An Echo State Network Imparts a Curve Fitting

G Manjunath
[1]Department of Mathematics and Applied Mathematics, University of Pretoria, 0002, Pretoria, South Africa

Recurrent neural networks are successfully employed in processing information from temporal data. Approaches to training such networks are varied, and reservoir computing based attainments such as the echo state network provides great ease in training. Akin to many machine learning algorithms rendering an interpolation function or fitting a curve, we observe that a driven system such as a recurrent neural network renders a continuous curve fitting if and only if it satisfies the echo state property. The domain of the learnt curve is an abstract space of left-infinite sequence of inputs and the codomain is the space of readout values. When the input originates from discrete-time dynamical systems, we find theoretical conditions under which a topological conjugacy between the input and reservoir dynamics can exist, and present some numerical results relating the linearity in the reservoir to the forecasting abilities of the echo state networks.

I. Introduction

Many machine learning problems can be formulated as a problem of finding an unknown interpolating function or fitting a curve through a set of data points. In the case of neural networks, feedforward networks were conceived as an input-output relationship according to some mathematical function ever since the idea of a perceptron was formulated. Although they could handle static patterns, recurrent neural networks are primarily designed to handle sequential or temporal data. We show that the dynamics of a class of recurrent neural networks that employ the reservoir computing methodology or, more generally, the dynamics of a class of input driven dynamical systems renders a curve fitting for the temporal data.

Temporal data is data that is obtained sequentially as streams. Examples are ubiquitous, and with the advancement of technology, our ability to record rich and complex data has increased profoundly. One way of processing temporal data is inspired by how the brain maps a stimulus onto its enormously higher-dimensional space by affecting the states of a vast number of neurons. Such a projection of data onto a higher-dimensional space is commonly found to exaggerate certain features of the data to render better separability of inputs. Inspired by such an idea, reservoir computing (RC) employs an artificially generated reservoir analogous to a collection of neurons to map the data into a state in a higher-dimensional space. In RC, the reservoir is realized almost randomly. A simple regression technique is then used to obtain linear observables of the higher-dimensional states that aim to approximate the desired function of the temporal signal. Such ease of training has made prediction, adaptive filtering, noise reduction, recognition of temporal data in several domains fructuous. Although the idea of RC can be traced back to much earlier times, the methodology has gained the state-of-the-art status since the work of two research groups involving recurrent neural networks (RNNs) – Jaeger et al. in [1], [2] as the Echo State Network (ESN) and Maass et al. in [3] as the Liquid State Machine.

Today, RC has also been realized with dedicated hardware comprising photonic chips, memristors, spintronic nanodevices. (e.g., [4], [5], [6]). Applications of RC are varied and include robotics, prosthetic control, and epileptic seizure detection besides classical application of prediction, filtering, and information processing of temporal data (e.g., [7] and references therein) and also potentially modeling of nonautonomous dynamical systems [8]. There is a compelling need to operate these systems under high-efficiency and hence the quest for the mathematical underpinnings in designing good reservoirs.

The principle behind training only an observable in the RC methodology is that the evolving reservoir's states for practical purposes could be made independent of its initial state and thus making its evolution primarily governed by the input. Such evolution where the reservoir's states are forgotten can be framed as an asymptotic state-contraction property described as the echo state property (ESP) in [1],[2] and recurrent neural networks designed to achieve this asymptotic state contraction are called echo state networks (ESNs). Also these evolutions are easily computable since they have some attracting properties, specifically, they are always nonautonomous pullback attractors [9], [10], [11], and uniform attractors [12] when the input space is compact. The evolution of the network or reservoir states x_n and the read-out or output y_n when $n \in \mathbb{Z}$ can be formulated more generally as a discrete-state-space model

$$x_{n+1} = g(u_n, x_n) \quad (1)$$
$$y_{n+1} = f(x_{n+1}), \quad (2)$$

where u_n belonging to an input space U and state x_n belonging to a compact space X together with a mapping $g : U \times X \to X$ constitutes a driven system in (1), and $f : X \to Y$ is a continuous (function) read-out observable in (2), with Y being the output space. An example of a discrete-state space model that has gained traction for applications is

Manuscript received October 7, 2020. Corresponding author: G Manjunath (email: manjunath.gandhi@up.ac.za).

that of a ESN of the form

$$g(u,x) = (1-a)x + a\,\overline{\tanh}(W_{in}u + Wx) \quad (3)$$
$$f(x) = W_{out}x, \quad (4)$$

where $\overline{\tanh}(*)$ is (the nonlinear activation) $\tanh$ performed component-wise on $*$, a is a constant often called the leak rate, W_{in} and W are matrices of appropriate dimensions. The read-out is linear mapping described by the matrix W_{out}. Such discrete state space models are suitable for multi-tasking since usually W_{in} and W are not necessarily task specific and only W_{out} is purposed to the given task.

For instance, in the task of predicting future of an input, we deem the output space Y to be U itself, and the read-out f is realized as a best (least-square) linear fit of the mappings $(x_n, u_{n-1}) \mapsto u_n$ for the available data, now a popular training procedure [1]. Since u_{n-1} is involved in the computation of x_n, we could treat the readout y_n as a function of only the reservoir state as in (2) above.

With a strong mathematical slant, several useful results have been obtained recently. A universal approximation property [13] shows that given an input-output causal relation, there is a discrete-state space model, which is an ESN that can approximate the causal relation to a given precision. Recently, there is also a result [14], which says that given a driven system g that has the RNN form, a read-out can be designed so that any input and output causal relation can be approximated to a given precision. Also recently, the author in [15] has shown that a large class of driven systems encompassing recurrent neural networks exhibit a stable response to input if and only if they have the input-specific echo-state property (see Section II).

Despite the mathematical results, we remark that, in practice, the performance of ESNs in prediction tasks involves a great degree of variability. This can be demonstrated in the context where the inputs originate from a dynamical system. For instance, ESNs can be employed with great success in predicting an input from the Mackey-Glass oscillator, while for predicting an input from the Lorenz system, one has to resort to feedback connections into the network, and prediction is satisfactory for only about four lobe switchings [1, Supplementary Online Information]. Although authors have often brushed aside this inadequacy due to the sensitive dependence on initial conditions in the learnt system, we remark that after a few iterations the phase portrait of the learnt system does not resemble that of the map used during training. We illustrate this in Fig. 1 where we observe that soon after 100 time-steps, there is no resemblance of the ESN output to the input originating from the full logistic map. Thus the phase portrait, i.e., plot of y_{n+1} vs. y_n (not shown here) cannot be expected to resemble the graph of the full logistic map. It is likely that the ESN in its autonomous mode of operation has a spurious attractor and the dynamics has veered off from the dynamics of the logistic map to such an attractor. We remark

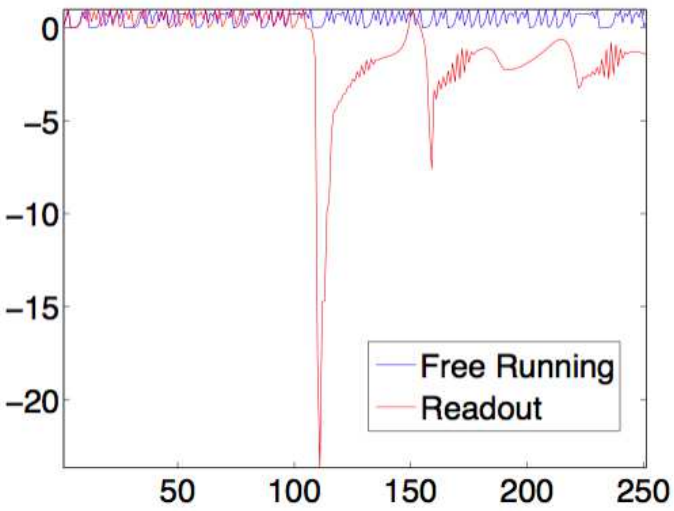

Fig. 1. A times series of the full logistic map (in blue) and its prediction from a linear read-out of a ESN (in red) plotted against time; ESN in (3) with 1000 neurons, $a = 0.2$ and a spectral radius 1.25 was used with a training length of 2000.

that in obtaining the plot in Fig. 1, we have used the same ESN network that predicts the Mackey-Glass attractor with great success. No significant improvements in forecasting was observed even when the number of neurons or the spectral radius was varied. In fact, the universal approximation property of ESNs [14] or the input-related stability property [15] of the ESNs is inadequate to explain this variability in performance in practice.

We also note that these mathematical results use the notion of product topology. The distance between sequences $\bar{x} = \{x_i\}$ and $\bar{y} = \{y_i\}$, $d(\bar{x}, \bar{y}) := \sum_{i=-\infty}^{\infty} d(x_i, y_i)/2^{|i|}$ generates the product topology. This metric is insensitive to differences in sequence tails and consequently there might be more profound reasons for the failure. With the notion of a *domain of learning* of a driven system, we point out in Section III that the reservoir dynamics could get intricately woven due to what is called an *indecomposable continua* that can make predictions prone to errors. One of this paper's goals is to initiate a line of research that could explain the variability in the ESN performance, keeping in mind the larger goal that it could lead to better RC architectures in the future.

Towards this end, we show that given a task, the model of the type (1)–(2) imparts a curve-fitting on a certain domain when g has the echo state property (see Section II). Denote the left-infinite product $(\cdots \times U \times U)$ by $\overleftarrow{U}$. The collection of all inputs drawn for a specific task forms a subspace $\overleftarrow{U}_{\text{task}} \subset \overleftarrow{U}$.

So when ESNs are trained for a task, the inputs for training the network are drawn from $\overleftarrow{U}_{\text{task}}$. Under the premise of the echo state property, we show that for every model of the type (1)–(2) there exists a state-independent continuous function Θ defined on the left-infinite inputs $\overleftarrow{U}_{\text{task}}$ and taking values in the output

space, i.e., $\Theta : \overleftarrow{u}^n \mapsto y_n$ where $\overleftarrow{u}^n := (\ldots, u_{n-2}, u_{n-1})$. Although Θ is defined for all possible left-infinite inputs, the domain of learning of Θ would be $\overleftarrow{U}_{\text{task}}$ and its graph fits a "curve" through the points $\{(\overleftarrow{u}^k, y_k), (\overleftarrow{u}^{k+1}, y_{k+1}), \ldots\}$. At the heart of defining the map Θ is the existence of another map called a *universal semi-conjugacy* $h : \overleftarrow{U} \to X$ that is independent of the task. The terminology of a universal semi-conjugacy is since h establishes a relationship between the shift map dynamics on the input space and and the evolution of the state values x_n as in (7) in Theorem 1. When the ESN is fed with inputs from a deterministic dynamical system, and if the network is such that the universal semi-conjugacy embeds $\overleftarrow{U}_{\text{task}}$ in X, then a reservoir (autonomous) map $R : h(\overleftarrow{u}^n) \mapsto h(\overleftarrow{u}^{n+1})$ exists (see Theorem 2) for all $\overleftarrow{u}$ in $\overleftarrow{U}_{\text{task}}$. We make use of this setting to explain the variability of the performance of ESNs while forecasting dynamical systems. In particular, we explain by discussing the role of the topological structure of $\overleftarrow{U}_{\text{task}}$ (see Section III).

Historically, data has been transformed into a different space for processing to gain some advantages. For example, according to Vapnik-Chervonenkis (VC) theory (e.g., [16]) mapping data into a higher dimension often provides much greater classification power. Mapping temporal data through a delay-coordinate mapping in Takens delay embedding uncovers hidden information [17]. In data-driven modeling employing Koopman's theory (e.g., [18], [19], [20]), approximating the dynamics of a dynamical system through a linear dynamical system in a higher dimensional space is often possible by mapping the data sample-wise into the higher dimensional space. In reservoir computing, the dimension of the state space of the reservoir is typically much larger than the input space while forecasting dynamical systems, and the map f in (2) is supposed to have much lesser functional complexity than the map that generates the data. This is not surprising since for simple tasks, a linear regression has been found to be efficient in evaluating the readout in ESNs. Here, inspired by the Koopman's theory, we explore the possibility of approximating the reservoir dynamics (or the map R if it exists) through a linear map (see Section IV). In the context of prediction of data from dynamical systems, we find from numerical simulations that whenever the resultant reservoir dynamics can be approximated by a linear mapping, the ESNs are known to perform very well for inputs from such dynamical systems. Also, again through numerical simulations, when the resultant reservoir dynamics cannot be approximated through a linear map, it is found that the prediction performance for inputs from such dynamical systems is known to be relatively poor.

The paper is organized as follows. In Section II we provide the mathematical insights behind the map Θ and the universal semi-conjugacy h – the terminology of conjugacy is defended in Section III, where we also give an idea of the potential complicated topological structure of $\overleftarrow{U}_{task}$. In Section IV, we present a linear approximation to the reservoir dynamics. In Section V we draw some conclusions from the results.

II. Curve fitting of Temporal Data

We recall some results from [15] that were made in the context where an input-specific ESP was considered, and then apply to it to the case where g has the ESP for all inputs. The aim is to show the existence of the state-independent continuous function Θ and the universal semi-conjugacy h when g has the ESP for all the inputs.

Preliminaries. A driven system would comprise an input space U, a state space X and a function $g : U \times X \to X$ where (U, d_U) is a metric space, and (X, d) is a compact metric space, and g is a continuous map. For brevity, we refer to g as a driven system with all entities quietly understood. If U is also compact, then we say g is a compactly driven system. A sequence $\bar{u} = \{u_n\}_{n\in\mathbb{Z}} \subset U$ which we call an input, induces a sequence of self-maps $\{g(u_n, \cdot)\}_{n\in\mathbb{Z}}$ defined on X and the dynamics on X is generated by the update equation $x_{n+1} = g(u_n, x_n)$.

Given a driven system g and an input $\bar{u}$, we call a sequence $\{x_n\}$ an entire-solution if it satisfies $x_{n+1} = g(u_n, x_n)$ for all $n \in \mathbb{Z}$. For notational purposes we denote a left-infinite sequence of U by $\overleftarrow{u}$ and bi-finite sequence by $\bar{u}$, and denote a solution obtained by $\bar{u}$ as $\{x_n(\bar{u})\}$.

If Y is a metric space then we denote the product space $\overleftarrow{Y} := \prod_{i=-\infty}^{-1} Z_i$ where $Z_i \equiv Y$ and equip this space with the product topology.

Suppose a driven system g has been fed input values $u_m, u_{m+1}, \ldots, u_{n-1}$ starting at time m. Then the map g transports a state-value $x \in X$ at time m to give a state-value $g_{u_{n-1}} \circ \cdots \circ g_{u_m}(x)$ at time n.

Formally, for every choice of $\bar{u} = (\ldots, u_{-1}, u_0, u_1, \ldots)$ we define for all pair of integers $m \leq n$, the function that 'transports' a system state at x at time m through the inputs $u_m, u_{m+1}, \ldots u_{n-1}$ to the state at time n given by a composition-operator called a process by several authors (e.g.,[9]); the composition operator is the map $\phi_{\bar{u}} : \mathbb{Z}^2_{\geq} \times X \to X$, where $\mathbb{Z}^2_{\geq} := \{(n, m) : n \geq m, n, m \in \mathbb{Z}\}$ and

$$\phi_{\bar{u}}(n, m, x) := \begin{cases} x & \text{if } n = m, \\ g_{u_{n-1}} \circ \cdots \circ g_{u_{m+1}} \circ g_{u_m}(x) & \text{if } m < n. \end{cases} \tag{5}$$

Since $g(u, x) : X \to X$, it easily follows that $\phi_{\bar{u}}(n, m-1, X) \subset \phi_{\bar{u}}(n, m, X)$ (see [10] [15]), and since ϕ is continuous in the variable x, these sets are all closed. Further, since X is compact the set

$$X_n(\bar{u}) := \bigcap_{m<n} \phi_{\bar{u}}(n, m, X) \tag{6}$$

is nonempty.

Note that only the left-infinite sequence $(\ldots, u_{n-2}, u_{n-1})$ influences $X_n(\bar{u})$, and $X_n(\bar{u})$ denotes the set of all attainable states at time n. Hence when $n = 0$, we denote $\mathcal{E}(\overleftarrow{u}) = X_n(\bar{u})$ and call it the encoding of the left-infinite sequence $\overleftarrow{u}$. We recall the following definitions from [10] and [15].

Definition II.1. A driven system g is said to be have the ESP w.r.t. an input $\overleftarrow{u}$ if $\mathcal{E}(\overleftarrow{u})$ is a singleton subset of X. A driven system g is said to have the ESP (w.r.t. the space U) if g has the echo state property w.r.t. every input $\overleftarrow{u}$ in $\overleftarrow{U}$.

Definition II.2. A driven system g is said to be contractible if there exists some input in $\overleftarrow{u} \in \overleftarrow{U}$ so that g has the ESP w.r.t. $\overleftarrow{u}$.

Definition II.3. A driven system is said to be an open driven system if there is no input value u that maps any uncountable set A in X to a single value, i.e., $g(u, A)$ is not a singleton when A is uncountable.

Definition II.1 that defines "g has the ESP w.r.t. a left-infinite input $\overleftarrow{u}$" is equivalent [15] to saying "g has the ESP w.r.t. an right-infinite input $\bar{u}$" if there exists exactly one entire solution for nonautonomous system $\{g(u_n, \cdot)\}_{n\in\mathbb{Z}}$, where the right-half of the bi-infinite sequence $\bar{u}$ can be chosen arbitrarily. Also, (see [12] [15]) one can easily show that when there is exactly one entire solution $\{x_n\}$ for an input $\bar{u}$ then $X_n(\bar{u})$ is the singleton subset containing x_n for all $n \in \mathbb{Z}$. Regarding Definition II.3, all recurrent neural networks are open driven systems [15].

We denote the collection of all nonempty closed subsets of X by H_X. On H_X we employ the Hausdorff metric defined by $d_H(A, B) := \max(dist(A, B), dist(B, A)) := \inf\{\epsilon : A \subset B_\epsilon(B) \ \& \ B \subset B_\epsilon(A)\}$, where $B_\epsilon(A) := \{x \in X : d(x, A) < \epsilon\}$ is the open ϵ-neighborhood of A. We could treat the encoding function $\mathcal{E}(\cdot)$ and describe its continuity by treating either as a multivalued function of $\overleftarrow{u}$ taking values in X or as a set-valued function taking values in H_X (see [15]). We borrow the following facts from [15, Section 3]: **(F1).** If $\mathcal{E}(\overleftarrow{v})$ is a singleton subset of X then $\mathcal{E}(\cdot)$ is continuous at $\overleftarrow{v}$. **(F2).** If g is open and contractible and $\mathcal{E}(\cdot)$ is continuous at $\overleftarrow{v}$ then g has the ESP w.r.t. $\overleftarrow{v}$. To state our theorem, we define the subspace of H_X that contains the singleton subsets of X by S_X, and define the mapping $i : (X, d) \to (\mathsf{S}_X, d_H)$ by $i(a) = \{a\}$. Clearly i is invertible.

Theorem 1. *Let g be a driven system and $f : X \to Y$ be a continuous observable. Suppose that g has the ESP then there exists a continuous map $\Theta : \overleftarrow{U} \to Y$ and $h : \overleftarrow{U} \to X$ so that*

$$\Theta(\overleftarrow{v}) = f(h(\overleftarrow{v})) = f(x_0(\overleftarrow{v}))$$

for all $\overleftarrow{v} \in \overleftarrow{U}$ and $x_0(\overleftarrow{v})$ is the component of the entire-solution obtained by any bi-infinite input $\bar{v}$ for which $(\ldots, v_{-2}, v_{-1}) = \overleftarrow{v}$. Conversely, if g is open and contractible and if $h = i^{-1} \circ \mathcal{E}$ is continuous then g has the ESP.

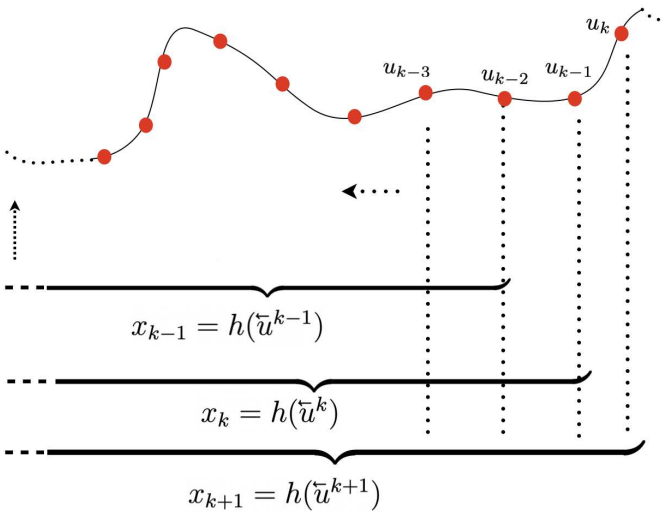

Fig. 2. Schematic to indicate the action of the universal semi-conjugacy h

Proof. Since g has the ESP it has the ESP w.r.t. all $\overleftarrow{u} \in \overleftarrow{U}$. Hence $\mathcal{E}(\overleftarrow{u})$ is a singleton subset of X for any $\overleftarrow{u}$ and by **(F1)**, $\mathcal{E}(\cdot)$ is continuous. Let $h = i^{-1} \circ \mathcal{E}$. Now, the mapping i is an isometry since

$$\begin{aligned} d_H(i(a), i(b)) &= \max\left(\sup_{a\in\{a\}} d(a, b), \sup_{b\in\{b\}} d(b, a)\right) \\ &= \max(d(a, b), d(b, a)) = d(a, b). \end{aligned}$$

We know if there is an isometry i between spaces, then both i and i^{-1} are continuous. Hence $h = i^{-1} \circ \mathcal{E}$ is continuous if and only if $\mathcal{E}$ is continuous. We already know $\mathcal{E}$ is continuous and thus h is also continuous. By definition of $\mathcal{E}$, $\mathcal{E}(\overleftarrow{u}) = \{x_0(\bar{u})\}$ where $\bar{u}$ is such that its left-infinite part $(\ldots, u_{-2}, u_{-1})$ is $\overleftarrow{u}$. Hence $f(h(\overleftarrow{u})) = f(x_0(\overleftarrow{u}))$. Since f and h are continuous, $\Theta = f \circ h$ is continuous.

Proof of the second (converse) statement: Since i is an isometry, $\mathcal{E}$ is continuous on $\overleftarrow{U}$. When g is contractible and open, by **(F2).** g has the ESP. ■

The map h in the above theorem would be called the universal semi-conjugacy of the driven system g. Given $\bar{u}$ one can obtain a left infinite sequence $\overleftarrow{u}^k = (\ldots, u_{k-2}, u_{k-1})$ for each $k \in \mathbb{Z}$. From the above theorem it follows easily that $h(\overleftarrow{u}^k) = x_k$ for each $k \in \mathbb{Z}$ (see Figure II).

III. Inverse Limit Space: The domain of Learning

We now consider the task of predicting an autonomous discrete dynamical system. Such systems are stand-alone models or also can be obtained by discretization of an ordinary differential equation. Predicting their dynamics are used as a benchmark for predicting complex temporal data. An autonomous discrete-time dynamical system comprises a tuple (U, T), where $T : U \to U$ is a surjective map and U is a

compact metric space. A sequence $\{u_n\} \subset U$ is an orbit of the dynamical system or of T if it satisfies $u_{n+1} = T(u_n)$ for $n \in \mathbb{Z}$. There could be another dynamical system (V, S) whose dynamics is equivalent to that of (U, T) if there exists a homeomorphism $\phi : U \to V$ so that $V \circ \phi = \phi \circ T$. In this case, we say that (V, S) is topologically conjugate to (U, T). Conjugacy means that there is a one-to-one correspondence between the two systems, and when $\phi : U \to V$ is a continuous map, we say that (V, S) is semi-conjugate to (U, T). Semi-conjugacy is a weaker notion since ϕ could be a many-to-one mapping in which case the system (V, S) is a coarse-grain description of (U, T).

Next consider the sequences in $\overleftarrow{U}$ given by $\widehat{U}_T := \{(\ldots, u_{-2}, u_{-1}) : T(u_n) = u_{n+1}\}$. Such a subspace $\widehat{U}_T$ is referred to as inverse limit space of (U, T) in the dynamical system literature (e.g. [21]), and is customarily written as a collection of right-infinite sequences but it is equivalent to $\widehat{U}_T$ that we have defined. The map T also induces a self-map $\widehat{T}$ on $\widehat{U}_T$ defined by $\widehat{T} : (\ldots, u_{-2}, u_{-1}) \mapsto (\ldots, u_{-2}, u_{-1}, T(u_{-1}))$. Hence $(\widehat{U}_T, \widehat{T})$ is a dynamical system in its own right and it is called the inverse limit system of (U, T). When g receives inputs as orbits of (U, T) then the inverse limit space $\widehat{U}_T$ is the domain of learning.

We next identify a subspace X_U of X that contains all possible solutions of the driven system g into which one can possibly topologically embed $\widehat{U}_T$. We recall that a space Z is embedded in X if there exists $Y \subset X$ and a function h so that $h : Z \to Y$ is a homeomorphism. For a driven system g, we define its *reachable set* to be the union of all the elements of all the solutions (entire-solutions), i.e.,

$$X_U := \Big\{ x \in X : \{x_n\} \text{ is a solution for some } \bar{u} \;\&\; x = x_k \text{ for some } k \in \mathbb{Z} \Big\}.$$

The above reachable set can be defined regardless of whether g satisfies the ESP or not. When g has the ESP, since every bi-infinite sequence $\bar{u} := \{u_n\}_{n\in\mathbb{Z}} \subset U$, and its left-infinite part $\overleftarrow{u}^n := (\ldots, u_{n-2}, u_{n-1})$ belongs to $\overleftarrow{U}$ regardless of $n \in \mathbb{Z}$, the map h defined in Theorem 1 satisfies $h(\overleftarrow{U}) = X_U$. Further when a new input value v appears at time n, i.e., $\overleftarrow{u}^n v := (\ldots, u_{n-2}, u_{n-1}, v)$ denotes symbolically the input up to time n, we define the mapping $\sigma_v : \overleftarrow{u}^n \mapsto \overleftarrow{u}^n v$. Thus, without reference to index n, the mapping σ_v defines a substitution of an input value v on to the right-end of every element $\overleftarrow{u} \in \overleftarrow{U}$. Theorem 2 shows that when g has the ESP, the map h establishes a topological semi-conjugacy between such v-substitution dynamics in the input space and the corresponding dynamics in the space X_U obtained through g.

Theorem 2. *Consider a compactly driven system g. When g has the ESP then the map $h : \overleftarrow{U} \to X_U$ defined by $h = i^{-1} \circ \mathcal{E}$ is such that it is continuous and surjective, and the following diagram commutes:*

$$\begin{array}{ccc} \overleftarrow{U} & \xrightarrow{\sigma_v} & \overleftarrow{U} \\ {\scriptstyle h}\downarrow & & \downarrow{\scriptstyle h} \\ X_U & \xrightarrow{g(v,\cdot)} & X_U. \end{array} \tag{7}$$

Further if (U, T) is a dynamical system and suppose h embeds the inverse limit space $\widehat{U}_T$ in X_U, then there exists a continuous map $R : h(\widehat{U}_T) \to h(\widehat{U}_T)$ defined by $R : h(\overleftarrow{u}^n) \mapsto h(\overleftarrow{u}^{n+1})$, and the system $(h(\widehat{U}_T), R)$ is topologically conjugate to $(\widehat{U}_T, \widehat{T})$.

Proof. Since $h(\overleftarrow{u}) = i^{-1}(\mathcal{E}(\overleftarrow{u}))$. By definition of $\mathcal{E}$, we find (the required commutativity in the diagram in (7)) through the deduction:

$$\begin{aligned} g_v \circ i^{-1} \circ \mathcal{E}(\overleftarrow{u}) &= i^{-1} \circ \mathcal{E}(\overleftarrow{u}v), \\ &= i^{-1} \circ \mathcal{E}(\sigma_v(\overleftarrow{u})). \end{aligned}$$

It remains to be shown that h is surjective and continuous. By definition of X_U, it follows that $h(\overleftarrow{U}) = X_U$ and hence h is surjective. Also h is continuous by Theorem 1.

When we restrict the input to $\widehat{U}_T$, then $\sigma_v(\overleftarrow{u}) = (\ldots, u_{-2}, u_{-1}, T(u_{-1}))$, and hence $\sigma_v(\overleftarrow{u}) = \widehat{T}(\overleftarrow{u})$. When h embeds $\widehat{U}_T$ in X_U, then from (7), we have $h_{|\widehat{U}_T} \circ \widehat{T} = g_{T(u_{-1})} \circ h_{|\widehat{U}_T}$, where $g_{T(u_{-1})}(\cdot) = g(T(u_{-1}), \cdot)$. Since $h_{|\widehat{U}_T}$ is a homeomorphism, $g_{T(u_{-1})} = h_{|\widehat{U}_T} \circ \widehat{T} \circ \lambda$ where $\lambda : h(\widehat{U}_T) \to \widehat{U}_T$ is the inverse of $h_{|\widehat{U}_T}$, i.e., $h_{|\widehat{U}_T} \circ \lambda$ is the identity map. Since $h_{|\widehat{U}_T} \circ \widehat{T} \circ \lambda$ is a well-defined mapping independent of u_{-1}, the map $R := g_{T(u_{-1})}$ is independent of u_{-1}. Also by (7), $R : h(\overleftarrow{u}^n) \mapsto h(\overleftarrow{u}^{n+1})$. Thus $h_{|\widehat{U}_T} \circ \widehat{T} = R \circ h_{|\widehat{U}_T}$, and hence $(h(\widehat{U}_T), R)$ is topologically conjugate to $(\widehat{U}_T, \widehat{T})$. ■

Remarkably, the result in Theorem 2 theorem can be strengthened (see [12]) to "(7) commutes if and only if g has the ESP". Suppose U is a manifold of dimension m, then $\overleftarrow{U}$ is infinite-dimensional and the universal conjugacy h would not embed $\overleftarrow{U}$ in X_U since X is a subspace of $\mathbb{R}^N$. However, if the component functions of h are sufficiently independent, it could embed a finite dimensional subspace of $\overleftarrow{U}$ in X_U, and in particular if h embeds an inverse limit space of $\widehat{U}_T$ of a dynamical system, then the reservoir dynamics determined by $(h(\widehat{U}_T), R)$ has all the information to reconstruct the dynamics of $(\widehat{U}_T, \hat{T})$. Although it is not possible to verify if h is an embedding, we note that by Whitney's embedding theorem, a generic map from a sufficiently smooth manifold m to a manifold of dimension $2m + 1$ is a differentiable embedding (e.g., [22]). Now, is $\widehat{U}_T$ also a manifold of dimension of m if U is a manifold of dimension m? The answer is affirmative when T is a homeomorphism. A more general question is: would $(\widehat{U}_T, \widehat{T})$ be topologically conjugate to (U, T) when T

is a homeomorphism? The answer to this question is also affirmative and can be explained as follows. By considering a map $\Lambda : \widehat{U}_T \to U$ by $\Lambda : (\ldots, u_{-2}, u_{-1}) \mapsto u_{-1}$, we observe that $T \circ \Lambda = \Lambda \circ \widehat{T}$ holds. Hence the dynamical system (U, T) is semi-conjugate to $(\widehat{U}, \widehat{T})$, and whenever, T is a homeomorphism, it follows that $(\widehat{U}, \widehat{T})$ is conjugate to (U, T), since the mapping $u_{-1} \mapsto (\ldots, T^{-2}u_{-1}, T^{-1}u_{-1}, u_{-1})$ is Λ^{-1} is a homeomorphism. Thus when T is a homeomorphism, $(\widehat{U}_T, \widehat{T})$ is topologically conjugate to (U, T). So from a practical viewpoint when h is an embedding, the reservoir dynamics does not introduce any new complexity than that is present in the input system (U, T) since $(h(\widehat{U}_T), R)$ is conjugate to $(\widehat{U}_T, \widehat{T})$. In the ESN literature, the forecasting results of invertible maps like the Henon map or that of the the time-one maps of the flows of ordinary differential equations are very encouraging. We now give a plausible explanation for this success. For a homeomorphism $T : U \to U$, every point u defines its left-infinite orbit and a unique point in $\widehat{U}_T$. An injective map taking values in a compact space is always an embedding. So if h restricted to $\widehat{U}_T$ is injective then h embeds $\widehat{U}_T$ in X_U since X_U is compact. Therefore, for $h(\overline{u}) \neq h(\overline{v})$ to hold when $\overline{u} \neq \overline{v}$, all h needs to do is to distinguish u_{-1} and v_{-1}, or more formally if there exists a map $\gamma : h(\widehat{U}_T) \to U$ so that $\gamma \circ h$ is injective, then h restricted to $\widehat{U}_T$ is injective.

We now argue that when $\widehat{U}_T$ has a complicated topological structure and T is chaotic and noninvertible, the universal semi-conjugacy h of an ESN may not be able to embed $\widehat{U}_T$ in X_U. To illustrate our point, we first note that when T is not invertible, the dynamics of (U, T) is only topologically semi-conjugate to $(\widehat{U}, \widehat{T})$, and hence the dynamics in $(\widehat{U}, \widehat{T})$ could have additional complexity. Next, we consider the topological structure of inverse limit spaces. The inverse limit spaces tell a lot about dynamical systems, and an excellent but less-technical treatment can be found in [23]. The topological description of the inverse limit system is complicated for a chaotic system. Here is a discussion on how complicated these spaces can be. The closed interval $X = [0, 1]$ which we know is compact and connected, can be written as a union of two proper closed and connected subsets $X = [0, 1/2] \cup [1/2, 1]$. Now, there are spaces that cannot be written as the union of two proper compact connected subsets. Such spaces are referred to as an indecomposable continua.

Formally, a continua is a nonempty compact connected metric space. A nonempty connected closed subset of a continuum is a continuum as well, and is called a subcontinuum. A continua is called an indecomposable if it cannot be written as the union of two proper subcontinua. The indecomposable continua have an interesting property in that it can be shown to be an uncountable collection of mutually disjoint connected sets, each of which is dense in the continuum. Kennedy et al. in [23] describe that the continuum consists of a collection of "highways", each close to any other but always separate. These highways are mathematically defined through the terminology of composants [23], but without going into finer mathematical details, we refer to Fig. 3 where a few such highways are

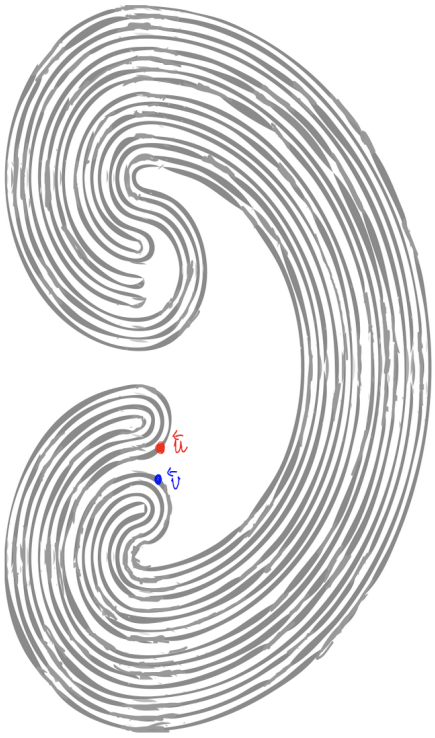

Fig. 3. Conceptual diagram of inverse-limit spaces: a sketch of some "highways" that are connected sets without (topological) interiors that make up an indecomposable continua. By breaking such a structure into two, at least one of them loses its connectedness property

sketched. Whenever we split the set described into Fig. 3, at least one of those pieces would not be connected. We refer the interested reader to [24] for computer generated pictures of continua that correspond to inverse limit spaces of different spaces.

There are also different subtypes of indecomposable continua, and they correspond to chaotic dynamical systems of a specific subtype as well [23]. In particular, the inverse limit spaces of interval maps always have a great deal of complexity in their inverse limit spaces [25]. We believe the ESNs determined in practice do not always produce a universal conjugacy h that embeds the inverse-limit spaces of chaotic interval maps (that are always noninvertible) in X_U since h may not distinguish two left-infinite sequences if only their tails are different. As numerical evidence, we have illustrated the inability of the ESN to forecast the logistic map $T_1(x) = 4x(1 - x)$ on $[0, 1]$ in Section I. During the course of the review of this paper, an anonymous referee suggested investigation into the role of the possible stochastic dynamics with regard to the poor performance of the ESNs for the logistic map. We address that next.

Complexity of deterministic systems can also be described macroscopically by describing how the density of an ensemble of initial conditions evolves upon iterating the map. An operator called the Perron Frobenius operator determines the

evolution of an initial density [26]. A core concept in studying such stochastic dynamics, is the notion of an invariant density p that is a fixed point of the Perron Frobenius operator that also can determine the visitation frequency of typical orbits in a subspace. We refer the reader to [26] for more details. The invariant density of the full logistic map $T_1(x) = 4x(1-x)$ is found to be $\frac{1}{\pi\sqrt{x(1-x)}}$, and one may suspect that since the invariant density explodes at the end points of $[0,1]$, it could lead to an ESN performing poorly while forecasting. To illustrate that the invariant density of the logistic map is not influencing the ESN performance, we consider two other symmetric interval maps $T_2 := \sqrt{1-|2x-1|}$ and T_3 (defined in (8)) defined on $U = [0,1]$ for our study on determining the performance of an ESN for forecasting data from interval maps.

$$T_3(x) := \begin{cases} \sqrt{2}x & \text{if } 0 \le x \le \frac{1}{\sqrt{8}}, \\ 1 - \sqrt{\frac{1}{2} - 2x^2} & \text{if } \frac{1}{\sqrt{8}} < x \le \frac{1}{2}, \\ 1 - \sqrt{\frac{1}{2} - 2(1-x)^2} & \text{if } \frac{1}{2} < x \le 1 - \frac{1}{\sqrt{8}}, \\ \sqrt{2}(1-x) & \text{if } 1 - \frac{1}{\sqrt{8}} < x \le 1. \end{cases} \tag{8}$$

The invariant densities of maps T_2 and T_3 are $p_2(x) = 2x$ and $p_3(x) = 2 - |2 - 4x|$ respectively (proof in [27]) do not have such an explosiion at the end points of $[0,1]$, and all of these maps are examples of chaotic maps as well. The graphs and invariant densities of T_1, T_2 and T_3 are plotted in Fig 4(a) and (b). We compare the performance of the forecasting for orbits of these maps with an identical RNN, with 200 neurons in the reservoir, and training with 2000 samples of the data.

In (c) and (d) of Fig. 4 we plot the phase portrait of the readout, i.e., $y(n+1)$ vs. $y(n)$ of the reconstructed orbit of the logistic map T_1 using 100 sample values of $y(n)$ after 100 time-steps (in (c)) and after 200 time-steps (in (d)). We note that barring for a small epoch where the concerned graph is indicated in a box the phase portrait did not resemble the graph of the logistic map. The plots in Fig. (e) and (f) correspond to similar phase portraits of T_2 while that in Fig. (f) and (g) correspond to T_3. As evident, the ESN performance is bad and T_3 performs the worst. The performance can be made only less worse by choosing different network sizes for the three maps, and we attribute this for h not being able to embed the inverse limit spaces of these maps in X_U, and hence the ESN in its autonomous mode operation develops spurious attractors.

Intuitively, regardless of whether T is invertible or not, by increasing the size of the recurrent network, and also by choosing a reservoir that is not very strongly connected one may expect that there are enough component functions of h that are independent so that it may embed $\widehat{U}_T$. However, there are no theoretical condition to guarantee an ESN has an h that embeds the inverse limit space of a dynamical system in X_U, and it may require some tuning or adjustments. One possible way to alter the domain of learning itself. With feedback connections, for instance in a RNN of the form $x(n+1) = \sigma(W_{in}\ u(n) + Wx(n) + \ W_{fb}\, W_{out}\ x(n))$, where W_{fb} is the feedback matrix and W_{out} is the output matrix (e.g., [1], [2]), the domain of learning on which h is defined would stand altered due to the feedback connections. Studying any advantage of this change in the domain of learning calls for further research.

IV. On the Linearity of the Reservoir Dynamics

We now describe the possibility of approximating the reservoir dynamics through a linear map. Recently, application of Koopman's theory (e.g., [18], [19], [20]) has shown that by mapping data arising from a dynamical system into a higher dimensional space one gains greater forecasting capacities since some aspect of the dynamics can be approximated through linear maps. Given a dynamical system (U,T) and a vector space V of observables f whose domain is U, the operator $\mathcal{K} : V \to V$ so that $\mathcal{K}f = f \circ T$ holds is called the Koopman operator. Often V is a Hilbert space containing complex-valued functions in forecasting applications. In this case a complex number $\lambda \in \mathbb{C}$ and $\phi \in V$ is called an eigenvalue and eigenfunction if $\mathcal{K}\phi = \lambda\phi$. In this case, the following diagram commutes [18], [19], [20]:

$$\begin{array}{ccc} U & \xrightarrow{\ T\ } & U \\ {\scriptstyle\phi}\downarrow & & \downarrow{\scriptstyle\phi} \\ \phi(U) & \xrightarrow{x \mapsto \lambda x} & \phi(U). \end{array} \tag{9}$$

When, the eigenfunction $\phi : U \to V$ is continuous then the dynamical system $(\phi(U), \lambda)$ is topologically semiconjugate to (U,T). So every eigenvalue captured gives some additional coarse-grain description of (U,T). Since finding non-constant eigenfunctions even if T is known is difficult, they are determined through data. Often a collection of observables $\mathbf{f} = \{f_1, \ldots, f_N\}$ are employed to approximate the Koopman operator through a linear transformation in $\mathbb{C}^N \to \mathbb{C}^N$. If the collection of observables are chosen such that its linear span is invariant under the Koopman operator, then one can find a linear approximation to the Koopman operator through the transformation determined by a matrix A that is guaranteed to capture a few eigenvalues and a few eigenvectors of $\mathcal{K}$. The interesting point in the data-driven theory is that the matrix A is constructed only through the data from the observables, i.e., $\{f_i(T^k(u))\}_{0 \le k \le m, 1 \le i \le N}$. Also, to capture a larger set of eigenvalues one has to use a larger set of observables. We now consider the component functions of h that a driven system g with the ESP determines as real-valued observables of the inverse limit system $(\widehat{U}_T, \widehat{T})$, and then find the matrix A.

With this matrix A, we approximate the dynamical relationship between $h(\overleftarrow{u}^n)$ and $h(\overleftarrow{u}^{n+1})$ regardless of whether there exists a well-defined reservoir mapping R just as in the Koopman operator based data-driven theory. The difference in this case and the Koopman's theory is that we do not have an analytical

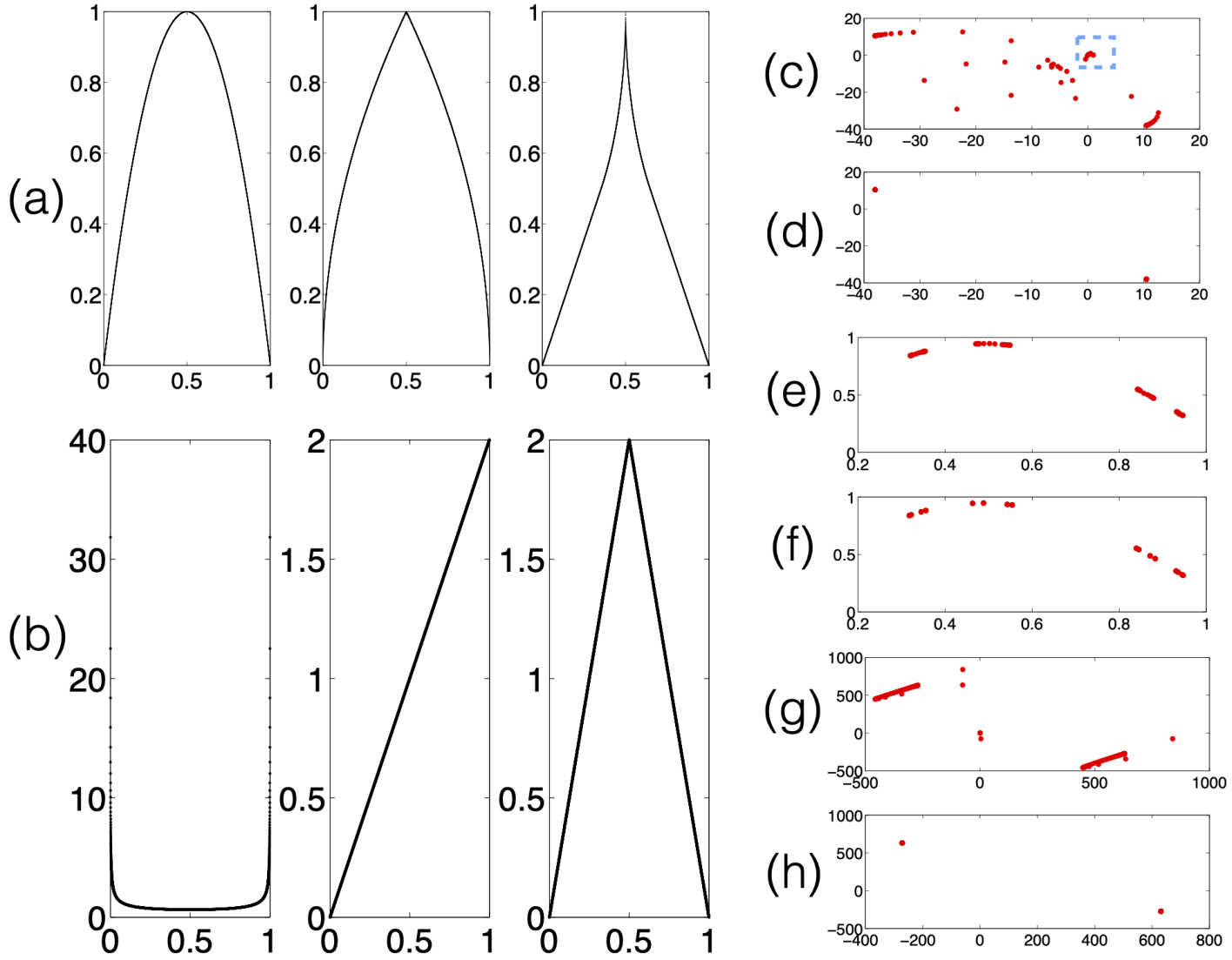

Fig. 4. ESN performance: Graphs of maps T_1, T_2 and T_3 in (a), and their invariant densities in (b) plotted from left to right. Phase portraits of the readouts after 100 time-steps in (c), (e) and (g), and after 200 time-steps in (d), (f) and (h) using 100 samples: ESN in (3) with 200 neurons, $a = 0.2$ and spectral radius of 1.25, and training length of 2000 was used; only a small portion of the phase portrait in (c) falls within the unit square as indicated in a box with a dotted boundary.

expression for the observables since they are components of the universal semi-conjugacy h. However, since the driven dynamical system implements the calculation of $h(\overline{u}^{n+1})$ from $h(\overline{u}^n)$ and u_n through $h(\overline{u}^{n+1}) = g(u_n, h(\overline{u}^n))$, we can compute the data required to find A.

Towards finding A, we consider $m+1$ successive states of the reservoir $(x_t, x_{t+1}, \ldots, x_{t+m-1}, x_{t+m})$ and then arrange them into two data matrices of dimension $m \times N$, where N is the reservoir size, and each x_{t+i} is treated as a row vector:

$$D := \begin{pmatrix} x_t \\ x_{t+1} \\ \vdots \\ x_{t+m-1} \end{pmatrix} \text{ and } D_\# := \begin{pmatrix} x_{t+1} \\ x_{t+2} \\ \vdots \\ x_{t+m} \end{pmatrix}.$$

To obtain a single linear approximation of the collection of mappings from $x_t \mapsto x_{t+1}$, $x_{t+1} \mapsto x_{t+2}, \ldots, x_{t+m-1} \mapsto x_{t+m}$, we consider the operator A that minimizes $\|DA - D_\#\|$ where $\|\ \|$ denotes the Frobenius norm ($\|B\| = \sqrt{\text{Trace}(BB^T)}$. Such an A is given by

$$A = D^+ D_\#, \tag{10}$$

where D^+ is the pseudo-inverse of D. The number of samples m is chosen so that the columns of the data matrix D are linearly independent. That is, its rank is N, where N is the reservoir size. Given a reservoir of size N, this can be achieved by increasing the number of dat points m, and of course with $m \geq N$. The motivation for this requirement is the following result. Suppose we have two matrices D and $D_\#$ of size $m \times N$, then $D_\# = DA$ holds if and only if the nullspace of $D_\#$ contains the nullspace of D where $A := D^+ D_\#$ (for e.g., [19, Theorem 2] – the data in the

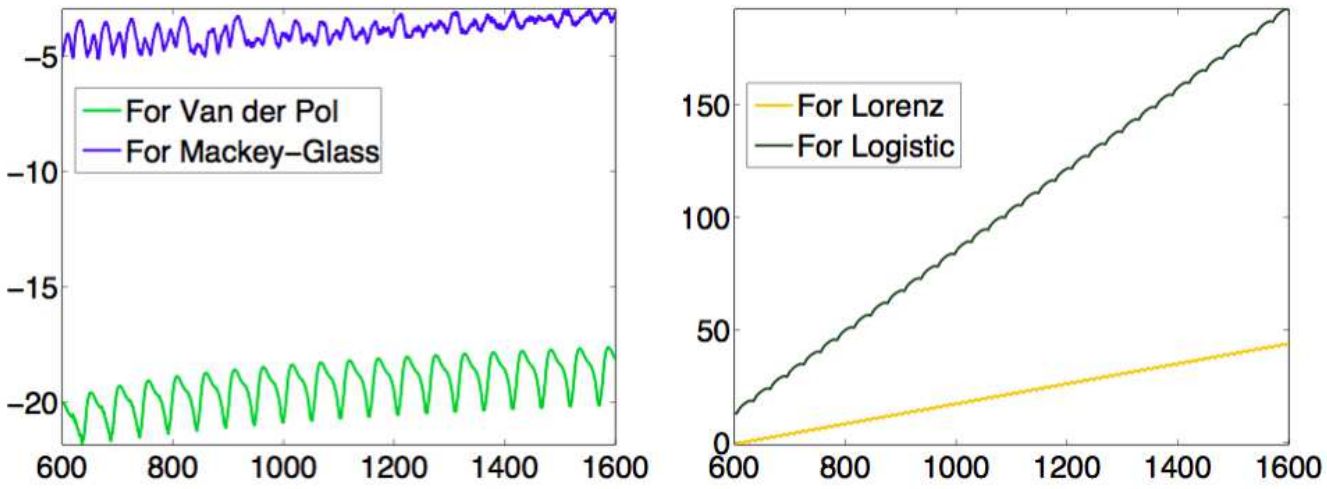

Fig. 5. Logarithm of the average absolute error while predicting the reservoir states using the same ESN network with N nodes: $\log(\frac{1}{N}\sum_{i=1}^{N}|x_n^i - \hat{x}_n^i|)$ plotted against time n where the inputs are generated from (i). Van der Pol Oscillator (ii). Mackey-Glass Oscillator (iii). Lorenz System (iv). Logistic map. ESN as in (3) with 200 neurons, $a = 1$ and W with a spectral radius of 1.1 was used; with $m = 220$, the data matrix D had full rank for all the inputs.

citation is arranged column-wise as against row-wise here). We also emphasize that the relationship in (10) does not mean that $D_\#$ is generated from D through the linear dynamics determined by the matrix A, but instead the row vector x_kA is intended to approximate the row vector x_{k+1}. We then consider the evolution of the reservoir states $(x_k, x_kA, x_kA^2, \ldots)$ determined by A neglecting the input totally. We remark that if the data $(x_t, x_{t+1}, \ldots)$ were considered column vectors and had been arranged as columns in the two data matrices, then $A = D_\#D^+$, and one would then have to consider the linear evolution $(x_k, A^Tx_k, \ldots)$ determined by the A^T. We obtain the linear evolution determined by A from reservoir data obtained from different input dynamical systems. If the actual reservoir state at time k is x_k for a particular input, we plot a distance between x_{k+n} and the linear evolution $\hat{x}_{k+n} := x_kA^n$ against n. To illustrate the effect of the linearity in the reservoir dynamics on prediction, we consider inputs originating from an unforced Van der Pol oscillator (defined by $x'' - 4(1-x^2)x' + x = 0$), Mackey-Glass oscillator (defined by $x'(t) = \frac{0.2x(t-\tau)}{1+x(t-\tau)^{10}} - 0.1x(t)$ with $\tau = 17$), both of which have been found in the ESN literature of rendering excellent prediction accuracy. We also consider the Lorenz system as in [2, Supporting Online Material] (defined with coefficients $\sigma = 10$, $\rho = 28$ and $\beta = 8/3$) for which we know ESN prediction fails after a few lobe switchings, and the logistic map where the prediction is actually more worse as reported earlier. In the plot of Fig. 5, we consider the same ESN for all the inputs and plot the logarithm of the average error between x_{m+n} and $\hat{x}_{m+n}$ defined by $\frac{1}{N}\sum_{i=1}^{N}|x_n^i - \hat{x}_n^i|$ against n where N is the number of nodes in the ESN. With $m = 220$ and $N = 200$, the data matrix D had full rank for all the inputs, and n is varied from 600 to 1000 in Fig. 5. Also, input data between 300 and 520 samples were used in the computation of A. We observe that the logarithm of the average error is much small for the Van der Pol oscillator for 1000 time steps, strongly suggesting that the mapping $h(\overline{u}^n) \mapsto h(\overline{u}^{n+1})$ is approximated well by a linear map. The case is similar to the Mackey-Glass attractor, although the approximation of the reservoir dynamics through the matrix A is slightly worse. For the Lorenz and the logistic map, the reservoir dynamics cannot be captured by the linear evolution determined by A, and for these systems ESN do not have long-term prediction capabilities.

V. Conclusions

The current experimental research on designing ESNs focuses on narrowing the randomness of the choice of a good reservoir and also offers some analytical measures of the reservoir that could enhance the performance. In spite of choosing a reservoir less arbitrarily, not a black art by any means, the nuance involved in training an echo state network has remained only a well-practiced craft, and it calls for research towards making training a science with practical guidelines. In this paper, we have shown that every echo state network yields a curve-fitting and the universal semi-conjugacy exists. Although the universal semi-conjugacy cannot be analytically expressed, its existence helps us understand the failure of the ESNs to approximate some chaotic dynamical systems. A future study should also find theoretical evidence to find for what class of maps or what class of ESNs, the universal semi-conjugacy could embed the inverse limit spaces into the reservoir state space.

When inputs originate from a chaotic dynamical system, the domain of learning could have a complex topology. Thus this could be a curse of memory in the network since an ESN fails to learn a simple polynomial function like the full logistic map. Feedback connections into the reservoir have shown to indicate better predictions for certain chaotic dynamical systems. Could

feedback connections change the the domain of learning from that of an indecomposable continua to a simpler object? Much more research is needed even to explain how the domain of learning changes heuristically.

We also observe that when inputs originating from dynamical systems provide reservoir dynamics that is tractable through a linear map, then their forecasting performance is known to be better. Of course, one needs to further understand why the reservoir dynamics turns linear and how such reservoirs or even driven system could be designed so that their evolution could be approximated by linear maps. We hope that the work in this paper would trigger research in this field. In summary, we believe that even though a random initialization of the reservoir could leave a grey area in the performance analysis of ESNs, other issues could also critically affect the performance of the network, and such issues deserve attention.

Acknowledgements. The author thanks the referees for the suggestions on improving the presentation. The author acknowledges the National Research Foundation, South Africa, for an incentive funding that supports research activities.

REFERENCES

[1] H. Jaeger, "The "echo state" approach to analysing and training recurrent neural networks-with an erratum note," *Bonn, Germany: German National Research Center for Information Technology GMD Technical Report*, vol. 148, no. 34, p. 13, 2001.

[2] H. Jaeger and H. Haas, "Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication," *science*, vol. 304, no. 5667, pp. 78–80, 2004.

[3] W. Maass, T. Natschläger, and H. Markram, "Real-time computing without stable states: A new framework for neural computation based on perturbations," *Neural computation*, vol. 14, no. 11, pp. 2531–2560, 2002.

[4] L. Appeltant, M. C. Soriano, G. Van der Sande, J. Danckaert, S. Massar, J. Dambre, B. Schrauwen, C. R. Mirasso, and I. Fischer, "Information processing using a single dynamical node as complex system," *Nature communications*, vol. 2, p. 468, 2011.

[5] D. Kudithipudi, Q. Saleh, C. Merkel, J. Thesing, and B. Wysocki, "Design and analysis of a neuromemristive reservoir computing architecture for biosignal processing," *Frontiers in neuroscience*, vol. 9, p. 502, 2016.

[6] K. Vandoorne, P. Mechet, T. Van Vaerenbergh, M. Fiers, G. Morthier, D. Verstraeten, B. Schrauwen, J. Dambre, and P. Bienstman, "Experimental demonstration of reservoir computing on a silicon photonics chip," *Nature communications*, vol. 5, p. 3541, 2014.

[7] M. Lukoševičius and H. Jaeger, "Reservoir computing approaches to recurrent neural network training," *Computer Science Review*, vol. 3, no. 3, pp. 127–149, 2009.

[8] G. Manjunath, "Evolving network model that almost regenerates epileptic data," *Neural computation*, vol. 29, no. 4, pp. 937–967, 2017.

[9] P. E. Kloeden and M. Rasmussen, *Nonautonomous dynamical systems*. American Mathematical Soc., 2011, no. 176.

[10] G. Manjunath and H. Jaeger, "Echo state property linked to an input: Exploring a fundamental characteristic of recurrent neural networks," *Neural computation*, vol. 25, no. 3, pp. 671–696, 2013.

[11] A. Ceni, P. Ashwin, L. Livi, and C. Postlethwaite, "The echo index and multistability in input-driven recurrent neural networks," *Physica D: Nonlinear Phenomena*, vol. 412, p. 132609, 2020.

[12] G. Manjunath, "Embedding information onto a dynamical system," *arXiv preprint arXiv:2105.10766*, 2021.

[13] L. Grigoryeva and J.-P. Ortega, "Echo state networks are universal," *Neural Networks*, vol. 108, pp. 495–508, 2018.

[14] L. Gonon, L. Grigoryeva, and J.-P. Ortega, "Approximation bounds for random neural networks and reservoir systems," *arXiv preprint arXiv:2002.05933*, 2020.

[15] G. Manjunath, "Stability and memory-loss go hand-in-hand: three results in dynamics and computation," *Proceedings of the Royal Society A*, vol. 476, no. 2242, p. 20200563, 2020.

[16] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.

[17] F. Takens, "Detecting strange attractors in turbulence," in *Dynamical systems and turbulence, Warwick 1980*. Springer, 1981, pp. 366–381.

[18] P. J. Schmid, "Dynamic mode decomposition of numerical and experimental data," *Journal of fluid mechanics*, vol. 656, pp. 5–28, 2010.

[19] J. H. Tu, C. W. Rowley, D. M. Luchtenburg, S. L. Brunton, and J. N. Kutz, "On dynamic mode decomposition: Theory and applications," *arXiv preprint arXiv:1312.0041*, 2013.

[20] M. O. Williams, I. G. Kevrekidis, and C. W. Rowley, "A data–driven approximation of the koopman operator: Extending dynamic mode decomposition," *Journal of Nonlinear Science*, vol. 25, no. 6, pp. 1307–1346, 2015.

[21] W. T. Ingram and W. S. Mahavier, *Inverse limits: From continua to Chaos*. Springer Science & Business Media, 2011, vol. 25.

[22] A. Hart, J. Hook, and J. Dawes, "Embedding and approximation theorems for echo state networks," *Neural Networks*, 2020.

[23] J. Kennedy, D. R. Stockman, and J. A. Yorke, "The inverse limits approach to chaos," *Journal of Mathematical Economics*, vol. 44, no. 5-6, pp. 423–444, 2008.

[24] W. Lewis and P. Minc, "Drawing the pseudo-arc," *Houston J. Math*, vol. 36, pp. 905–934, 2010.

[25] M. Barge and J. Martin, "Chaos, periodicity, and snakelike continua," *Transactions of the American Mathematical Society*, vol. 289, no. 1, pp. 355–365, 1985.

[26] A. Lasota and M. C. Mackey, *Chaos, fractals, and noise: stochastic aspects of dynamics*. Springer Science & Business Media, 2013, vol. 97.

[27] N. Wei, "Solutions of the inverse frobenius-perron problem," Ph.D. dissertation, Concordia University, 2015.