in\mathbb{R}^{N\times\bar{m}}, \qquad (106)
$$

then we have

$$
\begin{aligned}
c^\top e^{Vt}Me^{Ws}u=c^\top e^{Vt}P\cdot Qe^{Ws}u&=\sum_{n=1}^{N}\sum_{i=1}^{\bar{m}} e^{-\frac{i+1}{c_0}t}P_{in}\cdot\sum_{j=1}^{\bar{m}}Q_{nj}e^{-\frac{j+1}{c_0}s}\\
&=\sum_{n=1}^{N} e^{-\frac{t}{c_0}}\tilde{\varphi}_n(e^{-\frac{t}{c_0}})\cdot e^{-\frac{s}{c_0}}\tilde{\phi}_n(e^{-\frac{s}{c_0}}). \qquad (107)
\end{aligned}
$$

Plugging this into (102) gives

$$
\begin{aligned}
\|\boldsymbol{H}-\widehat{\boldsymbol{H}}\| &\le \|\rho-\hat{\rho}\|_{L^1} \qquad (108)\\
&=\left\|\rho(t,s)-c^\top e^{Vt}Me^{Ws}u\right\|_{L^1([0,\infty)^2)}\\
&\lesssim \frac{C(\alpha)\gamma}{\beta^2}\left\{\left(1+\sqrt{\tilde{N}_0-N}\right)\cdot\frac{1}{\bar{m}^\alpha}+\sqrt{\sum_{n=N+1}^{\tilde{N}_0}\lambda_n}+\sqrt{\sum_{n=N+1}^{\tilde{N}_0}\sqrt{\lambda_n}}\cdot\frac{1}{\bar{m}^{\alpha/2}}\right\}, \qquad (109)
\end{aligned}
$$

and the number of trainable parameters is $2N\bar{m}$. Together with Lemma D.1, the proof is completed. □

Extension to multi-dimensional inputs. The above results can be naturally extended to the general case where a d-dimensional input is given ($\forall d \in \mathbb{N}_+$). In fact, let

$$\left\|\rho_i(t,s) - c^\top e^{Vt} M_i e^{Ws} u\right\|_{L^1([0,\infty)^2)} \lesssim \epsilon, \quad i = 1, 2, \cdots, d, \tag{110}$$

for some $0 < \epsilon \ll 1$, then we take

$$M = (M_1, M_2, \cdots, M_d) \in \mathbb{R}^{\bar{m} \times d\bar{m}}, \tag{111}$$

$$W = \mathrm{diag}(W, W, \cdots, W) \in \mathbb{R}^{d\bar{m} \times d\bar{m}}, \quad U = \mathrm{diag}(u, u, \cdots, u) \in \mathbb{R}^{d\bar{m} \times d}, \tag{112}$$

and have

$$\begin{aligned} c^\top e^{Vt} M e^{Ws} U &= c^\top e^{Vt} \cdot (M_1, M_2, \cdots, M_d) \cdot \mathrm{diag}(e^{Ws}, e^{Ws}, \cdots, e^{Ws}) \cdot \mathrm{diag}(u, u, \cdots, u) \\ &= (c^\top e^{Vt} M_1, c^\top e^{Vt} M_2, \cdots, c^\top e^{Vt} M_d) \cdot \mathrm{diag}(e^{Ws} u, e^{Ws} u, \cdots, e^{Ws} u) \\ &= (c^\top e^{Vt} M_1 e^{Ws} u, c^\top e^{Vt} M_2 e^{Ws} u, \cdots, c^\top e^{Vt} M_d e^{Ws} u). \end{aligned} \tag{113}$$

If the form $PQ(= M)$ is required, it is sufficient to take $M_i = P_i Q_i$, $i = 1, 2, \cdots, d$, and

$$P = (P_1, P_2, \cdots, P_d) \in \mathbb{R}^{\bar{m} \times dN}, \ Q = \mathrm{diag}(Q_1, Q_2, \cdots, Q_d) \in \mathbb{R}^{dN \times d\bar{m}}. \tag{114}$$

Therefore, we obtain

$$\sum_{i=1}^{d} \left\|\rho_i(t,s) - \left[c^\top e^{Vt} M e^{Ws} U\right]_i\right\|_{L^1([0,\infty)^2)} \lesssim d\epsilon, \tag{115}$$

with the number of parameters increased by d-times compared to the corresponding one-dimensional setting.

D.3 Case analysis

Bounds under different settings. Now we make the comparison between (108) and (64). Recall that $\{\lambda_n\}_{n=1}^{N_0}$ is a positive decreased sequence (with $\lim_{n\to\infty} \lambda_n = 0$ and $\sum_{n=1}^{\infty} \lambda_n \le \|R\|^2_{L^2([0,1]^2)}$ by (87), if $N_0 = +\infty$), and $\tilde{N}_0 \le \bar{m}$, we have the following cases.

- if $\tilde{N}_0 = o(\bar{m})$, we set $N = \tilde{N}_0$ in (108) and get the same bound as (64), but the number of parameters is only $\mathcal{O}(\tilde{N}_0 \bar{m}) = o(\bar{m}^2)$;
- if $\tilde{N}_0 = \mathcal{O}(\bar{m})$, then (108) implies that

$$\begin{aligned} &\left\|\rho(t,s) - c^\top e^{Vt} M e^{Ws} u\right\|_{L^1([0,\infty)^2)} \\ \lesssim &\frac{C(\alpha)\gamma}{\beta^2} \left\{ \left(1 + \sqrt{\bar{m} - N}\right) \cdot \frac{1}{\bar{m}^\alpha} + \sqrt{\sum_{n=N+1}^{\bar{m}} \lambda_n} + \sqrt{\sum_{n=N+1}^{\bar{m}} \sqrt{\lambda_n}} \cdot \frac{1}{\bar{m}^{\alpha/2}} \right\} \\ \lesssim &\frac{C(\alpha)\gamma}{\beta^2} \left\{ \frac{1}{\bar{m}^{\alpha - \frac{1}{2}}} + \sqrt{\sum_{n=N+1}^{\bar{m}} \lambda_n} + \sqrt{\sum_{n=N+1}^{\bar{m}} \sqrt{\lambda_n}} \cdot \frac{1}{\bar{m}^{\alpha/2}} \right\}. \end{aligned} \tag{116}$$

  We are supposed to require that

$$\begin{aligned} &\frac{1}{\bar{m}^{\alpha - \frac{1}{2}}} \gtrsim \max\left\{ \sqrt{\sum_{n=N+1}^{\bar{m}} \lambda_n}, \sqrt{\sum_{n=N+1}^{\bar{m}} \sqrt{\lambda_n}} \cdot \frac{1}{\bar{m}^{\alpha/2}} \right\} \\ \Leftrightarrow &\sum_{n=N+1}^{\bar{m}} \lambda_n \lesssim \frac{1}{\bar{m}^{2\alpha - 1}}, \quad \sum_{n=N+1}^{\bar{m}} \sqrt{\lambda_n} \lesssim \frac{1}{\bar{m}^{\alpha - 1}}. \end{aligned} \tag{117}$$

  We give the following typical examples to illustrate sufficient conditions to guarantee (117).

– If $\lambda_n = \mathcal{O}(n^{-r})$ with $r > 2\alpha + 1 \geq 3$ ($\alpha \in \mathbb{N}_+$), since

$$\sum_{n=N+1}^{\bar{m}} n^{-r} \leq \int_N^{\bar{m}} x^{-r} dx = \frac{1}{r-1}\left(\frac{1}{N^{r-1}} - \frac{1}{\bar{m}^{r-1}}\right), \quad \forall r > 1, \tag{118}$$

we get by (117) that

$$\sum_{n=N+1}^{\bar{m}} \lambda_n \lesssim \sum_{n=N+1}^{\bar{m}} n^{-r} \lesssim \frac{1}{N^{r-1}} \lesssim \frac{1}{\bar{m}^{2\alpha-1}} \Leftrightarrow N \gtrsim \bar{m}^{\frac{2\alpha-1}{r-1}}, \tag{119}$$

$$\sum_{n=N+1}^{\bar{m}} \sqrt{\lambda_n} \lesssim \sum_{n=N+1}^{\bar{m}} n^{-\frac{r}{2}} \lesssim \frac{1}{N^{\frac{r}{2}-1}} \lesssim \frac{1}{\bar{m}^{\alpha-1}} \Leftrightarrow N \gtrsim \bar{m}^{\frac{\alpha-1}{\frac{r}{2}-1}}. \tag{120}$$

Assume that $N \sim \bar{m}^{\delta}$ with $\delta \in [0, 1)$, then by (119) and (120), we require $\delta \geq \max\{\frac{2\alpha-1}{r-1}, \frac{\alpha-1}{\frac{r}{2}-1}\}$, i.e.

$$r \geq \max\left\{\frac{2\alpha-1}{\delta} + 1,\, 2\left(\frac{\alpha-1}{\delta} + 1\right)\right\} = \frac{2\alpha-1}{\delta} + 1. \tag{121}$$

Meanwhile, the POD-estimate (116) achieves an accuracy scaling like $(1/\bar{m})^{\alpha-\frac{1}{2}}$ with $\mathcal{O}(\bar{m}^{1+\delta})$ parameters, while under the same capacity, the accuracy of (64) scales like $(1/\bar{m})^{\frac{\alpha(1+\delta)}{2}}$. The former beats the latter if $\alpha - \frac{1}{2} > \frac{\alpha(1+\delta)}{2}$, i.e. $\delta < 1 - 1/\alpha$ ($\alpha \geq 2$). When $\delta = 1 - 1/\alpha$, (121) becomes $r \geq \max\{\frac{2\alpha^2-1}{\alpha-1}, 2(\alpha+1)\} = \frac{2\alpha^2-1}{\alpha-1}$. That is to say, $r^* := \frac{2\alpha^2-1}{\alpha-1}$ can be viewed as an upper bound of the critical point where the two estimates are comparable. When $r > r^*$, the POD-estimate outperforms the non-POD-estimate, and this effect gets more notable with r increasing. In fact, when $r > r^*$, we take $N = \bar{m}^{\frac{2\alpha-1}{r-1}} > \bar{m}^{\frac{\alpha-1}{\frac{r}{2}-1}}$ (hence satisfying (119) and (120)), which gives an $\mathcal{O}((1/\bar{m})^{\alpha-\frac{1}{2}})$ approximation error with $\mathcal{O}(\bar{m}^{1+\frac{2\alpha-1}{r-1}})$ parameters for the POD-estimate, while $\mathcal{O}(\bar{m}^{2-\frac{1}{\alpha}})$ trainable parameters are needed to achieve the same accuracy using the non-POD-estimate. Since $\frac{2\alpha-1}{r-1} < 1 - \frac{1}{\alpha}$, we get that the POD-estimate outperforms the non-POD-estimate. When $r \to +\infty$, the number of trainable parameters for the POD-estimate is $\mathcal{O}(\bar{m})$, much better than the non-POD-estimate.

Remark D.2. *Let*

$$K(v, w) := \int_0^1 R(u, v)R(u, w)du, \quad v, w \in [0, 1], \tag{122}$$

we get

$$(\mathcal{K}\phi)(w) = \int_0^1 K(v, w)\phi(v)dv, \quad \phi \in L^2[0, 1]. \tag{123}$$

Recall that $R \in C^{\alpha}([0, 1]^2) \subset L^{\infty}([0, 1]^2)$, we get $K \in L^{\infty}([0, 1]^2) \subset L^2([0, 1]^2)$, and hence $\mathcal{K}$ can be also viewed as a Hilbert-Schmidt integral operator with the kernel K, where K is obviously symmetric as $K(v, w) = K(w, v)$, and positive definite since $0 \leq \langle \mathcal{K}\phi, \phi\rangle_{L^2[0,1]} = \int_0^1 \int_0^1 K(v, w)\phi(v)\phi(w)dvdw$, $\phi \in L^2[0, 1]$ by (75). Since the derivatives of R (up to α-order) are continuous on $[0, 1]^2$ (hence bounded and integrable), we get $K \in C^{\alpha,\alpha}([0, 1]^2)$ (α-differentiable for both arguments). According to Chang & Ha (1999) (Theorem 1), we have

$$\sum_{n=N+2\alpha+1}^{\infty} \lambda_n \lesssim N^{-\alpha}, \quad \forall N \in \mathbb{N}_+ \Rightarrow \lambda_n = \mathcal{O}(n^{-(\alpha+1)}), \quad n \to \infty. \tag{124}$$

That is to say, this general setting (only assume smoothness of the kernel) can not guarantee the sufficient condition provided here ($\lambda_n = \mathcal{O}(n^{-r})$ with $r > \frac{2\alpha^2-1}{\alpha-1}$), i.e. the point that POD-estimate outperforms non-POD-estimate requires a faster decay of the eigenvalues.

– If $\lambda_n = \mathcal{O}(e^{-\omega n})$ with $\omega > 0$, we get by (117) that

$$\sum_{n=N+1}^{\bar{m}} \lambda_n \lesssim \sum_{n=N+1}^{\infty} e^{-\omega n} = \frac{e^{-\omega(N+1)}}{1-e^{-\omega}} \lesssim \frac{1}{\bar{m}^{2\alpha-1}} \Leftrightarrow N \gtrsim \frac{1}{\omega} \ln\left(\frac{\bar{m}^{2\alpha-1}}{1-e^{-\omega}}\right) - 1, \tag{125}$$

$$\sum_{n=N+1}^{\bar{m}} \sqrt{\lambda_n} \lesssim \sum_{n=N+1}^{\infty} e^{-\frac{\omega}{2}n} = \frac{e^{-\frac{\omega}{2}(N+1)}}{1-e^{-\frac{\omega}{2}}} \lesssim \frac{1}{\bar{m}^{\alpha-1}} \Leftrightarrow N \gtrsim \frac{2}{\omega} \ln\left(\frac{\bar{m}^{\alpha-1}}{1-e^{-\frac{\omega}{2}}}\right) - 1. \tag{126}$$

That is to say, for any $\alpha \in \mathbb{N}_+$, $\omega > 0$, we have $N \sim (2\alpha - 1)\ln \bar{m}$. This implies an $\mathcal{O}((1/\bar{m})^{\alpha-\frac{1}{2}})$ approximation error with $\mathcal{O}(\bar{m}\ln\bar{m})$ parameters, while $\mathcal{O}(\bar{m}^{2-\frac{1}{\alpha}})$ parameters are needed to achieve the same accuracy using the non-POD-estimate.

– If $N_0 < +\infty$ (i.e. $\mathcal{K}$ is a finite rank operator by (77)), one can just take $N = N_0$ and get by (116) an $\mathcal{O}((1/\bar{m})^{\alpha-\frac{1}{2}})$ approximation error. For $\bar{m} \in \mathbb{N}_+$ sufficiently large such that $\bar{m} \sim N_0^{\kappa}$ for some $\kappa \gg 1$, the number of trainable parameters is $\mathcal{O}(N\bar{m}) = \mathcal{O}(N_0^{\kappa+1})$, while $\mathcal{O}\left(N_0^{\kappa(2-\frac{1}{\alpha})}\right)$ parameters are needed to achieve the same accuracy using the non-POD-estimate. Obviously, if $\alpha \geq 2$, we get $\kappa(2-\frac{1}{\alpha}) \geq \frac{3}{2}\kappa \gg \kappa + 1$.

Eigenvalue approximation. One can apply (97) in Lemma D.2 to estimate the eigenvalues $\{\lambda_n\}_{n=1}^{N_0}$, where $R_1 = R$ and R_2 is taken as some approximator of R, say $\hat{R}$. Let $\hat{\lambda} := \lambda^{\hat{R}}$, we recall (97) as

$$\left|\sqrt{\lambda_k} - \sqrt{\hat{\lambda}_k}\right| \leq \|R - \hat{R}\|_{L^2([0,1]^2)}. \tag{127}$$

We provide a naive method here. That is, one can take $\hat{R}$ as a piece-wise constant approximation of R. Fix any $n \in \mathbb{N}_+$. Let $\mathcal{I}_0 := \{0\}$, $\mathcal{I}_i := (\frac{i-1}{n}, \frac{i}{n}]$ for $i = 1, 2, \cdots, n$, then $\{\mathcal{I}_i \times \mathcal{I}_j\}_{i,j=0}^n$ is the uniform partition over $[0,1]^2$. Let $\hat{R}(u,v) := \sum_{i,j=0}^n R(\frac{i}{n}, \frac{j}{n}) \mathbf{1}_{\mathcal{I}_i \times \mathcal{I}_j}(u,v)$. Then for any $\phi \in L^2[0,1]$, if $w \in \mathcal{I}_k$, $k = 0, 1, \cdots, n$, we have

$$\begin{aligned}
(\mathcal{K}_{\hat{R}}\phi)(w) &= \int_0^1 \int_0^1 \hat{R}(u,v)\phi(v)dv \cdot \hat{R}(u,w)du \\
&= \sum_{i=0}^n \sum_{j=0}^n \sum_{i'=0}^n \sum_{j'=0}^n R\left(\frac{i}{n}, \frac{j}{n}\right) R\left(\frac{i'}{n}, \frac{j'}{n}\right) \int_0^1 \int_0^1 \mathbf{1}_{\mathcal{I}_i \times \mathcal{I}_j}(u,v)\phi(v)dv \cdot \mathbf{1}_{\mathcal{I}_{i'} \times \mathcal{I}_{j'}}(u,w)du \\
&= \sum_{i=0}^n \sum_{j=0}^n R\left(\frac{i}{n}, \frac{j}{n}\right) R\left(\frac{i}{n}, \frac{k}{n}\right) \int_{\mathcal{I}_i} \int_{\mathcal{I}_j} \mathbf{1}_{\mathcal{I}_i \times \mathcal{I}_j}(u,v)\phi(v)dv \cdot \mathbf{1}_{\mathcal{I}_i \times \mathcal{I}_k}(u,w)du \\
&= \sum_{i=0}^n \sum_{j=0}^n R\left(\frac{i}{n}, \frac{j}{n}\right) R\left(\frac{i}{n}, \frac{k}{n}\right) \cdot \frac{1}{n} \int_{\mathcal{I}_j} \phi(v)dv,
\end{aligned} \tag{128}$$

which is a constant only related to k. That is, $\mathcal{K}_{\hat{R}}\phi$ is a piece-wise constant function, i.e. $\mathcal{K}_{\hat{R}}\phi \in \operatorname{span}\{\mathbf{1}_{\mathcal{I}_k}\}_{k=0}^n$, or $\operatorname{range}(\mathcal{K}_{\hat{R}}) \subset \operatorname{span}\{\mathbf{1}_{\mathcal{I}_k}\}_{k=0}^n$. Obviously, $\{\mathbf{1}_{\mathcal{I}_k}\}_{k=0}^n$ is an orthogonal set, which implies that the operator $\mathcal{K}_{\hat{R}}$ is of finite rank at most $n+1$. Let $\hat{R}_{ij} := \frac{1}{n} R(\frac{i-1}{n}, \frac{j-1}{n})$, $i, j = 1, 2, \cdots, n+1$, $\{\sigma_k\}_{k=1}^{n+1}$ be the singular value of $\hat{R} := [\hat{R}_{ij}] \in \mathbb{R}^{(n+1)\times(n+1)}$ and $V \in \mathbb{R}^{(n+1)\times(n+1)}$ with columns as the corresponding right singular vectors. Set $[e_1, e_2, \cdots, e_{n+1}] := [\mathbf{1}_{\mathcal{I}_0}, \mathbf{1}_{\mathcal{I}_1}, \cdots, \mathbf{1}_{\mathcal{I}_n}]\, V$, then by (128), we get for $l = 1, 2, \cdots, n+1$,

$$\begin{aligned}
(\mathcal{K}_{\hat{R}} e_l)(w) &= \sum_{i=0}^n \sum_{j=0}^n R\left(\frac{i}{n}, \frac{j}{n}\right) R\left(\frac{i}{n}, \frac{k}{n}\right) \cdot \frac{1}{n^2} V_{j+1,l} \\
&= \sum_{i=1}^{n+1} \sum_{j=1}^{n+1} \hat{R}_{ij} \hat{R}_{i,k+1} V_{jl} = \left[\hat{R}^\top \hat{R} V_{:,l}\right]_{k+1},
\end{aligned} \tag{129}$$

which gives

$$\mathcal{K}_{\hat{R}} e_l = \sum_{k=0}^{n} \left[\hat{R}^\top \hat{R} V_{:,l}\right]_{k+1} \mathbf{1}_{\mathcal{I}_k} = \sigma_l^2 \sum_{k=0}^{n} V_{k+1,l} \mathbf{1}_{\mathcal{I}_k} = \sigma_l^2 e_l, \tag{130}$$

i.e. the set $\{\sigma_k^2\}_{k=1}^{n+1}$ collects the all the eigenvalues of $\mathcal{K}_{\hat{R}}$, which can be obtained by the SVD of $\hat{R}$.

For the error estimate, it is straightforward to have

$$\begin{aligned}
&\left\| R(u,v) - \hat{R}(u,v) \right\|_{L^2([0,1]^2)}^2 \\
=& \sum_{i=1}^{n} \sum_{j=1}^{n} \int_{\frac{i-1}{n}}^{\frac{i}{n}} \int_{\frac{j-1}{n}}^{\frac{j}{n}} \left| R(u,v) - R\left(\frac{i}{n}, \frac{j}{n}\right) \right|^2 dudv \\
\leq& \sum_{i=1}^{n} \sum_{j=1}^{n} \int_{\frac{i-1}{n}}^{\frac{i}{n}} \int_{\frac{j-1}{n}}^{\frac{j}{n}} \max_{(u,v)\in[0,1]^2} \|\nabla R(u,v)\|_2^2 \cdot \left\| \left(u - \frac{i}{n}, v - \frac{j}{n}\right) \right\|_2^2 dudv \\
\leq& \sum_{i=1}^{n} \sum_{j=1}^{n} \int_{\frac{i-1}{n}}^{\frac{i}{n}} \int_{\frac{j-1}{n}}^{\frac{j}{n}} 2C^2(\alpha)\gamma^2 \cdot \frac{2}{n^2} dudv = \frac{4C^2(\alpha)\gamma^2}{n^2},
\end{aligned} \tag{131}$$

hence by (127), $\left|\sqrt{\lambda_k} - \sigma_k\right| \leq \frac{2C(\alpha)\gamma}{n}$ for any $n \in \mathbb{N}_+$.

E Numerical settings

E.1 Settings of Figure 1

For the inputs, we generate 6 input sequences $\boldsymbol{x}$ using Gaussian i.i.d random variables with length 30. For the outputs (target), we set

$$H_t(\boldsymbol{x}) = \sum_{s=1}^{30} \rho(t,s) x_s. \tag{132}$$

The high rank target has the representation

$$\rho^{\text{high}}(t,s) = \begin{cases} \cos(t), & t = s, \\ 0, & t \neq s, \end{cases} \tag{133}$$

while the low rank target has the representation

$$\rho^{\text{low}}(t,s) = \sum_{n=0}^{99} \frac{1}{n+1} \cos(n\pi t)\cos(n\pi s). \tag{134}$$

E.2 Settings of Figure 2

Consider the target $\boldsymbol{H}$ with the following representation

$$\rho(t,s) = e^{-\frac{t}{c_0}} e^{-\frac{s}{c_0}} R(e^{-\frac{t}{c_0}}, e^{-\frac{s}{c_0}}), \tag{135}$$

where

$$R(u,v) = \sum_{n=1}^{\infty} \sigma_n \varphi_n(u) \phi_n(v), \tag{136}$$

with both $\{\varphi_k\}$ and $\{\phi_k\}$ as orthonormal bases. In this way, we construct a target with singular values $\{\sigma_k\}$.

We use the following linear, width-m RNN encoder-decoder with $m = m_D = m_E$:

$$\widehat{H}_t(\boldsymbol{x}) = \sum_{s=1}^{\tau} c^\top V^t P Q W^{s-1} U x(\tau - s), \tag{137}$$

where $P \in \mathbb{R}^{m\times N}$, $Q \in \mathbb{R}^{N\times m}$. Recall the approximation error is derived as

$$\|\boldsymbol{H} - \widehat{\boldsymbol{H}}\| \lesssim \|R - \hat{R}\|_{L^1([0,1]^2)} \leq \|R - \hat{R}\|_{L^2([0,1]^2)}, \tag{138}$$

where $\hat{R}(u,v) := \sum_{i=1}^{N_0} \tilde{\varphi}_i(u)\tilde{\phi}_i(v) \in \mathcal{P}_m^2$ with $\tilde{\varphi}_n, \tilde{\phi}_n \in \mathcal{P}_m$. In the numerical experiments, we first construct an $R(u,v)$, and then fit it with the polynomial $\hat{R}(u,v)$ using the method of least squares. The norm $\|R - \hat{R}\|_{L^2([0,1]^2)}$ is used to evaluate the approximation error.

In this particular example, we set $m = 128$, $\varphi_n(u) = \sqrt{2}\sin(n\pi u)$ and $\phi_n(v) = \sqrt{2}\sin(n\pi v)$. The singular values are taken as: (a) $\sigma_n = \begin{cases} n^{-\frac{1}{8}}, & n \leq N_0 \\ 0, & n > N_0 \end{cases}$; (b) $\sigma_n = \begin{cases} n^{-1}, & n \leq N_0 \\ 0, & n > N_0 \end{cases}$; (c) $\sigma_n = n^{-2}$, with $N_0 = 2, 4, 6, 8$. As an infinite sum, R is constructed under a finite truncation with the first 50 terms.