Neural Networks 128 (2020) 234–247

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Embedding and approximation theorems for echo state networks

Allen Hart [*], James Hook, Jonathan Dawes

Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK

ARTICLE INFO

Article history:
Received 14 August 2019
Received in revised form 6 May 2020
Accepted 11 May 2020
Available online 16 May 2020

Keywords:
Reservoir computing
Lorenz equations
Dynamical system
Delay embedding
Persistent homology
Recurrent neural networks

ABSTRACT

Echo State Networks (ESNs) are a class of single-layer recurrent neural networks that have enjoyed recent attention. In this paper we prove that a suitable ESN, trained on a series of measurements of an invertible dynamical system, induces a C^1 map from the dynamical system's phase space to the ESN's reservoir space. We call this the Echo State Map. We then prove that the Echo State Map is generically an embedding with positive probability.

Under additional mild assumptions, we further conjecture that the Echo State Map is almost surely an embedding. For sufficiently large, and specially structured, but still randomly generated ESNs, we prove that there exists a linear readout layer that allows the ESN to predict the next observation of a dynamical system arbitrarily well. Consequently, if the dynamical system under observation is structurally stable then the trained ESN will exhibit dynamics that are topologically conjugate to the future behaviour of the observed dynamical system.

Our theoretical results connect the theory of ESNs to the delay-embedding literature for dynamical systems, and are supported by numerical evidence from simulations of the traditional Lorenz equations. The simulations confirm that, from a one dimensional observation function, an ESN can accurately infer a range of geometric and topological features of the dynamics such as the eigenvalues of equilibrium points, Lyapunov exponents and homology groups.

1. Introduction

An Echo State Network (ESN) is a single-layer recurrent neural network composed of a trainable readout layer connected to a reservoir of randomly initialised, and randomly coupled, untrainable 'neurons'. This architecture has been investigated and used by many authors since the seminal papers by Jaeger (2001) and Maass, Natschläger, and Markram (2002). Tanaka et al. (2019) present a review of ESNs, among other recurrent neural network models, under the umbrella term *reservoir computing*.

The wide range of problems to which the ESN framework has been applied include speech recognition (Skowronski & Harris, 2007), learning grammatical structure (Tong, Bickett, Christiansen, & Cottrell, 2007), and financial time series prediction (Ilies et al., 2007; Lin, Yang, & Song, 2009). Several authors including Gürel and Egert (2010) have also discussed how the ESN is a plausible model for the information processing performed by biological neurons. Most ambitiously, Plöger, Arghir, Günther, and Hosseiny (2004) discuss ESNs in the context of *building by 2050, a team of fully autonomous humanoid robots to beat the human winning team of the FIFA Soccer World Cup*.

The ESN has associated to it a *reservoir state* denoted $r_k \in \mathbb{R}^n$ at time k. The structure of the recurrent layer is described by an $n \times n$ matrix A that is the weighted adjacency matrix of the system of n 'neurons'. If neuron i is not connected to neuron j then $A_{ij} = 0$, and if they are connected with some weight $a \in \mathbb{R}$ then $A_{ij} = a$. Connections need not be symmetric, so in general $A_{ij} \neq A_{ji}$. Typically, A is sparse and has approximately 1% of its entries nonzero. Connection weights are usually i.i.d. random variables, and typically are chosen to be either uniformly distributed on a fixed interval, or Gaussian. The ESN also contains an $(n \times m)$ input matrix W^{in}, where m is the dimension of the training data. Like the reservoir A, W^{in} is also populated with i.i.d random variables. Finally, the ESN has an activation function $\varphi : \mathbb{R}^n \to \mathbb{R}^n$, for which there are several standard choices, for example tanh (performed component-wise).

The operation of the ESN is divided into two phases: an initial training phase, followed by an autonomous phase. During the training phase, the ESN is trained on a given input time series denoted by vectors u_0, u_1, u_2... u_K each in $\mathbb{R}^m$. We will assume in this paper that the input sequence is bounded, though we note the recent work of Grigoryeva and Ortega (2019) establishes a framework that encompasses unbounded input sequences as well. We will also assume in this paper that $m = 1$, so that we consider a scalar input time series. The *reservoir state* at time k is defined by choosing an initial state e.g. $r_1 = (0, 0, \ldots, 0)^\top$ and

[*] Corresponding author.
E-mail addresses: a.hart@bath.ac.uk (A. Hart), j.l.hook@bath.ac.uk (J. Hook), j.h.p.dawes@bath.ac.uk (J. Dawes).

https://doi.org/10.1016/j.neunet.2020.05.013
0893-6080/© 2020 Elsevier Ltd. All rights reserved.

defining subsequent states recursively by

$$r_{k+1} = \varphi(Ar_k + W^{\text{in}}u_k).$$

Having computed the new reservoir states $r_1, r_2...r_K$, the output matrix W^{out} is fitted to solve the optimisation problem

$$\min_{W^{\text{out}}} \sum_{k=1}^{K} \|W^{\text{out}}r_k - a_k\|^2 + \lambda\|W^{\text{out}}\|_2^2,$$

where a_k is some known target sequence we want the ESN to mimic, often taken to be equal to the input sequence u_k, and $\lambda > 0$ is a regularisation parameter. Minimisation problems of this kind are often referred to as ridge regression, or Tikhonov, or L^2 regularisation. Having trained the output matrix W^{out} the reservoir states s_k can then be liberated from their reliance on the driving input u_k and evolve under the autonomous dynamical system defined by

$$v_{k+1} = W^{\text{out}}s_k,$$
$$s_{k+1} = \varphi(As_k + W^{\text{in}}v_{k+1}),$$

where $s_0 = r_K$. If the training has been successful, then the trained ESN should provide good predictions of the future time series $v_1 \approx u_{K+1}, v_2 \approx u_{K+2}$, etc, and future evolution of the reservoir state $s_1 \approx r_{K+1}, s_2 \approx r_{K+2}$ etc. The viewpoint we take here clearly distinguishes between the training phase of the ESN where it is an externally-driven dynamical system, and the 'test' phase where we consider it as an autonomous dynamical system in $\mathbb{R}^n$.

In complete generality the process defining u_k could be anything, including a realisation of a random process. However, importantly, throughout this paper, we will restrict our attention and assume that $u_0, u_1, u_2, \dots$ are one dimensional observations of an invertible discrete-time dynamical system with evolution operator $\phi \in \text{Diff}^1(M)$ observed via a function $\omega \in C^1(M, \mathbb{R})$ on a compact manifold M. In particular $u_0 = \omega(x), u_1 = \omega \circ \phi(x), u_2 = \omega \circ \phi^2(x), u_3 = \omega \circ \phi^3(x)$, etc. The model we have in mind is that ϕ is the evolution operator for a time Δt of a set of Lipschitz ordinary differential equations on M. Illustrations of the training and autonomous phases are shown in Fig. 1.

The idea to draw training data from a dynamical system was by Jaeger and Haas (2004) who drew observations from a trajectory through the Mackey–Glass attractor. We were attracted to the idea by a recent paper by Pathak, Lu, Hunt, Girvan, and Ott (2017) who trained ESNs on the Lorenz equations and the Kuramoto–Sivashinsky equation (in one spatial dimension). In particular, we conjecture that under the right technical conditions an ESN with random reservoir matrix and input matrix trained on a one dimensional observation of a dynamical system will embed the system into the reservoir space almost surely. We call this the ESN Embedding Conjecture (Conjecture 2.3.4). We believe this conjecture is true as a consequence of Takens (1981) theorem stating that a generic delay observation map is an embedding. This connection between Takens' delay embedding theorem and the ESN was remarked on by Jaeger (2001) and has been discussed in several later works including by Løkse, Bianchi, and Jenssen (2017), Schrauwen, Verstraeten, and Van Campenhout (2007), Shi and Han (2007), Vlachas et al. (2019), Xi, Shi, and Han (2005), Yeo (2019), and Yong, Yibin, Qun, and Caihong (2010). We go on to prove that our statement of the ESN Embedding Conjecture holds with probability $\alpha > 0$. We finally prove that when the ESN does successfully embed a structurally stable dynamical system into its reservoir, there exists a trainable readout layer such that the autonomous phase of the ESN will adopt the topology of the driving dynamical system. We call this the ESN Approximation Theorem (Theorem 2.4.13). This theorem complements the results of Grigoryeva and Ortega (2018) and Gonon, Grigoryeva, and Ortega (2020) stating that the ESN (with tunable and randomly initialised A and b respectively) is a universal approximator of discrete-time fading memory filters.

To demonstrate the theory we present numerical evidence that an ESN trained on a numerically integrated trajectory of the Lorenz system can replicate several of the Lorenz system's geometric and topological properties. In particular, we computed the Lyapunov exponents of the ESN autonomous phase and compared them to the known exponents of the Lorenz system. We also compared the eigenvalues of the system linearisation on the Lorenz system's fixed points to the eigenvalues of the linearisation on the fixed points belonging to the ESN autonomous phase. Finally we compared the homology of the driven and autonomous reservoir attractors to the Lorenz attractor using persistent homology. For the reader unfamiliar with persistent homology Ghrist (2008) offers an excellent primer.

The remainder of the paper is set out as follows. In Section 2 we present basic definitions and define a family of maps that captures the effect on the reservoir state of training with increasing amounts of data. In Section 2.2 we prove that the family converges to a C^1 map that we call the Echo State Map. We conjecture in Section 2.3 that generically the Echo State Map is an embedding. In Section 2.4 we prove an ESN Approximation Theorem that guarantees that the autonomous dynamics of the ESN is (in a suitable sense) conjugate via a diffeomorphism to the original dynamical system on which the ESN was trained. in Section 3 we present numerical results supporting the theory.

2. Theory of ESNs

Our analysis makes use of several different norms. In particular, if $x \in \mathbb{R}^m$ is a vector then $\|x\|$ is the Euclidean norm, and for A a matrix then $\|A\|_2$ is the matrix 2 norm. If f is a real valued function, then $\|f\|_\infty$ will denote the supremum norm and if f is continuously differentiable then we will use the C^1 norm $\|f\|_{C^1}$ defined by

$$\|f\|_{C^1} := \|f\|_\infty + \|Df\|_\infty$$

where D is the derivative operator.

2.1. The Echo State Network

We begin our summary of the background to Echo State Networks (ESNs) with a definition.

Definition 2.1.1 (*Echo State Network*)**.** Let the activation function σ be a function $\sigma \in C^1(\mathbb{R}, (-1, 1))$ that has its derivative take values in the range $(0, 1)$. Let $n \in \mathbb{N}$, A be a real $n \times n$ matrix, and W^{in} a real $n \times 1$ matrix. Let $b_i \in \mathbb{R}\ \forall i \in \{1, \dots, n\}$. Let $I_n := [-1, 1]^n$ and define the function $\varphi : \mathbb{R}^n \to I_n$ component-wise by

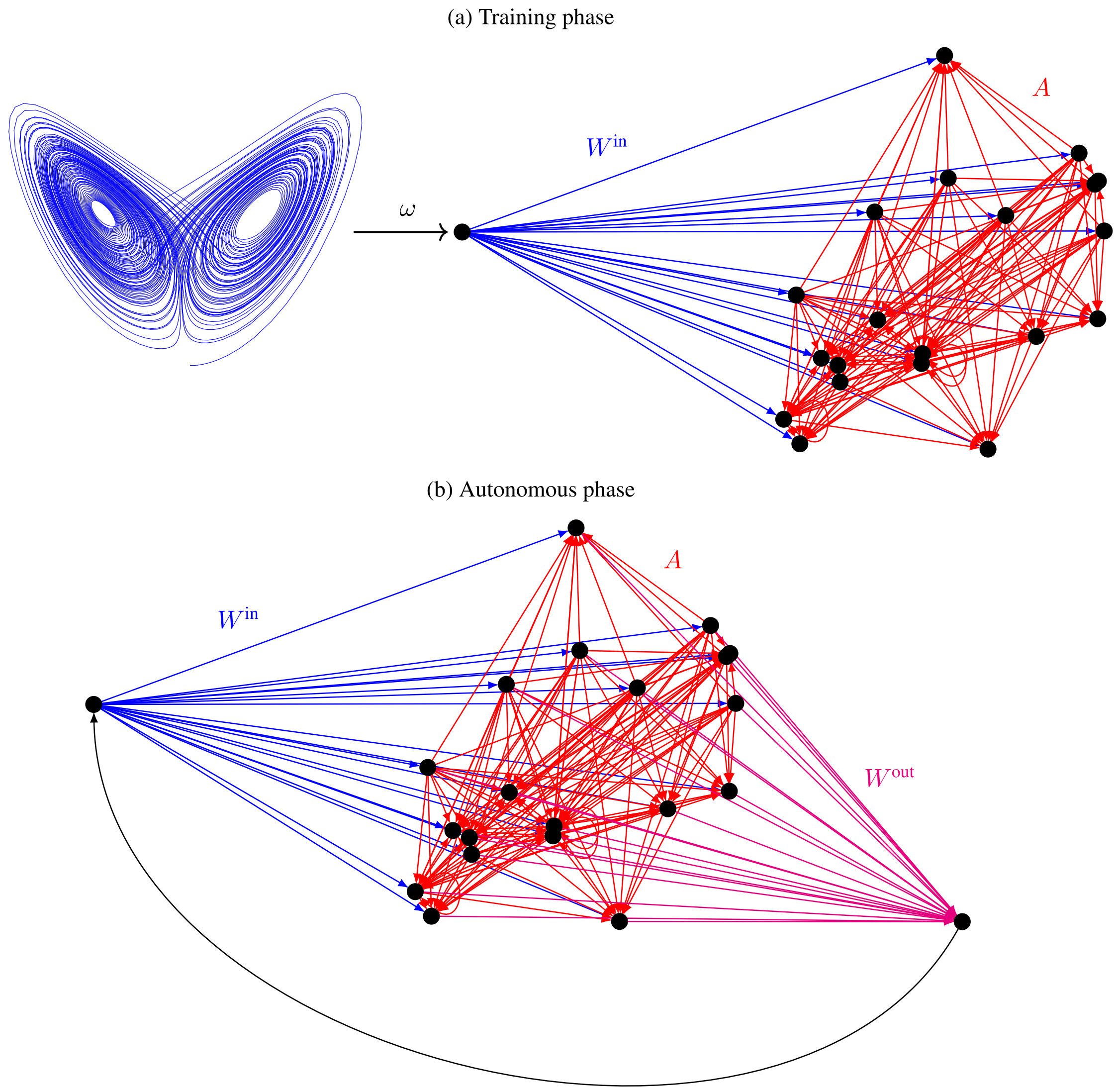
$$\varphi_i(r) = \sigma(r_i + b_i)\ \forall i \in \{1, \dots, n\}. \tag{1}$$

We then define an Echo State Network (ESN) of size n to be the triple $(\varphi, A, W^{\text{in}})$.

σ is often chosen to be the hyperbolic function tanh, though other choices of activation function abound in the machine learning literature. The conditions on these functions are sometimes less restrictive than those imposed above on σ; other common choices of activation function include the linear unit (also known as the identity map) and the rectified linear unit (often abbreviated relu) defined by

$$\text{relu}(r_i) = \begin{cases} r_i & \text{if } r_i > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Glorot, Bordes, and Bengio (2011) discuss how Recurrent Neural Networks supported by a relu activation function are less



(a) Training phase

(b) Autonomous phase

Fig. 1. (a) During the training phase the ESN observes a dynamical system via the function $\omega \in C^1(M, \mathbb{R})$; this sequence of observations is distributed into the nodes in the reservoir r by the linear map W^{in}. (b) After training, in the autonomous phase, the driving is replaced by the output created by the best-fit linear map W^{out}. These images were produced using the TikZ-network package developed by Hackl (2018).

prone to the *vanishing gradient problem* than sigmoidal activation functions. More exotic activation functions include radial basis functions, which take the shape of bell curves. Throughout this paper however, we will restrict ourselves to activation functions as defined above, i.e. functions $\sigma \in C^1(\mathbb{R}, (-1, 1))$ whose derivatives take values in $(0, 1)$.

2.2. The Echo State Map

We will begin by introducing a family of functions that describe the mapping between this time series of observations and the reservoir state; this will be of fundamental importance throughout the remainder of the paper.

Definition 2.2.1 (*Echo State Family*)**.** Let M be a compact m-manifold and $n \in \mathbb{N}$. Let A be an $n \times n$, and let W^{in} an $n \times 1$ matrix: let the triple $(\varphi, A, W^{\text{in}})$ be an ESN. Let the discrete dynamical system be $\phi \in \text{Diff}^1(M)$ and let the observation function $\omega \in C^1(M, \mathbb{R})$. Let the family of functions $F = \{f_k^{r_0} : M \to I_n : r_0 \in I_n,\ k \in \mathbb{N}_0\}$ be defined as follows:

$$f_0^{r_0}(x) = r_0$$
$$f_{k+1}^{r_0}(x) = \varphi(Af_k^{r_0} \circ \phi^{-1}(x) + W^{\text{in}}\omega(x)).$$

We call the set of functions F the *Echo State Family*.

To provide some intuition as to where this family came from, we observe that $f_k^{r_0}$ is the function that takes a point $x \in M$ and first applies the inverse evolution operator k times, yielding the past state $\phi^{-k}(x)$ of the dynamical system. A list of $k+1$ observations $\omega \circ \phi^{-k}(x),\ \omega \circ \phi^{1-k}(x),\ \omega \circ \phi^{2-k}(x),\ \ldots$ are then obtained, in sequence, forward from this point. An ESN with initial reservoir state r_0 is trained on this list of inputs, and its reservoir state is then exactly given by the value of $f_k^{r_0}(x)$. The function $f_k^{r_0}$ is therefore the map induced by $k+1$ steps of the training phase of the ESN, i.e. it sends a point $x \in M$ to reservoir space I_n according to its one dimensional history. Our plan for the upcoming section is to show that for any $r_0 \in I_n$

$$f^{r_0} := \lim_{k \to \infty} f_k^{r_0}$$

exists, and that $f^{r_0} = f^{s_0} =: f$ for any $r_0, s_0 \in I_n$. We will call f the Echo State Map, and show it is continuously differentiable, i.e. $f \in C^1(M, \mathbb{R}^n)$. These results will appear together and called the Echo State Mapping Theorem. Equivalently, we could say the Echo State Map f is the unique C^1 generalised synchronisation (in the sense described by Kocarev and Parlitz (1996)) between a pair of unidirectionally coupled systems, the dynamics given by ϕ and the driven ESN phase.

We will further conjecture that f is a C^1 embedding almost surely, and therefore (almost surely) it is a topology-preserving map from the manifold M to the reservoir space I_n. We will call this the ESN Embedding Conjecture, and go on to prove a partial result that f is a C^1 embedding with positive probability.

Theorem 2.2.2 (*Echo State Mapping Theorem*)**.** *With the notation and hypotheses of Definition 2.2.1, and the further assumption that $\|A\|_2 < \min(1, 1/\|D\phi^{-1}\|_\infty)$, there exists a unique solution $f \in C^1(M, \mathbb{R}^n)$ of the equation*

$$f = \varphi(Af \circ \phi^{-1} + W^{in}\omega)$$

such that for all $r_0 \in I_n$ the sequence $f_k^{r_0}$ converges in the C^1 topology to f as $k \to \infty$. We call f the Echo State Map.

Proof. Let $\tilde{\Psi} : C^1(M, \mathbb{R}^n) \to C^1(M, \mathbb{R}^n)$ be defined by

$$\tilde{\Psi}(f) = \varphi(Af \circ \phi^{-1} + W^{\text{in}}\omega)$$

then we can see that

$$f_k^{r_0} = \tilde{\Psi}(f_{k-1}^{r_0}) = \tilde{\Psi}^k(f_0^{r_0}).$$

Now, we will show that $\tilde{\Psi}$ is a contraction mapping and therefore has a unique fixed point $f \in C^1(M, \mathbb{R}^n)$ by the contraction mapping theorem (Banach, 1922). This will complete the proof.

$$\begin{aligned}
\|\tilde{\Psi}(f) - \tilde{\Psi}(g)\|_{C^1} &= \|\varphi(Af \circ \phi^{-1} + W^{\text{in}}\omega) \\
&\quad -\varphi(Ag \circ \phi^{-1} + W^{\text{in}}\omega)\|_{C^1} \\
&\leq \|Af \circ \phi^{-1} + W^{\text{in}}\omega - Af \circ \phi^{-1} \\
&\quad -W^{\text{in}}\omega\|_{C^1} \text{ because } \varphi \text{ is contracting in } C^1 \\
&= \|A(f \circ \phi^{-1} - g \circ \phi^{-1})\|_{C^1} \\
&\leq \|A\|_2 \|f \circ \phi^{-1} - g \circ \phi^{-1}\|_{C^1} \\
&= \|A\|_2(\|f \circ \phi^{-1} - g \circ \phi^{-1}\|_\infty \\
&\quad +\|Df \circ \phi^{-1} D\phi^{-1} - Dg \circ \phi^{-1} D\phi^{-1}\|_\infty) \\
&\leq \|A\|_2(\|f \circ \phi^{-1} - g \circ \phi^{-1}\|_\infty \\
&\quad +\|D\phi^{-1}\|_\infty \|Df \circ \phi^{-1} - Dg \circ \phi^{-1}\|_\infty) \\
&\leq \|A\|_2 \max(1, \|D\phi^{-1}\|_\infty)\|f - g\|_{C^1}
\end{aligned}$$

and $\|A\|_2 \max(1, \|D\phi^{-1}\|_\infty) < 1$, so we have that $\tilde{\Psi}$ is contracting. □

We remark here that if ϕ is obtained by the discretisation of a continuous time flow with a small time step, the evolution operator ϕ is close to the identity map, so $\|D\phi^{-1}\|_\infty$ is close to 1. Consequently, the condition $\|A\|_2 < \min(1, 1/\|D\phi^{-1}\|_\infty)$ is not much more restrictive than enforcing $\|A\|_2 < 1$.

2.3. The ESN embedding theorem

In this section we will discuss the conditions under which the Echo State Map $f \in C^1(M, \mathbb{R}^n)$ is a C^1 embedding (i.e. an injective immersion whose domain and image are diffeomorphic). We will also conjecture that for a generic observation function ω and random matrices A and W^{in}, the Echo State Map f is a C^1 embedding almost surely. To set the scene for these results, we first recall Whitney's Weak Embedding Theorem and Takens' Theorem for delay observation maps.

Theorem 2.3.1 (*Whitney's Weak Embedding Theorem*)**.** *Let M be a compact m-manifold and choose $n \in \mathbb{N}$ such that $n > 2m$. Then the set of C^r embeddings is generic in $C^r(M, \mathbb{R}^n)$ with respect to the Whitney C^1 topology (This is the topology on $C^1(M, \mathbb{R}^n)$ induced by the C^1-norm).*

Proof. Whitney (1944). □

Corollary 2.3.2. *Let M be a compact m-manifold and $n \in \mathbb{N}$ such that $n > 2m$. Let A be an $n \times n$ matrix for which $\|A\|_2 < min(1/\|D\phi^{-1}\|_\infty, 1)$. Let W^{in} be an $n \times 1$ matrix, and let the triple (φ, A, W^{in}) be an ESN. As usual, let $\phi \in Diff^1(M)$, and $\omega \in C^1(M, \mathbb{R})$. If $n > 2m$, then the ESM $f \in C^1(M, \mathbb{R}^n)$ is a limit point in the Whitney C^1 topology of the set of C^1 embeddings.*

Proof. $f \in C^1(M, \mathbb{R}^n)$ by Theorem 2.2.2 so, by the Weak Whitney Embedding Theorem (Theorem 2.3.1), f is a limit point of the C^1 embeddings with respect to the Whitney C^1 topology. □

From Corollary 2.3.2 it is clear that the Echo State Map f is always close to an embedding, but this says nothing about necessary or sufficient conditions for f to actually *be* an embedding. In fact f may never actually be an embedding. That said, since embeddings are generic in the space $C^1(M, \mathbb{R}^n)$ we expect heuristically that a function in $C^1(M, \mathbb{R}^n)$ that is assembled without explicitly desiring that it is not an embedding, is overwhelmingly likely actually to be an embedding. This suggests (heuristically) that a generic Echo State Map f is indeed an embedding. The first step we take towards proving this is to introduce Takens' Theorem.

Theorem 2.3.3 (*Huke's Formulation of Takens' Theorem*)**.** *Let M be a compact manifold of dimension m. Suppose $\phi \in Diff^2(M)$ has the following two properties:*

(1) ϕ has only finitely many periodic points with periods less than or equal to $2m$.

(2) If $x \in M$ is any periodic point with period $k < 2m$ then the eigenvalues of the derivative $D\phi^k$ at x are distinct.

Then for a generic C^2 observation function $\omega \in C^2(M, \mathbb{R})$ the $(2m + 1)$ delay observation map $\Phi_{(\phi,\omega)} : M \to \mathbb{R}^{2m+1}$ defined by

$$\Phi_{(\phi,\omega)}(x) := (\omega(x), \omega \circ \phi(x), \omega \circ \phi^2(x), \ldots, \omega \circ \phi^{2m}(x))$$

is a C^1 embedding.

Proof. Huke (2006). □

Huke's proof that $\Phi_{(\phi,\omega)}$ is a C^1 embedding for generic ω comprises two steps. First, he shows that $\Phi_{(\phi,\omega)}$ is a C^1 embedding for an open subset of C^2 observation functions, and second, he shows that $\Phi_{(\phi,\omega)}$ is an embedding for a dense subset of all C^2 observation functions. The first step (to prove openness) is fairly simple while the second (the proof of density) is long and delicate. A brief summary of the density part of the proof is as follows. An arbitrary C^2 observation function ω is carefully perturbed on each open set in a cover of the manifold M such that ω becomes immersive on each set. The observation function ω is then perturbed again on each open set in the cover in order to make ω injective, with care taken to ensure ω remains immersive on each open set. This procedure is applied separately to open sets which contain periodic points and open sets that do not. We believe it is possible to build on this result and modify the proof of Huke's Theorem in order to prove an ESN Embedding Conjecture in the form that we now state.

Conjecture 2.3.4 (*ESN Embedding Conjecture*)**.** *Let M be a compact m-manifold and $n \in \mathbb{N}$ such that $n > 2m$. Let A be an $n \times n$ matrix with $\|A\|_2 < min(1/\|D\phi^{-1}\|_\infty, 1)$, and W^{in} a $n \times 1$ matrix, and let the triple (φ, A, W^{in}) be an ESN. Let $\omega \in C^2(M, \mathbb{R})$ and let $\phi \in Diff^2(M)$ (and possibly requiring additional properties), and let A, W^{in} be generic matrices in the topology induced by the matrix 2-norm. Then the Echo State Map $f \in C^1(M, \mathbb{R}^n)$ is a C^1 embedding.*

We now summarise our partial success towards proving this conjecture. In particular we can establish the properties analogous to the first part of Huke's proof of Takens' Theorem: we will show that the set of triples $(A, W^{\text{in}}, \omega)$ of reservoir matrix, input matrix, and observation function for which f is a C^1 embedding, is open and non-empty. Consequently, for a generic observation function ω, and matrices A and W^{in} drawn from a distribution with full support (if the pdf is well defined, it is greater than 0 over its domain), f is a C^1 embedding with probability $\alpha > 0$. To prove the full ESN Embedding Conjecture, all that remains is to show that the triples $(A, W^{\text{in}}, \omega)$ for which f is an embedding are dense in the space of admissible triples, but this is no easy task, so we will be satisfied here with the proof of only openness and non-emptiness.

Lemma 2.3.5. *Let M be a compact m-manifold and $n \in \mathbb{N}$. Let A be an $n \times n$ matrix, and suppose that $\|A\|_2 < \min(1/\|D\phi^{-1}\|_\infty, 1)$. As usual let W^{in} a $n \times 1$ matrix, let the triple (φ, A, W^{in}) be an ESN, $\phi \in \mathit{Diff}^1(M)$ and $\omega \in C^1(M, \mathbb{R})$. Define the set $\Omega := \{(A, W^{in}, \omega) \mid f_{A,W^{in},\omega}$ is a C^1 embedding.$\}$. Then the set Ω is open in the C^1 topology.*

Proof. First we define the map Ψ that associates the ESM f to the triple $(A, W^{\text{in}}, \omega)$; let $\Psi : (A, W^{\text{in}}, \omega) \to C^1(M, \mathbb{R}^n)$ be defined by $\Psi(A, W^{\text{in}}, \omega) = f_{A,W^{\text{in}},\omega}$. We now argue as follows. Since C^1 embeddings form an open subset of $C^1(M, \mathbb{R})$, and the inverse image of a continuous map is open, it suffices to show that Ψ is continuous in order to then conclude that Ω is open. To show continuity of Ψ we must prove that if $(A_n, W_n^{\text{in}}, \omega_n)_{n\in\mathbb{N}} \to (A, W^{\text{in}}, \omega)$ then $\|\Psi(A_n, W_n^{\text{in}}, \omega_n) - \Psi(A, W^{\text{in}}, \omega)\|_{C^1} \to 0$.

To lighten the notation we will write f for $f_{A,W^{\text{in}},\omega}$ and f_n for $f_{A_n,W_n^{\text{in}},\omega_n}$. As a preliminary result we estimate as follows:

$$\begin{aligned}
\|A_nf_n \circ \phi^{-1} - Af \circ \phi^{-1}\|_{C^1} &= \|A_nf_n \circ \phi^{-1} - Af \circ \phi^{-1}\|_\infty \\
&\quad + \|A_nDf_n \circ \phi^{-1}D\phi^{-1} \\
&\quad - ADf \circ \phi^{-1}D\phi^{-1}\|_\infty \qquad (2) \\
&\qquad \text{by definition of the } C^1 \text{ norm} \\
&\leq \|A_nf_n \circ \phi^{-1} - Af \circ \phi^{-1}\|_\infty \\
&\quad + \|D\phi^{-1}\|_\infty \|A_nDf_n \circ \phi^{-1} \\
&\quad - ADf \circ \phi^{-1}\|_\infty \\
&\leq \|A_nf_n - Af\|_\infty + \|D\phi^{-1}\|_\infty \|A_nDf_n \\
&\quad - ADf\|_\infty \\
&\leq \max(1, \|D\phi^{-1}\|_\infty)(\|A_nf_n - Af\|_\infty \\
&\quad + \|A_nDf_n - ADf\|_\infty) \\
&\leq \max(1, \|D\phi^{-1}\|_\infty)\|A_nf_n - Af\|_{C^1} \\
&= \tau\|A_nf_n - Af\|_{C^1} \qquad (3)
\end{aligned}$$

where we have defined $\tau := \max(1, \|D\phi^{-1}\|_\infty)$. We will prove one more preliminary result: that $\|f_n\|_{C^1}$ is bounded. We can see that $\|f_n\|_\infty$ is bounded by boundedness of φ so all that remains is to bound $\|Df_n\|_\infty$. Since

$$f_n = \varphi(A_nf_n \circ \phi^{-1} + W_n^{\text{in}}\omega_n)$$

we compute directly that

$$Df_n = D\varphi(A_nf_n \circ \phi^{-1}W_n^{\text{in}}\omega_n)(A_nDf_n \circ \phi^{-1}D\phi^{-1} + W_n^{\text{in}}D\omega_n)$$

from which we can estimate that

$$\begin{aligned}
\|Df_n\|_\infty &= \|D\varphi(A_nf_n \circ \phi^{-1}W_n^{\text{in}}\omega_n) \\
&\quad \times (A_nDf_n \circ \phi^{-1}D\phi^{-1} + W_n^{\text{in}}D\omega_n)\|_\infty \\
&\leq \|A_nDf_n \circ \phi^{-1}D\phi^{-1} + W_n^{\text{in}}D\omega_n\|_\infty \\
&\leq \|A\|_2\|Df_n \circ \phi^{-1}\|_\infty\|D\phi^{-1}\|_\infty + \|W_n^{\text{in}}\omega_n\|_\infty \\
&< \bar{\rho}\|Df_n \circ \phi^{-1}\|_\infty\|D\phi^{-1}\|_\infty + \|W_n^{\text{in}}D\omega_n\|_\infty \\
&\qquad \text{where } \bar{\rho} = \sup_{n\in\mathbb{N}} \|A_n\|_2 < 1 \\
&= \bar{\rho}\|Df_n\|_\infty\|D\phi^{-1}\|_\infty + \|W_n^{\text{in}}D\omega_n\|_\infty \\
&< \bar{\rho}\|Df_n\|_\infty\|D\phi^{-1}\|_\infty + \nu
\end{aligned}$$

where ν is a bound for the sequence $\|W_n^{\text{in}}D\omega_n\|_\infty$, which we know exists because $\|W_n^{\text{in}}D\omega_n\|_\infty$ converges. Now upon rearrangement

$$\|Df_n\|_\infty < \frac{\nu}{1 - \bar{\rho}\|D\phi^{-1}\|_\infty},$$

hence we have bounded $\|Df_n\|_\infty$ and $\|f_n\|_\infty$ thus we have a bound for $\|f_n\|_{C^1}$, which we will call μ. Now, for all $\epsilon > 0$ there exists $n \in \mathbb{N}$ such that both

$$\|A_n - A\|_2 < \frac{\epsilon(1 - \tau\|A\|_2)}{2\tau\mu} \qquad (4)$$

and

$$\|W_n^{\text{in}}\omega_n - W^{\text{in}}\omega\|_{C^1} < \frac{\epsilon(1 - \tau\|A\|_2)}{2}. \qquad (5)$$

Armed with these estimates we can now compute that

$$\begin{aligned}
\|f_n - f\|_{C^1} &= \|\varphi(A_nf_n \circ \phi^{-1} + W_n^{\text{in}}\omega_n)\|. \\
&\quad \|. - \varphi(Af \circ \phi^{-1} + W^{\text{in}}\omega)\|_{C^1} \text{ by Theorem 2.2.2} \\
&\leq \|A_nf_n \circ \phi^{-1} + W_n^{\text{in}}\omega_n - Af \circ \phi^{-1} - W^{\text{in}}\omega\|_{C^1} \\
&\qquad \text{because } \varphi \text{ is contracting} \\
&\leq \|A_nf_n \circ \phi^{-1} - Af \circ \phi^{-1} + W_n^{\text{in}}\omega_n - W^{\text{in}}\omega\|_{C^1} \\
&\leq \|A_nf_n \circ \phi^{-1} - Af \circ \phi^{-1}\|_{C^1} + \|W_n^{\text{in}}\omega_n - W^{\text{in}}\omega\|_{C^1} \\
&\leq \tau\|A_nf_n - Af\|_{C^1} + \|W_n^{\text{in}}\omega_n - W^{\text{in}}\omega\|_{C^1} \\
&\qquad \text{by Eqs. (2)-(3)} \\
&\leq \tau\|A_nf_n - Af_n + Af_n - Af\|_{C^1} + \|W_n^{\text{in}}\omega_n - W^{\text{in}}\omega\|_{C^1} \\
&\leq \tau(\|Af_n - Af\|_{C^1} + \|A_nf_n - Af_n\|_{C^1}) \\
&\quad + \|W_n^{\text{in}}\omega_n - W^{\text{in}}\omega\|_{C^1} \\
&\leq \tau\|A\|_2\|f_n - f\|_{C^1} + \tau\|f_n\|_{C^1}\|A_n - A\|_2 \\
&\quad + \|W_n^{\text{in}}\omega_n - W^{\text{in}}\omega\|_{C^1} \\
&< \tau\|A\|_2\|f_n - f\|_{C^1} + \tau\mu\|A_n - A\|_2 \\
&\quad + \|W_n^{\text{in}}\omega_n - W^{\text{in}}\omega\|_{C^1} \\
&< \tau\|A\|_2\|f_n - f\|_{C^1} + \frac{\epsilon(1 - \tau\|A\|_2)}{2} + \frac{\epsilon(1 - \tau\|A\|_2)}{2} \\
&\qquad \text{by Eqs. (4) and (5)} \\
&< \tau\|A\|_2\|f_n - f\|_{C^1} + \epsilon(1 - \tau\|A\|_2).
\end{aligned}$$

Hence, rearranging we see that $\|f_n - f\|_{C^1}(1 - \tau\|A\|_2) < \epsilon(1 - \tau\|A\|_2)$ which implies $\|f_n - f\|_{C^1} < \epsilon$ as required. □

To prove non-emptiness we construct an explicit reservoir matrix A and input matrix W^{in} for which the Echo State Map f is an embedding, using a trick borrowed from Shi and Han (2007).

First, for a given observation function ω we define Λ_ω to be the subset of matrices A and W^{in} for which the associated map f is a C^1 embedding:

$$\Lambda_\omega := \{(A, W^{\text{in}}) \mid f_{A,W^{\text{in}},\omega} \text{ is a } C^1 \text{ embedding.}\} \qquad (6)$$

Lemma 2.3.6. *Let M be a compact m-manifold and $n \in \mathbb{N}$. Let A be an $n \times n$ matrix and suppose $\|A\|_2 < \min(1/\|D\phi^{-1}\|_\infty, 1)$. Let W^{in} be an $n \times 1$ matrix and let the triple (φ, A, W^{in}) be an ESN. Suppose that $\phi \in \mathit{Diff}^2(M)$ has the following two properties:*

(1) ϕ has only finitely many periodic points with periods less than or equal to $2m$.

(2) If $x \in M$ is any periodic point with period $k < 2m$ then the eigenvalues of the derivative $D\phi^k$ at x are distinct.

Then for a generic $\omega \in C^2(M, \mathbb{R})$, Λ_ω is non-empty.

Proof. Let

$$A = \frac{1}{2}\begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ & & & & \\ 0 & 0 & 0 & \ddots & 0 \end{bmatrix}$$

and $W^{\text{in}}_1 = 1,\ W^{\text{in}}_j = 0$ for $2 \le j \le n$. Then the ESM

$$f := \begin{bmatrix} \varphi_1 \circ \omega \\ \varphi_2 \circ 2^{-1}\varphi_1 \circ \omega \circ \phi^{-1} \\ \varphi_3 \circ 2^{-1}\varphi_2 \circ 2^{-1}\varphi_1 \circ \omega \circ \phi^{-2} \\ \vdots \\ \varphi_n \circ 2^{1-n}\varphi_{n-1} \dots 2^{-1}\varphi_1 \circ \omega \circ \phi^{-n+1} \end{bmatrix},$$

where $\varphi_i(r_i) = \sigma(r_i + b_i)$ is the ith component function of φ, as defined in (1), solves the equation

$$f = \varphi(Af \circ \phi^{-1} + W^{\text{in}}\omega).$$

We can see moreover that $f \equiv g \circ \Phi_{(\phi,\omega)}$ where

$$g := \begin{bmatrix} \varphi_1 \\ \varphi_2 \circ 2^{-1}\varphi_1 \\ \varphi_3 \circ 2^{-2}\varphi_2 \circ 2^{-1}\varphi_1 \\ \vdots \\ \varphi_n \circ 2^{1-n}\varphi_{n-1} \dots 2^{-1}\varphi_1 \end{bmatrix}$$

and $\Phi_{(\phi,\omega)}$ is the delay observation map

$$\Phi_{(\phi,\omega)}(x) = (\omega(x), \omega \circ \phi^{-1}(x), \omega \circ \phi^{-2}(x), \dots, \omega \circ \phi^{-n+1}(x)).$$

By design, each φ_i is a C^1 embedding hence g is a C^1 embedding. For generic $\omega \in C^2(M, \mathbb{R})$ the delay observation map $\Phi_{(\phi,\omega)}$ is also a C^1 embedding, thanks to Takens' Theorem. Noting that the composition of C^1 embeddings is a C^1 embedding completes the proof. □

Theorem 2.3.7 (*Weak ESN Embedding Theorem*)**.** *Let M be a compact m-manifold and $n \ge 2m + 1$. Let A be a random variable with a distribution that has full support on the space of $n \times n$ matrices for which $\|A\|_2 < \min(1/\|D\phi^{-1}\|_\infty, 1)$, and let W^{in} be a random variable with a distribution that has full support on the space of $n \times 1$ matrices, and let the triple (φ, A, W^{in}) be an ESN. Suppose $\phi \in \mathit{Diff}^2(M)$ has the following two properties:*

(1) ϕ has only finitely many periodic points with periods less than or equal to $2m$.

(2) If $x \in M$ is any periodic point with period $k < 2m$ then the eigenvalues of the derivative $D\phi^k$ at x are distinct.

Then for a generic observation function $\omega \in C^2(M, \mathbb{R})$ the Echo State Map f is a C^1 embedding with probability $\alpha > 0$.

Proof. The space of all observation functions $\omega \in C^2(M, \mathbb{R})$ such that the delay observation map $\Phi_{(\phi,\omega)}$ is an embedding is generic in $C^2(M, \mathbb{R})$, thanks to Takens' Theorem. For any one of these observation functions, the set Λ_ω defined in (6) is non-empty by Lemma 2.3.6 and open by Lemma 2.3.5. Since A, and W^{in} are random variables with full support, they take values in Λ with probability $\alpha > 0$. □

Remark 2.3.8. The Embedding Conjecture and Weak ESN Embedding Theorem state that under the right conditions f is an embedding. In practical examples we cannot compute f exactly because it is obtained in the limit of infinitely many past observations. In practice, if we have k observations the best we can do is to use all available observations and compute $f_k^{r_0}$. Fortunately, the set of C^1 embeddings is open in the C^1 topology, and $f_k^{r_0}$ converges to f in this topology, so there exists a sufficiently large number ℓ of previous observations such that for all $k > \ell$, $f_k^{r_0}$ is an embedding.

The ESN Embedding Conjecture also admits a biological interpretation. Consider an organism with a (primitive) nervous system ('brain') comprised of neurons. Neurons are connected to each other with random connection weights (including zero) representing the strength of the connection (or no connection). The adjacency matrix forms the reservoir matrix A. The reservoir state r is a vector representing the firing rate of every neuron. Suppose that the organism has a sensory organ connected to the brain which at any point in time senses a scalar measure of the environment, for example an average environmental light intensity. The connection weight from the sensory organ to the ith neuron is then the ith entry of W^{in}. Suppose that the light intensity depends on the state of the environment which evolves as a high dimensional dynamical system. Then the nervous system and sensory organ together operate as an ESN. Since the entries of A and W^{in} are random variables, the ESN Embedding Conjecture states that the dynamics of the environment are indeed embedded into the nervous system without the nervous system needing to possess any special structure provided by learning or natural selection. The embedding of the natural world into the brain is obtained 'for free'. This leaves cognition, defined as 'the art of performing computation on our representation of the environment', as the faculty that requires optimisation by natural selection or learning.

2.4. The ESN approximation theorem

In this section we will state and prove the ESN Approximation Theorem — that an ESN which successfully embeds a dynamical system into the reservoir space can approximate the system's dynamics during the autonomous phase, hence replicate the topology of a structurally stable dynamical system. We will use several preliminary results introduced over the proceeding subsections.

2.4.1. The universal approximation theorem

The first major result we will use to prove the ESN Approximation Theorem is the Universal Approximation Theorem. This theorem is highly celebrated in the literature on mathematical analysis of neural networks, and states that smooth functions and any number of their derivatives can be approximated by single layer neural network with sufficiently many neurons. In this section we recall this theorem and then present an extension suitable for ESNs, to take account of the fact that for an ESN the neural network weights v_i and biases b_i are randomly chosen but then fixed; only the output weights w_i can be chosen to give a good approximation to an input function f. We will use the Universal Approximation Theorem presented by Hornik, Stinchcombe, and White (1990), because it concerns smooth functions *and any number of their derivatives* while the earlier seminal paper by Cybenko (1989) does not.

Definition 2.4.1 (ℓ*-finite*)**.** Let $\ell \in \mathbb{N}_0$. Then we say an ℓ-times differentiable scalar function $\sigma \in C^\ell(\mathbb{R})$ is ℓ-finite if

$$0 < \int_{\mathbb{R}} \left| \frac{d^\ell \sigma}{dx^\ell} \right| dx < \infty.$$

Remark 2.4.2. The activation function $\sigma \in C^1(\mathbb{R}, (-1, 1))$ with derivative in the range $(0, 1)$ is 1-finite; meaning ℓ-finite with $\ell = 1$.

Theorem 2.4.3 (*Universal Approximation Theorem*)**.** *If the activation function σ is ℓ-finite, then for all $0 \leq m \leq \ell$ functions $g : I_n \to \mathbb{R}$ of the form*

$$g(x) = \sum_{j=1}^{N} w_j \sigma(v_j^\top x + b_j)$$

are dense in $C^m(I_n, \mathbb{R})$.

Proof. Hornik et al. (1990). □

The Universal Approximation Theorem essentially states that if we are interested in approximating a function f to some tolerance ϵ we can create a neural network of size N and modify the weights until the network approximates f to the tolerance ϵ. We want to slightly extend the theorem for our purposes. Recall that an ESN has random reservoir weights comprising the matrix A and random input weights comprising the matrix W^{in}, and it is only the output connection weights W^{out} that are trained. We therefore want to show that for any continuously differentiable function f and a sufficiently large neural network with random weights v_i and biases b_i, we can choose linear readout weights w_i such that the resulting neural network approximates f arbitrarily well with probability arbitrarily close to 1. We will call this the Random Universal Approximation Theorem (RUAT), and remark that the RUAT is highly related to Theorem 2.1 appearing in the seminal paper on Extreme Learning Machines by Huang, Zhu, and Siew (2006). We can also view the RUAT as a special case of Theorem 2.1 presented by Gonon et al. (2020), who prove a stronger result in the more general context of filters.

The idea behind the proof of the RUAT is as follows. First we note that there is a neural network $\hat{g}$ that approximates f by the Universal Approximation Theorem. Then, we create sample sequences of weights and biases v_i, b_i by repeated draws from appropriate random variables. There will eventually be some randomly generated samples v_j, b_j that are close to each of the weights and biases of the network $\hat{g}$. From this list of weights and biases in the sample sequences we select those that match closely, and so create a neural network g, choosing linear readout weights w_i either to match the respective weight in $\hat{g}$ or choosing to set $w_i = 0$ in order effectively to discard those samples v_i, b_i that not close to values in $\hat{g}$. Now by construction g is a good approximation to $\hat{g}$ which is itself a good approximation to f. The details are presented in the following lemma and theorem.

Lemma 2.4.4. *Let $(X_j)_{j \in \mathbb{N}}$ be a sequence of i.i.d. random variables and $S_1, \dots, S_\ell$ be a list of ℓ events, and suppose that for each i (and for any j since they are i.i.d.) there exists θ_i such that $\mathbb{P}(X_j \in S_i) = \theta_i > 0$. Then for all $\alpha \in (0, 1)$ there exists $N \in \mathbb{N}$ such that*

$$\mathbb{P}\big(\exists \text{ injective } \phi : \{1, \dots, \ell\} \to \{1, \dots, N\} : X_{\phi(i)} \in S_i, \\ \forall\, i \in \{1, \dots, \ell\}\big) > \alpha.$$

Proof. First, fix $\alpha \in (0, 1)$. Then define the set $\{n_0, \dots, n_\ell\}$ as follows. Set $n_0 = 0$ and for any $i \in \{1, \dots, \ell\}$ let

$$n_i - n_{i-1} := \text{ceil}\left(\frac{\log(1 - \alpha^{1/\ell})}{\log(1 - \theta_i)}\right) + 1.$$

Finally, set $N = n_\ell$. Then we can calculate that

$$\begin{aligned}
&\mathbb{P}\big(\exists \text{ injective } \phi : X_{\phi(i)} \in S_i \;\forall i \in \{1, \dots, \ell\}\big) \\
&\quad > \mathbb{P}\big(\forall\, i \in \{1, \dots, \ell\}\; \exists\, j \in \{1 + n_{i-1}, \dots, n_i\} : X_j \in S_i\big) \\
&\quad = \prod_{i=1}^{\ell} \mathbb{P}\big(\exists j \in \{1 + n_{i-1}, \dots, n_i\} : X_j \in S_i\big) \\
&\quad = \prod_{i=1}^{\ell} 1 - \mathbb{P}\big(X_j \notin S_i \;\forall j \in \{1 + n_{i-1}, \dots, n_i\}\big) \\
&\quad \geq \prod_{i=1}^{\ell} 1 - (1 - \theta_i)^{n_i - n_{i-1}} \\
&\quad = \prod_{i=1}^{\ell} 1 - (1 - \theta_i)^{\text{ceil}\left(\log(1-\alpha^{1/\ell})/\log(1-\theta_i)\right)+1} \\
&\quad > \prod_{i=1}^{\ell} 1 - (1 - \theta_i)^{\left(\log(1-\alpha^{1/\ell})/\log(1-\theta_i)\right)} \\
&\quad = \prod_{i=1}^{\ell} 1 - \exp\left(\frac{\log(1 - \alpha^{1/\ell})}{\log(1 - \theta_i)} \log(1 - \theta_i)\right) \\
&\quad = \prod_{i=1}^{\ell} 1 - (1 - \alpha^{1/\ell}) = \prod_{i=1}^{\ell} \alpha^{1/\ell} = \alpha. \quad \square
\end{aligned}$$

Theorem 2.4.5 (*Random Universal Approximation Theorem*)**.** *Let I_n denote the unit hypercube of dimension n and let $f \in C^1(I_n, \mathbb{R})$. Let $\sigma \in C^1(\mathbb{R})$ be 1-finite, and let $(b_j)_{j \in \mathbb{N}}$, $(v_j)_{j \in \mathbb{N}}$ be sequences of i.i.d. random variables with full support. Then for any $\alpha \in (0, 1)$ and $\epsilon > 0$ there exists some natural number $N \in \mathbb{N}$ such with, probability greater than α, there exist real numbers $w_1, \dots, w_N \in \mathbb{R}$ such that the* random neural network *$g : I_n \to \mathbb{R}$ defined by*

$$g(x) = \sum_{j=1}^{N} w_j \sigma(v_j^\top x + b_j)$$

satisfies

$$\|f - g\|_{C^1} < \epsilon.$$

Proof. First, by the Universal Approximation Theorem we know that for any $\epsilon > 0$ there exists a neural network $\hat{g} : I_n \to \mathbb{R}$ of size ℓ defined by

$$\hat{g}(x) = \sum_{i=1}^{\ell} \hat{w}_i \sigma(\hat{v}_i^\top x + \hat{b}_i)$$

such that

$$\|f - \hat{g}\|_{C^1} < \frac{\epsilon}{2}. \tag{7}$$

Now, consider two sequences of i.i.d. random variables $(b_j)_{j \in \mathbb{N}}$ and $(v_j)_{j \in \mathbb{N}}$ with full support, and let $X_j := (b_j, v_j)$. Fix $\epsilon > 0$ and define a collection of ℓ events $S_1, \dots, S_\ell$ by

$$S_i := \Bigg\{ (b, v) \in \mathbb{R} \times \mathbb{R}^n : \|\sigma(\hat{v}_i^\top \cdot + \hat{b}_i) - \sigma(v^\top \cdot + b)\|_{C^1} \\ < \frac{\epsilon}{2\ell \max_k(\hat{w}_k)} \Bigg\},$$

where the weights $\hat{w}_k$ are given by the form of the network $\hat{g}$. Observe that each of the S_i have strictly positive measure, so there exists $\theta_i > 0$ such that $\mathbb{P}(X_j \in S_i) > \theta_i > 0 \;\forall j \in \mathbb{N}$. Hence it follows by Lemma 2.4.4 that for all $\alpha \in (0, 1)$ there exists $N \in \mathbb{N}$ such that

$$\mathbb{P}\big(\exists \text{ injective } \phi : \{1, \dots, \ell\} \to \{1, \dots, N\} : X_{\phi(i)} \in S_i \\ \forall i \in \{1, \dots, \ell\}\big) > \alpha.$$

Now, on the event

$$\exists \text{ injective } \phi : \{1, \dots, \ell\} \to \{1, \dots, N\} : X_{\phi(i)} \in S_i \;\forall i \in \{1, \dots, \ell\}$$

we define

$$w_j := \begin{cases} \hat{w}_i \text{ if } \phi(i) = j \\ 0 \text{ otherwise} \end{cases}$$

for all $j \in \{1, \ldots, N\}$, and define the *random neural network* $g : I_n \to \mathbb{R}$ by

$$g(x) = \sum_{j=1}^{N} w_j \sigma(v_j^\top x + b_j).$$

Now observe

$$\begin{aligned} \|\hat{g} - g\|_{C^1} &= \left\| \sum_{i=1}^{\ell} \hat{w}_i \sigma(\hat{v}_i^\top \cdot + \hat{b}_i) - \sum_{j=1}^{N} w_j \sigma(v_j^\top \cdot + b_j) \right\|_{C^1} \\ &= \left\| \sum_{i=1}^{\ell} \hat{w}_i \big(\sigma(\hat{v}_i^\top \cdot + \hat{b}_i) - \sigma\big(v_{\phi(i)}^\top \cdot + b_{\phi(i)}\big)\big) \right\|_{C^1} \\ &\leq \sum_{i=1}^{\ell} \hat{w}_i \left\| \big(\sigma(\hat{v}_i^\top \cdot + \hat{b}_i) - \sigma\big(v_{\phi(i)}^\top \cdot + b_{\phi(i)}\big)\big) \right\|_{C^1} \\ &< \sum_{i=1}^{\ell} \frac{\hat{w}_i \epsilon}{2\ell \max_k(\hat{w}_k)} < \frac{\epsilon}{2}. \end{aligned}$$

Combining this with (7) and using the triangle inequality we obtain

$$\|f - g\|_{C^1} \leq \|f - \hat{g}\|_{C^1} + \|\hat{g} - g\|_{C^1} < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,$$

which completes the proof. □

2.4.2. The ESN approximation theorem

In this subsection we will state and prove the ESN Approximation Theorem which states that there exists a linear readout layer W^{out} giving rise to an autonomous ESN phase with a normally hyperbolic attracting m-submanifold on which the autonomous dynamics are topologically conjugate to a structurally stable ϕ. The idea behind the theorem is observe that the ESN looks enough like a single layer neural network that the Random Universal Approximation Theorem holds. Consequently we can choose linear readout weights stored in the matrix W^{out} to approximate any C^1 function. We will assume that f is an embedding, and therefore invertible on its image, and choose readout weights W^{out} such that the autonomous ESN approximates a C^1 dynamical system possessing an m dimensional normally hyperbolic attracting submanifold on which the dynamics approximate $f \circ \phi \circ f^{-1}$. We want the manifold to be normally hyperbolic and attracting to ensure that an autonomous trajectory that leaves the manifold by some small distance is attracted back towards the manifold, preventing an accumulation of errors from sending the trajectory too far away. Autonomous trajectories originating near the manifold therefore remain near, all the while approximating $f \circ \phi \circ f^{-1}$. To formalise these ideas, we will first define a normally hyperbolic attracting submanifold.

Definition 2.4.6 (*Normally Hyperbolic Attracting Submanifold*). Let $\phi \in \text{Diff}^1(M)$, then, a ϕ-invariant submanifold $\Lambda \subset M$ is a normally hyperbolic attracting submanifold if the restriction to Λ of the tangent bundle of M admits a splitting into a direct sum of two $D\phi$-invariant subbundles, the tangent bundle of Λ, and the stable bundle E^s. Furthermore, with respect to some Riemannian metric on M, the restriction of $D\phi$ to E^s must be a contraction, and must be relatively neutral on $T\Lambda$. Thus, there exist constants $0 < \lambda < \mu^{-1} < 1$ and $c > 0$ such that

$$T_\Lambda M = T\Lambda \oplus E^s$$
$$(D\phi)_x E_x^s = E_{\phi(x)}^s \;\; \forall x \in \Lambda$$
$$\|D\phi^k v\| \leq c\lambda^k \|v\| \;\; \forall v \in E^s, \;\; \forall k \in \mathbb{N}$$
$$\|D\phi^k v\| \leq c\mu^{|k|} \|v\|.$$

Before we present the ESN Approximation Theorem itself we will prove that there exists a C^1 evolution operator η defined on $\mathbb{R}^d$ that has a normally hyperbolic attracting submanifold on which the dynamics of η are conjugate to ϕ. The existence of this map η is guaranteed by standard topological machinery which we recall briefly here, and which is presented in detail by Warner (1971).

Definition 2.4.7 (*Cubic Centred Chart*). A chart (V, φ) belonging to a d-manifold is called a cubic chart if $\varphi(V)$ is an open cube centred about the origin in $\mathbb{R}^d$. If $x \in V$ and $\varphi(x) = 0$, then the chart (V, φ) is centred at x.

Definition 2.4.8 (*Slice Coordinates*). Suppose that (V, φ) is a chart on a d-manifold D with coordinate functions $x_1, \ldots, x_d$ and that m is an integer $0 \leq m \leq d$. Let $a \in \varphi(V)$ and let

$$S = \{q \in V \mid x_i(q) = a_i, i = m + 1, \ldots, d\}.$$

The subspace S of D together with coordinate maps $x|_S$ for $j = 1, \ldots, m$ forms a submanifold of D, called a slice of the chart (V, φ).

Lemma 2.4.9 (*Slice Lemma*). *Let M be a compact m-manifold, let $f : M \to \mathbb{R}^d$ be an immersion, and let $x \in M$. Then there exists a cubic centred chart (V, φ) about $f(x)$ and a neighbourhood U of x such that $f|_U$ is injective and $f(U)$ is a slice of (V, φ).*

Proof. Warner (1971) page 28 prop 1.35. □

Lemma 2.4.10. *Let $d > m$ and M be a compact m-manifold. Let $\phi \in \text{Diff}^1(M)$. Suppose $f \in C^1(M, \mathbb{R}^d)$ is a C^1 embedding. Then there is an open subset $\Omega \in \mathbb{R}^d$ and $\eta \in \text{Diff}^1(\Omega)$ with $f(M)$ a normally hyperbolic attracting submanifold such that $\eta|_{f(M)} = f \circ \phi \circ f^{-1}$ (where we have defined f^{-1} on the image of f).*

Proof. We will make a similar argument to Warner (1971) in the proof of his Proposition 1.36, on page 29. First let $x \in M$. Then by the Slice Lemma there exists a cubic centred chart (V_x, φ_x) about $f(x)$ and a neighbourhood U_x of x such that $f(U_x)$ is a slice (V_x, φ_x). Let $x_1, \ldots, x_m$ be the slice coordinates in the chart (V_x, φ_x) of points in $f(U_x)$. Then we can define a map $\eta_x \in \text{Diff}^1(V_x, \mathbb{R}^d)$ applying the map $f \circ \phi \circ f^{-1}$ on the slice co-ordinates and dividing the remaining co-ordinates by 2. We can make this argument for every $x \in M$ hence define a collection of maps $\{\eta_x\}$ over a collection of open sets $\{V_x\}$ which cover $f(M)$. Now we let $\{\alpha_j \mid j \in \mathbb{N}\}$ form a partition of unity subordinate to the cover $\{V_x\}$. We take a subsequence $\{\alpha_k\}$ such that $\text{supp}(\alpha_k) \cap f(M) \neq \emptyset$ and denote the collection of sets to which $\{\alpha_k\}$ is subordinate by $\{V_k\}$. We then define a map η on a neighbourhood $\Omega := \cup_k V_k$ of $f(M)$ by

$$\eta = \sum_k \alpha_k \eta_x.$$

By construction, $\eta|_{f(M)} = f \circ \phi \circ f^{-1}$ and η has a normally hyperbolic attracting submanifold $f(M)$. □

Not only does the dynamical system η exist, but importantly, its normally hyperbolic attracting submanifold is preserved by any sufficiently good approximation. This is made formal in the Invariant Manifold Theorem, which we will use in the proof of the ESN Approximation Theorem.

Theorem 2.4.11 (*Invariant Manifold Theorem*)**.** *Let K be a compact manifold and $\eta \in Diff^1(K)$ with normally hyperbolic attracting submanifold Λ. Then, $\exists \epsilon > 0$ such that for any $u \in Diff^1(K)$ with $\|\eta - u\|_{C^1} < \epsilon$, the diffeomorphism u has a normally hyperbolic attracting submanifold U such that $\|U - \Lambda\|_{C^1} < \epsilon$.*

Proof. Hirsch, Pugh, and Shub (1977). □

With these preliminaries established we are ready to prove our ESN Approximation Theorem. Our strategy involves imposing a special structure on the reservoir matrix A in order to obtain sufficiently many neurons for the Random Universal Approximation Theorem to hold while controlling the dimension of the codomain of the Echo State Map. The structure of A is made clear in the statement of the ESN Approximation Theorem and illustrated in Fig. 2, where we call the connections represented by the matrix A 'strongly recurrent' and those represented by X 'weakly recurrent'. The weakly recurrent neurons and the vector Y of inputs are introduced in the proof of the ESN Approximation Theorem in order to satisfy the conditions of the Random Universal Approximation Theorem.

Definition 2.4.12 (*ESN Autonomous Phase*)**.** The ESN autonomous phase with parameters $(A, W^{\text{in}}, W^{\text{out}}, \varphi)$ is a discrete time autonomous dynamical system $\psi \in C^1(\mathbb{R}^n)$ defined by

$$\psi(s) = \varphi\big((A + W^{\text{in}}W^{\text{out}})s\big).$$

Theorem 2.4.13 (*ESN Approximation Theorem*)**.** *Let M be a compact m-manifold and $n \in \mathbb{N}$ such that $n > 2m$. Let A be an $n \times n$ matrix where $\|A\|_2 < min(1/\|D\phi^{-1}\|_\infty, 1)$, and W^{in} an $n \times 1$ matrix, and let the triple (φ, A, W^{in}) be an ESN. Let $\phi \in Diff^1(M)$ be structurally stable, and let $\omega \in C^1(M, \mathbb{R})$. Suppose the Echo State Map $f \in C^1(M, \mathbb{R}^n)$ is a C^1 embedding. Let $(x_j)_{j\in\mathbb{N}}$, $(y_j)_{j\in\mathbb{N}}$, and $(b_j)_{j\in\mathbb{N}}$ be sequences of i.i.d. $\mathbb{R}^n$, $\mathbb{R}$, and $\mathbb{R}$-valued random variables, respectively, with full support. Let $\alpha \in (0, 1)$. Then, with probability α, there exists $d \in \mathbb{N}$ with $d > n$, a $d \times 1$ matrix W^{out}, a $d \times d$ matrix $\tilde{A}$, and a $d \times 1$ matrix $\tilde{W}^{in}$ assembled from the $n \times n$ matrix A, the $(d - n) \times n$ matrix X with jth row x_j, and the $(d - n) \times 1$ matrix Y with jth row y_j, like so:*

$$\tilde{A} = \begin{bmatrix} A & 0 \\ X & 0 \end{bmatrix} \qquad \text{and} \qquad \tilde{W}^{in} = \begin{bmatrix} W^{in} \\ Y \end{bmatrix},$$

and an activation function

$$\tilde{\varphi}_i(r) = \sigma(r_i + b_i) \quad \forall\, i \in \{1, \dots, d\}$$

such that the autonomous ESN $\psi \in C^1(\mathbb{R}^d)$ with parameters $(\tilde{A}, \tilde{W}^{in}, W^{out}, \tilde{\varphi})$ has a normally hyperbolic attracting submanifold on which ψ is topologically conjugate to ϕ.

Proof. By assumption, the Echo State Map f defined for the ESN $(\varphi, A, W^{\text{in}})$ with respect to (ϕ, ω) is an embedding, so the Echo State Map $\tilde{f}$ defined for $(\tilde{\varphi}, \tilde{A}, \tilde{W}^{\text{in}})$ with respect to (ϕ, ω) is also an embedding. For the remainder of the proof we will restrict the codomain of $\tilde{f}$ to its image in order to yield a C^1 diffeomorphism. Before we proceed, we will establish some preliminary results. First we define $y : M \to y(M) \subset \mathbb{R}^{n+1}$ by

$$y_1(x) = \omega(x) \qquad \text{and} \qquad \begin{bmatrix} y_2(x) \\ y_3(x) \\ \vdots \\ y_{n+1}(x) \end{bmatrix} = f \circ \phi^{-1}(x).$$

Furthermore we will define maps

$$\mathcal{F} : C^1(M, \mathbb{R}^d) \to C^1(\tilde{f}(M), \mathbb{R}^d) \qquad \text{by} \qquad \mathcal{F}(g) = g \circ \tilde{f}^{-1}$$

and

$$\mathcal{Y} : C^1(y(M), \mathbb{R}) \to C^1(M, \mathbb{R}) \qquad \text{by} \qquad \mathcal{Y}(g) = g \circ y.$$

Next we will show that $\mathcal{F}$ and $\mathcal{Y}$ are Lipschitz continuous. To see that $\mathcal{F}$ is Lipschitz continuous observe

$$\begin{aligned} \|\mathcal{F}(g) - \mathcal{F}(h)\|_{C^1} &= \|g \circ \tilde{f}^{-1} - h \circ \tilde{f}^{-1}\|_{C^1} \\ &= \|g \circ \tilde{f}^{-1} - h \circ \tilde{f}^{-1}\|_\infty + \|Dg \circ \tilde{f}^{-1}D\tilde{f}^{-1} \\ &\quad -Dh \circ \tilde{f}^{-1}D\tilde{f}^{-1}\|_\infty \\ &\leq \|g \circ \tilde{f}^{-1} - h \circ \tilde{f}^{-1}\|_\infty + \|D\tilde{f}^{-1}\|_\infty\|Dg \circ \tilde{f}^{-1} \\ &\quad -Dh \circ \tilde{f}^{-1}\|_\infty \\ &= \|g - h\|_\infty + \|D\tilde{f}^{-1}\|_\infty\|Dg - Dh\|_\infty \\ &\leq \max(1, \|D\tilde{f}^{-1}\|_\infty)(\|g - h\|_\infty + \|Dg - Dh\|_\infty) \\ &= \max(1, \|D\tilde{f}^{-1}\|_\infty)\|g - h\|_{C^1}. \end{aligned}$$

We can make an almost identical argument to show that $\mathcal{Y}$ is Lipschitz continuous. We will denote the Lipschitz constants for $\mathcal{F}$ and $\mathcal{Y}$ by L and M respectively. We are now ready to proceed with the proof.

By Lemma 2.4.10, there exists an open subset $\Omega \in \mathbb{R}^d$ containing $\tilde{f}(M)$ and $\eta \in \text{Diff}^1(\Omega)$ with $\tilde{f}(M)$ a normally hyperbolic attracting submanifold such that

$$\eta|_{\tilde{f}(M)} = \tilde{f} \circ \phi \circ \tilde{f}^{-1}.$$

Now let $K \subset \Omega$ be a compact manifold containing $\tilde{f}(M)$. Normally hyperbolic invariant submanifolds persist under small perturbations, by the Invariant Manifold Theorem, so $\exists\, \epsilon > 0$ such that any $u \in \text{Diff}^1(K)$ which satisfies $\|u - \eta|_K\|_{C^1} < \epsilon$ is topologically conjugate to η. For any given value $\alpha \in (0, 1)$, by the Random Universal Approximation Theorem, there exists a $d \in \mathbb{N}$ and a $d \times 1$ matrix W^{out} such that $g \in C^1(\mathbb{R}^{n+1}, \mathbb{R})$ defined by

$$g(z) = \sum_{i=1}^{d} W_i^{\text{out}}\sigma\left(\begin{bmatrix}\tilde{W}^{\text{in}} & \tilde{A}\end{bmatrix}_i z + b_i\right) \tag{8}$$

satisfies

$$\|g - \omega \circ \phi \circ y^{-1}\|_{C^1} < \frac{\epsilon}{LM\|W^{\text{in}}\|} \tag{9}$$

where $[\tilde{W}^{\text{in}}, \tilde{A}]_i$ is a $1 \times (n + 1)$ matrix with 1st entry $\tilde{W}_i^{\text{in}}$ and $(j + 1)$th entry $\tilde{A}_{ij}$. Now

$$\begin{aligned} &\|\psi|_{\tilde{f}(M)} - \eta|_{\tilde{f}(M)}\|_{C^1} \\ &\quad \leq L\|\psi \circ \tilde{f} - \eta \circ \tilde{f}\|_{C^1} \\ &\quad = L\|\psi \circ \tilde{f} - \tilde{f} \circ \phi\|_{C^1} \text{ because } \eta|_{\tilde{f}(M)} = \tilde{f} \circ \phi \circ \tilde{f}^{-1} \\ &\quad = L\|\tilde{\varphi}(\tilde{A}\tilde{f} + \tilde{W}^{\text{in}}W^{\text{out}}\tilde{f}) - \tilde{\varphi}(\tilde{A}\tilde{f} + \tilde{W}^{\text{in}}\omega \circ \phi)\|_{C^1} \\ &\qquad \text{by definition of } \psi \\ &\quad \leq L\|(\tilde{A}\tilde{f} + \tilde{W}^{\text{in}}W^{\text{out}}\tilde{f}) - (\tilde{A}\tilde{f} + \tilde{W}^{\text{in}}\omega \circ \phi)\|_{C^1} \\ &\qquad \text{because } \tilde{\varphi} \text{ is contracting} \\ &\quad = L\|(\tilde{W}^{\text{in}}W^{\text{out}}\tilde{f}) - (\tilde{W}^{\text{in}}\omega \circ \phi)\|_{C^1} \text{ because } \tilde{A}\tilde{f} - \tilde{A}\tilde{f} = 0 \\ &\quad \leq L\|\tilde{W}^{\text{in}}\|_2\|(W^{\text{out}}\tilde{f} - \omega \circ \phi)\|_{C^1} \text{ by factoring out } W^{\text{in}} \\ &\quad = L\|\tilde{W}^{\text{in}}\|_2\|W^{\text{out}}\tilde{\varphi}(\tilde{A}\tilde{f} \circ \phi^{-1} + \tilde{W}^{\text{in}}\omega) - \omega \circ \phi\|_{C^1} \\ &\qquad \text{by Theorem 2.2.2} \\ &\quad = L\|\tilde{W}^{\text{in}}\|_2\left\|\sum_{i=1}^{d} W_i^{\text{out}}\sigma\left(\begin{bmatrix}\tilde{W}^{\text{in}} & \tilde{A}\end{bmatrix}_i y + b_i\right) - \omega \circ \phi\right\|_{C^1} \\ &\qquad \text{by definition of } \tilde{\varphi} \text{ and } y \\ &\quad = L\|\tilde{W}^{\text{in}}\|_2\|g \circ y - \omega \circ \phi\|_{C^1} \text{ by (8)} \\ &\quad \leq LM\|\tilde{W}^{\text{in}}\|_2\|g|_{y(M)} - \omega \circ \phi \circ y^{-1}\|_{C^1} \end{aligned}$$

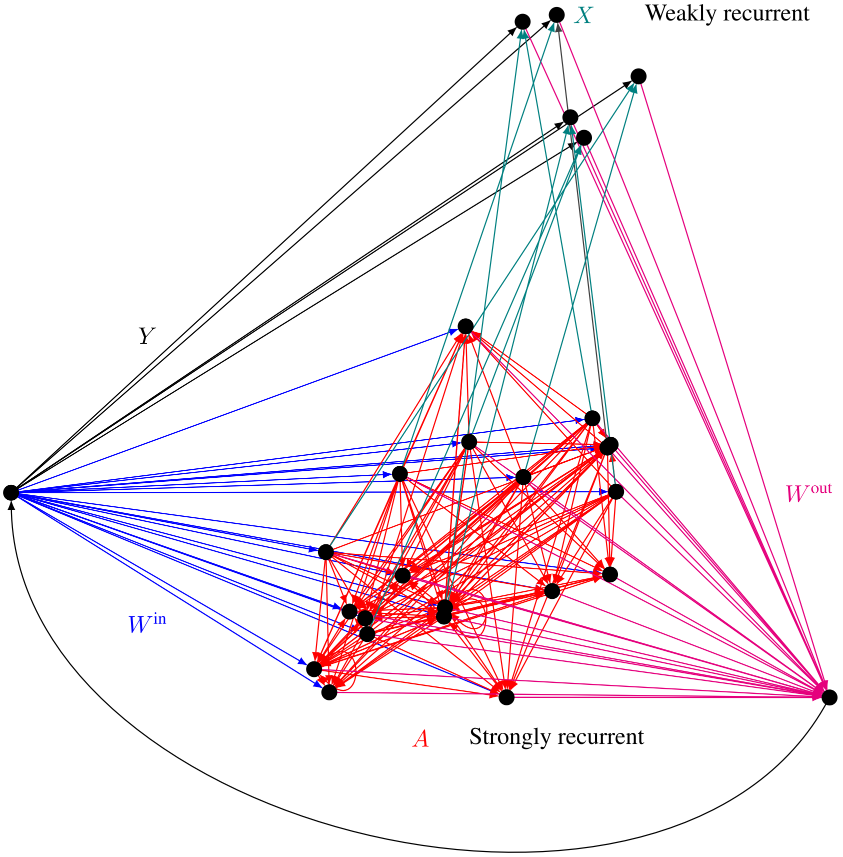


Fig. 2. The ESN with sparsity structure imposed on A so that we can prove the ESN Approximation Theorem. The matrix X and vector Y are defined in the statement of the ESN Approximation Theorem.

$$\begin{tikzcd}
M \arrow[r, "\tilde{f}"] \arrow[d, "\phi"] & \tilde{f}(M) \arrow[r, "h"] \arrow[d, "\eta"] & h \circ \tilde{f}(M) \arrow[d, "\psi"] \\
M \arrow[r, "\tilde{f}"] & \tilde{f}(M) \arrow[r, "h"] & h \circ \tilde{f}(M)
\end{tikzcd}$$

Fig. 3. A commuting diagram representing the ESN Approximation Theorem where the terms are defined throughout the theorem's proof.

$$
\begin{aligned}
&< LM\|\tilde{W}^{\text{in}}\|_2 \frac{\epsilon}{LM\|\tilde{W}^{\text{in}}\|_2} \quad \text{by (9)} \\
&= \epsilon
\end{aligned}
$$

hence there is some open set $\tilde{\Omega} \subset K$ containing $\tilde{f}(M)$ such that

$$\|\psi|_{\tilde{\Omega}} - \eta|_{\tilde{\Omega}}\|_{C^1} < \epsilon$$

so $\psi|_{\tilde{\Omega}}$ is conjugate to $\eta|_{\tilde{\Omega}}$. Consequently, there exists an $h \in \text{Diff}^1(\tilde{\Omega})$ such that $\psi|_{\tilde{\Omega}} = h \circ \eta|_{\tilde{\Omega}} \circ h^{-1}$. Now $\tilde{f}(M)$ is a normally hyperbolic attracting submanifold of η where $\eta|_{\tilde{f}(M)} = \tilde{f} \circ \phi \circ \tilde{f}^{-1}$ so $h \circ \tilde{f}(M)$ is a normally hyperbolic attracting submanifold of ψ on which

$$\psi = h \circ \eta \circ h^{-1} = h \circ \tilde{f} \circ \phi \circ \tilde{f}^{-1} \circ h^{-1} \cong \phi. \quad \square$$

Remark 2.4.14. A consequence of the ESN Approximation Theorem is that the diagram shown in Fig. 3 commutes.

3. Numerical experiments with ESNs

In the previous section we showed that for a given structurally stable dynamical system and a sufficiently large ESN there exists a linear output matrix W^{out} that gives rise to an autonomous ESN with dynamics that are topologically conjugate to those of the given dynamical system.

To test whether these results hold in practice we took a 1D observation of a numerically integrated trajectory of the Lorenz system, fed this into an ESN implemented on a commercial laptop, and sought to discover whether the autonomous phase of the ESN would adopt dynamics topologically conjugate to the Lorenz system. In particular we computed several topological invariants of the ESN autonomous phase including the Lyapunov exponents, fixed point eigenvalues, and homology, then compared these to the known invariants of the Lorenz system. This work was inspired by a paper by Pathak et al. (2017) who trained an ESN on a full 3D trajectory of the Lorenz system, rather than a 1D observation, and compared the Lyapunov exponents of the autonomous phase to the known exponents of the Lorenz system. In a more recent work, Vlachas et al. (2019) train an ESN on 1D observations of the Lorenz-96 system, and also compare the Lyapunov exponents of the autonomous phase to the known exponents of the Lorenz system. Chattopadhyay, Hassanzadeh, Palem, and Subramanian (2019) also train an ESN on observations of the Lorenz-96 system and evaluate the accuracy of future prediction for reservoirs of different size.

We used MATLAB's ODE45 to integrate a trajectory of the Lorenz (1963) system

$$
\begin{aligned}
\dot{x} &= \sigma(y - x) \\
\dot{y} &= x(\rho - z) - y \\
\dot{z} &= xy - \beta z
\end{aligned}
\tag{10}
$$

with parameters $\sigma = 10, \beta = 8/3, \rho = 28$ chosen so the system produces the celebrated Lorenz attractor shown in Fig. 4.

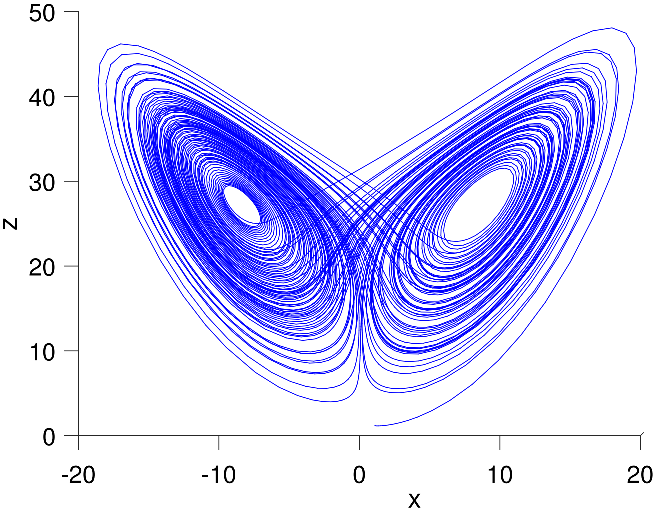


Fig. 4. A picture of the famous Lorenz attractor. Here the trajectory was initialised at (1, 1, 1) and quickly converges to the attractor.

We then observed the x component of the trajectory by choosing the observation function $\omega(x, y, z) = x$ to create a 1 dimensional time series. We fed this time series into an ESN with the following parameters: spectral radius $\rho = 1$, reservoir size $n = 300$, and activation function $\varphi = \tanh$. The reservoir matrix A is an Erdős-Rényi matrix with mean 6 and connection weights (where they are non-zero) i.i.d Gaussian, re-scaled such that $\rho = 1$. The keen reader will notice that the structure of A does not conform to the reservoir matrix $\tilde{A}$ described in the statement of the ESN Approximation Theorem. The fact that our numerical experiments produce good results despite this suggests this weakly connected $\tilde{A}$ is unnecessary, but rather a decision we made to make the ESN Approximation Theorem easier to prove. Furthermore, insisting that $\rho < 1$ is not sufficient in to ensure that $\|A\|_2 < 1$, but this is a common choice in practical applications. The matrix W^{out} is populated with i.i.d Gaussian weights $\sim \mathcal{N}(0, 1)$ which are then scaled by a 'strength parameter' $p = 0.1$. We choose a regularisation parameter $\lambda = 10^{-6}$ to solve the regularised least squares problem

$$\min_{W^{\text{out}}} \sum_{k=1}^{K} \|W^{\text{out}} r_k - u_k\|^2 + \lambda \|W^{\text{out}}\|_2^2$$

using the SVD method presented by Hansen, Nagy, and O'leary (2006). We will note here that the linear output layer W^{out} obtained by this procedure is not necessarily the same as that guaranteed by the ESN Approximation Theorem. These parameters were carefully hand tuned so that the autonomous phase appeared (by eye) to match the driven phase. The question of how to systematically choose good parameters is discussed by Yperman and Becker (2016) who searched through parameter space using Bayesian optimisation, and used cross validation to test the goodness of fit. Now, with W^{out} obtained, we ran the autonomous ESN and plotted the future observations v_i in Fig. 5. We can see from this Figure that the ESN seems to predict the qualitative features of the future trajectory very well.

Since the Lorenz system is defined on a 3-manifold, we can usefully plot trajectories of the entire system. To check by eye whether the reservoir dynamics of both the driven phase and autonomous phase are topologically conjugate to the Lorenz dynamics, we projected the driven and autonomous dynamics onto the first 3 principal components of the driven trajectory and present them in Fig. 6.

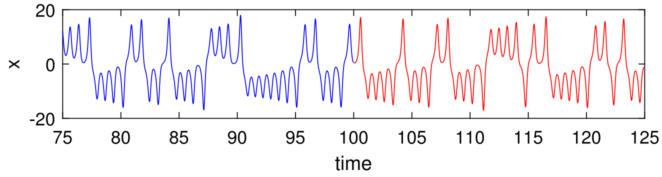


Fig. 5. Here the 1D observations are shown in blue (up to time 100 for those of you reading in black and white) and future predictions shown in red (onwards from time 100). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

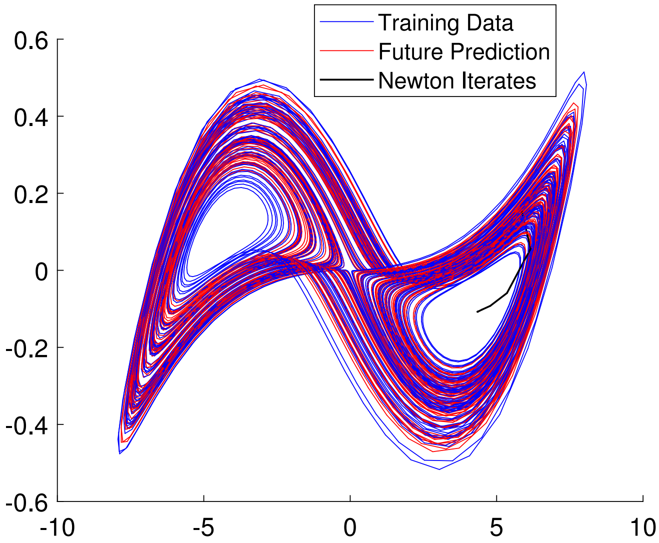


Fig. 6. The driven reservoir dynamics are plotted in blue and autonomous dynamics are plotted in red. Both were projected onto the first three principal components of the driven dynamics, then the axes are rotated such that the projection appears on the first 2 components. The black line indicates the iterates of Newton's method, used to locate a fixed point — the method eventually converges to a fixed point in the middle of the right wing of the figure. We can see by eye that the reservoir dynamics appear by eye to be topologically conjugate to the Lorenz system. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.1. *Locating fixed points and determining their eigenvalues*

If the ESM f is an embedding, then f will embed the fixed points of the Lorenz system into the reservoir space. Moreover if the autonomous ESN approximates the embedded Lorenz system on a neighbourhood of the embedded fixed points sufficiently well, the autonomous dynamics will contain fixed points very close to those of the embedded Lorenz system. To verify this, we searched for the autonomous ESN's fixed points using Newton's method, and found them, as illustrated in Fig. 6.

Further, if the ESM f is a C^1 embedding of the original dynamics, we expect f to preserve the stability of fixed points, i.e. we expect the eigenvalues of the linearisation of the autonomous phase to be preserved at every fixed point. Now, comparing the eigenvalues of the linearisation of the Lorenz system and autonomous phase at the respective fixed points requires some subtlety, because the Lorenz system is a continuous time flow, while the autonomous phase is a discrete time map. So, we began by considering one of the known fixed points found in the Lorenz attractor's wings

$$x^* = (\sqrt{\beta(\rho-1)}, \sqrt{\beta(\rho-1)}, \rho - 1),$$

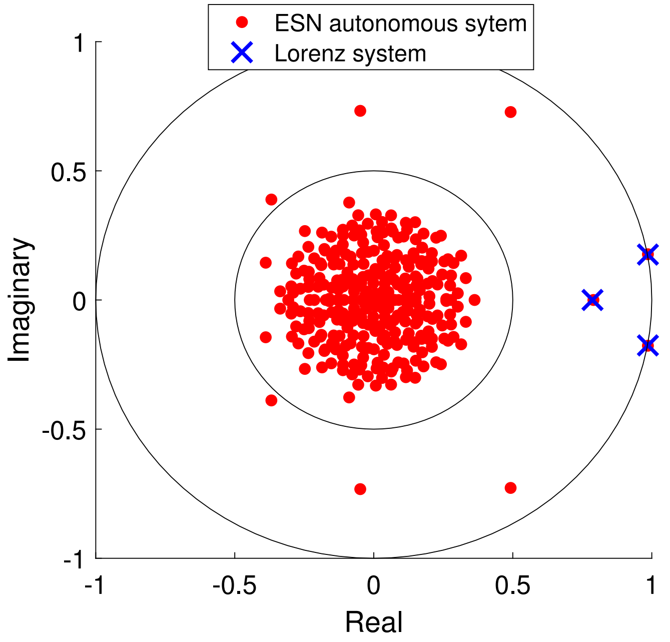


Fig. 7. Here the 3 eigenvalues of the linearisation of the Lorenz system on the fixed point inside one of the Lorenz attractor's wings are represented by blue crosses. The 300 eigenvalues of the linearisation of the ESN autonomous system at the fixed point found with Newton's method are represented by red dots.

and noted the Jacobian J of the continuous time Lorenz system evaluated at the fixed point x^* is therefore

$$J\Big|_{x^*} = \begin{bmatrix} -\sigma & \sigma & 0 \\ 1 & -1 & -\sqrt{\beta(\rho-1)} \\ \sqrt{\beta(\rho-1)} & \sqrt{\beta(\rho-1)} & -\beta \end{bmatrix}.$$

Now we can discretise the Lorenz system $\dot{x} = s(x)$ with the following map

$$x_{k+1} = x_k + \int_{t_k}^{t_{k+1}} s \circ x(t)dt,$$

hence the discrete time linearisation about the fixed point x^* is

$$x_{k+1} = \exp\left(J\Big|_{x^*}(t_{k+1} - t_k)\right)x_k,$$

which has 3 eigenvalues, which we have compared with the ESN autonomous eigenvalues in Fig. 7. If the ESM f is indeed a C^1 embedding, the dynamics of the autonomous phase are topologically conjugate to the discrete time Lorenz system on some 3-submanifold. This manifold is spanned by 3 eigenvectors, each with an associated eigenvalue, which will coincide with the eigenvalues of the linearisation of the Lorenz system on the fixed point. Fig. 7 appears to show 3 overlapping eigenvalues, suggesting that the autonomous phase is diffeomorphic to the Lorenz system (at least in a neighbourhood of x^*) in this simulation. This is particularly remarkable because x^* is distant from the training data. The ESN has successfully inferred the existence, position and eigenvalues of a fixed point from training data, which contains no fixed points. In the machine learning parlance, the ESN has generalised patterns in the training data to an unseen region of the phase space.

3.2. Comparison of Lyapunov spectra

Another topological invariant of the Lorenz system is the Lyapunov spectrum, which captures how quickly very close trajectories diverge from each other, and is used as a measure of chaos.

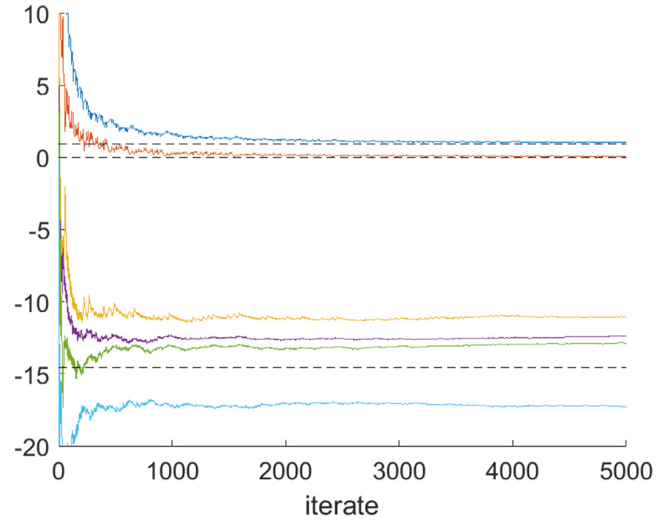


Fig. 8. The Lyapunov spectrum of the autonomous phase as the iterates increase is shown. The true Lyapunov exponents of the autonomous phase are given by the limit of these exponents as the iterations tend to infinity. These autonomous exponents are compared to the black dotted lines representing the 3 exponents of the Lorenz system.

To define the spectrum, let J be the Jacobian of the evolution operator of a continuous time dynamical system. Let Y be the solution of the ODE $\dot{Y} = JY$ with initial condition $Y(0) = x_0$. Then the Lyapunov Spectrum of the invariant set containing x_0 is the spectrum of the matrix Λ defined

$$\Lambda = \lim_{t\to\infty} \frac{1}{2t} YY^\top.$$

Each eigenvalue in the spectrum is called a *Lyapunov exponent* to signify that two initially close trajectories diverge or converge exponentially fast with exponentiation constant in the direction of each eigenvector of J given by a Lyapunov exponent. Details are discussed by Darbyshire and Broomhead (1996). The Lyapunov spectrum for the Lorenz system was estimated by Sprott (2003) as 0.9056, 0, -14.5723. In order to compare the Lorenz spectrum to the spectrum of the autonomous ESN, we computed the autonomous system's spectrum using the discrete time QR method discussed in Darbyshire and Broomhead (1996) and plotted each Lyapunov exponent against the known exponents of the Lorenz system in Fig. 8. We found the largest 2 in good agreement while there was significant error in the smallest, which is a common problem also encountered by Pathak et al. (2017).

3.3. Persistent homology

We compared the homology groups of the Lorenz attractor to the persistent homology groups of the autonomous and driven attractors. We followed the lead of Garland, Bradley, and Meiss (2016) who computed the persistent homology of the Lorenz system reconstructed from a sequence of 1D observations of a Lorenz trajectory using the delay observation map described in Takens' Theorem. The authors used the open source software Javaplex created by Tausz, Vejdemo-Johansson, and Adams (2014) to find the Witness Complex for the delay embedded Lorenz attractor and computed the homology of the complex. They discuss a few subtleties that arise, in particular that the Lorenz attractor is a fractal, whose structure cannot be reconstructed exactly from any finite number of sample points. The authors therefore satisfied themselves by approximating the Lorenz attractor with a branched manifold model presented by Williams (1979) which

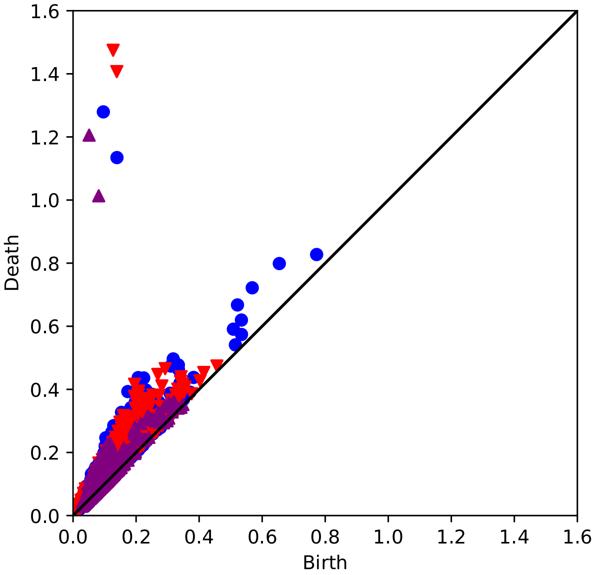


Fig. 9. We have plotted the H_1 persistence diagrams of the driven ESN dynamics, autonomous ESN dynamics, and Lorenz dynamics as blue circles, red downward triangles, and purple upward triangles. We can see that each of these 3 objects has a pair of points floating well above the diagonal, suggesting each has 2 holes. This is consistent with our expectation that all three adopt the topology of Fig. 8.

has the homology of Fig. 8. We made the same approximation, and expected to find that the application of persistent homology to the Lorenz system, driven ESN dynamics, and autonomous ESN dynamics would reveal that all three have Fig. 8 homology groups. In particular the persistence diagrams of these three systems would exhibit a pair of H_1 persistent homology groups floating well above the diagonal. To verify this, we produced persistence diagrams using the open source software Ripser produced by Tralie, Saul, and Bar-On (2018) and plotted the results in Fig. 9.

The reader may wonder why we would use persistent homology to show that the Lorenz system, driven ESN dynamics, and autonomous ESN dynamics all have the homology of Fig. 8 when this can clearly be seen in Figs. 4 and 6. The homology of a 3D system is usually apparent from a plot, but persistent homology can reveal the holes, voids and higher dimensional hypervoids of high dimensional systems that cannot be easily visualised. For example Muldoon, MacKay, Huke, and Broomhead (1993) computed the homology of a delay embedded time series from a fluid dynamics experiment, which could in general be of much higher dimension.

4. Conclusions and outlook

In this paper, we showed that an Echo State Network driven by a sequence of one dimensional observations of a dynamical system, evolving on a manifold M, induces a map $f \in C^1(M, \mathbb{R}^n)$, which we called the Echo State Map. We proved that for a randomly initialled ESN and generic observation function ω, that f is an embedding with positive probability, and called this the weak ESN Embedding Theorem. We conjectured that the theorem holds with probability 1, by analogy to Takens' Theorem. We went on to show that a randomly initialised ESN has a universal approximation property and called this the Random Universal Approximation Theorem (RUAT). Finally, we used both the RUAT and Embedding Theorem to prove that for an ESN trained a sequence of scalar observations of a structurally stable dynamical system, there is a choice of linear readout weights W^{out} for which the autonomous ESN has dynamics that are topologically conjugate to the input dynamical system, and we called this the ESN Approximation Theorem.

The theory presented here leaves some questions unanswered. In practice we use regularised least squares regression to learn an output matrix from the one-dimensional and finite training trajectory, but currently, we have no guarantee that this will result in an autonomous phase ESN that is topologically conjugate to the underlying dynamical system. This is analogous to the case of the Universal Approximation Theorem for feed forwards neural networks, where the theoretical result proves the existence of suitable set of weights but does not guarantee that a particular learning algorithm will be able to find them or how much training data may be required. It may be that imposing extra conditions on the target dynamical system, like ergodicity, allows us to prove that W^{out} obtained by least squares regression results in an arbitrarily good approximation. This seems to be supported by the experiments in Section 3.

Furthermore, it seems worthwhile to prove the ESN Embedding Conjecture, or some modification of it that is actually correct, by carefully modifying the proof of Takens' Theorem provided by Huke (2006). A sceptical reader may wonder why we would bother using an ESN to embed the trajectory in the first place, when a delay embedding would do. The reason being that it seems the ESN's learning and predictive powers are much more resilient to noise than the simple delay embedding presented by Takens. Heuristically it seems as an observed trajectory passes through the ESN, the noise cancels itself out by taking a nonlinear combination of positive and negative noise. We could therefore view the ESN as a nonlinear filter, generalising the linear filters discussed by Sauer, Yorke, and Casdagli (1991) in the context of *embedology* - the art building delay observation maps with special features, which include being more resultant to noise than Takens' original map. Understanding the noise cancelling benefits of the ESN could be a fruitful direction of future work.

Many of the assumptions we made throughout this paper are likely stronger than they need to be. For example Sauer et al. (1991) prove versions of Takens' Theorem for dynamics on a compact invariant set with real box counting dimension — generalising dynamics on a manifold with integer dimension. This is particularly worthwhile because chaotic attractors of interest often lie on invariant sets with non-integer dimension, with the Lorenz attractor serving as a perfect example. We also create a strangely shaped reservoir $\tilde{A}$ in our proof of the ESN Approximation Theorem, which numerical experiments suggest is unnecessary.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We offer gracious thanks to the anonymous reviewers for their detailed suggestions. In particular, we thank Juan-Pablo Ortega and Lyudmila Grigoryeva for thorough and fruitful discussions at Universität Sankt Gallen which lead to significant improvements to the manuscript.

We also thank Allen Hart's PhD confirmation examiners Alastair Spence and Chris Guiver who helpfully criticised and improved the content of this paper.

We acknowledge that Allen Hart is supported by a scholarship from the ESPRC Centre for Doctoral Training in Statistical Applied Mathematics at Bath (SAMBa), under the project EP/L015684/1.

References

Banach, S. (1922). Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fundamenta Mathematicae*, *22*, 133–181.

Chattopadhyay, A., Hassanzadeh, P., Palem, K., & Subramanian, D. (2019). Data-driven prediction of a multi-scale lorenz 96 chaotic system using a hierarchy of deep learning methods: Reservoir computing, ann, and rnn-lstm. arXiv:1906.08829.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, *2*(4), 303–314.

Darbyshire, A., & Broomhead, D. (1996). Robust estimation of tangent maps and liapunov spectra. *Physica D: Nonlinear Phenomena*, *89*(3), 287–305.

Garland, J., Bradley, E., & Meiss, J. D. (2016). Exploring the topology of dynamical reconstructions. *Physica D: Nonlinear Phenomena*, *334*, 49–59, Topology in Dynamics, Differential Equations, and Data..

Ghrist, R. (2008). Barcodes: The persistent topology of data. *American Mathematical Society. Bulletin*, *45*, 61–75.

Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep sparse rectifier neural networks. In G. Gordon, D. Dunson, & M. Dudík (Eds.), *Proceedings of machine learning research*: *vol. 15*, *Proceedings of the fourteenth international conference on artificial intelligence and statistics* (pp. 315–323). Fort Lauderdale, FL, USA: PMLR.

Gonon, L., Grigoryeva, L., & Ortega, J.-P. (2020). Approximation bounds for random neural networks and reservoir systems. arXiv:2002.05933.

Grigoryeva, L., & Ortega, J. (2018). Echo state networks are universal. *Neural Networks*, *108*, 495–508.

Grigoryeva, L., & Ortega, J.-P. (2019). Differentiable reservoir computing. *Journal of Machine Learning Research (JMLR)*, *20*(179), 1–62, URL http://jmlr.org/papers/v20/19-150.html.

Gürel, T., & Egert, S. R. U. (2010). Functional identification of biological neural networks using reservoir adaptation for point processes. *Journal of Computational Neuroscience*, *27*, 9–299.

Hackl, J. (2018). Tikz-network manual. arXiv:1709.06005.

Hansen, P. C., Nagy, J. G., & O'leary, D. (2006). *Deblurring images: Matrices, spectra, and filtering*. SIAM.

Hirsch, M. W., Pugh, C. C., & Shub, M. (1977). *Invariant manifolds*. Springer-Verlag.

Hornik, K., Stinchcombe, M., & White, H. (1990). Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Networks*, *3*(5), 551–560.

Huang, G.-B., Zhu, Q.-Y., & Siew, C.-K. (2006). Extreme learning machine: Theory and applications. *Neurocomputing*, *70*(1), 489–501, Neural Networks.

Huke, J. (2006). Embedding nonlinear dynamical systems: A guide to Takens' theorem.

Ilies, I., Jaeger, H., Kosuchinas, O., Rincon, M., Sakenas, V., & Vaskevicius, N. (2007). Stepping forward through echoes of the past: forecasting with echo state networks.

Jaeger, H. (2001). The echo state approach to analysing and training recurrent neural networks.

Jaeger, H., & Haas, H. (2004). Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science*, *304*(5667), 78–80.

Kocarev, L., & Parlitz, U. (1996). Generalized synchronization, predictability, and equivalence of unidirectionally coupled dynamical systems. *Physical Review Letters*, *76*, 1816–1819.

Lin, X., Yang, Z., & Song, Y. (2009). Short-term stock price prediction based on echo state networks. *Expert Systems with Applications*, *36*(3, Part 2), 7313–7317.

Løkse, S., Bianchi, F. M., & Jenssen, R. (2017). Training echo state networks with regularization through dimensionality reduction. *Cognitive Computation*, *9*(3), 364–378.

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, *20*(2), 130–141.

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

Muldoon, M., MacKay, R., Huke, J., & Broomhead, D. (1993). Topology from time series. *Physica D: Nonlinear Phenomena*, *65*(1), 1–16.

Pathak, J., Lu, Z., Hunt, B. R., Girvan, M., & Ott, E. (2017). Using machine learning to replicate chaotic attractors and calculate lyapunov exponents from data. *Chaos*, (27).

Plöger, P. G., Arghir, A., Günther, T., & Hosseiny, R. (2004). Echo state networks for mobile robot modeling and control. In D. Polani, B. Browning, A. Bonarini, & K. Yoshida (Eds.), *RoboCup 2003: Robot soccer world cup VII* (pp. 157–168). Berlin, Heidelberg: Springer Berlin Heidelberg.

Sauer, T., Yorke, J. A., & Casdagli, M. (1991). Embedology. *Journal of Statistical Physics*, *65*(3), 579–616.

Schrauwen, B., Verstraeten, D., & Van Campenhout, J. (2007). An overview of reservoir computing: theory, applications and implementations. In *Proceedings of the 15th European symposium on artificial neural networks. 471-482 2007* (pp 471–482).

Shi, Z., & Han, M. (2007). Support vector echo-state machine for chaotic time-series prediction. *IEEE Transactions on Neural Networks*, *18*(2), 359–372.

Skowronski, M. D., & Harris, J. G. (2007). Automatic speech recognition using a predictive echo state network classifier. *Neural Networks*, *20*(3), 414–423, Echo State Networks and Liquid State Machines.

Sprott, J. C. (2003). *Chaos and time-series analysis*. Oxford University Press.

Takens, F. (1981). Detecting strange attractors in turbulence. In *Lecture notes in mathematics*: *vol. 898*, (p. 366). Berlin: Springer Verlag.

Tanaka, G., Yamane, T., Héroux, J. B., Nakane, R., Kanazawa, N., Takeda, S., et al. (2019). Recent advances in physical reservoir computing: A review. *Neural Networks*, *115*, 100–123.

Tausz, A., Vejdemo-Johansson, M., & Adams, H. (2014). JavaPlex: A research software package for persistent (co)homology. In H. Hong, & C. Yap (Eds.), *Lecture notes in computer science*: *vol. 8592*, *Proceedings of ICMS 2014* (pp. 129–136).

Tong, M. H., Bickett, A. D., Christiansen, E. M., & Cottrell, G. W. (2007). Learning grammatical structure with echo state networks. *Neural Networks*, *20*(3), 424–432, Echo State Networks and Liquid State Machines.

Tralie, C., Saul, N., & Bar-On, R. (2018). Ripser.py: A lean persistent homology library for python. *The Journal of Open Source Software*, *3*(29), 925.

Vlachas, P. R., Pathak, J., Hunt, B. R., Sapsis, T. P., Girvan, M., Ott, E., et al. (2019). Forecasting of spatio-temporal chaotic dynamics with recurrent neural networks: a comparative study of reservoir computing and backpropagation algorithms. arXiv:1910.05266.

Warner, F. W. (1971). *Foundations of differentiable manifolds and lie groups, scott*. Foresman and Co.

Whitney, H. (1944). The self-intersections of a smooth n-manifold in 2n-space. *Annals of Mathematics*, *45*(2), 220–246.

Williams, R. F. (1979). The structure of lorenz attractors. *Publications Mathématiques de l'IHÉS*, *50*, 73–99.

Xi, J., Shi, Z., & Han, M. (2005). Analyzing the state space property of echo state networks for chaotic system prediction. In *Proceedings. 2005 IEEE international joint conference on neural networks, 2005.', Vol. 3* (pp. 1412–1417).

Yeo, K. (2019). Data-driven reconstruction of nonlinear dynamics from sparse observation. *Journal of Computational Physics*, *395*, 671–689.

Yong, Song, Yibin, Li, Qun, Wang, & Caihong, Li (2010). Multi-steps prediction of chaotic time series based on echo state network. In *2010 IEEE fifth international conference on bio-inspired computing: Theories and applications (BIC-TA)* (pp. 669–672).

Yperman, J., & Becker, T. (2016). Bayesian optimization of hyper-parameters in reservoir computing. arXiv:1611.05193.