Journal of Machine Learning Research 22 (2021) 1-48 Submitted 6/20; Revised 1/21; Published 1/21

Understanding Recurrent Neural Networks Using Nonequilibrium Response Theory

Soon Hoe Lim SOON.HOE.LIM@SU.SE
Nordita
KTH Royal Institute of Technology and Stockholm University
Stockholm 106 91, Sweden

Editor: Sayan Mukherjee

Abstract

Recurrent neural networks (RNNs) are brain-inspired models widely used in machine learning for analyzing sequential data. The present work is a contribution towards a deeper understanding of how RNNs process input signals using the response theory from nonequilibrium statistical mechanics. For a class of continuous-time stochastic RNNs (SRNNs) driven by an input signal, we derive a Volterra type series representation for their output. This representation is interpretable and disentangles the input signal from the SRNN architecture. The kernels of the series are certain recursively defined correlation functions with respect to the unperturbed dynamics that completely determine the output. Exploiting connections of this representation and its implications to rough paths theory, we identify a universal feature – the *response feature*, which turns out to be the signature of tensor product of the input signal and a natural support basis. In particular, we show that SRNNs, with only the weights in the readout layer optimized and the weights in the hidden layer kept fixed and not optimized, can be viewed as kernel machines operating on a reproducing kernel Hilbert space associated with the response feature.

Keywords: Recurrent Neural Networks, Nonequilibrium Response Theory, Volterra Series, Path Signature, Kernel Machines

Contents

1. Introduction

Sequential data arise in a wide range of settings, from time series analysis to natural language processing. In the absence of a mathematical model, it is important to extract useful information from the data to learn the data generating system.

Recurrent neural networks (RNNs) (Hopfield, 1984; McClelland et al., 1986; Elman, 1990) constitute a class of brain-inspired models that are specially designed for and widely used for learning sequential data, in fields ranging from the physical sciences to finance. RNNs are networks of neurons with feedback connections and are arguably biologically more plausible than other adaptive models. In particular, RNNs can use their hidden state (memory) to process variable length sequences of inputs. They are universal approximators of dynamical systems (Funahashi and Nakamura, 1993; Schäfer and Zimmermann, 2006; Hanson and Raginsky, 2020) and can themselves be viewed as a class of open dynamical systems (Sherstinsky, 2020).

Despite their recent innovations and tremendous empirical success in reservoir computing (Herbert, 2001; Maass et al., 2002; Tanaka et al., 2019), deep learning (Sutskever, 2013; Hochreiter and Schmidhuber, 1997; Goodfellow et al., 2016) and neurobiology (Barak, 2017), few studies have focused on the theoretical basis underlying the working mechanism of RNNs. The lack of rigorous analysis limits the usefulness of RNNs in addressing scientific problems and potentially hinders systematic design of the next generation of networks. Therefore, a deep understanding of the mechanism is pivotal to shed light on the properties of large and adaptive architectures, and to revolutionize our understanding of these systems.

In particular, two natural yet fundamental questions that one may ask are:

(Q1) *How does the output produced by RNNs respond to a driving input signal over time?*

(Q2) *Is there a universal mechanism underlying their response?*

One of the main goals of the present work is to address the above questions, using the nonlinear response theory from nonequilibrium statistical mechanics as a starting point, for a stochastic version of continuous-time RNNs (Pineda, 1987; Beer, 1995; Zhang et al., 2014), abbreviated SRNNs, in which the hidden states are injected with a Gaussian white noise. Our approach is cross-disciplinary and adds refreshing perspectives to the existing theory of RNNs.

This paper is organized as follows. In Section 2 we introduce our SRNN model, discuss the related work, and summarize our main contributions. Section 3 contains some preliminaries and core ideas of the paper. There we derive one of the main results of the paper in an informal manner to aid understanding and to gain intution. We present a mathematical formulation of the main results and other results in Section 4. We conclude the paper in Section 5. We postpone the technical details, proofs and further remarks to **SM**.

2. Stochastic Recurrent Neural Networks (SRNNs)

Throughout the paper, we fix a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$, $\mathbb{E}$ denotes expectation with respect to $\mathbb{P}$ and $T > 0$. $C(E, F)$ denotes the Banach space of continuous mappings from E to F, where E and F are Banach spaces. $C_b(\mathbb{R}^n)$ denotes the space of all bounded continuous functions on $\mathbb{R}^n$. $\mathbb{N} := \{0, 1, 2, \dots\}$, $\mathbb{Z}_+ := \{1, 2, \dots\}$ and $\mathbb{R}_+ := [0, \infty)$. The superscript ^{T} denotes transposition and * denotes adjoint.

2.1 Model

We consider the following model for our SRNNs. By an *activation function*, we mean a real-valued function that is non-constant, Lipschitz continuous and bounded. Examples of activation function include sigmoid functions such as hyperbolic tangent, commonly used in practice.

Definition 2.1 *(Continuous-time SRNNs) Let $t \in [0, T]$ and $\boldsymbol{u} \in C([0, T], \mathbb{R}^m)$ be a deterministic input signal. A continuous-time SRNN is described by the following state-space model:*

$$d\boldsymbol{h}_t = \boldsymbol{\phi}(\boldsymbol{h}_t, t)dt + \boldsymbol{\sigma} d\boldsymbol{W}_t, \tag{1}$$

$$\boldsymbol{y}_t = \boldsymbol{f}(\boldsymbol{h}_t). \tag{2}$$

In the above, Eq. (1) is a stochastic differential equation (SDE) for the hidden states $\boldsymbol{h} = (\boldsymbol{h}_t)_{t\in[0,T]}$, with the drift coefficient $\boldsymbol{\phi} : \mathbb{R}^n \times [0, T] \to \mathbb{R}^n$, noise coefficient $\boldsymbol{\sigma} \in \mathbb{R}^{n\times r}$, and $\boldsymbol{W} = (\boldsymbol{W}_t)_{t\geq 0}$ is an r-dimensional Wiener process defined on $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$, whereas Eq. (2) defines an observable with $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^p$ an activation function.

We consider an input-affine[1] *version of the SRNNs, in which:*

$$\boldsymbol{\phi}(\boldsymbol{h}_t, t) = -\boldsymbol{\Gamma}\boldsymbol{h}_t + \boldsymbol{a}(\boldsymbol{W}\boldsymbol{h}_t + \boldsymbol{b}) + \boldsymbol{C}\boldsymbol{u}_t, \tag{3}$$

1. We refer to Theorem C.1 in (Kidger et al., 2020) for a rigorous justification of considering input-affine continuous-time RNN models. See also Subsection 2.2, as well as Section IV in (Bengio et al., 1994) and the footnote on the first page of (Pascanu et al., 2013b) for discrete-time models.

where $\boldsymbol{\Gamma} \in \mathbb{R}^{n\times n}$ *is positive stable,* $\boldsymbol{a} : \mathbb{R}^n \to \mathbb{R}^n$ *is an activation function,* $\boldsymbol{W} \in \mathbb{R}^{n\times n}$ *and* $\boldsymbol{b} \in \mathbb{R}^n$ *are constants, and* $\boldsymbol{C} \in \mathbb{R}^{n\times m}$ *is a constant matrix that transforms the input signal.*

From now on, we refer to SRNN as the system defined by (1)-(3). The hidden states of a SRNN describe a nonautonomous stochastic dynamical system processing an input signal (c.f. (Ganguli et al., 2008; Dambre et al., 2012; Tino, 2020)). The constants $\boldsymbol{\Gamma}, \boldsymbol{W}, \boldsymbol{b}, \boldsymbol{C}, \boldsymbol{\sigma}$ and the parameters (if any) in $\boldsymbol{f}$ are the (learnable) parameters or weights defining the (architecture of) SRNN. For $T > 0$, associated with the SRNN is the output functional $F_T : C([0,T], \mathbb{R}^m) \to \mathbb{R}^p$ defined as the expectation (ensemble average) of the observable $\boldsymbol{f}$:

$$F_T[\boldsymbol{u}] := \mathbb{E}\boldsymbol{f}(\boldsymbol{h}_T), \tag{4}$$

which will be of interest to us.

2.2 Related Work

Our work is in line with the recently promoted approach of "formulate first, then discretize" in machine learning. Such approach is popularized in (Weinan, 2017), inspiring subsequent work (Haber and Ruthotto, 2017; Chen et al., 2018; Rubanova et al., 2019; Benning et al., 2019; E et al., 2019). Following the approach, here our SRNN model is formulated in the continuous time.

There are several benefits of adopting this approach. At the level of formulation, sampling from these RNNs gives the discrete-time RNNs, including randomized RNNs (Herbert, 2001; Grigoryeva and Ortega, 2018; Gallicchio and Scardapane, 2020) and fully trained RNNs (Bengio et al., 2013; Goodfellow et al., 2016), commonly encountered in applications. More importantly, the continuous-time SDE formulation gives us a guided principle and flexibility in designing RNN architectures, in particular those that are capable of adapting to the nature of data (e.g., those irregularly sampled (De Brouwer et al., 2019; Kidger et al., 2020; Morrill et al., 2020)) on hand, going beyond existing architectures. Recent work such as (Chang et al., 2019; Chen et al., 2019; Niu et al., 2019; Erichson et al., 2020; Rusch and Mishra, 2020) exploits these benefits and designs novel recurrent architectures with desirable stability properties by appropriately discretizing ordinary differential equations. Moreover, in situations where the input data are generated by continuous-time dynamical systems, it is desirable to consider learning models which are also continuous in time. From theoretical analysis point of view, a rich set of tools and techniques from the continuous-time theory can be borrowed to simplify analysis and to gain useful insights.

The noise injection can be viewed as a regularization scheme or introducing noise in input data in the context of our SRNNs. Generally, noise injection can be viewed as a stochastic learning strategy used to improve robustness of the learning model against data perturbations. We refer to, the demonstration of these benefits in, for instance, (Jim et al., 1996) for RNNs, (Sun et al., 2018) for deep residual networks and (Liu et al., 2020) for Neural SDEs. On the other hand, both the continuous-time setting and noise injection are natural assumptions in modelling biological neural networks (Cessac and Samuelides, 2007; Touboul, 2008; Cessac, 2019). Our study here belongs to the paradigm of "formulate first" (in continuous-time) and in fact covers both artificial and biological RNNs.

We now discuss in detail an example of how sampling from SRNNs gives rise to a class of discrete-time RNNs. Consider the following SRNN:

$$d\boldsymbol{h}_t = -\gamma \boldsymbol{h}_t dt + \boldsymbol{a}(\boldsymbol{W}\boldsymbol{h}_t + \boldsymbol{b})dt + \boldsymbol{u}_t dt + \sigma d\boldsymbol{W}_t, \tag{5}$$

$$\boldsymbol{y}_t = \boldsymbol{f}(\boldsymbol{h}_t), \tag{6}$$

where $\gamma, \sigma > 0$ are constants. The Euler-Mayurama approximations $(\hat{\boldsymbol{h}})_{t=0,1,\ldots,T}$, with a uniform step size Δt, to the solution of the SDE (5) are given by (Kloeden and Platen, 2013):

$$\hat{\boldsymbol{h}}_{t+1} = \alpha \hat{\boldsymbol{h}}_t + \beta(\boldsymbol{a}(\boldsymbol{W}\hat{\boldsymbol{h}}_t + \boldsymbol{b}) + \boldsymbol{u}_t) + \theta \boldsymbol{\xi}_t, \tag{7}$$

for $t = 0, 1, \ldots, T-1$, where $\hat{\boldsymbol{h}}_0 = \boldsymbol{h}_0$, $\alpha = 1 - \gamma\Delta t$, $\beta = \Delta t$, $\theta = \sqrt{\Delta t}\sigma$ and the $\boldsymbol{\xi}_t$ are i.i.d. standard Gaussian random variables. In particular, setting $\gamma = \Delta t = 1$ gives:

$$\hat{\boldsymbol{h}}_{t+1} = \boldsymbol{a}(\boldsymbol{W}\hat{\boldsymbol{h}}_t + \boldsymbol{b}) + \boldsymbol{u}_t + \sigma \boldsymbol{\xi}_t. \tag{8}$$

Now, by taking $\hat{\boldsymbol{h}}_t = (\boldsymbol{x}_t, \boldsymbol{z}_t)$, $\boldsymbol{a}(\boldsymbol{W}\hat{\boldsymbol{h}}_t + \boldsymbol{b}) = (\tanh(\boldsymbol{W}'\boldsymbol{x}_t + \boldsymbol{U}'\boldsymbol{z}_t + \boldsymbol{b}'), \boldsymbol{0})$ and $\boldsymbol{u}_t = (\boldsymbol{0}, \boldsymbol{y}_t)$ for some matrices $\boldsymbol{W}'$ and $\boldsymbol{U}'$, some vector $\boldsymbol{b}'$ and input data $(\boldsymbol{y}_t)_{t=0,1,\ldots,T-1}$, we have $\boldsymbol{z}_t = \boldsymbol{y}_t$ and

$$\boldsymbol{x}_{t+1} = tanh(\boldsymbol{W}'\boldsymbol{x}_t + \boldsymbol{U}'\boldsymbol{y}_t + \boldsymbol{b}') + \sigma \boldsymbol{\xi}_t, \tag{9}$$

which is precisely the update equation of a standard discrete-time RNN (Bengio et al., 1994) whose hidden states $\boldsymbol{x}$ are injected by the Gaussian white noise $\sigma\boldsymbol{\xi}$. Note that the above derivation also shows that discrete-time RNNs can be transformed into ones whose update equations are linear in input by simply introducing additional state variables. In general, different numerical approximations (e.g., those with possibly adaptive step sizes (Fang et al., 2020)) to the SRNN hidden states give rise to RNN architectures with different properties. Motivated by the above considerations, we are going to introduce a class of noisy RNNs which are obtained by discretizing SDEs and study these benefits, both theoretically and experimentally, for our RNNs in a forthcoming work.

Lastly, we discuss some related work that touches upon connections between (discrete-time) RNNs and kernel methods. Although connections between non-recurrent neural networks such as two-layer neural networks with random weights and kernel methods are well studied, there are few rigorous studies connecting RNNs and kernel methods. We mention two recent studies here. In the context of reservoir computing, (Tino, 2020) analyzes the similarity between reservoir computers and kernel machines and shed some insights on which network topology gives rise to rich dynamical feature representations. However, the analysis was done for only linear non-noisy RNNs. In the context of deep learning, it is worth mentioning that (Alemohammad et al., 2020) kernelizes RNNs by introducing the Recurrent Neural Tangent Kernel (RNTK) to study the behavior of overparametrized RNNs during their training by gradient descent. The RNTK provides a rigorous connection between the inference performed by infinite width ($n = \infty$) RNNs and that performed by kernel methods, suggesting that the performance of large RNNs can be replicated by kernel methods for properly chosen kernels. In particular, the derivation of RNTK at initialization is based on the correspondence between randomly initialized infinite width neural networks and Gaussian Processes (Neal, 1996). Understanding the precise learning behavior of finite width RNNs remains an open problem in the field.

2.3 Main Contributions

Our main contributions in this paper are in the following two directions:

(1) We establish the relationship between the output functional of SRNNs and deterministic driving input signal using the response theory. In particular, we derive two series representations for the output functional of SRNNs and their deep version. The first one is a Volterra series representation with the kernels expressed as certain correlation (in time) functions that solely depend on the unperturbed dynamics of SRNNs (see Theorem 4.1). The second one is in terms of a series of iterated integrals of a transformed input signal (see Theorem 4.2). These representations are interpretable and allow us to gain insights into the working mechanism of SRNNs.

(2) Building on our understanding in (1), we identify a universal feature, called the *response feature*, potentially useful for learning temporal series. This feature turns out to be the signature of tensor product of the input signal and a vector whose components are orthogonal polynomials (see Theorem 4.3). We then show that SRNNs, with only the weights in the readout layer optimized and the weights in the hidden layer kept fixed and not optimized, are essentially kernel machines operating in a reproducing kernel Hilbert space associated with the response feature (see Theorem 4.4). This result characterizes precisely the idea that SRNNs with hidden-layer weights that are fixed and not optimized can be viewed as kernel methods.

In short, we focus on studying representations for output functionals of SRNNs and relating SRNNs to certain kernel machines in this work. To achieve (1) we develop and make rigorous nonlinear response theory for the SRNNs (1)-(3), driven by a small deterministic input signal (see **SM**). This makes rigorous the results of existing works on response theory in the physics literature and also extend the recent rigorous work of (Chen and Jia, 2020) beyond the linear response regime, which may be of independent interest.

For simplicity we have chosen to work with the SRNNs under a set of rather restricted but reasonable assumptions, i.e., Assumption 4.1, in this paper. Also, one could possibly work in the more general setting where the driving input signal is a rough path (Lyons and Qian, 2002). Relaxing the assumptions here and working in the more general setting come at a higher cost of technicality and risk burying intuitions, which we avoid in the present work.

3. Nonequilibrium Response Theory of SRNNs

3.1 Preliminaries and Notation

In this subsection we briefly recall preliminaries on Markov processes (Karatzas and Shreve, 1998; Pavliotis, 2014) and introduce some of our notations.

Let $t \in [0,T]$, $\gamma(t) := |\boldsymbol{C}\boldsymbol{u}_t| > 0$ and $\boldsymbol{U}_t := \boldsymbol{C}\boldsymbol{u}_t/|\boldsymbol{C}\boldsymbol{u}_t|$ be a normalized input signal (i.e., $|\boldsymbol{U}_t| = 1$). In the SRNN (1)-(3), we consider the signal $\boldsymbol{C}\boldsymbol{u} = (\boldsymbol{C}\boldsymbol{u}_t)_{t\in[0,T]}$ to be a perturbation with small amplitude $\gamma(t)$ driving the SDE:

$$d\boldsymbol{h}_t = \boldsymbol{\phi}(\boldsymbol{h}_t, t)dt + \boldsymbol{\sigma} d\boldsymbol{W}_t. \tag{10}$$

The unperturbed SDE is the system with $\boldsymbol{Cu}$ set to zero:

$$d\overline{\boldsymbol{h}}_t = \overline{\boldsymbol{\phi}}(\overline{\boldsymbol{h}}_t)dt + \boldsymbol{\sigma} d\boldsymbol{W}_t. \tag{11}$$

In the above, $\overline{\boldsymbol{\phi}}(\boldsymbol{h}_t) = -\boldsymbol{\Gamma}\boldsymbol{h}_t + \boldsymbol{a}(\boldsymbol{W}\boldsymbol{h}_t + \boldsymbol{b})$ and $\boldsymbol{\phi}(\boldsymbol{h}_t, t) = \overline{\boldsymbol{\phi}}(\boldsymbol{h}_t) + \gamma(t)\boldsymbol{U}_t$. The process $\boldsymbol{h}$ is a perturbation of the time-homogeneous Markov process $\overline{\boldsymbol{h}}$, which is not necessarily stationary.

The diffusion process $\boldsymbol{h}$ and $\overline{\boldsymbol{h}}$ are associated with a family of infinitesimal generators $(\mathcal{L}_t)_{t\geq 0}$ and $\mathcal{L}^0$ respectively, which are second-order elliptic operators defined by:

$$\mathcal{L}_t f = \phi^i(\boldsymbol{h}, t)\frac{\partial}{\partial h^i} f + \frac{1}{2}\Sigma^{ij}\frac{\partial^2}{\partial h^i \partial h^j} f, \tag{12}$$

$$\mathcal{L}^0 f = \bar{\phi}^i(\overline{\boldsymbol{h}})\frac{\partial}{\partial \overline{h}^i} f + \frac{1}{2}\Sigma^{ij}\frac{\partial^2}{\partial \overline{h}^i \partial \overline{h}^j} f, \tag{13}$$

for any observable $f \in C_b(\mathbb{R}^n)$, where $\boldsymbol{\Sigma} := \boldsymbol{\sigma}\boldsymbol{\sigma}^T > 0$. We define the transition operator $(P_{s,t})_{s\in[0,t]}$ associated with $\boldsymbol{h}$ as:

$$P_{s,t} f(\boldsymbol{h}) = \mathbb{E}[f(\boldsymbol{h}_t)|\boldsymbol{h}_s = \boldsymbol{h}], \tag{14}$$

for $f \in C_b(\mathbb{R}^n)$, and similarly for the transition operator $(P^0_{s,t})_{s\in[0,t]}$ (which is a Markov semigroup) associated with $\overline{\boldsymbol{h}}$.

Moreover, one can define the L^2-adjoint of the above generators and transition operators on the space of probability measures. We denote the adjoint generator associated to $\boldsymbol{h}$ and $\overline{\boldsymbol{h}}$ by $\mathcal{A}_t$ and $\mathcal{A}^0$ respectively, and the adjoint transition operator associated to $\boldsymbol{h}$ and $\overline{\boldsymbol{h}}$ by $(P^*_{s,t})_{s\in[0,t]}$ and $((P^0_{s,t})^*)_{s\in[0,t]}$ respectively. We assume that the initial measure and the law of the processes have a density with respect to Lebesgue measure. Denoting the initial density as $\rho(\boldsymbol{h}, t = 0) = \rho_{init}(\boldsymbol{h})$, $\rho(\boldsymbol{h}, t) = P^*_{0,t}\rho_{init}(\boldsymbol{h})$ satisfies a forward Kolmogorov equation (FKE) associated with $\mathcal{A}_t$.

We take the natural assumption that both perturbed and unperturbed process have the same initial distribution ρ_{init}, which is generally not the invariant distribution ρ_∞ of the unperturbed dynamics.

3.2 Key Ideas and Formal Derivations

This subsection serves to provide a *formal* derivation of one of the core ideas of the paper in an explicit manner to aid understanding.

First, we are going to derive a representation for the output functional of SRNN in terms of driving input signal. Our approach is based on the response theory originated from nonequilibrium statistical mechanics. For a brief overview of this theory, we refer to Chapter 9 in (Pavliotis, 2014) (see also (Kubo, 1957; Peterson, 1967; Hanggi et al., 1978; Baiesi and Maes, 2013)). In the following, we assume that any infinite series is well-defined and any interchange between summations and integrals is justified.

Fix a $T > 0$. Let $\epsilon := \sup_{t\in[0,T]} |\gamma(t)| > 0$ be sufficiently small and

$$\tilde{\boldsymbol{U}}_t := \frac{\boldsymbol{C}\boldsymbol{u}_t}{\sup_{t\in[0,T]} |\boldsymbol{C}\boldsymbol{u}_t|}. \tag{15}$$

To begin with, note that the FKE for the probability density $\rho(\boldsymbol{h}, t)$ is:

$$\frac{\partial \rho}{\partial t} = \mathcal{A}_t^\epsilon \rho, \quad \rho(\boldsymbol{h}, 0) = \rho_{init}(\boldsymbol{h}), \tag{16}$$

where $\mathcal{A}_t^\epsilon = \mathcal{A}^0 + \epsilon \mathcal{A}_t^1$, with:

$$\mathcal{A}^0 \cdot = -\frac{\partial}{\partial h^i}\left(\bar{\phi}^i(\boldsymbol{h})\cdot\right) + \frac{1}{2}\Sigma^{ij}\frac{\partial^2}{\partial h^i \partial h^j}\cdot, \quad \text{and} \quad \mathcal{A}_t^1 \cdot = -\tilde{U}_t^i \frac{\partial}{\partial h^i}\cdot. \tag{17}$$

The key idea is that since $\epsilon > 0$ is small we seek a perturbative expansion for ρ of the form

$$\rho = \rho_0 + \epsilon \rho_1 + \epsilon^2 \rho_2 + \dots. \tag{18}$$

Plugging this into the FKE and matching orders in ϵ, we obtain the following hierachy of equations:

$$\frac{\partial \rho_0}{\partial t} = \mathcal{A}^0 \rho_0, \quad \rho_0(\boldsymbol{h}, t = 0) = \rho_{init}(\boldsymbol{h}); \tag{19}$$

$$\frac{\partial \rho_n}{\partial t} = \mathcal{A}^0 \rho_n + \mathcal{A}_t^1 \rho_{n-1}, \quad n = 1, 2, \dots. \tag{20}$$

The formal solution to the ρ_n can be obtained iteratively as follows. Formally, we write $\rho_0(\boldsymbol{h}, t) = e^{\mathcal{A}^0 t}\rho_{init}(\boldsymbol{h})$. In the special case when the invariant distribution is stationary, $\rho_0(\boldsymbol{h}, t) = \rho_{init}(\boldsymbol{h}) = \rho_\infty(\boldsymbol{h})$ is independent of time.

Noting that $\rho_n(\boldsymbol{h}, 0) = 0$ for $n \geq 2$, the solutions ρ_n are related recursively via:

$$\rho_n(\boldsymbol{h}, t) = \int_0^t e^{\mathcal{A}^0 (t-s)} \mathcal{A}_s^1 \rho_{n-1}(\boldsymbol{h}, s) ds, \tag{21}$$

for $n \geq 2$. Therefore, provided that the infinite series below converges absolutely, we have:

$$\rho(\boldsymbol{h}, t) = \rho_0(\boldsymbol{h}, t) + \sum_{n=1}^{\infty} \epsilon^n \rho_n(\boldsymbol{h}, t). \tag{22}$$

Next we consider a scalar-valued observable[2], $\mathcal{F}(t) := f(\boldsymbol{h}_t)$, of the hidden dynamics of the SRNN and study the deviation of average of this observable caused by the perturbation of input signal:

$$\mathbb{E}\mathcal{F}(t) - \mathbb{E}^0\mathcal{F}(t) := \int f(\boldsymbol{h})\rho(\boldsymbol{h}, t) d\boldsymbol{h} - \int f(\boldsymbol{h})\rho_0(\boldsymbol{h}, t) d\boldsymbol{h}. \tag{23}$$

Using (22), the average of the observable with respect to the perturbed dynamics can be written as:

$$\mathbb{E}\mathcal{F}(t) = \mathbb{E}^0\mathcal{F}(t) + \sum_{n=1}^{\infty} \epsilon^n \int f(\boldsymbol{h})\rho_n(\boldsymbol{h}, t) d\boldsymbol{h}. \tag{24}$$

2. The extension to the vector-valued case is straightforward.

Without loss of generality, we take $\mathbb{E}^0\mathcal{F}(t) = e^{\mathcal{A}^0 t}\int \rho_{init}(\boldsymbol{h})f(\boldsymbol{h})d\boldsymbol{h} = 0$ in the following, i.e., $f(\boldsymbol{h})$ is taken to be mean-zero (with respect to ρ_{init}).

We have:

$$\int f(\boldsymbol{h})\rho_1(\boldsymbol{h},t)d\boldsymbol{h} = \int d\boldsymbol{h}\ f(\boldsymbol{h})\int_0^t ds\ e^{\mathcal{A}^0(t-s)}\mathcal{A}^1_s e^{\mathcal{A}^0 s}\rho_{init}(\boldsymbol{h}) \tag{25}$$

$$= -\int_0^t ds\ \int d\boldsymbol{h}\ f(\boldsymbol{h})e^{\mathcal{A}^0(t-s)}\tilde{U}^j_s\frac{\partial}{\partial h^j}\left(e^{\mathcal{A}^0 s}\rho_{init}(\boldsymbol{h})\right) \tag{26}$$

$$=: \int_0^t \mathcal{K}^k_{\mathcal{A}^0,\mathcal{F}}(t,s)\tilde{U}^k_s ds, \tag{27}$$

where the

$$\mathcal{K}^k_{\mathcal{A}^0,\mathcal{F}}(t,s) = -\int d\boldsymbol{h}\ f(\boldsymbol{h})e^{\mathcal{A}^0(t-s)}\frac{\partial}{\partial h^k}\left(e^{\mathcal{A}^0 s}\rho_{init}(\boldsymbol{h})\right) \tag{28}$$

$$= -\left\langle e^{\mathcal{L}^0(t-s)}f(\boldsymbol{h})\frac{\partial}{\partial h^k}\left(e^{\mathcal{A}^0 s}\rho_{init}(\boldsymbol{h})\right)\rho^{-1}_{init}(\boldsymbol{h})\right\rangle_{\rho_{init}}, \tag{29}$$

are the *first-order response kernels*, which are averages, with respect to ρ_{init}, of a functional of only the unperturbed dynamics. Note that in order to obtain the last line above we have integrated by parts and assumed that $\rho_{init} > 0$.

Formula (29) expresses the *nonequilibrium fluctuation-dissipation relation* of Agarwal type (Agarwal, 1972). In the case of stationary invariant distribution, we recover the well-known equilibrium fluctuation-dissipation relation in statistical mechanics, with the (vector-valued) response kernel:

$$\boldsymbol{\mathcal{K}}_{\mathcal{A}^0,\mathcal{F}}(t,s) = \boldsymbol{\mathcal{K}}_{\mathcal{A}^0,\mathcal{F}}(t-s) = \langle f(\boldsymbol{h}_{t-s})\boldsymbol{\nabla}_{\boldsymbol{h}}L(\boldsymbol{h}_{t-s})\rangle_{\rho^\infty}, \tag{30}$$

where $L(\boldsymbol{h}) = -\log\rho_\infty(\boldsymbol{h})$. In the special case of linear SRNN (i.e., $\boldsymbol{\phi}(\boldsymbol{h},t)$ linear in $\boldsymbol{h}$) and $f(\boldsymbol{h}) = \boldsymbol{h}$, this essentially reduces to the covariance function (with respect to ρ^∞) of $\boldsymbol{h}_{t-s}$.

So far we have looked at the linear response regime, where the response depends linearly on the input. We now go beyond this regime by extending the above derivations to the case of $n \geq 2$. Denoting $s_0 := t$ and applying (21), we derive:

$$\int f(\boldsymbol{h})\rho_n(\boldsymbol{h},t)d\boldsymbol{h} = \int_0^t ds_1\tilde{U}^{k_1}_{s_1}\int_0^{s_1} ds_2\tilde{U}^{k_2}_{s_2}\cdots\int_0^{s_{n-1}} ds_n\tilde{U}^{k_n}_{s_n}\mathcal{K}^{\boldsymbol{k}^{(n)}}(s_0,s_1,\ldots,s_n), \tag{31}$$

where $\boldsymbol{k}^{(n)} := (k_1,\ldots,k_n)$, and the $\mathcal{K}^{\boldsymbol{k}^{(n)}}$ are the *nth order response kernels*:

$$\mathcal{K}^{\boldsymbol{k}^{(n)}}(s_0,s_1,\ldots,s_n) = (-1)^n\left\langle f(\boldsymbol{h})\rho^{-1}_{init}(\boldsymbol{h})R^{\boldsymbol{k}^{(n)}}(s_0,s_1,\ldots,s_n)e^{\mathcal{A}^0_{s_n}}\rho_{init}(\boldsymbol{h})\right\rangle_{\rho_{init}}, \tag{32}$$

with

$$R^{k_1}(s_0,s_1)\cdot = e^{\mathcal{A}^0(s_0-s_1)}\frac{\partial}{\partial h^{k_1}}\cdot, \tag{33}$$

$$R^{\boldsymbol{k}^{(n)}}(s_0,s_1,\ldots,s_n)\cdot = R^{\boldsymbol{k}^{(n-1)}}(s_0,s_1,\ldots,s_{n-1})\cdot\left(e^{\mathcal{A}^0(s_{n-1}-s_n)}\frac{\partial}{\partial h^{k_n}}\cdot\right), \tag{34}$$

for $n = 2, 3, \ldots$. Note that these higher order response kernels, similar to the first order ones, are averages, with respect to ρ_{init}, of some functional of only the unperturbed dynamics.

Collecting the above results, (24) becomes a series of generalized convolution integrals, given by:

$$\mathbb{E}f(\boldsymbol{h}_t) = \sum_{n=1}^{\infty} \epsilon^n \int_0^t ds_1 \tilde{U}_{s_1}^{k_1} \int_0^{s_1} ds_2 \tilde{U}_{s_2}^{k_2} \cdots \int_0^{s_{n-1}} ds_n \tilde{U}_{s_n}^{k_n} \mathcal{K}^{\boldsymbol{k}^{(n)}}(s_0, s_1, \ldots, s_n), \tag{35}$$

with the time-dependent kernels $\mathcal{K}^{\boldsymbol{k}^{(n)}}$ defined recursively via (32)-(34). *More importantly, these kernels are completely determined in an explicit manner by the unperturbed dynamics of the SRNN.* Therefore, the output functional of SRNN can be written (in fact, uniquely) as a series of the above form. This statement is formulated precise in Theorem 4.1, thereby addressing (Q1).

We now address (Q2). By means of expansion techniques, one can derive (see Section B.5):

$$\mathbb{E}f(\boldsymbol{h}_t) = \sum_{n=1}^{\infty} \epsilon^n Q_{\boldsymbol{p}^{(n)}}^{\boldsymbol{k}^{(n)}} \left(s_0^{p_0} \int_0^{s_0} ds_1 s_1^{p_1} \tilde{U}_{s_1}^{k_1} \cdots \int_0^{s_{n-1}} ds_n s_n^{p_n} \tilde{U}_{s_n}^{k_n} \right), \tag{36}$$

where the $Q_{\boldsymbol{p}^{(n)}}^{\boldsymbol{k}^{(n)}}$ are constants independent of time and the signal $\tilde{\boldsymbol{U}}$. This expression disentangles the driving input signal from the SRNN architecture in a systematic manner. Roughly speaking, it tells us that the response of SRNN to the input signal can be obtained by adding up products of two components, one of which describes the unperturbed part of SRNN (the $Q_{\boldsymbol{p}^{(n)}}^{\boldsymbol{k}^{(n)}}$ terms in (36)) and the other one is an iterated integral of a time-transformed input signal (the terms in parenthesis in (36)). This statement is made precise in Theorem 4.2, which is the starting point to addressing (Q2). See further results and discussions in Section 4.

4. Main Results

4.1 Assumptions

For simplicity and intuitive appeal we work with the following rather restrictive assumptions on the SRNNs (1)-(3). These assumptions can be either relaxed at an increased cost of technicality (which we do not pursue here) or justified by the approximation result in Section C.

Recall that we are working with a deterministic input signal $\boldsymbol{u} \in C([0,T], \mathbb{R}^m)$.

Assumption 4.1 *Fix a $T > 0$ and let U be an open set in $\mathbb{R}^n$.*
(a) $\gamma(t) := |\boldsymbol{C}\boldsymbol{u}_t| > 0$ is sufficiently small for all $t \in [0,T]$.
(b) $\boldsymbol{h}_t, \overline{\boldsymbol{h}}_t \in U$ for all $t \in [0,T]$, and, with probability one, there exists a compact set $K \subset U$ such that, for all $\gamma(t)$, $\boldsymbol{h}_t, \overline{\boldsymbol{h}}_t \in K$ for all $t \in [0,T]$.
(c) The coefficients $\boldsymbol{a} : \mathbb{R}^n \to \mathbb{R}^n$ and $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^p$ are analytic functions.
(d) $\boldsymbol{\Sigma} := \boldsymbol{\sigma}\boldsymbol{\sigma}^T \in \mathbb{R}^{n \times n}$ is positive definite and $\boldsymbol{\Gamma} \in \mathbb{R}^{n \times n}$ is positive stable (i.e., the real part of all eigenvalues of $\boldsymbol{\Gamma}$ is positive).
(e) The initial state $\boldsymbol{h}_0 = \overline{\boldsymbol{h}}_0$ is a random variable distributed according to the probability density ρ_{init}.

Assumption 4.1 (a) implies that we work with input signals with sufficiently small amplitude. This is important to ensure that certain infinite series are absolutely convergent with a radius of convergence that is sufficiently large (see the statements after Definition A.5). (b) and (c) ensure some desirable regularity and boundedness properties. In particular, they imply that $\boldsymbol{a}, \boldsymbol{f}$ and all their partial derivatives are bounded[3] and Lipschitz continuous in $\boldsymbol{h}_t$ and $\overline{\boldsymbol{h}}_t$ for all $t \in [0,T]$. (d) implies that the system is damped and driven by a nondegenerate noise, ensuring that the unperturbed system could be exponentially stable. (e) is a natural assumption for our analysis since $\boldsymbol{h}$ is a perturbation of $\overline{\boldsymbol{h}}$.

Assumption 4.1 is implicitly assumed throughout the paper unless stated otherwise.

Further Notation. We now provide a list of the spaces and their notation that we will need from now on in the main paper and the **SM**:

- $L(E_1, E_2)$: the Banach space of bounded linear operators from E_1 to E_2 (with $\|\cdot\|$ denoting norms on appropriate spaces)
- $C_c^n(0,t)$, $n \in \mathbb{Z}_+ \cup \{\infty\}$, $t > 0$: the space of real-valued functions of class $C^n(0,t)$ with compact support
- $C_b^n(0,t)$, $n \in \mathbb{Z}_+ \cup \{\infty\}$, $t > 0$: the space of bounded real-valued functions of class $C^n(0,t)$
- $B(\mathbb{R}_+^n)$: the space of bounded absolutely continuous measures on $\mathbb{R}_+^n$, with $|\mu| = \int d|\mu| = \int |\rho(x)|dx$, where ρ denotes the density of the measure μ
- $L^p(\rho)$, $p > 1$: the ρ-weighted L^p space, i.e., the space of functions f such that $\|f\|_{L^p(\rho)} := \int |f(x)|^p \rho(x) dx < \infty$, where ρ is a weighting function

4.2 Representations for Output Functionals of SRNNs

Without loss of generality, we are going to take $p = 1$ and assume that $\int f(\boldsymbol{h})\rho_{init}(\boldsymbol{h})d\boldsymbol{h} = 0$ in the following.

First, we define response functions of an observable, extending the one formulated in (Chen and Jia, 2020) for the linear response regime. Recall that for $t > 0$, $\gamma := (\gamma(s) := |\boldsymbol{C}\boldsymbol{u}_s|)_{s\in[0,t]}$.

Definition 4.1 *(Response functions) Let $f : \mathbb{R}^n \to \mathbb{R}$ be a bounded observable. For $t \in [0,T]$, let F_t be the functional on $C([0,t],\mathbb{R})$ defined as $F_t[\gamma] = \mathbb{E}f(\boldsymbol{h}_t)$ and $D^n F_t[\gamma] := \delta^n F_t/\delta\gamma(s_1)\cdots\delta\gamma(s_n)$ denote the nth order functional derivative of F_t with respect to γ (see Definition A.4). For $n \in \mathbb{Z}_+$, if there exists a locally integrable function $R_f^{(n)}(t,\cdot)$ such that*

$$\begin{aligned}&\int_{[0,t]^n} ds_1\cdots ds_n \frac{1}{n!} D^n F_t\big|_{\gamma=0}\phi(s_1)\cdots\phi(s_n)\\ &= \int_{[0,t]^n} ds_1\cdots ds_n R_f^{(n)}(t,s_1,\ldots,s_n)\phi(s_1)\cdots\phi(s_n),\end{aligned} \tag{37}$$

3. Boundedness of these coefficients will be important when deriving the estimates here.

for all test functions $\phi \in C_c^\infty(0,t)$, then $R_f^{(n)}(t,\cdot)$ is called the nth order response function of the observable f.

Note that since the derivatives in Eq. (37) are symmetric mappings (see Subsection A.1), the response functions $R_f^{(n)}$ are symmetric in $s_1, \ldots, s_n$.

We can gain some intuition on the response functions by first looking at $R_f^{(1)}$. Taking $\phi(s) = \delta(x-s) = \delta_s(x)$, we have, formally, $R_f^{(1)}(t,s) = \int_0^t R_f^{(1)}(t,u)\delta(s-u)du = \int_0^t du \frac{\delta F_t}{\delta \gamma}\big|_{\gamma=0}\delta(s-u) = \lim_{\epsilon\to 0}\frac{1}{\epsilon}(F_t[\epsilon\delta_s] - F_t[0])$. This tells us that $R_f^{(1)}$ represents the rate of change of the functional F_t at time t subject to the small impulsive perturbation to $\boldsymbol{h}$ at time s. Similarly, the $R_f^{(n)}$ give higher order rates of change for $n > 1$. Summing up these rates of change allows us to quantify the full effect of the perturbation on the functional F_t order by order.

We now show that, under certain assumptions, these rates of change are well-defined and compute them. The following proposition provides explicit expressions for the nth order response function for a class of observables of the SRNN. In the first order case, it was shown in (Chen and Jia, 2020) that the response function can be expressed as a correlation function of the observable and a unique conjugate observable with respect to the unperturbed dynamics, thereby providing a mathematically rigorous version of the Agarwal-type fluctuation-dissipation relation (A-FDT). We are going to show that a similar statement can be drawn for the higher order cases.

In the following, let $f \in C_b^\infty(\mathbb{R}^n)$ be any observable and $\Delta\mathcal{L}_t := \mathcal{L}_t - \mathcal{L}^0$, for $t \in [0,T]$.

Proposition 4.1 *(Explicit expressions for response functions) For $n \in \mathbb{Z}_+$, let $R_f^{(n)}$ be the nth-order response function of f. Then, for $0 < s_n < s_{n-1} < \cdots < s_0 := t \le T$:*
(a)

$$R_f^{(n)}(t,s_1,\ldots,s_n) = \mathbb{E}P^0_{0,s_n}\Delta\mathcal{L}_{s_n}P^0_{s_n,s_{n-1}}\Delta\mathcal{L}_{s_{n-1}}\cdots P^0_{s_1,t}f(\boldsymbol{h}_0). \tag{38}$$

(b) (Higher-order A-FDTs) If, in addition, ρ_{init} is positive, then

$$R_f^{(n)}(t,s_1,\ldots,s_n) = \mathbb{E}f(\boldsymbol{h}_0)v^{(n)}_{t,s_1,\ldots,s_n}(\boldsymbol{h}_{s_1},\ldots,\boldsymbol{h}_{s_n}), \tag{39}$$

where

$$v^{(n)}_{t,s_1,\ldots,s_n}(\boldsymbol{h}_1,\ldots,\boldsymbol{h}_n) = \frac{(-1)^n}{\rho_{init}}(P^0_{s_1,t})^*\boldsymbol{\nabla}^T_{\boldsymbol{h}_1}[\boldsymbol{U}_{s_1}\cdots(P^0_{s_n,s_{n-1}})^*\boldsymbol{\nabla}^T_{\boldsymbol{h}_n}[\boldsymbol{U}_{s_n}p_{s_n}(\boldsymbol{h})]]. \tag{40}$$

We make a few remarks on the above results.

Remark 4.1 *In the linear response regime with $\overline{\boldsymbol{h}}$ stationary, if one further restricts to reversible diffusion (when detailed balance holds), in which case the drift coefficient in the SRNN can be expressed as negative gradient of a potential function, then Eq.* (39) *reduces to the equilibrium FDT (see, for instance, (Kubo, 1966) or (Chetrite and Gawedzki, 2008)), a cornerstone of nonequilibrium statistical mechanics. In the one-dimensional case, explicit calculation and richer insight can be obtained. See, for instance, Example 9.5 in (Pavliotis, 2014) for a formal analysis of stochastic resonance using linear response theory.*

Remark 4.2 *The observables $v^{(n)}_{t,s_1,\dots,s_n}(\boldsymbol{h}_{s_1},\dots,\boldsymbol{h}_{s_n})$ in Proposition 4.1(b) can be viewed as nonlinear counterparts of the conjugate observable obtained in (Chen and Jia, 2020) in the linear response regime. Indeed, when $n=1$, $v^{(1)}_{t,s_1}(t,s_1)$ is exactly the conjugate observable in (Chen and Jia, 2020). Hence, it is natural to call them higher order conjugate observables. Proposition 4.1(b) tells us that any higher order response of the observable f to a small input signal can be represented as correlation function of f and the associated higher order conjugate observable.*

Moreover, the conjugate observables $v^{(n)}_{t,s_1,\dots,s_n}$ are uniquely determined in the following sense.

Corollary 4.1 *Let $n\in\mathbb{Z}_+$ and $0<s_n<\cdots<s_1<s_0:=t\le T$. Assume that there is another function $\tilde{v}^{(n)}_{t,s_1,\dots,s_n}\in L^1(\rho_{init})$ on $\mathbb{R}^n\times\cdots\times\mathbb{R}^n$ such that*

$$\mathbb{E}f(\boldsymbol{h})v^{(n)}_{t,s_1,\dots,s_n}(\boldsymbol{h}_{s_1},\dots,\boldsymbol{h}_{s_n})=\mathbb{E}f(\boldsymbol{h})\tilde{v}^{(n)}_{t,s_1,\dots,s_n}(\boldsymbol{h}_{s_1},\dots,\boldsymbol{h}_{s_n}),\tag{41}$$

for all $f\in C^\infty_c(\mathbb{R}^n)$. Then $v^{(n)}_{t,s_1,\dots,s_n}=\tilde{v}^{(n)}_{t,s_1,\dots,s_n}$ almost everywhere.

We now have the ingredients for the first main result addressing (Q1). The following theorem provides a series representation for the output functional of a SRNN driven by the deterministic signal γ. It says that infinite series of the form (35) are in fact (or can be made sense as) Volterra series (Volterra, 1959; Boyd and Chua, 1985) (see Subsection A.1 for definition of Volterra series and related remarks).

Theorem 4.1 *(Memory representation) Let $t\in[0,T]$. The output functional, $\mathbb{E}f(\boldsymbol{h}_t)$, of the SRNN (1)-(3) is the limit as $N\to\infty$ of*

$$F^{(N)}_t[\gamma]=\sum_{n=1}^{N}\int_{[0,t]^n}ds_1\cdots ds_nR^{(n)}_f(t,s_1,\dots,s_n)\gamma(s_1)\cdots\gamma(s_n),\tag{42}$$

where the $R^{(n)}_f$ are given in Proposition 4.1. The limit exists and is a unique convergent Volterra series. If G_t is another such series with the response functions $Q^{(n)}_f$, then $F_t=G_t$.

Using Theorem 4.1 we can obtain another representation (c.f. Eq. (36)) of the output functional, provided that additional (but reasonable) assumptions are imposed. Recall that we are using Einstein's summation notation for repeated indices.

Theorem 4.2 *(Memoryless representation) Assume that the operator $\mathcal{A}^0$ admits a well-defined eigenfunction expansion. Then, the output functional $\mathbb{E}f(\boldsymbol{h}_t)$ of the SRNN (1)-(3) admits a convergent series expansion, which is the limit as $N,M\to\infty$ of:*

$$F^{(N,M)}_t[\gamma]=\sum_{n=1}^{N}a_{p_0,\dots,p_n,l_1,\dots,l_n}t^{p_0}\int_0^t ds_1s_1^{p_1}u_{s_1}^{l_1}\int_0^{s_1}ds_2s_2^{p_2}u_{s_2}^{l_2}\cdots\int_0^{s_{n-1}}ds_ns_n^{p_n}u_{s_n}^{l_n},\tag{43}$$

where the $a_{p_0,\dots,p_n,l_1,\dots,l_n}$ are constant coefficients that depend on the p_i, the l_i, the eigenvalues and eigenfunctions of $\mathcal{A}^0$, f and ρ_{init} but independent of the input signal and time. Here, $p_i\in\{0,1,\dots,M\}$ and $l_i\in\{1,2,\dots,m\}$.

This representation abstracts away the details of memory induced by the SRNN (thus the name memoryless) so that any parameters defining the SRNN are encapsulated in the coefficients $a_{p_0,\ldots,p_n,l_1,\ldots,l_n}$ (see Eq. (157) for their explicit expression). The parameters are either learnable (e.g., learned via gradient descent using backpropagation) or fixed (e.g., as in the case of reservoir computing). The representation tells us that the iterated integrals in (43) are building blocks used by SRNNs to extract the information in the input signal. See also Remark B.2 and Remark B.3 in **SM**. This result can be seen as a stochastic analog of the one in Chapter 3 of (Isidori, 2013). It can also be shown that Eq. (43) is a generating series in the sense of Chen-Fliess (Fliess, 1981).

Composition of multiple SRNNs (i.e., a deep SRNN), preserves the form of the above two representations. More importantly, as in applications, composition of truncated (finite) series increases the number of nonlinear terms and gives a richer set of response functions and features in the resulting series representation. We make this precise for composition of two SRNNs in the following proposition. Extension to composition of more than two SRNNs is straightforward.

Proposition 4.2 *(Representations for certain deep SRNNs) Let F_t and G_t be the output functional of two SRNNs, with the associated truncated Volterra series having the response kernels $R_f^{(n)}$ and $R_g^{(m)}$, $n = 1,\ldots,N$, $m = 1,\ldots,M$, respectively. Then $(F \circ G)_t[\gamma] = F_t[G_t[\gamma]]$ is a truncated Volterra series with the $N + M$ response kernels:*

$$\begin{aligned} &R_{fg}^{(r)}(t,t_1,\ldots,t_r) \\ &= \sum_{k=1}^{r} \sum_{\mathcal{C}_r} \int_{[0,t]^k} R_f^{(k)}(t,s_1,\ldots,s_k) \\ &\quad \times R_g^{(i_1)}(t-s_1,t_1-s_1,\ldots,t_{i_1}-s_1)\cdots R_g^{(i_k)}(t-s_k,t_{r-i_k+1}-s_k,\ldots,t_r-s_k)ds_1\cdots ds_k, \end{aligned} \tag{44}$$

for $r = 1,\ldots,N+M$, where

$$\mathcal{C}_r = \{(i_1,\ldots,i_k) : i_1,\ldots,i_k \geq 1, 1 \leq k \leq r, i_1 + \cdots + i_k = r\}. \tag{45}$$

If F_t and G_t are Volterra series (i.e., $N, M = \infty$), then $(F \circ G)_t[\gamma]$ is a Volterra series (whenever it is well-defined) with the response kernels $R_{fg}^{(r)}$ above, for $r = 1, 2, \ldots$.

Moreover, the statements in Theorem 4.2 apply to $(F \circ G)_t$, i.e., $(F \circ G)_t$ admits a convergent series expansion of the form (43) *under the assumption in Theorem 4.2.*

Alternatively, the response kernels (44) can be expressed in terms of the exponential Bell polynomials (Bell, 1927) (see Subsection B.6 in **SM**).

Remark 4.3 *It is the combinatorial structure and properties (such as convolution identity and recurrence relations) of the Bell polynomials that underlie the richness of the memory representation of a deep SRNN. To the best of our knowledge, this seems to be the first time where a connection between Bell polynomial and deep RNNs is made. It may be potentially fruitful to further explore this connection to study expressivity and memory capacity of different variants of deep RNNs (Pascanu et al., 2013a).*

Next, we focus on (Q2). The key idea is to link the above representations for the output functional to the notion and properties of path signature (Lyons et al., 2007; Lyons, 2014; Levin et al., 2013; Liao et al., 2019), by lifting the input signal to a tensor algebra space.

Fix a Banach space E in the following. Denote by

$$T((E)) = \{\boldsymbol{a} := (a_0, a_1, \dots) : \forall n \geq 0,\ a_n \in E^{\otimes n}\}, \tag{46}$$

where $E^{\otimes n}$ denotes the n-fold tensor product of E ($E^{\otimes 0} := \mathbb{R}$). This is the space of formal series of tensors of E and can be identified with the free Fock space $T_0((E)) = \bigoplus_{n=0}^{\infty} E^{\otimes n}$ when E is a Hilbert space (Parthasarathy, 2012).

Definition 4.2 *(Signature of a path) Let $\boldsymbol{X} \in C([0,T], E)$ be a path of bounded variation (see Definition A.6). The signature of $\boldsymbol{X}$ is the element S of $T((E))$, defined as*

$$S(\boldsymbol{X}) = (1, \boldsymbol{X}^1, \boldsymbol{X}^2, \dots), \tag{47}$$

where

$$\boldsymbol{X}^n = \int_{\Delta_T^n} d\boldsymbol{X}_{s_1} \otimes \cdots \otimes d\boldsymbol{X}_{s_n} \in E^{\otimes n}, \tag{48}$$

for $n \in \mathbb{Z}_+$, with $\Delta_T^n := \{0 \leq s_1 \leq \cdots \leq s_n \leq T\}$.

Let $(e_{i_1} \otimes \cdots \otimes e_{i_n})_{(i_1,\dots,i_n)\in\{1,\dots,m\}^n}$ be the canonical basis of $E^{\otimes n}$, then we have:

$$S(\boldsymbol{X}) = 1 + \sum_{n=1}^{\infty} \sum_{i_1,\dots,i_n} \left(\int_{\Delta_T^n} dX_{s_1}^{i_1} \cdots dX_{s_n}^{i_n} \right) (e_{i_1} \otimes \cdots \otimes e_{i_n}) \in T((E)). \tag{49}$$

Denoting by $\langle \cdot, \cdot \rangle$ the dual pairing, we have

$$\langle S(\boldsymbol{X}), e_{i_1} \otimes \cdots \otimes e_{i_n} \rangle = \int_{\Delta_T^n} dX_{s_1}^{i_1} \cdots dX_{s_n}^{i_n}. \tag{50}$$

Theorem 4.3 *(Memoryless representation in terms of signature) Let p be a positive integer and assume that the input signal $\boldsymbol{u}$ is a path of bounded variation. Then the output functional F_t of a SRNN is the limit as $p \to \infty$ of $F_t^{(p)}$, which are linear functionals of the signature of the path, $\boldsymbol{X}^{(p)} = \boldsymbol{u} \otimes \boldsymbol{\psi}^{(p)} \in \mathbb{R}^{m\times p}$ (which can be identified with $\mathbb{R}^{mp}$ via vectorizaton), where $\boldsymbol{\psi}^{(p)} = (1, t, t^2, \dots, t^{p-1}) \in \mathbb{R}^p$, i.e.,*

$$F_t^{(p)}[\boldsymbol{u}] = \sum_n b_n(t) \langle S(\boldsymbol{X}_t^{(p)}), e_{i_1} \otimes \cdots \otimes e_{i_n} \rangle, \tag{51}$$

where the $b_n(t)$ are coefficients that only depend on t.

It was shown in (Lyons, 2014) that the signature is a universal feature set, in the sense that any continuous map can be approximated by a linear functional of signature of the input signal (see also Remark A.3). This result is a direct consequence of the Stone-Weierstrass theorem. On the other hand, Theorem 4.3 implies that the output functional of SRNNs admits a linear representation in terms of signature of a time-augmented input signal, a lift of the original input signal to higher dimension to account for time. We shall call the signature of this time-augmented input signal the *response feature*, which is richer as a feature set than the signature of the input signal and may be incorporated in a machine learning framework for learning sequential data.

Remark 4.4 *Note that our SRNN* (1)-(3) *can be interpreted as a controlled differential equation (CDE) (Lyons et al., 2007). We emphasize that while the connection between signature and CDEs are well known within the rough paths community, it is established explicitly using local Taylor approximations (see, for instance, (Boedihardjo et al., 2015) or Section 4 of (Liao et al., 2019)) that ignore any (non-local) memory effects, whereas here we establish such connection explicitly by deriving a signature-based representation from the memory representation that takes into account the memory effects (albeit under more restrictive assumptions). It is precisely because of the global (in time) nature of our approximation that the resulting representation for our particular output functionals is in terms of signature of a time-augmented signal and not simply in terms of signature of the signal itself. Our response theory based approach offers alternative, arguably more intuitive, perspectives on how signature arise in SRNNs. This may be useful for readers not familiar with rough paths theory.*

4.3 Formulating SRNNs as Kernel Machines

We now consider a supervised learning (regression or classification) setting where we are given N training input-output pairs $(\boldsymbol{u}_n, y_n)_{n=1,\dots,N}$, where the $\boldsymbol{u}_n \in \chi$, the space of paths in $C([0,T],\mathbb{R}^m)$ with bounded variation, and $y_n \in \mathbb{R}$, such that $y_n = F_T[\boldsymbol{u}_n]$ for all n. Here F_T is a continuous target mapping.

Consider the optimization problem:

$$\min_{\hat{F}\in\mathcal{G}}\{L(\{(\boldsymbol{u}_n, y_n, \hat{F}[\boldsymbol{u}_n])\}_{n=1,\dots,N}) + R(\|\hat{F}\|_{\mathcal{G}})\}, \tag{52}$$

where $\mathcal{G}$ is a hypothesis (Banach) space with norm $\|\cdot\|_{\mathcal{G}}$, $L : (\chi \times \mathbb{R}^2)^N \to \mathbb{R} \cup \{\infty\}$ is a loss functional and $R(x)$ is a strictly increasing real-valued function in x.

Inspired by Theorem 4.3 (viewing $\mathcal{G}$ as a hypothesis space induced by the SRNNs), we are going to show that the solution to this problem can be expressed as kernel expansion over the training examples (c.f. (Evgeniou et al., 2000; Schölkopf et al., 2002; Hofmann et al., 2008)).

In the following, consider the Hilbert space

$$\mathcal{H} := \mathcal{P} \otimes \mathbb{R}^m, \tag{53}$$

where $\mathcal{P}$ is the appropriately weighted l^2 space of sequences of the form $(P_0(t), P_1(t), \cdots)$ with the $P_n(t)$ orthogonal polynomials on $[0,T]$. Let $T_s((\mathcal{H}))$ denote the symmetric Fock space over $\mathcal{H}$ (see Subsection B.8 for definition) and $T_s^{\otimes L}((\mathcal{H}))$ denote L-fold tensor product of $T_s((\mathcal{H}))$ for $L \in \mathbb{Z}_+$.

Proposition 4.3 *Let $L \in \mathbb{Z}_+$. Consider the map $K : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$, defined by*

$$K(\boldsymbol{v}, \boldsymbol{w}) = \langle S(\boldsymbol{v}), S(\boldsymbol{w})\rangle_{T_s^{\otimes L}((\mathcal{H}))}. \tag{54}$$

Then K is a kernel over $\mathcal{H}$ and there exists a unique RKHS, denoted $\mathcal{R}_L$ with the norm $\|\cdot\|_{\mathcal{R}_L}$, for which K is a reproducing kernel.

One can view SRNNs, with only the weights in the readout layer optimized and the weights in the hidden layer kept fixed and not optimized, as kernel machines operating on a RKHS associated with the response feature. In particular, we can see the continuous-time version of echo state networks in reservoir computing (Lukoševičius and Jaeger, 2009) as a kernel method. This is captured precisely in the following theorem.

Theorem 4.4 *(Representer theorem) Consider the time-augmented paths $\boldsymbol{X}_n = \boldsymbol{v} \otimes \boldsymbol{u}_n$, where the $\boldsymbol{u}_n$ are $\mathbb{R}^m$-valued input paths in χ and $\boldsymbol{v}$ is a $\mathbb{R}^{\mathbb{Z}_+}$-valued vector in $\mathcal{P}$. Then:*
(a) Any solution to the optimization problem (52) *with the hypothesis space of $\mathcal{G} := \mathcal{R}_1$ admits a representation of the form:*

$$\hat{F}_t^* = \sum_{n=1}^{N} c_n \langle S(\boldsymbol{X}_n), S(\boldsymbol{X}) \rangle_{T_s((\mathcal{H}))}, \tag{55}$$

where the $c_n \in \mathbb{R}$ and N is the number of training input-output pairs.

(b) Let $L \in \mathbb{Z}_+$. If we instead consider the paths, denoted $\tilde{\boldsymbol{X}}_n$, obtained by linear interpolating on the $L+1$ data points $(\boldsymbol{X}_n(t_i))_{i=0,1,\ldots,L}$ sampled at time $t_i \in [0,T]$, then any solution to the corresponding optimization problem (52) *with the hypothesis space of $\mathcal{G} := \mathcal{R}_L$ admits a representation of the form:*

$$\hat{F}_t^* = \sum_{n=1}^{N} \alpha_n \prod_{l=1}^{L} \exp\left(\left\langle \Delta \boldsymbol{X}_n^{(l)}, \Delta \boldsymbol{X}^{(l)} \right\rangle_{\mathcal{H}} \right), \tag{56}$$

where the $\alpha_n \in \mathbb{R}$ and $\Delta \boldsymbol{X}^{(l)} := \boldsymbol{X}(t_l) - \boldsymbol{X}(t_{l-1})$ for $l = 1, \ldots, L$.

We emphasize that the idea of representing Volterra series as elements of a RKHS is certainly not new (see, e.g., (Zyla and deFigueiredo, 1983; Franz and Schölkopf, 2006)). However, it is the use of the response feature here that makes our construction differs from those in previous works (c.f. (Király and Oberhauser, 2019; Toth and Oberhauser, 2020)). The representer theorem here is obtained for a wide class of optimization problems using an alternative approach in the SRNN setting. The significance of the theorem obtained here is not only that the optimal solution will live in a subspace with dimension no greater than the number of training examples but also how the solution depends on the number of samples in the training data stream. Note that the appearance of orthogonal polynomials here is not too surprising given their connection to nonlinear dynamical systems (Kowalski, 1997; Mauroy et al., 2020). Lastly, we remark that the precise connections between (finite width) RNNs whose all weights are optimized (via gradient descent) in deep learning and kernel methods remain an open problem and we shall leave it to future work.

5. Conclusion

In this paper we have addressed two fundamental questions concerning a class of stochastic recurrent neural networks (SRNNs), which can be models for artificial or biological networks, using the nonlinear response theory from nonequilibrium statistical mechanics as

the starting point. In particular, we are able to characterize, in a systematic and order-by-order manner, the response of the SRNNs to a perturbing deterministic input signal by deriving two types of series representation for the output functional of these SRNNs and a deep variant in terms of the driving input signal. This provides insights into the nature of both memory and memoryless representation induced by these driven networks. Moreover, by relating these representations to the notion of path signature, we find that the set of response feature is the building block in which SRNNs extract information from when processing an input signal, uncovering the universal mechanism underlying the operation of SRNNs. In particular, we have shown, via a representer theorem, that SRNNs can be viewed as kernel machines operating on a reproducing kernel Hilbert space associated with the response feature.

We end with a few final remarks. From mathematical point of view, it would be interesting to relax the assumptions here and work in a general setting where the driving input signal is a rough path, where regularity of the input signal could play an important role. One could also study how SRNNs respond to perturbations in the input signal and in the driving noise (regularization) by adapting the techniques developed here. So far we have focused on the "formulate first" approach mentioned in Introduction. The results obtained here suggest that one could study the "discretize next" part by devising efficient algorithm to exploit the use of discretized response features and related features in machine learning tasks involving temporal data, such as predicting time series generated by complex dynamical systems arising in science and engineering.

Acknowledgments

The author is grateful to the support provided by the Nordita Fellowship 2018-2021.

Supplementary Material (SM)

Appendix A. Preliminaries and Mathematical Formulation

A.1 Differential Calculus on Banach Spaces

In this subsection, we present elements of differential calculus on Banach spaces, introducing our notation and terminology along the way. We refer to the classic book of Cartan (Cartan, 1983) for more details (see also (Abraham et al., 2012)).

We will need the notion of functional derivatives before we dive into response functions. Functional derivatives are generalization of ordinary derivatives to functionals. At a formal level, they can be defined via the variation δF of the functional $F[u]$ which results from variation of u by δu, i.e., $\delta F = F[u + \delta u] - F[u]$. The technique used to evaluate δF is a Taylor expansion of the functional $F[u+\delta u] = F[u+\epsilon\eta]$ in powers of δu or ϵ. The functional $F[u + \epsilon\eta]$ is an ordinary function of ϵ. This implies that the expansion in terms of powers of ϵ is a standard Taylor expansion, i.e.,

$$F[u + \epsilon\eta] = F[u] + \frac{dF[u + \epsilon\eta]}{d\epsilon}\bigg|_{\epsilon=0} \epsilon + \frac{1}{2!}\frac{d^2F[u + \epsilon\eta]}{d\epsilon^2}\bigg|_{\epsilon=0} \epsilon^2 + \cdots, \tag{57}$$

provided that the "derivatives" above can be made sense of. We first define such "derivatives".

Recall that for a function f of a real variable, the derivative of f is defined by

$$f'(x) = \lim_{h\to 0} \frac{f(x + h) - f(x)}{h}, \tag{58}$$

provided that the limit exists. This definition becomes absolete when f is a function of vector variable since then division by a vector is meaningless. Therefore, to define a derivative for mappings from Banach space into a Banach space, one needs to revise the above definition. This leads to the notion of Fréchet differentiability, generalizing the notion of slope of the line tangent to the graph of the function at some point.

In the following, let E_1, E_2 be Banach spaces over $\mathbb{R}$, $T : E_1 \to E_2$ be a given mapping, and $\|\cdot\|$ represents the norm on appropriate space.

Definition A.1 *(Fréchet differentiabiliy) Fix an open subset U of a Banach space E_1 and let $u_0 \in U$. We say that the mapping $T : U \to E_2$ is Fréchet differentiable at the point u_0 if there is a bounded linear map $DT(u_0) : E_1 \to E_2$ such that for every $\epsilon > 0$, there is a $\delta > 0$ such that*

$$\frac{\|T(u_0 + h) - T(u_0) - DT(u_0) \cdot h\|}{\|h\|} < \epsilon \tag{59}$$

whenever $\|h\| \in (0, \delta)$, where $DT(u_0) \cdot e$ denotes the evaluation of $DT(u_0)$ on $e \in E_1$. This can also be written as

$$\lim_{\|h\|\to 0} \frac{\|T(u_0 + h) - T(u_0) - DT(u_0) \cdot h\|}{\|h\|} = 0. \tag{60}$$

Note that this is equivalent to the existence of a linear map $DT(u_0) \in L(E_1, E_2)$ such that

$$T(u_0 + h) - T(u_0) = DT(u_0) \cdot h + e(h) \tag{61}$$

where

$$\lim_{\|h\|\to 0} \frac{\|e(h)\|}{\|h\|} = 0, \tag{62}$$

i.e., $e(h) = o(\|h\|)$ *as* $\|h\| \to 0$.

Definition A.2 *(Fréchet derivative) If the mapping* T *is Fréchet differentiable at each* $u_0 \in U$*, the map* $DT : U \to L(E_1, E_2)$*,* $u \mapsto DT(u)$*, is called the (first order) Fréchet derivative of* T*. Moreover, if* DT *is a norm continuous map, we say that* T *is of class* C^1*. We define, inductively, for* $r \geq 2$*, the* r*-th order derivative* $D^rT := D(D^{r-1}T) : U \subset E_1 \to L^{(r)}(E_1, E_2) := L(E_1, L^{(r-1)}(E_1, E_2))$*, with* $L^{(1)}(E_1, E_2) := L(E_1, E_2)$*, whenever it exists. If* D^rT *exists and is norm continuous, we say that* T *is of class* C^r*.*

A few remarks follow. A weaker notion of differentiability, generalizing the idea of directional derivative in finite dimensional spaces, is provided by Gateaux.

Definition A.3 *(Gateaux differentiability and derivative) Fix* $u_0 \in E_1$*. The mapping* $T : E_1 \to E_2$ *is Gateaux differentiable at* u_0 *if there exists a continuous linear operator* A *such that*

$$\lim_{\epsilon\to 0} \left\| \frac{T(u_0 + \epsilon h) - T(u_0)}{\epsilon} - A(h) \right\| = 0 \tag{63}$$

for every $h \in E_1$*, where* $\epsilon \to 0$ *in* $\mathbb{R}$*. The operator* A *is called the Gateaux derivative of* T *at* u_0 *and its value at* h *is denoted by* $A(h)$*. The higher order Gateaux derivatives can be defined by proceeding inductively.*

Remark A.1 *It is a standard result that if the Gateaux derivative exists, then it is unique (similarly for Fréchet derivative). Gateaux differentiability is weaker than that of Fréchet in the sense that if a mapping has the Fréchet derivative at a point* u_0*, then it has the Gateaux derivative at* u_0 *and both derivatives are equal, in which case* $A(h) = DT(u_0) \cdot h$ *using the notations in the above definitions. The converse does not generally hold. For basic properties of these derivatives, we refer to the earlier references. From now on, we will work with the more general notion of Fréchet differentiability.*

We now introduce the notion of functional derivative for the special case of mappings that are real-valued functionals. Let $E_2 = \mathbb{R}$, so that T is a functional on E_1. When T is Fréchet differentiable at some $u \in E_1$, its derivative is a bounded linear functional on E_1, i.e., $DT[u] \in L(E_1, \mathbb{R}) =: E_1^*$. If E_1 is a Hilbert space, then by the Riesz representation theorem[4], there exists a unique element $y \in E_1$ such that $DT[u] \cdot e = \langle y, e\rangle$ for every $e \in E_1$. The derivative $DT[u]$ can thus be identified with y, which we sometimes call the gradient of T. In the case when E_1 is a functional space, we call the derivative the *functional derivative* of T with respect to u, denoted $\delta T/\delta u$:

$$\langle \delta T/\delta u, e\rangle = DT[u] \cdot e. \tag{64}$$

Moreover, T is also Gateaux differentiable at $u \in E_1$ and we have $DT[u] \cdot e = \frac{d}{dt}\big|_{t=0} T[u + te]$.

4. Note that we cannot apply this fact when defining the response functions in Definition 4.1 since the space of test functions is not a Hilbert space.

The higher order derivatives can be defined inductively. For instance, if T is a twice differentiable functional on a real Hilbert space E_1, then for all $u \in E_1$, $D^2T[u] \cdot (e_1, e_2) = D((DT)(\cdot) \cdot e_2)[u] \cdot e_1$. More generally, if $e_1, \ldots, e_{n-1} \in E_1$, $T : U \to E_2$ is n times differentiable, then the map $Q : U \to E_2$ defined by $Q[u] = D^{n-1}T[u] \cdot (e_1, \ldots, e_{n-1})$ is differentiable and $DQ[u] \cdot e = D^nT[u] \cdot (e, e_1, \ldots, e_{n-1})$. Moreover, $D^nT[u]$ is multi-linear symmetric, i.e.,

$$D^nT[u] \cdot (e_0, e_1, \ldots, e_{n-1}) = D^nT[u] \cdot (e_{\sigma(0)}, e_{\sigma(1)}, \ldots, e_{\sigma(n-1)}), \tag{65}$$

where σ is any permutation of $\{0, 1, \ldots, n\}$ (see Theorem 5.3.1 in (Cartan, 1983)). In this case, we have

$$\lim_{\epsilon \to 0} \left\| \frac{(D^{n-1}T[u + \epsilon h] - D^{n-1}T[u]) \cdot (h_1, \ldots, h_{n-1})}{\epsilon} - D^nT[u] \cdot (h, h_1, \ldots, h_{n-1}) \right\| = 0 \tag{66}$$

and we write

$$D^nT[u] \cdot (h, h_1, \ldots, h_{n-1}) = \lim_{\epsilon \to 0} \frac{(D^{n-1}T[u + \epsilon h] - D^{n-1}T[u]) \cdot (h_1, \ldots, h_{n-1})}{\epsilon}. \tag{67}$$

The notion of functional derivatives that we will need is defined in the following.

Definition A.4 *(Functional derivatives) Fix $t > 0$. Let F be a functional on $C([0, t], \mathbb{R})$ and $u \in C([0, t], \mathbb{R})$. Then for $n \in \mathbb{Z}_+$, the nth order functional derivative of F with respect to u is a functional $\delta^n F / \delta \boldsymbol{u}^{(n)}$, with $\delta \boldsymbol{u}^{(n)} := \delta u(s_1) \cdots \delta u(s_n)$, on $C_c^\infty(0, t)$, defined as:*

$$\int_{(0,t)^n} ds_1 \cdots ds_n \frac{\delta^n F[u]}{\delta \boldsymbol{u}^{(n)}} \phi(s_1) \cdots \phi(s_n) = D^nF[u] \cdot (\phi, \phi, \ldots, \phi), \tag{68}$$

whenever the derivative exists, for any $\phi \in C_c^\infty(0, t)$.

Recall the following Taylor's theorem for a mapping $T : U \to E_2$.

Theorem A.1 *(Taylor's theorem – Theorem 5.6.3 in (Cartan, 1983)) Let T be an $n - 1$ times differentiable mapping. Suppose that T is n times differentiable at the point $u \in U$. Then, denoting $(h)^n := (h, \ldots, h)$ (n times),*

$$\left\| T(u + h) - T(u) - DT(u) \cdot h - \cdots - \frac{1}{n!} D^nT(u) \cdot (h)^n \right\| = o(\|h\|^n). \tag{69}$$

By *Taylor series* of a mapping T at a point $u \in U$, we mean the series of homogeneous polynomials given by

$$T(u + h) = T(u) + DT(u) \cdot h + \cdots + \frac{1}{n!} D^nT(u) \cdot (h)^n + \ldots \tag{70}$$

that is absolutely convergent.

An important example of Taylor series of a mapping is given by the Volterra series, which is widely studied in systems and control theory (Boyd and Chua, 1985; Brockett, 1976; Fliess and Hazewinkel, 2012).

Definition A.5 *(Volterra series (Boyd et al., 1984)) Let $u \in C([0,t],\mathbb{R})$. A Volterra series operator is an operator N given by:*

$$N_t[u] = h_0 + \sum_{n=1}^{\infty} \int_{[0,t]^n} h_n(s_1,\dots,s_n)u(t-s_1)\cdots u(t-s_n)ds_1\cdots ds_n, \tag{71}$$

satisfying $h_0 \in \mathbb{R}$, $h_n \in B(\mathbb{R}_+^n)$, and $a := \limsup_{n\to\infty} \|h_n\|_\infty^{1/n} < \infty$, i.e., $\{\|h_n\|_\infty^{1/n}\}_{n\in\mathbb{Z}_+}$ is bounded.

It can be shown that the integrals and sum above converge absolutely for inputs with $\|u\|_\infty < \rho := 1/a$, i.e., for u belonging to the ball of radius ρ, denoted B_ρ, in L^∞. Moreover, $\|N_t[u]\|_\infty \le g(\|u\|_\infty) := |h_0| + \sum_{n=1}^{\infty} \|h_n\|_\infty \|u\|_\infty^n$. Also, N_t is a continuous map from B_ρ into L^∞ and has bounded continuous Fréchet derivatives of all orders, with (Boyd and Chua, 1985)

$$\begin{aligned} &D^k N_t[u]\cdot(u_1,\dots,u_k) \\ &= \sum_{n=k}^{\infty} n(n-1)\dots(n-k+1) \int_{[0,t]^n} SYM h_n(s_1,\dots,s_n) \prod_{i=1}^{k} u_i(t-s_i)ds_i \prod_{i=k+1}^{n} u(t-s_i)ds_i, \end{aligned} \tag{72}$$

where $SYM h_n(s_1,\dots,s_n) := \frac{1}{n!}\sum_{\sigma\in S_n} h_n(s_{\sigma(1)},\dots,s_{\sigma(n)})$, with S_n the group of all permutations of the set $\{1,\dots,n\}$.

When evaluated at $u = 0$, these derivatives can be associated with the nth term of the Volterra series and are given by:

$$\frac{1}{n!} D^n N_t[0]\cdot(u_1,\dots,u_n) = \int_{[0,t]^n} h_n(s_1,\dots,s_n)u_1(t-s_1)\cdots u_n(t-s_n)ds_1\cdots ds_n. \tag{73}$$

Therefore, the Volterra series above are in fact Taylor series of operators from L^∞ to L^∞.

A.2 Signature of a Path

We provide some background on the signature of a path. The signature is an object arising in the rough paths theory, which provide an elegant yet robust nonlinear extension of the classical theory of differential equations driven by irregular signals (such as the Brownian paths). In particular, the theory allows for deterministic treatment of stochastic differential equations (SDEs).

For our purpose, we do not need the full machinery of the theory but only a very special case. We refer to (Lyons and Qian, 2002; Lyons et al., 2007; Friz and Victoir, 2010; Friz and Hairer, 2014) for full details. In particular, we will only consider signature for paths with bounded variation. The following materials are borrowed from (Lyons et al., 2007).

Fix a Banach space E (with norm $|\cdot|$) in the following. We first recall the notion of paths with bounded variation.

Definition A.6 *(Path of bounded variation) Let $I = [0,T]$ be an interval. A continuous path $\boldsymbol{X} : I \to E$ is of bounded variation (or has finite 1-variation) if*

$$\|\boldsymbol{X}\|_{1-var} := \sup_{\mathcal{D}\subset I} \sum_{i=0}^{n-1} |\boldsymbol{X}_{t_{i+1}} - \boldsymbol{X}_{t_i}| < \infty, \tag{74}$$

where $\mathcal{D}$ is any finite partition of I, i.e., an increasing sequence $(t_0, t_1, \dots, t_n)$ such that $0 \le t_0 < t_1 < \cdots < t_n \le T$.

When equipped with the norm $\|\boldsymbol{X}\|_{BV} := \|\boldsymbol{X}\|_{1-\text{var}} + \|\boldsymbol{X}\|_\infty$, the space of continuous paths of bounded variation with values in E becomes a Banach space.

The following lemma will be useful later.

Lemma A.1 *Let $E = \mathbb{R}$. If the path X is increasing on $[0,T]$, then X is of bounded variation on $[0,T]$. Moreover, if Y and Z are paths of bounded variation on $[0,T]$ and k is a constant, then $X+Y$, $X-Y$, XY and kX are paths of bounded variation on $[0,T]$.*

Proof The first statement is an easy consequence of the telescoping nature of the sum in (74) for any finite partitions of $[0,T]$. In fact, $\|X\|_{1-\text{var}} = X_T - X_0$. For the proof for the last statement, we refer to Theorem 2.4 in (Grady, 2009). ■

The signature of a continuous path with bounded variation lives in some tensor algebra space, which we now elaborate on.

Denote by $T((E)) = \{\boldsymbol{a} := (a_0, a_1, \dots) : \forall n \in \mathbb{N},\ a_n \in E^{\otimes n}\}$ a tensor algebra space, where $E^{\otimes n}$ denotes the n-fold tensor product of E ($E^{\otimes 0} := \mathbb{R}$), which can be identified with the space of homogeneous noncommuting polynomials of degree n. $T((E))$ is the space of formal series of tensors of E and, when endowed with suitable operations, becomes a real non-commutative algebra (with unit $\Omega = (1, 0, 0, \dots)$). When E is a Hilbert space, it can be identified with the free Fock space, $T_0((E)) := \bigoplus_{n=0}^\infty E^{\otimes n}$ (Parthasarathy, 2012).

As an example/reminder, consider the case when $E = \mathbb{R}^m$. In this case, if $(e_1, \dots, e_m)$ is a basis of E, then any element $\boldsymbol{x}_n \in E^{\otimes n}$ can be written as the expansion:

$$\boldsymbol{x}_n = \sum_{I=(i_1,\dots,i_n)\subset\{1,\dots,m\}^n} a_I (e_{i_1} \otimes \cdots \otimes e_{i_n}) \tag{75}$$

where the a_I are scalar coefficients. For any nonnegative integer n, the tensor space $E^{\otimes n}$ can be endowed with the inner product $\langle \cdot, \cdot \rangle_{E^{\otimes n}}$ (defined in the usual way (Parthasarathy, 2012)). Then, for $\boldsymbol{a} = (a_0, a_1, \dots)$ and $\boldsymbol{b} = (b_0, b_1, \dots)$ in $T((E))$, we can define an inner product in $T((E))$ by:

$$\langle \boldsymbol{a}, \boldsymbol{b} \rangle_{T((E))} = \sum_{n \ge 0} \langle a_n, b_n \rangle_{E^{\otimes n}}. \tag{76}$$

We now define the signature of a path as an element in $T((E))$.

Definition A.7 *(Definition 4.2 in the main paper) Let $\boldsymbol{X} \in C([0,T], E)$ be a path of bounded variation. The signature of $\boldsymbol{X}$ is the element S of $T((E))$, defined as*

$$S(\boldsymbol{X}) = (1, \boldsymbol{X}^1, \boldsymbol{X}^2, \dots), \tag{77}$$

where, for $n \ge 1$, $\Delta_T^n := \{0 \le s_1 \le \cdots \le s_n \le T\}$, and

$$\boldsymbol{X}^n = \int_{\Delta_T^n} d\boldsymbol{X}_{s_1} \otimes \cdots \otimes d\boldsymbol{X}_{s_n} \in E^{\otimes n}. \tag{78}$$

Note that in the definition above since $\boldsymbol{X}$ is a path of bounded variation, the integrals are well-defined as Riemann-Stieljes integrals. The signature in fact determines the path completely (up to tree-like equivalence); see Theorem 2.29 in (Lyons et al., 2007).

Let $(e_{i_1} \otimes \cdots \otimes e_{i_n})_{(i_1,\ldots,i_n)\in\{1,\ldots,m\}^n}$ be the canonical basis of $E^{\otimes n}$, then we have:

$$S(\boldsymbol{X}) = 1 + \sum_{n=1}^{\infty} \sum_{i_1,\ldots,i_n} \left(\int_{\Delta_T^n} dX_{s_1}^{i_1} \cdots dX_{s_n}^{i_n} \right) (e_{i_1} \otimes \cdots \otimes e_{i_n}) \in T((E)). \tag{79}$$

Denoting by $\langle \cdot, \cdot \rangle$ the dual pairing, we have $\langle S(\boldsymbol{X}), e_{i_1} \otimes \cdots \otimes e_{i_n} \rangle = \int_{\Delta_T^n} dX_{s_1}^{i_1} \cdots dX_{s_n}^{i_n}$.

Next, we discuss a few special cases where the signature can be computed explicitly. If $\boldsymbol{X} := X$ is a one-dimensional path, then we can compute the nth level signature to be $\boldsymbol{X}^n = (X_T - X_0)^n/n!$. If $\boldsymbol{X}$ is an E-valued linear path, then $\boldsymbol{X}^n = (\boldsymbol{X}_T - \boldsymbol{X}_0)^{\otimes n}/n!$ and $S(\boldsymbol{X}) = \exp((\boldsymbol{X}_T - \boldsymbol{X}_0))$.

We will need the following fact, which is useful in itself for numerical computation in practice, later.

Lemma A.2 *If $\boldsymbol{X}$ is a E-valued piecewise linear path, i.e., $\boldsymbol{X}$ is obtained by concatenating L linear paths, $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(L)}$, such that $\boldsymbol{X} = \boldsymbol{X}^{(1)} \star \cdots \star \boldsymbol{X}^{(L)}$ (with $\star$ denoting path concatenation), then:*

$$S(\boldsymbol{X}) = \bigotimes_{l=1}^{L} S(\boldsymbol{X}^{(l)}) = \bigotimes_{l=1}^{L} \exp\left(\Delta \boldsymbol{X}^{(l)}\right), \tag{80}$$

where the $\Delta \boldsymbol{X}^{(l)}$ denotes the increment of the path $\boldsymbol{X}^{(l)}$ between its endpoints.

Proof The lemma follows from applying Chen's identity (see Theorem 2.9 in (Lyons et al., 2007)) iteratively and exploiting linearity of the paths $\boldsymbol{X}^{(l)}$. ■

A few remarks are now in order.

Remark A.2 *By definition, the signature is a collection of definite iterated integrals of the path (Chen et al., 1977). It appears naturally as the basis to represent the solution to a linear controlled differential equation via the Picard iteration (c.f. Lemma 2.10 in (Lyons et al., 2007)). The signature of the path is in fact involved in a lot of rich algebraic and geometric structures. In particular, the first level signature, $\boldsymbol{X}^1$, is the increment of the path, i.e., $\boldsymbol{X}_T - \boldsymbol{X}_0$, the second level signature, $\boldsymbol{X}^2$, represents the signed area enclosed by the path and the cord connecting the ending and starting point of the path. For further properties of the signature, we refer to the earlier references. Interestingly, iterated integrals are also objects of interest in quantum field theory and renormalization (Brown, 2013; Kreimer, 1999).*

Remark A.3 *From the point of view of machine learning, the signature is an efficient summary of the information contained in the path and leads to invention of the Signature Method for ordered data (see, for instance, (Chevyrev and Kormilitzin, 2016) for a tutorial and the related works by the group of Terry Lyons). More importantly, the signature is*

a universal feature, in the sense that any continuous functionals on compact sets of paths can be approximated arbitrarily well by a linear functional on the signature. This is stated precisely in Theorem 3.1 in (Levin et al., 2013) (see also Theorem B.1 in (Morrill et al., 2020) with log-signature as the feature instead).

Appendix B. Proof of Main Results and Further Remarks

B.1 Auxiliary Lemmas

Recall that we assume Assumption 4.1 to hold throughout the paper unless stated otherwise. The following lemma on boundedness and continuity of amplitude of the input perturbation will be essential.

Lemma B.1 *Let $\gamma = (\gamma(t))_{t\in[0,T]}$, i.e., the path measuring the amplitude of $\boldsymbol{C}\boldsymbol{u}_t$ driving the SRNN. Then $\gamma \in C([0,T],\mathbb{R})$ and thus bounded.*

Proof Note that using boundedness of $\boldsymbol{C}$, for $0 \le s < t \le T$:

$$||\boldsymbol{C}\boldsymbol{u}_t| - |\boldsymbol{C}\boldsymbol{u}_s|| \le |\boldsymbol{C}\boldsymbol{u}_t - \boldsymbol{C}\boldsymbol{u}_s| \le |\boldsymbol{C}(\boldsymbol{u}_t - \boldsymbol{u}_s)| \le C|\boldsymbol{u}_t - \boldsymbol{u}_s|, \tag{81}$$

where $C > 0$ is a contant. Therefore, continuity of γ follows from continuity of $\boldsymbol{u}$. That γ is bounded follows from the fact that continuous functions on compact sets are bounded. ■

In the sequel, for two sets A and B, $A\backslash B$ denotes the set difference of A and B, i.e., the set of elements in A but not in B.

The following lemma will be useful later.

Lemma B.2 *Let a_i, b_i be operators, for $i = 1,\dots,N$. Then, for $N \ge 2$,*
(a)

$$\prod_{n=1}^{N} a_n - \prod_{m=1}^{N} b_m = \sum_{k=1}^{N} \left(\prod_{l=1}^{k-1} a_l\right)(a_k - b_k)\left(\prod_{p=k+1}^{N} b_p\right), \tag{82}$$

(b)

$$\begin{aligned}
&\prod_{n=1}^{N} a_n - \prod_{m=1}^{N} b_m \\
&= \sum_{k=1}^{N} \left(\prod_{l=1}^{k-1} b_l\right)(a_k - b_k)\left(\prod_{p=k+1}^{N} b_p\right) + \sum_{k=1}^{N}\Bigg(\sum_{p_1,\dots,p_{k-1}\in\Omega} b_1^{p_1}(a_1 - b_1)^{1-p_1} \\
&\quad \times \cdots b_{k-1}^{p_{k-1}}(a_{k-1} - b_{k-1})^{1-p_{k-1}}\Bigg)(a_k - b_k)\left(\prod_{p=k+1}^{N} b_p\right).
\end{aligned} \tag{83}$$

whenever the additions and multiplications are well-defined on appropriate domain. In the above,

$$\Omega := \{p_1,\dots,p_{k-1} \in \{0,1\}\}\backslash\{p_1 = \cdots = p_{k-1} = 1\}, \tag{84}$$

and we have used the convention that $\prod_{l=1}^{0} a_l := I$ *and* $\prod_{p=N+1}^{N} b_p := I$*, where* I *denotes identity.*

Proof
(a) We prove by induction. For the base case of $N = 2$, we have:

$$a_1a_2 - b_1b_2 = a_1(a_2 - b_2) + (a_1 - b_1)b_2, \tag{85}$$

and so (82) follows for $N = 2$. Now, assume that (82) is true for $M > 2$. Then,

$$\prod_{n=1}^{M+1} a_n - \prod_{m=1}^{M+1} b_m = (a_1 \cdots a_M)(a_{M+1} - b_{M+1}) + \left(\prod_{n=1}^{M} a_n - \prod_{m=1}^{M} b_m\right) b_{M+1} \tag{86}$$

$$= (a_1 \cdots a_M)(a_{M+1} - b_{M+1}) + \sum_{k=1}^{M} \left(\prod_{l=1}^{k-1} a_l\right)(a_k - b_k)\left(\prod_{p=k+1}^{M} b_p\right) b_{M+1} \tag{87}$$

$$= \sum_{k=1}^{M+1} \left(\prod_{l=1}^{k-1} a_l\right)(a_k - b_k)\left(\prod_{p=k+1}^{M+1} b_p\right). \tag{88}$$

Therefore, the formula (82) holds for all $N \geq 2$.

(b) By part (a), we have:

$$\prod_{n=1}^{N} a_n - \prod_{m=1}^{N} b_m = \sum_{k=1}^{N} \left(\prod_{l=1}^{k-1} a_l\right)(a_k - b_k)\left(\prod_{p=k+1}^{N} b_p\right) \tag{89}$$

$$= \sum_{k=1}^{N} \left(\prod_{l=1}^{k-1} (b_l + (a_l - b_l))\right)(a_k - b_k)\left(\prod_{p=k+1}^{N} b_p\right). \tag{90}$$

Note that

$$\prod_{l=1}^{k-1} (b_l + (a_l - b_l))$$

$$= \sum_{p_1=0}^{1} \cdots \sum_{p_{k-1}=0}^{1} b_1^{p_1}(a_1 - b_1)^{1-p_1} \cdots b_{k-1}^{p_{k-1}}(a_{k-1} - b_{k-1})^{1-p_{k-1}} \tag{91}$$

$$= \prod_{l=1}^{k-1} b_l + \sum_{p_1,\ldots,p_{k-1} \in \Omega} b_1^{p_1}(a_1 - b_1)^{1-p_1} \cdots b_{k-1}^{p_{k-1}}(a_{k-1} - b_{k-1})^{1-p_{k-1}}, \tag{92}$$

where $\Omega = \{p_1, \ldots, p_{k-1} \in \{0, 1\}\} \backslash \{p_1 = \cdots = p_{k-1} = 1\}$. Eq. (82) then follows from (90) and (92). ■

B.2 Proof of Proposition 3.1

Our proof is adapted from and built on that in (Chen and Jia, 2020), which provides a rigorous justification of nonequilibrium FDTs for a wide class of diffusion processes and nonlinear input perturbations in the linear response regime (see also the more abstract approach in (Hairer and Majda, 2010; Dembo and Deuschel, 2010)). We are going to extend the techniques in (Chen and Jia, 2020) to study the fully nonlinear response regime in the context of our SRNNs.

First, note that Assumption 4.1 implies that the processes $\boldsymbol{h}$ and $\overline{\boldsymbol{h}}$ (for all γ) automatically satisfy the regular conditions (Definition 2.2) in (Chen and Jia, 2020). Therefore, it follows from Proposition 2.5 in (Chen and Jia, 2020) that the weak solution of the SDE (5) exists up to time T and is unique in law. In particular, $\boldsymbol{h}$ and $\overline{\boldsymbol{h}}$ are nonexplosive up to time T. Moreover, if $n = r$ and $\boldsymbol{\sigma} > 0$, the strong solution of the SDE exists up to time T and is pathwise unique.

We collect some intermediate results that we will need later. Recall that $\Delta\mathcal{L}_t := \mathcal{L}_t - \mathcal{L}^0$, for $t \in [0, T]$, where the infinitesimal generators $\mathcal{L}_t$ and $\mathcal{L}^0$ are defined in (12) and (13) in the main paper respectively. Also, we are using Einstein's summation convention.

Lemma B.3 *For any $f \in C_b(\mathbb{R}^n)$ and $0 \le s \le t \le T$, the function $v(\boldsymbol{h}, s) = P^0_{s,t} f(\boldsymbol{h}) \in C_b(\mathbb{R}^n \times [0, t]) \cap C^{1,2}(\mathbb{R}^n \times [0, t])$ is the unique bounded classical solution to the (parabolic) backward Kolmogorov equation (BKE):*

$$\frac{\partial v}{\partial s} = -\mathcal{L}^0 v, \quad 0 \le s < t, \tag{93}$$

$$v(\boldsymbol{h}, t) = f(\boldsymbol{h}). \tag{94}$$

If instead $f \in C^\infty_b(\mathbb{R}^n)$, then the above statement holds with the BKE defined on $0 \le s \le t$ (i.e., with the endpoint t included), in which case $v \in C^{1,2}(\mathbb{R}^n \times [0, t])$.

Proof The statements are straightforward applications of Lemma 3.1 and Remark 3.2 in (Chen and Jia, 2020) to our setting, since Assumption 4.1 ensures that the regular conditions there hold. See the proof in (Chen and Jia, 2020) for details. The idea is that upon imposing some regular conditions, one can borrow the results from (Lorenzi, 2011) to prove the existence and uniqueness of a bounded classical solution to the BKE (93)-(94). That $v(\boldsymbol{h}, s) = P^0_{s,t} f(\boldsymbol{h}) = \mathbb{E}[f(\boldsymbol{h}_t) | \boldsymbol{h}_s = \boldsymbol{h}]$ satisfies the BKE and the terminal condition (94) follows from an application of Itô's formula and standard arguments of stochastic analysis (Karatzas and Shreve, 1998). ■

Lemma B.4 *Let $\gamma \in C([0, T], \mathbb{R})$ and denote $P^\gamma_{s,t} := P_{s,t}$. Then:*
(a) For $f \in C^2_b(\mathbb{R}^n)$ and $0 \le s \le t \le T$,

$$P^\gamma_{s,t} f(\boldsymbol{h}) - P^0_{s,t} f(\boldsymbol{h}) = \int_s^t du \gamma(u) U^i_u P^0_{s,u} \frac{\partial}{\partial h^i} P^\gamma_{u,t} f(\boldsymbol{h}). \tag{95}$$

(b) For all $f \in C^\infty_b(\mathbb{R}^n)$, $\phi \in C([0, T], \mathbb{R})$, and $0 \le s \le t \le T$,

$$\lim_{\epsilon \to 0} \frac{1}{\epsilon} (P^{\epsilon\phi}_{0,t} f(\boldsymbol{h}) - P^0_{0,t} f(\boldsymbol{h}) = \int_0^t \phi(s) P^0_{0,s} \Delta\mathcal{L}_s P^0_{s,t} f(\boldsymbol{h}) ds, \tag{96}$$

where $\Delta\mathcal{L}_s\cdot = (\mathcal{L}_s - \mathcal{L}_s^0)\cdot = U^i\frac{\partial}{\partial h^i}\cdot$.

Proof The following proof is based on and adapted from that of Lemma 3.5 and Lemma 3.7 in (Chen and Jia, 2020).

(a) As noted earlier, due to Assumption 4.1, both $\boldsymbol{h}$ and $\overline{\boldsymbol{h}}$ satisfy the regular conditions (Definition 2.2) in (Chen and Jia, 2020). Therefore, we can apply Lemma B.3 and so their transition operators satisfy the BKE (93). This implies that $u(\boldsymbol{h}, s) := P_{s,t}^{\gamma} f(\boldsymbol{h}) - P_{s,t}^{0} f(\boldsymbol{h})$ is the bounded classical solution to:

$$\frac{\partial u(\boldsymbol{h}, s)}{\partial s} = -\mathcal{L}^0 u(\boldsymbol{h}, s) - \gamma(s)\Delta\mathcal{L}_s P_{s,t}^{\gamma} f(\boldsymbol{h}), \quad 0 \le s < t, \tag{97}$$

$$u(\boldsymbol{h}, t) = 0. \tag{98}$$

Applying Itô's formula gives

$$u(\overline{\boldsymbol{h}}_s, s) = \int_0^s \frac{\partial}{\partial r} u(\overline{\boldsymbol{h}}_r, r)dr + \int_0^s \mathcal{L}^0 u(\overline{\boldsymbol{h}}_r, r)dr + \int_0^s \nabla u(\overline{\boldsymbol{h}}_r, r)^T \boldsymbol{\sigma} d\boldsymbol{W}_r \tag{99}$$

$$= -\int_0^s \gamma(r)\Delta\mathcal{L}_r P_{r,t}^{\gamma} f(\overline{\boldsymbol{h}}_r)dr + \int_0^s \nabla u(\overline{\boldsymbol{h}}_r, r)^T \boldsymbol{\sigma} d\boldsymbol{W}_r. \tag{100}$$

For $R > 0$, let $B_R := \{\boldsymbol{h} \in \mathbb{R}^n : |\boldsymbol{h}| < R\}$. Define $\tau_R := \inf\{t \ge 0 : \overline{\boldsymbol{h}}_t \in \partial B_R\}$, the hitting time of the sphere ∂B_R by $\overline{\boldsymbol{h}}$, and the explosion time $\tau = \lim_{R\to\infty} \tau_R$. Note that it follows from Assumption 4.1 and an earlier note that $\tau > T$.

Then, if $|\boldsymbol{h}| < R$ and $s < r < t$,

$$u(\boldsymbol{h}, s) = \mathbb{E}[u(\overline{\boldsymbol{h}}_{r\wedge\tau_R}, r \wedge \tau_R)|\overline{\boldsymbol{h}}_s = \boldsymbol{h}] + \mathbb{E}\left[\int_s^r \gamma(u)\Delta\mathcal{L}_u P_{u,t}^{\gamma} f(\overline{\boldsymbol{h}}_u) 1_{u\le\tau_R} du \middle| \overline{\boldsymbol{h}}_s = \boldsymbol{h}\right]. \tag{101}$$

Since $\overline{\boldsymbol{h}}$ is nonexplosive up to time T and $u \in C_b(\mathbb{R}^n \times [0, T])$, we have:

$$\lim_{r\to t}\lim_{R\to\infty} \mathbb{E}[u(\overline{\boldsymbol{h}}_{r\wedge\tau_R}, r \wedge \tau_R)|\overline{\boldsymbol{h}}_s = \boldsymbol{h}] = \lim_{r\to t} \mathbb{E}[u(\overline{\boldsymbol{h}}_r, r)|\overline{\boldsymbol{h}}_s = \boldsymbol{h}] = \mathbb{E}[u(\overline{\boldsymbol{h}}_t, t)|\overline{\boldsymbol{h}}_s = \boldsymbol{h}] = 0. \tag{102}$$

Note that $(\Delta\mathcal{L}_s P_{s,t}^{\gamma} f(\boldsymbol{h}))_{s\in[0,t]} \in C_b(\mathbb{R}^n \times [0, t])$ for any $f \in C_b^2(\mathbb{R}^n)$ by Theorem 3.4 in (Chen and Jia, 2020) and recall that $\gamma \in C([0, t], \mathbb{R})$ for $t \in [0, T]$. Therefore, it follows from the above and an application of dominated convergence theorem that

$$u(\boldsymbol{h}, s) = \int_s^t \gamma(u)\mathbb{E}[\Delta\mathcal{L}_u P_{u,t}^{\gamma} f(\overline{\boldsymbol{h}}_u)|\overline{\boldsymbol{h}}_s = \boldsymbol{h}]du, \tag{103}$$

from which (95) follows.

(b) By (103) (with $\gamma := \epsilon\phi$ and $s := 0$ there), we have, for $\epsilon > 0$,

$$\frac{1}{\epsilon}(P_{0,t}^{\epsilon\phi} f(\boldsymbol{h}) - P_{0,t}^{0} f(\boldsymbol{h})) = \int_0^t \phi(s)\mathbb{E}[\Delta\mathcal{L}_s P_{s,t}^{\epsilon\phi} f(\overline{\boldsymbol{h}}_s)|\overline{\boldsymbol{h}}_0 = \boldsymbol{h}]ds. \tag{104}$$

We denote $g^\epsilon(\boldsymbol{h}, s) := \Delta\mathcal{L}_s P^{\epsilon\phi}_{s,t} f(\boldsymbol{h})$. Then, it follows from Assumption 4.1 that $\|g^\epsilon\|_\infty < \infty$, and

$$\sup_{s\in[0,t]} |\epsilon\phi(s) g^\epsilon(\cdot, s)| \leq \epsilon \|\phi\|_\infty \|g^\epsilon\|_\infty \to 0, \tag{105}$$

as $\epsilon \to 0$.

Now, let $t \in [0, T]$. For any $\alpha \in C_b^\infty(\mathbb{R}^n \times [0, t])$, consider the following Cauchy problem:

$$\frac{\partial u(\boldsymbol{h}, s)}{\partial s} = -\mathcal{L}^0 u(\boldsymbol{h}, s) - \alpha(\boldsymbol{h}, s), \quad 0 \leq s \leq t, \tag{106}$$

$$u(\boldsymbol{h}, t) = 0. \tag{107}$$

By Theorem 2.7 in (Lorenzi, 2011), the above equation has a unique bounded classical solution. Moreover, there exists a constant $C > 0$ such that

$$\|u\|_\infty \leq C \|\alpha\|_\infty. \tag{108}$$

Therefore, (97)-(98) together with (105)-(108) give:

$$\sup_{s\in[0,t]} |P^{\epsilon\phi}_{s,t} f - P_{s,t} f| \to 0, \tag{109}$$

as $\epsilon \to 0$. Thus, we have $g^\epsilon(\boldsymbol{h}, s) \to \Delta\mathcal{L}_s P^0_{s,t} f(\boldsymbol{h})$ as $\epsilon \to 0$.

Finally, using all these and the dominated convergence theorem, we have:

$$\lim_{\epsilon\to 0} \frac{1}{\epsilon}(P^{\epsilon\phi}_{0,t} f(\boldsymbol{h}) - P^0_{0,t} f(\boldsymbol{h})) = \int_0^t \phi(s) \mathbb{E}\left[\lim_{\epsilon\to 0} g^\epsilon(\overline{\boldsymbol{h}}_s, s) \middle| \overline{\boldsymbol{h}}_0 = \boldsymbol{h}\right] ds \tag{110}$$

$$= \int_0^t \phi(s) \mathbb{E}\left[\Delta\mathcal{L}_s P^0_{s,t}(\overline{\boldsymbol{h}}_s) \middle| \overline{\boldsymbol{h}}_0 = \boldsymbol{h}\right] ds. \tag{111}$$

■

We now prove Proposition 3.1.

Proof (Proof of Proposition 3.1 (a)) Recall that it follows from our assumptions that $f(\boldsymbol{h}_t) \in C_b^\infty(\mathbb{R}^n)$ for all $t \in [0, T]$.

We proceed by induction. For the base case of $n = 1$, we have, for $0 \leq t \leq T$,

$$\mathbb{E} f(\boldsymbol{h}_t) = \int P^0_{0,t} f(\boldsymbol{h}) \rho_{init}(\boldsymbol{h}) d\boldsymbol{h} = \mathbb{E} P^0_{0,t} f(\boldsymbol{h}_0). \tag{112}$$

Then, for $u_0 \in C([0,t],\mathbb{R})$ and any $\phi \in C_c^\infty(0,t)$:

$$\int_0^t \frac{\delta F_s}{\delta u}\Big|_{u=u_0} \phi(s)ds = DF_t[u_0]\cdot\phi \tag{113}$$

$$= \lim_{\epsilon\to 0}\frac{1}{\epsilon}(F_t[u_0+\epsilon\phi] - F_t[u_0]) \tag{114}$$

$$= \lim_{\epsilon\to 0}\frac{1}{\epsilon}(\mathbb{E}f(\boldsymbol{h}_t^{u_0+\epsilon\phi}) - \mathbb{E}f(\boldsymbol{h}_t^{u_0})) \tag{115}$$

$$= \lim_{\epsilon\to 0}\mathbb{E}\left[\frac{1}{\epsilon}(P_{0,t}^{u_0+\epsilon\phi}f(\boldsymbol{h}_0) - P_{0,t}^{u_0}f(\boldsymbol{h}_0))\right] \tag{116}$$

$$= \int_0^t ds\phi(s)\mathbb{E}P_{0,s}^{u_0}\Delta L_s P_{s,t}^{u_0}f(\boldsymbol{h}_0), \tag{117}$$

where the last equality follows from Lemma B.4 (b). Therefore, the result for the base case follows upon setting $u_0 = 0$.

Now assume that

$$D^{n-1}F_t[u_0]\cdot(\phi)^{n-1} = (n-1)!\int_{[0,t]^{n-1}} ds_1\cdots ds_{n-1}\phi(s_1)\cdots\phi(s_{n-1}) \times \mathbb{E}P_{0,s_{n-1}}^{u_0}\Delta\mathcal{L}_{s_{n-1}}P_{s_{n-1},s_{n-2}}^{u_0}\Delta\mathcal{L}_{s_{n-2}}\cdots P_{s_1,t}^{u_0}f(\boldsymbol{h}_0) \tag{118}$$

holds for any $\phi \in C_c^\infty(0,t)$, for $n > 1$. Then, for any $\phi \in C_c^\infty(0,t)$:

$$D^nF_t[0]\cdot(\phi)^n \tag{119}$$

$$= \lim_{\epsilon\to 0}\frac{(D^{n-1}F_t[\epsilon\phi] - D^{n-1}F_t[0])\cdot(\phi)^{n-1}}{\epsilon} \tag{120}$$

$$= (n-1)!\lim_{\epsilon\to 0}\frac{1}{\epsilon}\Bigg(\int_{[0,t]^{n-1}} ds_1\cdots ds_{n-1}\phi(s_1)\cdots\phi(s_{n-1})\big(\mathbb{E}P_{0,s_{n-1}}^{\epsilon\phi}\Delta\mathcal{L}_{s_{n-1}} P_{s_{n-1},s_{n-2}}^{\epsilon\phi}\Delta\mathcal{L}_{s_{n-2}}\cdots P_{s_1,t}^{\epsilon\phi}f(\boldsymbol{h}_0) - \mathbb{E}P_{0,s_{n-1}}^{0}\Delta\mathcal{L}_{s_{n-1}}P_{s_{n-1},s_{n-2}}^{0}\Delta\mathcal{L}_{s_{n-2}}\cdots P_{s_1,t}^{0}f(\boldsymbol{h}_0)\big)\Bigg). \tag{121}$$

Note that by use of Lemma B.2(b) and that the limit of products of two or more terms of the form $P_{s,s'}^{\epsilon\phi} - P_{s,s'}^0$ (with $s \le s'$), when multiplied by $(1/\epsilon)$, vanishes as $\epsilon \to 0$, we have

$$\frac{1}{\epsilon}\big(\mathbb{E}P_{0,s_{n-1}}^{\epsilon\phi}\Delta\mathcal{L}_{s_{n-1}}P_{s_{n-1},s_{n-2}}^{\epsilon\phi}\Delta\mathcal{L}_{s_{n-2}}\cdots P_{s_1,t}^{\epsilon\phi}f(\boldsymbol{h}_0) - \mathbb{E}P_{0,s_{n-1}}^{0}\Delta\mathcal{L}_{s_{n-1}}P_{s_{n-1},s_{n-2}}^{0}\Delta\mathcal{L}_{s_{n-2}}\cdots P_{s_1,t}^{0}f(\boldsymbol{h}_0)\big) \tag{122}$$

$$= \frac{1}{\epsilon}\mathbb{E}\sum_{k=1}^{n-1}\left(\prod_{l=1}^{k-1}\Delta\mathcal{L}_{s_l}P_{s_l,s_{l-1}}^0\right)(\Delta\mathcal{L}_{s_k}P_{s_k,s_{k-1}}^{\epsilon\phi} - \Delta\mathcal{L}_{s_k}P_{s_k,s_{k-1}}^0)\left(\prod_{p=k+1}^{n}\Delta\mathcal{L}_{s_p}P_{s_p,s_{p-1}}^0\right)f(\boldsymbol{h}_0) + e(\epsilon), \tag{123}$$

where $e(\epsilon) = o(\epsilon)$ as $\epsilon \to 0$, and we have set $s_0 := t$, $s_n := 0$ and $\Delta\mathcal{L}_{s_n} := 1$.

Moreover,

$$\lim_{\epsilon\to 0}\frac{1}{\epsilon}\big(\mathbb{E}P^{\epsilon\phi}_{0,s_{n-1}}\Delta\mathcal{L}_{s_{n-1}}P^{\epsilon\phi}_{s_{n-1},s_{n-2}}\Delta\mathcal{L}_{s_{n-2}}\cdots P^{\epsilon\phi}_{s_1,t}f(\boldsymbol{h}_0) - \mathbb{E}P^{0}_{0,s_{n-1}}\Delta\mathcal{L}_{s_{n-1}}P^{0}_{s_{n-1},s_{n-2}}\Delta\mathcal{L}_{s_{n-2}}\cdots P^{0}_{s_1,t}f(\boldsymbol{h}_0)\big) \tag{124}$$

$$= \mathbb{E}\sum_{k=1}^{n}\left(\prod_{l=1}^{k-1}\Delta\mathcal{L}_{s_l}P^0_{s_l,s_{l-1}}\right)\Delta\mathcal{L}_{s_k}\frac{1}{\epsilon}\lim_{\epsilon\to 0}(P^{\epsilon\phi}_{s_k,s_{k-1}}-P^0_{s_k,s_{k-1}})\left(\prod_{p=k+1}^{n}\Delta\mathcal{L}_{s_p}P^0_{s_p,s_{p-1}}\right)f(\boldsymbol{h}_0) \tag{125}$$

$$= \mathbb{E}\sum_{k=1}^{n}\left(\prod_{l=1}^{k-1}\Delta\mathcal{L}_{s_l}P^0_{s_l,s_{l-1}}\right)\Delta\mathcal{L}_{s_k}\int_0^t ds\phi(s)P^0_{s_k,s}\Delta\mathcal{L}_s P^0_{s,s_{k-1}}\left(\prod_{p=k+1}^{n}\Delta\mathcal{L}_{s_p}P^0_{s_p,s_{p-1}}\right)f(\boldsymbol{h}_0), \tag{126}$$

where we have applied Lemma B.4(b) in the last line.

Hence, using the above expression and symmetry of the mapping associated to derivative, we have

$$D^nF_t[0]\cdot(\phi)^n = n(n-1)!\int_{[0,t]^n} ds_1\cdots ds_{n-1}ds\phi(s)\phi(s_1)\cdots\phi(s_{n-1}) \times \mathbb{E}P^0_{0,s}\Delta\mathcal{L}_sP^0_{s,s_{n-1}}\Delta\mathcal{L}_{s_{n-1}}P^0_{s_{n-1},s_{n-2}}\Delta\mathcal{L}_{s_{n-2}}\cdots P^0_{s_1,t}f(\boldsymbol{h}_0). \tag{127}$$

Therefore, (a) holds for all $n\geq 2$. ■

Remark B.1 *Note that here it is crucial to have (b)-(c) in Assumption 4.1 to ensure that all derivatives of the form* (127) *are bounded and Lipschitz continuous. It may be possible to relax the assumptions on f at an increased cost of technicality but we choose not to pursue this direction. Had we been only interested in the linear response regime (i.e., $n=1$ case), then one can indeed relax the assumption on f substantially (see (Chen and Jia, 2020)).*

Proof (Proof of Proposition 3.1 (b)) We proceed by an induction argument. We use the notation $(f,g):=\int_{\mathbb{R}^n}f(\boldsymbol{h})g(\boldsymbol{h})d\boldsymbol{h}$ for $f,g\in C_b(\mathbb{R}^n)$, in the following. Let p_t denote the probability density of $\overline{\boldsymbol{h}}_t$, $t\in[0,T]$, and recall that $p_0=\rho_{init}$.

For the base case of $n=1$, first note that, using the properties of expectation,

$$\mathbb{E}[P^0_{0,s_1}\Delta\mathcal{L}_{s_1}P^0_{s_1,t}f(\boldsymbol{h}_0)] = \mathbb{E}[\mathbb{E}[\Delta\mathcal{L}_{s_1}P^0_{s_1,t}f(\boldsymbol{h}_{s_1})|\boldsymbol{h}_0=\boldsymbol{h}_0]] = \mathbb{E}[\Delta\mathcal{L}_{s_1}P^0_{s_1,t}f(\boldsymbol{h}_{s_1})], \tag{128}$$

for any $s_1 < t$. Therefore, by part (a) and applying integration by parts:

$$R_f^{(1)}(t,s_1) = \mathbb{E}[\Delta\mathcal{L}_{s_1} P^0_{s_1,t} f(\boldsymbol{h}_{s_1})] \tag{129}$$
$$= \int_{\mathbb{R}^n} \Delta\mathcal{L}_{s_1} P^0_{s_1,t} f(\boldsymbol{h}) p_{s_1}(\boldsymbol{h}) d\boldsymbol{h} \tag{130}$$
$$= \left(\Delta\mathcal{L}_{s_1} P^0_{s_1,t} f, p_{s_1}\right) \tag{131}$$
$$= \left(P^0_{s_1,t} f, \Delta\mathcal{A}_{s_1} p_{s_1}\right) \tag{132}$$
$$= \left(f, ((P^0_{s_1,t})^* \Delta\mathcal{A}_{s_1} p_{s_1})\right) \tag{133}$$
$$= \left(f, [((P^0_{s_1,t})^* \Delta\mathcal{A}_{s_1} p_{s_1})\rho_{init}^{-1}]\rho_{init}\right) \tag{134}$$
$$= \int_{\mathbb{R}^n} f(\boldsymbol{h}) v^{(1)}_{t,s_1}(\boldsymbol{h}_{s_1}) p_0(\boldsymbol{h}) d\boldsymbol{h} \tag{135}$$

where the second last line is well-defined since $\rho_{init}(\boldsymbol{h}) = p_0(\boldsymbol{h}) > 0$. Therefore,

$$R_f^{(1)}(t,s_1) = \mathbb{E} f(\boldsymbol{h}_0) v^{(1)}_{t,s_1}(\boldsymbol{h}_{s_1}). \tag{136}$$

Now, assume that (39)-(40) in the main paper holds for $n = k$. Note that

$$\mathbb{E}[P^0_{0,s_n} \Delta\mathcal{L}_{s_n} P^0_{s_n,s_{n-1}} \Delta\mathcal{L}_{s_{n-1}} \cdots P^0_{s_1,t} f(\boldsymbol{h}_0)] \tag{137}$$
$$= \mathbb{E}[\mathbb{E}[\Delta\mathcal{L}_{s_n} P^0_{s_n,s_{n-1}} \Delta\mathcal{L}_{s_{n-1}} \cdots P^0_{s_1,t} f(\boldsymbol{h}_{s_n}) | \boldsymbol{h}_0 = \boldsymbol{h}_0]] \tag{138}$$
$$= \mathbb{E}[\Delta\mathcal{L}_{s_n} P^0_{s_n,s_{n-1}} \Delta\mathcal{L}_{s_{n-1}} \cdots P^0_{s_1,t} f(\boldsymbol{h}_{s_n})] \tag{139}$$

for any n.

Then, by part (a),

$$R_f^{(k+1)}(t,s_1,\ldots,s_{k+1}) = \mathbb{E}[\Delta\mathcal{L}_{s_{k+1}} P^0_{s_{k+1},s_k} \Delta\mathcal{L}_{s_k} \cdots P^0_{s_1,t} f(\boldsymbol{h}_{s_{k+1}})] \tag{140}$$
$$= \int_{\mathbb{R}^n} \Delta\mathcal{L}_{s_{k+1}} P^0_{s_{k+1},s_k} \Delta\mathcal{L}_{s_k} \cdots P^0_{s_1,t} f(\boldsymbol{h}) p_{s_{k+1}}(\boldsymbol{h}) d\boldsymbol{h} \tag{141}$$
$$= \left(\Delta\mathcal{L}_{s_{k+1}} P^0_{s_{k+1},s_k} \Delta\mathcal{L}_{s_k} \cdots P^0_{s_1,t} f, p_{s_{k+1}}\right) \tag{142}$$
$$= \left(f, (P^0_{s_1,t})^* \Delta\mathcal{A}_{s_1} (P^0_{s_2,s_1})^* \cdots \Delta\mathcal{A}_{s_k} (P^0_{s_{k+1},s_k})^* \Delta\mathcal{A}_{s_{k+1}} p_{s_{k+1}}\right) \tag{143}$$
$$= \int_{\mathbb{R}^n} f(\boldsymbol{h}) v^{(k+1)}_{t,s_1,\ldots,s_{k+1}}(\boldsymbol{h}_{s_1},\ldots,\boldsymbol{h}_{s_{k+1}}) p_0(\boldsymbol{h}) d\boldsymbol{h}, \tag{144}$$

where $v^{(k+1)}_{t,s_1,\ldots,s_{k+1}}(\boldsymbol{h}_{s_1},\ldots,\boldsymbol{h}_{s_{k+1}})$ is given by (40) in the main paper. Note that we have applied integration by parts multiple times to get the second last line above. Therefore,

$$R_f^{(k+1)}(t,s_1,\ldots,s_{k+1}) = \mathbb{E} f(\boldsymbol{h}_0) v^{(k+1)}_{t,s_1,\ldots,s_{k+1}}(\boldsymbol{h}_{s_1},\ldots,\boldsymbol{h}_{s_{k+1}}). \tag{145}$$

The proof is done.

■

B.3 Proof of Corollary 3.1

Proof (Proof of Corollary 3.1) It follows from (144) that for $n \geq 1$, $f \in C_c^\infty$,

$$\int_{\mathbb{R}^n} f(\boldsymbol{h})(v^{(n)}_{t,s_1,\ldots,s_n}(\boldsymbol{h}_{s_1},\ldots,\boldsymbol{h}_{s_n}) - \tilde{v}^{(n)}_{t,s_1,\ldots,s_n}(\boldsymbol{h}_{s_1},\ldots,\boldsymbol{h}_{s_n}))\rho_{init}(\boldsymbol{h})d\boldsymbol{h} = 0. \tag{146}$$

Since f is arbitrary and $\rho_{init} > 0$, the result follows. ■

B.4 Proof of Theorem 3.1

Proof (Proof of Theorem 3.1) Recall that $\gamma := (\gamma(s) := |\boldsymbol{C}\boldsymbol{u}_s|)_{s\in[0,t]} \in C([0,t],\mathbb{R})$ for $t \in [0,T]$ by Lemma B.1.

Associated with $\mathbb{E}f(\boldsymbol{h}_t)$ is the mapping $F_t : C([0,t],\mathbb{R}) \to \mathbb{R}$, $\gamma \to \mathbb{E}f(\boldsymbol{h}_t)$, where $\boldsymbol{h}$ is the hidden state of the SRNN. Since by our assumptions the $D^nF_t[0]\cdot(\gamma,\ldots,\gamma)$ are well-defined for $n \in \mathbb{Z}_+$, the mapping F_t admits an absolutely convergent Taylor series at the point 0 for sufficiently small γ:

$$F_t[\gamma] = \sum_{n=1}^{\infty} \frac{1}{n!} D^nF_t[0]\cdot(\gamma)^n, \tag{147}$$

where $(\gamma)^n := (\gamma,\ldots,\gamma)$ (n times). Moreover, the derivatives are bounded and Lipschitz continuous, therefore integrable on compact sets. They can be identified with the response kernels $R_f^{(n)}(t,\cdot)$ given in Proposition 3.1 in the main paper. The resulting series is a Volterra series in the sense of Definition A.5. The uniqueness follows from the symmetry of the derivative mappings. ■

B.5 Proof of Theorem 3.2

We start with the proof and then give a few remarks.

Proof (Proof of Theorem 3.2) By assumption, an eigenfunction expansion of the operator $\mathcal{A}^0$ exists and is well-defined. We consider the eigenvalue-eigenfunction pairs $(-\lambda_m,\phi_m)_{m\in\mathbb{Z}_+}$, i.e., $\mathcal{A}^0\phi_m = \lambda_m\phi_m$ (for $m \in \mathbb{Z}_+$), where the $\lambda_m \in \mathbb{C}$ and the $\phi_m \in L^2(\rho_{init})$ are orthonormal eigenfunctions that span $L^2(\rho_{init})$ (see also Remark B.2).

In this case, we have $e^{\mathcal{A}^0t}\phi_m = e^{-\lambda_m t}\phi_m$, which implies that for any $f \in L^2(\rho_{init})$ we have $e^{\mathcal{A}^0t}f(\boldsymbol{h}) = \sum_n \alpha_n(\boldsymbol{h})e^{-\lambda_n t}$, where $\alpha_n(\boldsymbol{h}) = \langle\phi_n, f\rangle\phi_n(\boldsymbol{h})$.

We first derive formula (36). Applying the above representation in (32)-(34) in the main paper, we arrive at the following formula for the response kernels (recall that we are using Einstein's summation notation for repeated indices):

$$\mathcal{K}^{\boldsymbol{k}^{(n)}}(s_0,s_1,\ldots,s_n) = e^{-\lambda_{m_n}s_n}e^{-\lambda_{l_1}(s_0-s_1)}\cdots e^{-\lambda_{l_n}(s_{n-1}-s_n)}Z^{\boldsymbol{k}^{(n)}}_{l_1,\ldots,l_n,m_n}, \tag{148}$$

for $n \in \mathbb{Z}_+$, where

$$Z^{\boldsymbol{k}^{(n)}}_{l_1,\ldots,l_n,m_n} = (-1)^n \int d\boldsymbol{h} f(\boldsymbol{h})\alpha_{l_1}(\boldsymbol{h})\frac{\partial}{\partial h^{k_1}}\left[\cdots\alpha_{l_n}(\boldsymbol{h})\frac{\partial}{\partial h^{k_n}}\left[\alpha_{m_n}(\boldsymbol{h})\rho_{init}(\boldsymbol{h})\right]\right], \tag{149}$$

which can be written as average of a functional with respect to ρ_{init}. In the above, $\alpha_m = \langle \phi_m, \rho_{init} \rangle \phi_m$, and

$$\alpha_{l_1} = \left\langle \phi_{l_1}, \frac{\partial}{\partial h^{k_1}}(\alpha_{m_n}(\boldsymbol{h}) \right\rangle, \tag{150}$$

$$\alpha_{l_n} = \left\langle \phi_{l_1}, \frac{\partial}{\partial h^{k_1}}(\phi_{l_1}(\boldsymbol{h}) \cdots \frac{\partial}{\partial h^{k_{n-1}}}(\phi_{l_{n-1}}(\boldsymbol{h}) \frac{\partial}{\partial h^{k_n}}(\alpha_{m_n}(\boldsymbol{h})))) \right\rangle, \tag{151}$$

for $n = 2, 3, \ldots$.

Plugging in the above expressions into (35) in the main paper, we cast $\mathbb{E}f(\boldsymbol{h}_t)$ into a series of generalized convolution integrals with exponential weights:

$$\mathbb{E}f(\boldsymbol{h}_t) = \sum_{n=1}^{\infty} \epsilon^n Z^{\boldsymbol{k}^{(n)}}_{l_1,\ldots,l_n,m_n} e^{-\lambda_{l_1} t} \tag{152}$$

$$\times \int_0^t ds_1 \tilde{U}^{k_1}_{s_1} \cdots \int_0^{s_{n-1}} ds_n \tilde{U}^{k_n}_{s_n} e^{-\lambda_{m_n} s_n} e^{-\lambda_{l_1}(s_0 - s_1)} \cdots e^{-\lambda_{l_n}(s_{n-1} - s_n)} \tag{153}$$

$$= \sum_{n=1}^{\infty} \epsilon^n Z^{\boldsymbol{k}^{(n)}}_{l_1,\ldots,l_n,m_n} e^{-\lambda_{l_1} t} \int_0^t ds_1 \tilde{U}^{k_1}_{s_1} e^{-(\lambda_{l_2} - \lambda_{l_1}) s_1} \cdots \int_0^{s_{n-1}} ds_n \tilde{U}^{k_n}_{s_n} e^{-(\lambda_{m_n} - \lambda_{l_n}) s_n} \tag{154}$$

(with the λ_{m_n} equal zero in the case of stationary invariant distribution). Note that the expression above is obtained after performing interchanges between integrals and summations, which are justified by Fubini's theorem.

To isolate the unperturbed part of SRNN from $\tilde{U}$ completely, we expand the exponentials in Eq. (154) in power series to obtain:

$$\begin{aligned}
&\mathbb{E}f(\boldsymbol{h}_t) \\
&= \sum_{n=1}^{\infty} \epsilon^n Z^{\boldsymbol{k}^{(n)}}_{l_1,\ldots,l_n,m_n} (-1)^{p_0 + \cdots + p_n} \frac{(\lambda_{l_1})^{p_0}}{p_0!} \frac{(\lambda_{l_2} - \lambda_{l_1})^{p_1}}{p_1!} \cdots \frac{(\lambda_{m_n} - \lambda_{l_n})^{p_n}}{p_n!} \\
&\quad \times \left(t^{p_0} \int_0^t ds_1 (s_1)^{p_1} \tilde{U}^{k_1}_{s_1} \cdots \int_0^{s_{n-1}} ds_n (s_n)^{p_n} \tilde{U}^{k_n}_{s_n} \right) \qquad (155) \\
&=: \sum_{n=1}^{\infty} \epsilon^n Q^{\boldsymbol{k}^{(n)}}_{\boldsymbol{p}^{(n)}} \left(t^{p_0} \int_0^t ds_1 s_1^{p_1} \tilde{U}^{k_1}_{s_1} \cdots \int_0^{s_{n-1}} ds_n s_n^{p_n} \tilde{U}^{k_n}_{s_n} \right). \qquad (156)
\end{aligned}$$

This is the formula (36) in the main paper. The series representation in (43) in the main paper and its convergence then follows from the above result and Assumption 4.1. In particular, the constant coefficients $a_{p_0,\ldots,p_n,l_1,\ldots,l_n}$ in (43) are given by:

$$a_{p_0,\ldots,p_n,l_1,\ldots,l_n} = Q^{\boldsymbol{k}^{(n)}}_{\boldsymbol{p}^{(n)}} C^{k_1 l_1} \cdots C^{k_n l_n}, \tag{157}$$

where the $Q^{\boldsymbol{k}^{(n)}}_{\boldsymbol{p}^{(n)}}$ are defined in (156) (recall that we are using Einstein's summation notation for repeated indices). ■

Remark B.2 *If the operator $\mathcal{A}^0$ is symmetric, then such eigenfunction expansion exists and is unique, the $\lambda_m \in \mathbb{R}$, and, moreover, the real parts of λ_m are positive for exponentially stable SRNNs (Pavliotis, 2014). Working with the eigenfunction basis allows us to "linearize" the SRNN dynamics, thereby identifying the dominant directions and time scales of the unperturbed part of the SRNNs. If, in addition, the unperturbed SRNN is stationary and ergodic, one can, using Birkhoff's ergodic theorem, estimate the spatial average in the response functions with time averages.*

Remark B.3 *Eq.* (148) *disentangles the time-dependent component from the static component described by the $Z^{\boldsymbol{k}^{(n)}}_{l_1,\ldots,l_n,m_n}$. The time-dependent component is solely determined by the eigenvalues λ_i's, which give us the set of memory time scales on which the response kernels evolve. The static component is dependent on the eigenfunctions ϕ_n's, as well as on the initial distribution ρ_{init} of the hidden state, and the activation function f. The eigenvalues and eigenfunctions are determined by the choice of activation function (and their parameters) and the noise in the hidden states of SRNN. In practice, the multiple infinite series above can be approximated by a finite one by keeping only the dominant contributions, thereby giving us a feasible way to visualize and control the internal dynamics of SRNN by manipulating the response functions.*

B.6 Proof of Proposition 3.3

The computation involved in deriving Eq. (36) in the main paper essentially comes from the following combinatorial result.

Lemma B.5 *Consider the formal power series $a(x) = \sum_{i=1}^{\infty} a_i x^i$ and $b(x) = \sum_{i=1}^{\infty} b_i x^i$ in one symbol (indeterminate) with coefficients in a field. Their composition $a(b(x))$ is again a formal power series, given by:*

$$a(b(x)) = \sum_{n=1}^{\infty} c_n, \tag{158}$$

with

$$c_n = \sum_{\mathcal{C}_n} a_k b_{i_1} \cdots b_{i_k}, \tag{159}$$

where $\mathcal{C}_n = \{(i_1, \cdots, i_k) : 1 \le k \le n,\ i_1 + \cdots + i_k = n\}$. If $a_i := \alpha_i / i!$ and $b_i := \beta_i / i!$ in the above, then the c_n can be expressed in terms of the exponential Bell polynomials $B_{n,k}$ (Bell, 1927):

$$c_n = \frac{1}{n!} \sum_{k=1}^{n} \alpha_k B_{n,k}(\beta_1, \ldots, \beta_{n-k+1}), \tag{160}$$

where

$$\begin{aligned} &B_{n,k}(\beta_1, \ldots, \beta_{n-k+1}) \\ &= \sum_{c_1, c_2, \ldots, c_{n-k+1} \in \mathbb{Z}^+} \frac{n!}{c_1! c_2! \cdots c_{n-k+1}!} \left(\frac{\beta_1}{1!}\right)^{c_1} \left(\frac{\beta_2}{2!}\right)^{c_2} \cdots \left(\frac{\beta_{n-k+1}}{(n-k+1)!}\right)^{c_{n-k+1}} \end{aligned} \tag{161}$$

are homogeneous polynomials of degree k such that $c_1 + 2c_2 + 3c_3 + \cdots + (n-k+1)c_{n-k+1} = n$ and $c_1 + c_2 + c_3 + \cdots + c_{n-k+1} = k$.

Proof See, for instance, Theorem A in Section 3.4 in (Comtet, 2012) (see also Theorem 5.1.4 in (Stanley and Fomin, 1999)). ■

For a statement concerning the radius of convergence of composition of formal series, see Proposition 5.1 in Section 1.2.5 in (Cartan, 1995). See also Theorem C in Section 3.4 in (Comtet, 2012) for relation between the above result with the nth order derivative of $a(b(x))$ when a, b are smooth functions of a real variable.

Proof (Proof of Proposition 3.3) Since Volterra series are power series (in the input variable) of operators from L^∞ to L^∞ (see the remarks after Definition A.5), the idea is to apply Lemma B.5, from which (36) in the main paper follows.

Denote $\|\cdot\| := \|\cdot\|_\infty$ in the following. Note that for any finite positive r:

$$\|R_{fg}^{(r)}\| \leq \sum_{k=1}^{r} \sum_{\mathcal{C}_r} \|R_f^{(k)}\| \|R_g^{(i_1)}\| \cdots \|R_g^{(i_k)}\| < \infty \tag{162}$$

since the response functions in the finite series above are bounded by our assumptions. This justify any interchange of integrals and summations in the truncated Volterra series that involves during computation and therefore we can proceed and apply Lemma B.5.

For the case of (infinite) Volterra series, to ensure absolute convergence of the series we suppose[5] that $a_F(a_G(\|u\|)) < \infty$, where $a_F(x) = \sum_{n=1}^{\infty} \|R_f^{(n)}\| x^n$ and $a_G(x) = \sum_{n=1}^{\infty} \|R_g^{(n)}\| x^n$. Then, $G_t[u]$ is well-defined as a Volterra series and $\|G_t[u]\| \leq a_G(\|u\|)$. Also, $F_t[G_t[u]]$ is well-defined as a Volterra series by our assumption. In particular, it follows from (162) that $a_{F\circ G}(\|u\|) \leq a_F(a_G(\|u\|))$, where $a_{F\circ G}(x) = \sum_{n=1}^{\infty} \|R_{fg}^{(n)}\| x^n$. Thus, if ρ_F, ρ_G, $\rho_{F\circ G}$ denote the radius of convergence of F_t, G_t and $(F \circ G)_t$ respectively, then $\rho_{F\cdot G} \geq \min(\rho_G, a_G^{-1}(\rho_F))$. The last statement in the proposition follows from the above results.

■

B.7 Proof of Theorem 3.4

Proof (Proof of Theorem 3.4 – primal formulation) Let $p \in \mathbb{Z}_+$. First, note that by Lemma A.1 any increasing function on $[0,T]$ has finite one-variation on $[0,T]$. Since components of $\boldsymbol{\psi}^{(p)}$, being real-valued monomials on $[0,T]$, are increasing functions on $[0,T]$ and therefore they are of bounded variation on $[0,T]$. Since product of functions of bounded variation on $[0,T]$ is also of bounded variation on $[0,T]$, all entries of the matrix-valued paths $\boldsymbol{X}^{(p)} = \boldsymbol{u} \otimes \boldsymbol{\psi}^{(p)}$ are of bounded variation on $[0,T]$, and thus the signature of the $\boldsymbol{X}^{(p)}$ is well-defined. The result then essentially follows from Theorem 3.2 and Definition 3.2 in the main paper. ■

5. If the resulting formal series is not convergent, then one needs to treat it as an asymptotic expansion.

B.8 Proof of Proposition 3.2

First, we recall the definition of symmetric Fock space, which has played an important role in stochastic processes (Guichardet, 2006), quantum probability (Parthasarathy, 2012), quantum field theory (Goldfarb et al., 2013), and systems identification (Zyla and deFigueiredo, 1983).

In the following, all Hilbert spaces are real.

Definition B.1 *(Symmetric Fock space) Let $\mathcal{H}$ be a Hilbert space and $n \geq 2$ be any integer. For $h_1, \ldots, h_n \in \mathcal{H}$, we define the symmetric tensor product:*

$$h_1 \circ \cdots \circ h_n = \frac{1}{n!} \sum_{\sigma \in S_n} h_{\sigma(1)} \otimes \cdots \otimes h_{\sigma(n)}, \tag{163}$$

where S_n denotes the group of all permutations of the set $\{1, 2, \ldots, n\}$. By n-fold symmetric tensor product of $\mathcal{H}$, denoted $\mathcal{H}^{\circ n}$, we mean the closed subspace of $\mathcal{H}^{\otimes n}$ generated by the $h_1, \ldots, h_n$. It is equipped with the inner product:

$$\langle u_1 \circ \cdots \circ u_n, v_1 \circ \cdots \circ v_n \rangle_{\mathcal{H}^{\circ n}} = Per(\langle u_i, v_j \rangle)_{ij}, \tag{164}$$

where Per denotes the permanent (i.e., the determinant without the minus sign) of the matrix. Note that $\|u_1 \circ \cdots \circ u_n\|^2_{\mathcal{H}^{\circ n}} = n! \|u_1 \circ \cdots \circ u_n\|^2_{\mathcal{H}^{\otimes n}}$.

The symmetric Fock space over $\mathcal{H}$ is defined as $T_s((\mathcal{H})) = \bigoplus_{n=0}^{\infty} \mathcal{H}^{\circ n}$, with $\mathcal{H}^{\circ 0} := \mathbb{R}$, $\mathcal{H}^{\circ 1} := \mathcal{H}$. It is equipped with the inner product

$$\langle H, K \rangle_{T_s((\mathcal{H}))} = \sum_{n=0}^{\infty} n! \langle h_n, k_n \rangle_{\mathcal{H}^{\otimes n}}, \tag{165}$$

for elements $H := (h_m)_{m \in \mathbb{N}}$, $K := (k_m)_{m \in \mathbb{N}}$ in $T_s((\mathcal{H}))$, i.e., $h_m, k_m \in \mathcal{H}^{\circ m}$ for $m \in \mathbb{N}$, and $\|H\|^2_{T_s((\mathcal{H}))}, \|K\|^2_{T_s((\mathcal{H}))} < \infty$.

It can be shown that the symmetric Fock space can be obtained from a free Fock space by applying appropriate projection. One advantage of working with the symmetric Fock space instead of the free one is that the symmetric space enjoys a functorial property that the free space does not have (see (Parthasarathy, 2012)). Moreover, it satisfies the following property that we will need later.

Lemma B.6 *(Exponential property) Let $h \in \mathcal{H}$ and define the element (c.f. the so-called coherent state in (Parthasarathy, 2012)) $e(h) := \bigoplus_{n=0}^{\infty} \frac{1}{n!} h^{\otimes n}$ in $T_s((\mathcal{H})) \subset T_0((\mathcal{H}))$ (with $h^{\otimes 0} := 1$ and $0! := 1$). Then we have:*

$$\langle e(h_1), e(h_2) \rangle_{T_s((\mathcal{H}))} = \exp(\langle h_1, h_2 \rangle_{\mathcal{H}}). \tag{166}$$

Proof This is a straightforward computation. ■

Next we recall the definition of a kernel. Denote by $\mathbb{F}$ a field (e.g., $\mathbb{R}$ and $\mathbb{C}$).

Definition B.2 *(Kernel) Let χ be a nonempty set. Then a function $K : \chi \times \chi \to \mathbb{F}$ is called a kernel on χ if there exists a $\mathbb{F}$-Hilbert space $\mathcal{H}$ and a map $\phi : \chi \to \mathcal{H}$ such that for all $x, x' \in \chi$, we have $K(x, x') = \langle \phi(x'), \phi(x) \rangle_{\mathcal{H}}$. We call ϕ a feature map and $\mathcal{H}$ a feature space of K.*

By definition, kernels are positive definite.

Proof (Proof of Proposition 3.2) Let $L \in \mathbb{Z}_+$. Viewing $\mathcal{H} := \mathcal{P} \otimes \mathbb{R}^m$ as a set, it follows from Theorem 4.16 in (Steinwart and Christmann, 2008) that $K(\boldsymbol{v}, \boldsymbol{w}) = \langle S(\boldsymbol{v}), S(\boldsymbol{w}) \rangle_{T_s^{\otimes L}((\mathcal{H}))}$ is a kernel over $\mathcal{H}$ with the feature space $T_s^{\otimes L}((\mathcal{H}))$ since it is an inner product on the L-fold symmetric Fock space $T_s^{\otimes L}((\mathcal{H}))$, which is a $\mathbb{R}$-Hilbert space. The associated feature map is $\phi(\boldsymbol{v}) = S(\boldsymbol{v})$ for $\boldsymbol{v} \in \mathcal{H}$. The last statement in the proposition then follows from Theorem 4.21 in (Steinwart and Christmann, 2008). ■

B.9 Proof of Theorem 3.5

Proof (Proof of Theorem 3.5 – dual formulation)
(a) First, note that by Lemma A.1 any increasing function on $[0, T]$ has finite one-variation on $[0, T]$. Since components of $\boldsymbol{v}$, being real-valued polynomials on $[0, T]$, are linear combinations of monomials and linear combinations of functions of bounded variation on $[0, T]$ is also of bounded variation on $[0, T]$, the components of $\boldsymbol{v}$ are also of bounded variation on $[0, T]$. Since product of functions of bounded variation on $[0, T]$ is also of bounded variation on $[0, T]$, the $\boldsymbol{X}_n = \boldsymbol{v} \otimes \boldsymbol{u}_n$ are of bounded variation on $[0, T]$, and thus the signature of the $\boldsymbol{X}_n$ is well-defined.

By Proposition 3.2 in the main text, $\langle S(\boldsymbol{X}_n), S(\boldsymbol{X}) \rangle_{T_s((\mathcal{H}))}$ is a kernel on $\mathcal{H} := \mathcal{P} \otimes \mathbb{R}^m$ with the RKHS $\mathcal{R}_1$. Therefore, the result in (a) follows from a straightforward application of Theorem 1 in (Schölkopf et al., 2001).

(b) For $L \in \mathbb{Z}_+$ we compute:

$$\langle S(\tilde{\boldsymbol{X}}_n), S(\tilde{\boldsymbol{X}}) \rangle_{T_s^{\otimes L}((\mathcal{H}))} = \left\langle \bigotimes_{l=1}^{L} \exp(\Delta \boldsymbol{X}_n^{(l)}), \bigotimes_{l=1}^{L} \exp(\Delta \boldsymbol{X}^{(l)}) \right\rangle_{T_s^{\otimes L}((\mathcal{H}))} \tag{167}$$

$$= \prod_{l=1}^{L} \left\langle e(\Delta \boldsymbol{X}_n^{(l)}), e(\Delta \boldsymbol{X}^{(l)}) \right\rangle_{T_s((\mathcal{H}))} \tag{168}$$

$$= \prod_{l=1}^{L} \exp(\langle \Delta \boldsymbol{X}_n^{(l)}, \Delta \boldsymbol{X}^{(l)} \rangle_{\mathcal{H}}), \tag{169}$$

where we have used Lemma A.2 in the first line and Lemma B.6 in the last line above. The proof is done. ■

Appendix C. An Approximation Result for SRNNs

In this section we justify why Assumption 4.1, in particular the analyticity of the coefficients defining the SRNNs, is in some sense not too restrictive. We need the following extension of Carleman's theorem (Gaier, 1987) to multivariate functions taking values in multi-dimensional space. Note that Carleman's theorem itself can be viewed as an extension of Stone-Weierstrass theorem to non-compact intervals.

Theorem C.1 *If $\epsilon(\boldsymbol{x}) > 0$ and $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ are arbitrary continuous functions, then there is an entire function $\boldsymbol{g} : \mathbb{C}^n \to \mathbb{C}^m$ such that for all real $\boldsymbol{x} \in \mathbb{R}^n$ (or, real part of $\boldsymbol{z} \in \mathbb{C}^n$),*

$$|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{x})| < \epsilon(\boldsymbol{x}).$$

Proof This is a straightforward extension of the main result in (Scheinberg, 1976) to $\mathbb{R}^m$-valued functions. ■

The above theorem says that any $\mathbb{R}^m$-valued continuous function on $\mathbb{R}^n$ can be approximated arbitrarily closely by the restriction of an entire function of n complex variables. In the univariate one-dimensional case, an entire function g has an everywhere convergent Maclaurin series $g(z) = \sum_n a_n z^n$. The restriction of g to $\mathbb{R}$, $\tilde{g} = g|_{\mathbb{R}}$ therefore also admits an everywhere convergent series $\tilde{g}(x) = \sum_n a_n x^n$, thus analytic. Similar statements in the multi-dimensional and multivariate case hold.

Let $(\boldsymbol{h}'_t, \boldsymbol{y}'_t)$ be the hidden state and output of the SRNN defined by

$$d\boldsymbol{h}'_t = -\boldsymbol{\Gamma}\boldsymbol{h}'_t dt + \boldsymbol{a}'(\boldsymbol{W}\boldsymbol{h}'_t + \boldsymbol{b})dt + \boldsymbol{C}\boldsymbol{u}_t dt + \boldsymbol{\sigma} d\boldsymbol{W}_t, \tag{170}$$

$$\boldsymbol{y}'_t = \boldsymbol{f}'(\boldsymbol{h}'_t), \tag{171}$$

and satisfying Assumption 4.1 with $\boldsymbol{h}$, $\overline{\boldsymbol{h}}$ replaced by $\boldsymbol{h}'$, $\overline{\boldsymbol{h}}'$ respectively, the functions $\boldsymbol{a}$, $\boldsymbol{f}$ replaced by $\boldsymbol{a}'$, $\boldsymbol{f}'$ respectively, and with (c) there replaced by:

(c') The coefficients $\boldsymbol{a}' : \mathbb{R}^n \to \mathbb{R}$ and $\boldsymbol{f}' : \mathbb{R}^n \to \mathbb{R}^p$ are activation functions, i.e., bounded, non-constant, Lipschitz continuous functions.

We call this SRNN a *regular SRNN*.

On the other hand, let $(\boldsymbol{h}_t, \boldsymbol{y}_t)$ be the hidden state and output of the SRNN defined in (1)-(3) in the main paper and satisfying Assumption 4.1, with $\boldsymbol{h}_0 = \boldsymbol{h}'_0$. We call such SRNN an *analytic SRNN*.

Theorem C.2 *(Approximation of a regular SRNN by analytic SRNNs)*
Let $p > 0$. Given a regular SRNN, there exists an analytic SRNN such that for $\epsilon > 0$ arbitrarily small, $\sup_{t\in[0,T]} \mathbb{E}|\boldsymbol{y}_t - \boldsymbol{y}'_t|^p < \epsilon$.

Proof Let $\epsilon > 0$ be given and the regular SRNN be defined by (170)-(171).

By Theorem C.1, for any $\eta_i(\boldsymbol{h}) > 0$ $(i = 1, 2)$, there exists analytic functions $\boldsymbol{a}$ and $\boldsymbol{f}$ (restricted to $\mathbb{R}^n$) such that for all $\boldsymbol{h} \in \mathbb{R}^n$,

$$|\boldsymbol{a}(\boldsymbol{h}) - \boldsymbol{a}'(\boldsymbol{h})| < \eta_1(\boldsymbol{h}), \tag{172}$$
$$|\boldsymbol{f}(\boldsymbol{h}) - \boldsymbol{f}'(\boldsymbol{h})| < \eta_2(\boldsymbol{h}). \tag{173}$$

Therefore, by Lipschitz continuity of $\boldsymbol{a}'$ and $\boldsymbol{f}'$, we have:

$$|\boldsymbol{a}(\boldsymbol{h}) - \boldsymbol{a}'(\boldsymbol{h}')| \leq |\boldsymbol{a}(\boldsymbol{h}) - \boldsymbol{a}'(\boldsymbol{h})| + |\boldsymbol{a}'(\boldsymbol{h}) - \boldsymbol{a}'(\boldsymbol{h}')| < \eta_1(\boldsymbol{h}) + L_1|\boldsymbol{h} - \boldsymbol{h}'|, \tag{174}$$
$$|\boldsymbol{f}(\boldsymbol{h}) - \boldsymbol{f}'(\boldsymbol{h}')| \leq |\boldsymbol{f}(\boldsymbol{h}) - \boldsymbol{f}'(\boldsymbol{h})| + |\boldsymbol{f}'(\boldsymbol{h}) - \boldsymbol{f}'(\boldsymbol{h}')| < \eta_2(\boldsymbol{h}) + L_2|\boldsymbol{h} - \boldsymbol{h}'|, \tag{175}$$

where $L_1, L_2 > 0$ are the Lipschitz constants of $\boldsymbol{a}'$ and $\boldsymbol{f}'$ respectively.

From (1)-(3) and (170)-(171), we have, almost surely,

$$\boldsymbol{h}_t = \boldsymbol{h}_0 - \boldsymbol{\Gamma}\int_0^t \boldsymbol{h}_s ds + \int_0^t \boldsymbol{a}(\boldsymbol{W}\boldsymbol{h}_s + \boldsymbol{b})ds + \int_0^t \boldsymbol{C}\boldsymbol{u}_s ds + \boldsymbol{\sigma}\boldsymbol{W}_t, \tag{176}$$
$$\boldsymbol{h}'_t = \boldsymbol{h}_0 - \boldsymbol{\Gamma}\int_0^t \boldsymbol{h}'_s ds + \int_0^t \boldsymbol{a}'(\boldsymbol{W}\boldsymbol{h}'_s + \boldsymbol{b})ds + \int_0^t \boldsymbol{C}\boldsymbol{u}_s ds + \boldsymbol{\sigma}\boldsymbol{W}_t. \tag{177}$$

Let $p \geq 2$ first in the following. For $t \in [0, T]$,

$$|\boldsymbol{h}_t - \boldsymbol{h}'_t|^p \leq 2^{p-1}\left(\int_0^t |\boldsymbol{\Gamma}(\boldsymbol{h}_s - \boldsymbol{h}'_s)|^p ds + \int_0^t |\boldsymbol{a}(\boldsymbol{h}_s) - \boldsymbol{a}'(\boldsymbol{h}'_s)|^p ds\right). \tag{178}$$

Using boundedness of $\boldsymbol{\Gamma}$ and (174), we estimate

$$\mathbb{E}|\boldsymbol{h}_t - \boldsymbol{h}'_t|^p \leq C_1(p, T)\eta(T) + C_2(L_1, p)\int_0^t \mathbb{E}|\boldsymbol{h}_s - \boldsymbol{h}'_s|^p ds, \tag{179}$$

where $\eta(T) = \mathbb{E}\sup_{s\in[0,T]}(|\eta_1(\boldsymbol{h}_s)|^p)$, $C_1(p, T) > 0$ is a constant depending on p and T, and $C_2(L_1, p) > 0$ is a constant depending on L_1 and p. Therefore,

$$\sup_{t\in[0,T]} \mathbb{E}|\boldsymbol{h}_t - \boldsymbol{h}'_t|^p \leq C_1(p, T)\eta(T) + C_2(L_1, p)\int_0^T \sup_{u\in[0,s]} \mathbb{E}|\boldsymbol{h}_u - \boldsymbol{h}'_u|^p ds. \tag{180}$$

Applying Gronwall's lemma,

$$\sup_{t\in[0,T]} \mathbb{E}|\boldsymbol{h}_t - \boldsymbol{h}'_t|^p \leq C_1(p, T)\eta(T)e^{C_2(L_1,p)T}. \tag{181}$$

Thus, using (175) and (181) one can estimate:

$$\sup_{t\in[0,T]} \mathbb{E}|\boldsymbol{y}_t - \boldsymbol{y}'_t|^p \leq 2^{p-1}L_2^p \sup_{t\in[0,T]} \mathbb{E}|\boldsymbol{h}_t - \boldsymbol{h}'_t|^p + 2^{p-1}\mathbb{E}\sup_{s\in[0,T]} |\eta_2(\boldsymbol{h}_s)|^p \tag{182}$$
$$\leq C_3(L_2, p, T)\eta(T)e^{C_2(L_1,p)T} + 2^{p-1}\mathbb{E}\sup_{s\in[0,T]} |\eta_2(\boldsymbol{h}_s)|^p \tag{183}$$
$$\leq C_4(L_1, L_2, p, T)\beta(T), \tag{184}$$

where $\beta(T) = \mathbb{E}\sup_{s\in[0,T]}\left(\sum_{i=1}^{2}|\eta_i(\boldsymbol{h}_s)|^p\right)$ and the $C_i > 0$ are constants depending only on their arguments.

Since the $\eta_i > 0$ are arbitrary, we can choose them so that $\beta(T) < \epsilon/C_4$. This concludes the proof for the case of $p \geq 2$.

The case of $p \in (0, 2)$ then follows by an application of Hölder's inequality: taking $q > 2$ so that $p/q < 1$,

$$\sup_{t\in[0,T]}\mathbb{E}|\boldsymbol{y}_t - \boldsymbol{y}'_t|^p \leq \sup_{t\in[0,T]}(\mathbb{E}(|\boldsymbol{y}_t - \boldsymbol{y}'_t|^p)^{q/p})^{p/q} = \left(\sup_{t\in[0,T]}\mathbb{E}(|\boldsymbol{y}_t - \boldsymbol{y}'_t|^q)\right)^{p/q} \tag{185}$$

$$\leq C_5(L_1, L_2, p, q, T)\beta^{p/q}(T), \tag{186}$$

for some constant $C_5 > 0$, and choose $\beta(T)$ such that $\beta(T) < \left(\frac{\epsilon}{C_5}\right)^{q/p}$.

■

References

Ralph Abraham, Jerrold E Marsden, and Tudor Ratiu. *Manifolds, Tensor Analysis, and Applications*, volume 75. Springer Science & Business Media, 2012.

Girish Saran Agarwal. Fluctuation-dissipation theorems for systems in non-thermal equilibrium and applications. *Zeitschrift für Physik A Hadrons and nuclei*, 252(1):25–38, 1972.

Sina Alemohammad, Zichao Wang, Randall Balestriero, and Richard Baraniuk. The recurrent neural tangent kernel. *arXiv preprint arXiv:2006.10246*, 2020.

Marco Baiesi and Christian Maes. An update on the nonequilibrium linear response. *New Journal of Physics*, 15(1):013004, 2013.

Omri Barak. Recurrent neural networks as versatile tools of neuroscience research. *Current Opinion in Neurobiology*, 46:1–6, 2017.

Randall D Beer. On the dynamics of small continuous-time recurrent neural networks. *Adaptive Behavior*, 3(4):469–509, 1995.

Eric Temple Bell. Partition polynomials. *Annals of Mathematics*, pages 38–46, 1927.

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

Yoshua Bengio, Nicolas Boulanger-Lewandowski, and Razvan Pascanu. Advances in optimizing recurrent networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8624–8628. IEEE, 2013.

Martin Benning, Elena Celledoni, Matthias J Ehrhardt, Brynjulf Owren, and Carola-Bibiane Schönlieb. Deep learning as optimal control problems: Models and numerical methods. *Journal of Computational Dynamics*, 6(2):171, 2019.

Horatio Boedihardjo, Terry Lyons, Danyu Yang, et al. Uniform factorial decay estimates for controlled differential equations. *Electronic Communications in Probability*, 20, 2015.

Stephen Boyd and Leon Chua. Fading memory and the problem of approximating nonlinear operators with Volterra series. *IEEE Transactions on Circuits and Systems*, 32(11):1150–1161, 1985.

Stephen Boyd, Leon O Chua, and Charles A Desoer. Analytical foundations of Volterra series. *IMA Journal of Mathematical Control and Information*, 1(3):243–282, 1984.

Roger W Brockett. Volterra series and geometric control theory. *Automatica*, 12(2):167–176, 1976.

Francis Brown. Iterated integrals in quantum field theory. In *6th Summer School on Geometric and Topological Methods for Quantum Field Theory*, pages 188–240, 2013. doi: 10.1017/CBO9781139208642.006.

H. Cartan. *Differential Calculus*. Hermann, 1983. ISBN 9780901665140.

Henri Cartan. *Elementary Theory of Analytic Functions of One or Several Complex Variables*. Courier Corporation, 1995.

Bruno Cessac. Linear response in neuronal networks: from neurons dynamics to collective response. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(10):103105, 2019.

Bruno Cessac and Manuel Samuelides. From neuron to neural networks dynamics. *The European Physical Journal Special Topics*, 142(1):7–88, 2007.

Bo Chang, Minmin Chen, Eldad Haber, and Ed H Chi. AntisymmetricRNN: A dynamical system view on recurrent neural networks. *arXiv preprint arXiv:1902.09689*, 2019.

Kuo-Tsai Chen et al. Iterated path integrals. *Bulletin of the American Mathematical Society*, 83(5):831–879, 1977.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, pages 6571–6583, 2018.

Xian Chen and Chen Jia. Mathematical foundation of nonequilibrium fluctuation–dissipation theorems for inhomogeneous diffusion processes with unbounded coefficients. *Stochastic Processes and their Applications*, 130(1):171–202, 2020.

Zhengdao Chen, Jianyu Zhang, Martin Arjovsky, and Léon Bottou. Symplectic recurrent neural networks. *arXiv preprint arXiv:1909.13334*, 2019.

Raphaël Chetrite and Krzysztof Gawedzki. Fluctuation relations for diffusion processes. *Communications in Mathematical Physics*, 282(2):469–518, 2008.

Ilya Chevyrev and Andrey Kormilitzin. A primer on the signature method in machine learning. *arXiv:1603.03788*, 2016.

L. Comtet. *Advanced Combinatorics: The Art of Finite and Infinite Expansions*. Springer Netherlands, 2012.

Joni Dambre, David Verstraeten, Benjamin Schrauwen, and Serge Massar. Information processing capacity of dynamical systems. *Scientific Reports*, 2(1):1–7, 2012.

Edward De Brouwer, Jaak Simm, Adam Arany, and Yves Moreau. GRU-ODE-Bayes: Continuous modeling of sporadically-observed time series. In *Advances in Neural Information Processing Systems*, pages 7379–7390, 2019.

Amir Dembo and Jean-Dominique Deuschel. Markovian perturbation, response and fluctuation dissipation theorem. In *Annales de l'IHP Probabilités et statistiques*, volume 46, pages 822–852, 2010.

Weinan E, Chao Ma, and Lei Wu. Machine learning from a continuous viewpoint. *arXiv preprint arXiv:1912.12777*, 2019.

Jeffrey L Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990.

N Benjamin Erichson, Omri Azencot, Alejandro Queiruga, and Michael W Mahoney. Lipschitz recurrent neural networks. *arXiv preprint arXiv:2006.12070*, 2020.

Theodoros Evgeniou, Massimiliano Pontil, and Tomaso Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13(1):1, 2000.

Wei Fang, Michael B Giles, et al. Adaptive Euler–Maruyama method for SDEs with nonglobally Lipschitz drift. *Annals of Applied Probability*, 30(2):526–560, 2020.

Michel Fliess. Fonctionnelles causales non linéaires et indéterminées non commutatives. *Bulletin de la société mathématique de France*, 109:3–40, 1981.

Michel Fliess and Michiel Hazewinkel. *Algebraic and Geometric Methods in Nonlinear Control Theory*, volume 29. Springer Science & Business Media, 2012.

Matthias O Franz and Bernhard Schölkopf. A unifying view of Wiener and Volterra theory and polynomial kernel regression. *Neural Computation*, 18(12):3097–3118, 2006.

Peter K Friz and Martin Hairer. *A Course on Rough Paths. Universitext.* Springer, 2014.

Peter K Friz and Nicolas B Victoir. *Multidimensional Stochastic Processes as Rough Paths: Theory and Applications*, volume 120. Cambridge University Press, 2010.

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

Dieter Gaier. *Lectures on Complex Approximation*, volume 188. Springer, 1987.

Claudio Gallicchio and Simone Scardapane. Deep randomized neural networks. In *Recent Trends in Learning From Data*, pages 43–68. Springer, 2020.

Surya Ganguli, Dongsung Huh, and Haim Sompolinsky. Memory traces in dynamical systems. *Proceedings of the National Academy of Sciences*, 105(48):18970–18975, 2008.

S. Goldfarb, P.A. Martin, A.N. Jordan, F. Rothen, and S. Leach. *Many-Body Problems and Quantum Field Theory: An Introduction.* Theoretical and Mathematical Physics. Springer Berlin Heidelberg, 2013. ISBN 9783662084908.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning.* MIT Press, 2016.

Noella Grady. Functions of bounded variation. *Dostopno prek: https://www. whitman. edu/Documents/Academics/Mathematics/gra dy. pdf (Dostopano: 7.2. 2017)*, 2009.

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

Alain Guichardet. *Symmetric Hilbert spaces and related topics: Infinitely divisible positive definite functions. Continuous products and tensor products. Gaussian and Poissonian stochastic processes*, volume 261. Springer, 2006.

Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. *Inverse Problems*, 34(1):014004, 2017.

Martin Hairer and Andrew J Majda. A simple framework to justify linear response theory. *Nonlinearity*, 23(4):909, 2010.

P Hanggi et al. Stochastic processes. II. Response theory and fluctuation theorems. 1978.

Joshua Hanson and Maxim Raginsky. Universal simulation of stable dynamical systems by recurrent neural nets. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, pages 384–392. PMLR, 10–11 Jun 2020.

Jaeger Herbert. The echo state approach to analysing and training recurrent neural networks-with an erratum note. In *Bonn, Germany: German National Research Center for Information Technology GMD Technical Report*, volume 148, page 34. 2001.

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

Thomas Hofmann, Bernhard Schölkopf, and Alexander J Smola. Kernel methods in machine learning. *The Annals of Statistics*, pages 1171–1220, 2008.

J J Hopfield. Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences*, 81(10): 3088–3092, 1984. ISSN 0027-8424. doi: 10.1073/pnas.81.10.3088.

A. Isidori. *Nonlinear Control Systems*. Communications and Control Engineering. Springer London, 2013. ISBN 9781846286155.

Kam-Chuen Jim, C Lee Giles, and Bill G Horne. An analysis of noise in recurrent neural networks: convergence and generalization. *IEEE Transactions on Neural Networks*, 7(6): 1424–1438, 1996.

Ioannis Karatzas and Steven E Shreve. *Brownian Motion*. Springer, 1998.

Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. *arXiv preprint arXiv:2005.08926*, 2020.

Franz J Király and Harald Oberhauser. Kernels for sequentially ordered data. *Journal of Machine Learning Research*, 20, 2019.

Peter E Kloeden and Eckhard Platen. *Numerical Solution of Stochastic Differential Equations*, volume 23. Springer Science & Business Media, 2013.

Krzysztof Kowalski. Nonlinear dynamical systems and classical orthogonal polynomials. *Journal of Mathematical Physics*, 38(5):2483–2505, 1997.

D Kreimer. Chen's iterated integral represents the operator product expansion. *Advances in Theoretical and Mathematical Physics*, 3(3):627–670, 1999.

R Kubo. The fluctuation-dissipation theorem. *Reports on Progress in Physics*, 29(1):255, 1966.

Ryogo Kubo. Statistical-mechanical theory of irreversible processes. I. General theory and simple applications to magnetic and conduction problems. *Journal of the Physical Society of Japan*, 12(6):570–586, 1957.

Daniel Levin, Terry Lyons, and Hao Ni. Learning from the past, predicting the statistics for the future, learning an evolving system. *arXiv preprint arXiv:1309.0260*, 2013.

Shujian Liao, Terry Lyons, Weixin Yang, and Hao Ni. Learning stochastic differential equations using RNN with log signature features. *arXiv preprint arXiv:1908.08286*, 2019.

X. Liu, T. Xiao, S. Si, Q. Cao, S. Kumar, and C. J. Hsieh. How does noise help robustness? Explanation and exploration under the Neural SDE framework. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 279–287, 2020. doi: 10.1109/CVPR42600.2020.00036.

Luca Lorenzi. Optimal Hölder regularity for nonautonomous Kolmogorov equations. *Discrete & Continuous Dynamical Systems-S*, 4(1):169–191, 2011.

Mantas Lukoševičius and Herbert Jaeger. Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, 3(3):127–149, 2009.

T. Lyons and Z. Qian. *System Control and Rough Paths*. Oxford Mathematical Monographs. Clarendon Press, 2002. ISBN 9780198506485.

Terry Lyons. Rough paths, signatures and the modelling of functions on streams. *arXiv preprint arXiv:1405.4537*, 2014.

Terry J Lyons, Michael Caruana, and Thierry Lévy. *Differential Equations Driven by Rough Paths*. Springer, 2007.

Wolfgang Maass, Thomas Natschläger, and Henry Markram. Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural Computation*, 14(11):2531–2560, 2002.

A. Mauroy, I. Mezić, and Y. Susuki. *The Koopman Operator in Systems and Control: Concepts, Methodologies, and Applications*. Lecture Notes in Control and Information Sciences Series. Springer International Publishing AG, 2020. ISBN 9783030357139.

James L McClelland, David E Rumelhart, PDP Research Group, et al. Parallel distributed processing. *Explorations in the Microstructure of Cognition*, 2:216–271, 1986.

James Morrill, Patrick Kidger, Cristopher Salvi, James Foster, and Terry Lyons. Neural CDEs for long time series via the log-ODE method. *arXiv preprint arXiv:2009.08295*, 2020.

Radford M Neal. Priors for infinite networks. In *Bayesian Learning for Neural Networks*, pages 29–53. Springer, 1996.

Murphy Yuezhen Niu, Lior Horesh, and Isaac Chuang. Recurrent neural networks in the eye of differential equations. *arXiv preprint arXiv:1904.12933*, 2019.

Kalyanapuram R Parthasarathy. *An Introduction to Quantum Stochastic Calculus*, volume 85. Birkhäuser, 2012.

Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. How to construct deep recurrent neural networks. *arXiv preprint arXiv:1312.6026*, 2013a.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning*, pages 1310–1318, 2013b.

Grigorios A Pavliotis. *Stochastic Processes and Applications: Diffusion Processes, the Fokker-Planck and Langevin Equations*, volume 60. Springer, 2014.

Robert L Peterson. Formal theory of nonlinear response. *Reviews of Modern Physics*, 39 (1):69, 1967.

Fernando J Pineda. Generalization of back-propagation to recurrent neural networks. *Physical Review Letters*, 59(19):2229, 1987.

Yulia Rubanova, Ricky TQ Chen, and David K Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. In *Advances in Neural Information Processing Systems*, pages 5320–5330, 2019.

T Konstantin Rusch and Siddhartha Mishra. Coupled oscillatory recurrent neural network (coRNN): An accurate and (gradient) stable architecture for learning long time dependencies. *arXiv preprint arXiv:2010.00951*, 2020.

Anton Maximilian Schäfer and Hans Georg Zimmermann. Recurrent neural networks are universal approximators. In *International Conference on Artificial Neural Networks*, pages 632–640. Springer, 2006.

Stephen Scheinberg. Uniform approximation by entire functions. *Journal d'Analyse Mathématique*, 29(1):16–18, 1976.

Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. A generalized representer theorem. In *International Conference on Computational Learning Theory*, pages 416–426. Springer, 2001.

Bernhard Schölkopf, Alexander J Smola, Francis Bach, et al. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT press, 2002.

Alex Sherstinsky. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena*, 404:132306, 2020.

R.P. Stanley and S. Fomin. *Enumerative Combinatorics: Volume 2*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 1999. ISBN 9780521560696.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Science & Business Media, 2008.

Qi Sun, Yunzhe Tao, and Qiang Du. Stochastic training of residual networks: a differential equation viewpoint. *arXiv preprint arXiv:1812.00174*, 2018.

Ilya Sutskever. *Training recurrent neural networks*. University of Toronto, Ontario, Canada, 2013.

Gouhei Tanaka, Toshiyuki Yamane, Jean Benoit Héroux, Ryosho Nakane, Naoki Kanazawa, Seiji Takeda, Hidetoshi Numata, Daiju Nakano, and Akira Hirose. Recent advances in physical reservoir computing: A review. *Neural Networks*, 2019.

Peter Tino. Dynamical systems as temporal feature spaces. *Journal of Machine Learning Research*, 21(44):1–42, 2020.

Csaba Toth and Harald Oberhauser. Bayesian learning from sequential data using Gaussian processes with signature covariances. In *International Conference on Machine Learning*, pages 9548–9560. PMLR, 2020.

Jonathan Touboul. *Nonlinear and stochastic methods in neurosciences*. PhD thesis, 2008.

Vito Volterra. *Theory of Functionals and of Integral and Integro-Differential Equations*. Dover, 1959.

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

Huaguang Zhang, Zhanshan Wang, and Derong Liu. A comprehensive review of stability analysis of continuous-time recurrent neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 25(7):1229–1262, 2014.

LV Zyla and Rui JP deFigueiredo. Nonlinear system identification based on a Fock space framework. *SIAM Journal on Control and Optimization*, 21(6):931–939, 1983.