Physica D 421 (2021) 132882

Contents lists available at ScienceDirect

Physica D

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

Echo State Networks trained by Tikhonov least squares are $L^2(\mu)$ approximators of ergodic dynamical systems

Allen G. Hart *, James L. Hook, Jonathan H.P. Dawes

University of Bath, UK

ARTICLE INFO

Article history:
Received 10 May 2020
Received in revised form 18 February 2021
Accepted 19 February 2021
Available online 2 March 2021
Communicated by T. Sauer

Keywords:
Reservoir computing
Liquid state machine
Time series analysis
Lorenz equations
Delay embedding
Recurrent neural networks

ABSTRACT

Echo State Networks (ESNs) are a class of single-layer recurrent neural networks with randomly generated internal weights, and a single layer of tuneable outer weights, which are usually trained by regularised linear least squares regression. Remarkably, ESNs still enjoy the universal approximation property despite the training procedure being entirely linear. In this paper, we prove that an ESN trained on a sequence of observations from an ergodic dynamical system (with invariant measure μ) using Tikhonov least squares regression against a set of targets, will approximate the target function in the $L^2(\mu)$ norm. In the special case that the targets are future observations, the ESN is learning the next step map, which allows time series forecasting. We demonstrate the theory numerically by training an ESN using Tikhonov least squares on a sequence of scalar observations of the Lorenz system.

1. Introduction

Echo state networks (ESNs) are a class of single layer recurrent neural networks introduced at the turn on the millennium independently by Jaeger [1] and Maass et al. [2]. These relatively simple neural networks have been used to solve a range of machine problems where the input data is a time series, including speech recognition [3], learning the rules of grammar [4], financial time series prediction [5,6], short term traffic forecasting [7], placing UAV base stations [8] and learning about the behaviour of seals [7]. ESNs are also a plausible model for the information processing of biological neurons [9]. In this paper, we will present just enough definitions and theory to make sense of our results, but encourage the interested reader to read the recent review paper by Tanaka et al. [10] who cover recent developments and open questions in the field of *reservoir computing*, a field of which ESN comprise a subset. The ESN is defined by the recursion relation

$$x_{k+1} = \sigma(Ax_k + Cz_k + b)$$

where the x_k are T dimensional state vectors, $\sigma : \mathbb{R}^T \to \mathbb{R}^T$ is the activation function, A is the $T \times T$ *reservoir matrix*, representing the connection weights between neurons, C is the $T \times d$ *input matrix* connecting the d-dimensional inputs z_k to the reservoir matrix A, and $b \in \mathbb{R}^T$ is a bias vector. The reservoir matrix A, input matrix C and bias vector b are initialised randomly and remain unchanged. The ESN can be trained to approximate a sequence of target scalars u_k by solving the regularised linear squares problem

$$\min_W \sum_{k=0}^{\ell-1} \|W^\top x_k - u_k\|^2 + \lambda \|W\|^2$$

where $\lambda > 0$ is the Tikhonov regularisation parameter. If the target scalars u_k are equal to the observations z_k, then the ESN is being trained to predict the future. To see this, we can set up a sequence of scalars v_k defined by the recurrent relation

$$\begin{aligned} v_{k+1} &= W^\top s_k \\ s_{k+1} &= \sigma(As_k + Cv_{k+1} + b) \end{aligned} \tag{1}$$

and we then hope that $v_k \approx u_k$ for sufficiently many future values of k. We can view s_k as the state of a discrete time autonomous dynamical system which we will call the ESN autonomous phase. In this paper, we will suppose z_k are a sequence of sequential observations from an ergodic dynamical system, with invariant measure μ. We will go on to prove that ESNs trained by least squares can approximate arbitrary target functions (including one that returns future observations) of the ergodic dynamical system in the $L^2(\mu)$ norm. This theorem is closely related to recent work by Verzelli et al. [11] discussing the connection between ergodic dynamical systems and feasible learning. The result also explains the remarkable success of ESNs trained on dynamical systems explored numerically by, for example, Jaeger [1], Xi et al.

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

https://doi.org/10.1016/j.physd.2021.132882
0167-2789/

[12], Schrauwen et al. [13], Shi and Han [14], Song et al. [15], Yildiz et al. [16], Pathak et al. [17], Løkse et al. [18], Yeo [19], Chattopadhyay et al. [20], Vlachas et al. [21] and Hart et al. [22].

The remainder of the paper is organised as follows. In Section 2 we define an ergodic dynamical system and present Birkhoff's ergodic theorem. Next, in Section 3, we introduce the major result of this paper (Theorem 3.6), stating that an ESN trained on a sequence of observations from an ergodic dynamical system using Tikhonov least squares will $L^2(\mu)$ approximate an arbitrary target function. This arbitrary target function could be the *next step map* used for forecasting the future of the time series. Furthermore, we discuss the *central limit theorem for ergodic dynamical systems* in connection with the number of data points that are required for a good approximation.

In Section 4 we present the work of Luzzatto et al. [23] culminating in a proof that the Lorenz attractor is stably mixing, hence ergodic — explaining the success of so many authors using an ESN to forecast the trajectory of the Lorenz system.

In Section 5 we numerically simulate a trajectory of the Lorenz system. We observed the x-component of the system (which we called ξ to avoid notational clash) and assigned the z components (which we denote ζ) as targets. We explored how the approximation of the targets ζ_k given the observations ξ_k improved as the number of data points (ξ_k, ζ_k) grew. Finally, in Section 6 we summarise the paper and discuss ideas for future work.

2. Preliminaries on ergodic theory

We require that the underlying dynamical system is ergodic so that minimising the mean square differences between observations and targets does not create a bias towards areas with lots of training data. The ergodicity ensures that training data generated from a trajectory initialised at almost any point $m_0 \in M$ will represent all dynamics on M. To make this formal, we will introduce the definition of ergodicity and the celebrated ergodic theorem.

Definition 2.1 (*Generic Point*)**.** Suppose $\phi : M \to M$ is a measure preserving map with respect to the measure space (M, Σ, μ). Then $m_0 \in M$ is called a generic point if the orbit of m_0 is uniformly distributed over M according to the measure μ.

Proposition 2.2. *Suppose $\phi : M \to M$ is a measure preserving map with respect to the probability space (M, Σ, μ) and $s \in L^1(\mu)$. Suppose m_0 is a generic point in M then*

$$\lim_{\ell\to\infty} \frac{1}{\ell} \sum_{k=0}^{\ell-1} s \circ \phi^k(m_0) = \int_M s \, d\mu.$$

Definition 2.3 (*Ergodic*)**.** Let $\phi : M \to M$ be a measure preserving transformation on the probability space (M, Σ, μ). Then ϕ is ergodic if for every $\sigma \in \Sigma$ with $\phi^{-1}(\sigma) = \sigma$ either $\mu(\sigma) = 0$ or $\mu(\sigma) = 1$.

Theorem 2.4 (*Ergodic Theorem [24]*)**.** *Suppose $\phi : M \to M$ is ergodic with respect to the probability space (M, Σ, μ) and $s \in L^1(\mu)$. Then μ-almost all $m_0 \in M$ are generic hence for μ-almost all $m_0 \in M$*

$$\lim_{\ell\to\infty} \frac{1}{\ell} \sum_{k=0}^{\ell-1} s \circ \phi^k(m_0) = \int_M s \, d\mu. \tag{2}$$

The left hand side of (2) is called the *time average* taken from initial point $m_0 \in M$, and the right hand side called the *space average*. The ergodic theorem then states that the time average taken from almost all initial points equals the space average.

3. A training theorem for echo state networks

3.1. Preliminaries

Suppose we have an ergodic dynamical system $\phi : M \to M$, and we can observe the dynamics via an observation map $g : M \to \mathbb{R}^T$ and target map $u : M \to \mathbb{R}$. A trajectory originating from a generic point $m_0 \in M$ will ergodically explore the space M and yield a sequence of observations $g \circ \phi^k(m_0)$ and targets $u \circ \phi^k(m_0)$ for $k = 0, 1, 2, \ldots, \ell$.

Suppose we compute the vectors $W_\ell \in \mathbb{R}^T$ minimising the regularised least squares difference between the mapping of the observations $W^\top g \circ \phi^k(m_0)$ and the targets $u \circ \phi^k(m_0)$. We prove in the next lemma that as the number of data points ℓ grows large, the least squares solution W_ℓ minimises the ergodic average difference between the mapping on the observations $W^\top g \circ \phi^k(m_0)$ and the targets $u \circ \phi^k(m_0)$.

Lemma 3.1. *Let (M, Σ) be a measurable space, and suppose that $\phi : M \to M$ is ergodic with invariant measure μ. Let m_0 be a generic point in M. Let $g \in L^2(\mu)(M, \mathbb{R}^T)$ be an observation function and suppose that $u \in L^2(\mu)(M, \mathbb{R})$ is a target function we wish to approximate.*

Let $\lambda > 0$. Define the sequence $(W_\ell)_{\ell\in\mathbb{N}}$ such that, for each $\ell \in \mathbb{N}$, the vector $W_\ell \in \mathbb{R}^T$ is the unique minimiser of the regularised least squares difference

$$\frac{1}{\ell}\left(\sum_{k=0}^{\ell-1} \|W^\top g \circ \phi^k(m_0) - u \circ \phi^k(m_0)\|^2 + \lambda\|W\|^2\right).$$

Then, the sequence $(W_\ell)_{\ell\in\mathbb{N}}$ converges to

$$W_\infty = \left(\int_M g(m)g(m)^\top \, d\mu(m) + \lambda I\right)^{-1} \times \int_M u(m)g(m) \, d\mu(m)$$

which is the unique minimiser of

$$\|W^\top g - u\|^2_{L^2(\mu)} + \lambda\|W\|^2.$$

Proof. Consider the map $\Psi : \mathbb{R}^T \to \mathbb{R}$ defined

$$\Psi(W) = \|W^\top g - u\|^2_{L^2(\mu)} + \lambda\|W\|^2 = \int_M \|W^\top g(m) - u(m)\|^2 \, d\mu(m) + \lambda\|W\|^2.$$

The minimiser of Ψ satisfies $D\Psi = 0$ where D is the derivative operator, so we consider

$$\begin{aligned}
0 &= (D\Psi)(W) \\
&= D\left(\int_M \|W^\top g(m) - u(m)\|^2 \, d\mu(m) + \lambda\|W\|^2\right) \\
&= \int_M D\|W^\top g(m) - u(m)\|^2 \, d\mu(m) + \lambda D\|W\|^2 \\
&= \int_M 2(W^\top g(m) - u(m))g(m)^\top \, d\mu(m) + 2\lambda W^\top \\
&= \int_M (W^\top g(m) - u(m))g(m)^\top \, d\mu(m) + \lambda W^\top \\
&= W^\top \int_M g(m)g(m)^\top \, d\mu(m) - \int_M u(m)g(m)^\top \, d\mu(m) \\
&\quad + \lambda W^\top I \\
&= W^\top \left(\int_M g(m)g(m)^\top \, d\mu(m) + \lambda I\right) \\
&\quad - \int_M u(m)g(m)^\top \, d\mu(m),
\end{aligned}$$

which upon rearrangement yields

$$W = \left(\int_M g(m)g(m)^\top \, d\mu(m) + \lambda I\right)^{-1} \times \int_M u(m)g(m)\, d\mu(m).$$

Since this is the unique solution to $0 = D\Psi(W)$, this stationary point is unique, and we will denote it W_∞. We can see it is a minimum because the Hessian $H\Psi$ is positive definite. Next, define the map

$$\Phi : \{y \in C^1(\mathbb{R}^T, \mathbb{R}) \mid y \text{ has a unique minimum}\} \to \mathbb{R}^T$$

as the mapping on the C^1 functions with a unique minimum that returns their unique minimum. We can see that Φ is continuous with respect to the C^1 topology and standard topology on $\mathbb{R}$ respectively. We consider the family of functions $y_\ell \in \{y \in C^1(\mathbb{R}^T, \mathbb{R}) \mid y \text{ has a unique minimum}\}$

$$y_\ell(W) = \frac{1}{\ell}\left(\sum_{k=0}^{\ell-1} \|W^\top g \circ \phi^k(m_0) - u \circ \phi^k(m_0)\|^2 + \lambda\|W\|^2\right),$$

so that by definition $W_\ell = \Phi(y_\ell(W))$ and hence

$$\begin{aligned}\lim_{\ell\to\infty} W_\ell &= \lim_{\ell\to\infty} \Phi(y_\ell(W)) \\ &= \Phi\left(\lim_{\ell\to\infty} y_\ell(W)\right) \quad = \Phi\left(\|W^\top g - u\|^2_{L^2(\mu)} + \lambda\|W\|^2\right) \\ &= W_\infty.\end{aligned}$$

where we have used, respectively, continuity of Φ and the Ergodic Theorem. □

3.2. Echo state networks

An Echo State Network is a special case of a more general system called a state space system, or reservoir system. These are maps of the form $F : \mathbb{R}^N \times \mathbb{R}^d \to \mathbb{R}^N$, which admit an ESN as a special case when

$$F(x, z) = \sigma(Ax + Cz + b).$$

If a state space system is contracting in the state variable, i.e there exists a $c \in [0, 1)$ such that

$$\|F(x, z) - F(y, z)\| \leq c\|x - y\|,$$

and the inputs u_k are the observations of a dynamical system i.e $u_k = \omega \circ \phi^k(m_0)$ then there is a continuous map $f \in C^0(M, \mathbb{R}^N)$ synchronising the dynamics of ϕ on M to the dynamics of the reservoir states x_k. The map f is called a state synchronisation map (SSM) and is a generalised synchronisation in the sense described by Kocarev and Parlitz [25]. We can guarantee that an ESN is state contracting by bounding the 2-norm of the reservoir matrix $\|A\|_2 < 1$. An important existence result for SSMs is the following theorem, due to [26].

Theorem 3.2 ([26]). *Let M be a topological space, $\phi \in Hom(M)$ be a dynamical system, and $\omega \in C^0(M, \mathbb{R}^d)$ an observation function. Suppose that the state space system $F : \mathbb{R}^N \times \mathbb{R}^d \to \mathbb{R}^N$ is state contracting, i.e there exists a $c \in [0, 1)$ such that*

$$\|F(x, u) - F(y, u)\| \leq c\|x - y\|.$$

Then there exists a unique $f \in C^0(M, \mathbb{R}^N)$ called the state synchronisation map (SSM) such that, for any $m_0 \in M$ and $x_0 \in \mathbb{R}^N$ the sequence

$$x_{k+1} = F(x_k, \omega \circ \phi^k(m_0))$$

originating at x_0 converges to $f \circ \phi^k(m_0)$ as $k \to \infty$.

In order to approximate the arbitrary dynamics of ϕ via the observation function ω using state space systems, we require that the state space maps F possess some sort of universal approximation property. Thus, we will define a class of *linear universal approximators* with respect to an arbitrary complete norm $\|\cdot\|$. Every class of linear universal approximators contains maps, which after composition with another suitable map, forms a state map.

Definition 3.3. Let $\mathcal{F}$ be a sequence of maps $\{F_T\} : \mathbb{R}^N \times \mathbb{R}^d \to \mathbb{R}^T$. Let $C \subset \mathbb{R}^N$ and $K \subset \mathbb{R}^d$ be vectors and let $\Omega(C \times K, \mathbb{R})$ be a Banach space of real valued functions on $C \times K$, with norm denoted $\|\cdot\|_\Omega$. If, for any $g \in \Omega(C \times K, \mathbb{R})$ and any $\epsilon > 0$ there exists an $T_0 \in \mathbb{N}$ such that for any $T > T_0$ there exists a $W_* \in \mathbb{R}^T$ such that

$$\|W_*^\top F_T - g\|_\Omega < \epsilon$$

then we say that $\mathcal{F}$ is a class of *linear universal approximators* on $\Omega(C \times K, \mathbb{R}^N \times \mathbb{R}^d)$.

A widely used class of linear universal approximators is the class of Echo State Networks with randomly initialised internal weights, as shown by the following result.

Theorem 3.4. *Let $\mathcal{F}$ denote the sequence of maps $\{F_T\} : \mathbb{R}^N \times \mathbb{R}^d \to \mathbb{R}^T$ defined by*

$$F_T(x, z) = \sigma(Ax + Cz + b)$$

where

- *$\sigma \in C^1(\mathbb{R})$ is 1-finite (see [27] for the definition of ℓ-finite)*
- *A is a $T \times N$ random matrix, where $T > N$ and the first N rows of A form an $N \times N$ random submatrix with 2-norm less than 1 almost surely. The jth row of A (where $j > N$), denoted A_j, is a random variable with full support on $(\mathbb{R}^N)^\top$*
- *C is a $T \times d$ random matrix with jth row C_j, a random variable with full support on $(\mathbb{R}^d)^\top$*
- *b is a random T-vector with jth entry b_j, a random variable with full support on $\mathbb{R}$.*

Let $C \times K$ be an arbitrary compact subset of $\mathbb{R}^N \times \mathbb{R}^d$. Then, almost surely, $\mathcal{F}$ is a class of linear universal approximators on $L^2(C \times K, \mathbb{R})$.

Proof. Fix $g \in L^2(C \times K, \mathbb{R})$ and $\epsilon > 0$. Then for any $\alpha \in (0, 1)$, it follows from the Random Universal Approximation Theorem [22, Theorem 2.4.5.]) that there exists a $T_0 \in \mathbb{N}$ such that for any $T > T_0$, with probability at least α,

$$\|W^\top F_T - g\|_{L^2} < \epsilon,$$

hence $\mathcal{F}$ is a class of linear universal approximators. Since $\mathcal{F}$ is a class of linear universal approximators for any $\alpha \in (0, 1)$, $\mathcal{F}$ is almost surely a class of linear universal approximators. □

To construct such an ESN in practice, we create a reservoir system $F : \mathbb{R}^T \times \mathbb{R}^d \to \mathbb{R}^T$ by defining

$$F(x, z) = \sigma\left([A, 0]x + Cz + b\right)$$

where $[A, 0]$ is the $T \times T$ matrix where the first N columns form the matrix A and the remaining columns are 0. Suppose we truncate at N the state vectors $x \in \mathbb{R}^T$ by applying the canonical projection $\pi : \mathbb{R}^T \to \mathbb{R}^N$, and denote the truncation $\pi(x) = \bar{x} \in \mathbb{R}^N$. The dynamics of the truncated vectors $\bar{x}$ are given by the (state contracting) state space system $\pi \circ F_T : \mathbb{R}^N \times \mathbb{R}^d \to \mathbb{R}^N$, which is also an ESN as is defined by

$$\pi \circ F_T(\bar{x}, z) = \sigma(\bar{A}\bar{x} + \bar{C}z + \bar{b}).$$

Here, the $N \times N$ reservoir matrix $\bar{A}$ is created by truncating at N the rows and columns of A. The $N \times d$ input matrix $\bar{C}$ is

created by truncating at N the rows of C. The N-vector $\bar{b}$ is created by truncating at N the entries of b. We conclude that Echo State Networks with (appropriately chosen) randomly generated internal weights are a class of linear universal approximators that each give rise to a state synchronisation map.

We demanded that the $T \times T$ reservoir matrix take the form $[A, 0]$, whereas in practice, the reservoir matrix does not have this structure. We imposed this condition to simplify the proofs, but we believe, based on numerical evidence in the literature, that this choice of shape is not necessary.

There is one more technical lemma we will include here before presenting the main theorem (Theorem 3.6) of the paper. Recall that topological spaces have a natural Borel sigma algebra and are therefore measurable spaces. On such spaces we can integrate real valued functions. If A and B are homoeomorphic topological spaces, then integration on A is essentially the same as integration on B. We use this observation in Theorem 3.6 to move between integration on the topological space M to integration on the image $f(M)$. This demands the highly non-trivial assumption that the SSM f is a homeomorphism. The observation is made formal in the following lemma.

Lemma 3.5 (*Change of Variables*)**.** *Let A, B be homoeomorphic topological spaces and suppose $y \in Hom(A, B)$. The topologies on A, B induce Borel Sigma algebras $\mathscr{A}, \mathscr{B}$ on A, B respectively. Let μ_A be a measure on A and μ_B a measure on B (called the pushforward measure) defined $\mu_B(b) = \mu_A(y^{-1}(b))$ for all $b \in \mathscr{B}$. Then for any μ_B measurable function $g : B \to \mathbb{R}$*

$$\int_A g \circ y \, d\mu_A = \int_B g \, d\mu_B.$$

Proof. This is a special case of Theorem 3.6.1 in [28]. □

3.3. A training theorem for ESNs

Before we finally plunge into the statement and proof of the main theorem, we will describe the result in words. Suppose we have an ergodic dynamical system $\phi : M \to M$, which we observe via the function $\omega : M \to \mathbb{R}^d$ and that our goal is to approximate a target function $u : M \to \mathbb{R}$. Suppose we have at our disposal a class $\mathcal{F}$ of linear universal approximating state maps. For example, $\mathcal{F}$ could be a collection of arbitrarily high dimensional ESNs. Make the additional (and non trivial) assumption that the state maps give rise to an SSM that is homoeomorphic onto its image. Suppose then that the state map F is driven with observations of a trajectory $z_k = \omega \circ \phi^k(m_0)$ originating from a generic point m_0. This creates a sequence of reservoir states x_k that satisfy

$$x_{k+1} = F(x_k, z_k).$$

We also assemble a sequence of scalar targets $u \circ \phi^k(m_0)$.

Suppose we use regularised least squares regression to minimise the difference between the linear mapping on the observations $W^\top x_k$ and the targets $u \circ \phi^k(m_0)$. Then we can conclude that the ergodic average difference between the mapping on the data and the target map u can be made smaller than the arbitrary threshold ϵ. This requires that the trajectory length ℓ and state map dimension T are sufficiently large, while ensuring the regularisation parameter $\lambda > 0$ is sufficiently small.

We remark that a notable weakness of Theorem 3.6 is its non-constructive nature, because the actual values for ℓ, T and λ are not computed in terms of ϵ.

Theorem 3.6. *Let M be a topological space, and suppose that $\phi \in Hom(M)$ is ergodic with invariant measure μ. Let m_0 be a generic point in M. Let $\omega \in C^0(M, \mathbb{R}^d)$ be the observation function and suppose that $u \in L^2(\mu)(M, \mathbb{R})$ is a target function we wish to approximate.*

Suppose that $\mathcal{F}$ is a class of linear universal approximators on $L^2(C \times K, \mathbb{R})$ on every compact $C \subset \mathbb{R}^N$, $K \subset \mathbb{R}^d$. Let $(s_T)_{T\in\mathbb{N}} : \mathbb{R}^T \to \mathbb{R}^N$ be a sequence of maps. Suppose (for each large enough T) the state map $s_T \circ F_T : \mathbb{R}^N \times \mathbb{R}^d \to \mathbb{R}^N$ admits an SSM $f \in Hom(M, f(M))$. For each $T, \ell \in \mathbb{N}$, and $\lambda > 0$ let $W_\ell \in \mathbb{R}^T$ be the vector obtained by minimising the regularised least squares difference

$$\sum_{k=0}^{\ell} \|W^\top F_T(f \circ \phi^{k-1}(m_0), \omega \circ \phi^k(m_0)) - u \circ \phi^k(m_0)\|^2 + \lambda \|W\|^2.$$

Then, for any $\epsilon > 0$, there exists $\lambda^* > 0$ and $\ell_0, T_0 \in \mathbb{N}$ such that for all $\lambda \in (0, \lambda^*)$ and $\ell > \ell_0, T > T_0$

$$\|W_\ell^\top F_T(f \circ \phi^{-1}, \omega) - u\|^2_{L^2(\mu)} < \epsilon.$$

Proof. Let $y : M \to y(M) \subset (\mathbb{R}^N \times \mathbb{R}^d)$ be defined by

$$y(m) = (f \circ \phi^{-1}(m), \omega(m)) \; \forall m \in M$$

and note that $F_T(f \circ \phi^{-1}, \omega) = F_T \circ y$ and that $y \in \mathrm{Hom}(M, y(M))$ because $f \in \mathrm{Hom}(M, f(M))$. Now fix $\epsilon > 0$. Let μ' be a measure defined on $y(M) \subset (\mathbb{R}^N \times \mathbb{R}^d)$ by $\mu'(\sigma) = \mu(y^{-1}(\sigma))$ for all measurable subsets σ of $f(M)$. Using the assumption that $\mathcal{F}$ is a class of linear universal approximators, we can choose T_0 sufficiently large that for any $T > T_0$ there exists $W_* \in \mathbb{R}^T$ such that

$$\|W_*^\top F_T - u \circ y^{-1}\|^2_{L^2(\mu')} < \frac{\epsilon}{3},$$

hence (by Lemma 3.5)

$$\|W_*^\top F_T \circ y - u\|^2_{L^2(\mu)} = \|W_*^\top F_T - u \circ y^{-1}\|^2_{L^2(\mu')} < \frac{\epsilon}{3}.$$

Now let

$$\lambda^* = \frac{\epsilon}{3\|W_*\|^2}$$

and $\lambda \in (0, \lambda^*)$. Define the sequence $(W_\ell)_{\ell\in\mathbb{N}}$ such that, for each $\ell \in \mathbb{N}$, the vector $W_\ell \in \mathbb{R}^T$ is the unique minimiser of the regularised least squares difference

$$\frac{1}{\ell}\left(\sum_{k=0}^{\ell-1} \|W^\top F_T(f \circ \phi^{k-1}(m_0), \omega \circ \phi^k(m_0)) - u \circ \phi^k(m_0)\|^2 + \lambda \|W\|^2\right).$$

By Lemma 3.1, $(W_\ell)_{\ell\in\mathbb{N}}$ converges as $\ell \to \infty$ to W_∞ which minimises

$$\|W^\top F_T(f \circ \phi^{-1}, \omega) - u\|^2_{L^2(\mu)} + \lambda \|W\|^2.$$

Now we choose ℓ_0 such that for all $\ell > \ell_0$

$$\|W_\ell^T F_T(f \circ \phi^{-1}, \omega) - W_\infty^T F_T(f \circ \phi^{-1}, \omega)\|^2_{L^2(\mu)} < \frac{\epsilon}{3}.$$

Now the proof proceeds directly

$$\begin{aligned}
&\|W_\ell^T F_T(f \circ \phi^{-1}, \omega) - u\|^2_{L^2(\mu)} \\
=&\|W_\ell^T F_T(f \circ \phi^{-1}, \omega) - W_\infty^T F_T(f \circ \phi^{-1}, \omega) \\
&+ W_\infty^T F_T(f \circ \phi^{-1}, \omega) - u\|^2_{L^2(\mu)} \\
\leq&\|W_\ell^T F_T(f \circ \phi^{-1}, \omega) - W_\infty^T F_T(f \circ \phi^{-1}, \omega)\|^2_{L^2(\mu)} \\
&+ \|W_\infty^T F_T(f \circ \phi^{-1}, \omega) - u\|^2_{L^2(\mu)} \\
<&\frac{\epsilon}{3} + \|W_\infty^T F_T(f \circ \phi^{-1}, \omega) - u\|^2_{L^2(\mu)}
\end{aligned}$$

$$\leq \frac{\epsilon}{3} + \|W_\infty^T F_T(f \circ \phi^{-1}, \omega) - u\|_{L^2(\mu)}^2 + \lambda \|W_\infty\|^2$$
$$\leq \frac{\epsilon}{3} + \|W_*^T F_T(f \circ \phi^{-1}, \omega) - u\|_{L^2(\mu)}^2 + \lambda \|W_*\|^2$$
$$< \frac{\epsilon}{3} + \frac{\epsilon}{3} + \|W_*^T F_T(f \circ \phi^{-1}, \omega) - u\|_{L^2(\mu)}^2$$
$$= \frac{\epsilon}{3} + \frac{\epsilon}{3} + \|W_*^T F_T \circ y - u\|_{L^2(\mu)}^2$$
$$< \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon. \quad \square$$

Theorem 3.6 guarantees an approximation in the $L^2(\mu)$ norm, which is sadly weaker than the C^1 norm. That is to say, a sequence which converges in C^1 also converges in $L^2(\mu)$, but the converse does not hold in general. This distinction is particularly relevant when the problem is chaotic time series forecasting. In this case, the target function is the *next step map* $u = \omega \circ \phi$, and we recursively feed predictions into the state space map to create a trajectory into the future. An example is the ESN autonomous phase (Eq. (1)). A weakness of using ESN autonomous dynamics for time series forecasting is that small approximation errors accumulate resulting in a predicted trajectory that diverges from the true trajectory in the far future. That said, Hart et al. [22] show that under certain conditions (crucially that the next step map $u = \omega \circ \phi$ is well approximated in the C^1 norm) the ESN autonomous phase will adopt dynamics that are topologically conjugate to the original dynamical system.

We must conclude that least squares regression does not guarantee a topologically conjugate autonomous phase, but we note that real data sets are contaminated by noise and finite precision arithmetic where an $L^2(\mu)$ approximation may be most suitable. Moreover, computing the (regularised) least squares solution using the SVD decomposition, or some other algorithm, is much faster than minimising the maximal pointwise distance, which may be necessary to yield a good C^1 approximation. Indeed, despite the theoretical limitations of the regularised least squares approach it seems to work well in practice. In fact we can interpret bad C^1 approximations in the parlance of machine learning as overfitted solutions, as they fit the training data well, in exactly the terms that we define a good fit, but may fail to make good predictions about the unseen future.

3.4. Convergence rate of the time average to the space average

Theorem 3.6 guarantees, under appropriate conditions, that with sufficiently many neurons T and a sufficiently many training data ℓ we can obtain an arbitrarily good $L^2(\mu)$ approximation of a target function u. It is natural to wonder how many training data is required to achieve a given $L^2(\mu)$ approximation. To answer this, we turn our attention to the convergence rate of the time average to the space average

$$\lim_{\ell \to \infty} \frac{1}{\ell} \sum_{k=0}^{\ell-1} s \circ \phi^k(m_0) = \int_M s \, d\mu \tag{2}$$

as the timespan over which training data is collected grows. We want a uniform estimate for the rate of convergence for s over all ergodic maps ϕ. Unfortunately, no such estimate can possibly exist. Kachurovskii [29] presents negative results that (in the author's words) *leave no hope that estimates of the rate of convergence depending only on the averaged function s can be obtained in ergodic theorems*. The negative results presented by Kachurovskii [29] prove that the amount of training data required is strictly dependent on the dynamical system.

Though we cannot say exactly how many data points we need for a good $L^2(\mu)$ approximation, the central limit theorem for ergodic dynamical systems suggests that for an initial point chosen uniformly over the invariant measure of ϕ, the difference between the finite time average and space average converges to a mean 0 normal distribution with standard deviation $1/\sqrt{\ell}$. This is made precise by the central limit theorem for ergodic dynamical systems. Before we state the theorem, we recall the definition of Hölder continuity.

Definition 3.7 (*Hölder Continuous*)**.** Let (M, d) be a metric space. A map $s : M \to \mathbb{R}$ is called *Hölder continuous* if there exist constants $p \in (0, 1]$ and $K > 0$ such that

$$\|s(m) - s(m')\| \leq K d(m, m')^p$$

for all $m, m' \in M$.

Theorem 3.8 (*Central Limit Theorem for Ergodic Dynamical Systems*)**.** *Let $\phi : M \to M$ be ergodic with respect to the probability space (M, Σ, μ). Let X_0 be a uniform random variable with respect to the space (M, Σ, μ). Let $s \in L^1(\mu)(M, \mathbb{R})$ be Hölder continuous and denote the space average of s by*

$$\mathbb{E}[s] := \int_M s \, d\mu.$$

Let the random variables $X_j := s \circ \phi^j(X_0)$ for $j = 0, \ldots, \ell - 1$ and denote the partial sum $S_\ell = X_0 + \cdots + X_{\ell-1}$. Then, for some $\sigma > 0$, the partial sum S_ℓ satisfies the central limit theorem:

$$\lim_{\ell \to \infty} \mu\left(\left\{ \frac{S_\ell - \ell \mathbb{E}[s]}{\sqrt{\ell}} \leq z \right\} \right) = \frac{1}{2\pi\sigma} \int_{-\infty}^{z} \mathrm{e}^{-\frac{\tau^2}{2\sigma^2}} \, d\tau$$

almost surely, or in other words $(S_\ell - \ell \mathbb{E}[s])/\sqrt{\ell}$ converges in law to $\mathcal{N}(0, \sigma^2)$.

Proof. Camí [30]. $\square$

To see the connection between the central limit theorem and the work in this paper, suppose we choose a map s that returns the matrix vector pair

$$s(m_0) = \left(\left[f(m_0) f^\top(m_0) + I\lambda \right], f(m_0) u(m_0) \right)$$
$$=: (\Sigma_0, v_0),$$

and define a sequence of pairs with ℓth pair

$$(\Sigma_\ell, v_\ell) := \frac{1}{\ell} \sum_{k=0}^{\ell-1} s \circ \phi^k(m_0).$$

Then it follows that

$$W_\ell = \Sigma_\ell^{-1} v_\ell$$

is the linear readout layer obtained by regularised least squares regression using ℓ data points. Furthermore, it follows from the central limit theorem that for random initial points m_0 (distributed uniformly with respect to the invariant measure μ) the sequence $(\Sigma_\ell, v_\ell)_{\ell \in \mathbb{N}}$ converges in law to a (multivariate) normal distribution, with variance converging with order $1/\ell$, and mean (Σ, v) which satisfies

$$W_\infty = \Sigma^{-1} v.$$

We note that the convergence of $(\Sigma_\ell, v_\ell)_{\ell \in \mathbb{N}}$ to (Σ, v) with order $1/\sqrt{\ell}$ does not necessarily imply that $(W_\ell)_{\ell \in \mathbb{N}}$ converges to W_∞ at the same rate.

4. The Lorenz attractor is stably mixing

We have shown that we can approximate, in the $L^2(\mu)$ sense, any target function on an ergodic dynamical system using an ESN

and Tikhonov least squares. This partially explains the success enjoyed by Jaeger [1], Xi et al. [12], Schrauwen et al. [13], Shi and Han [14], Song et al. [15], Pathak et al. [17], Løkse et al. [18], Yeo [19], Chattopadhyay et al. [20], Vlachas et al. [21], and Hart et al. [22]. Many authors including Chattopadhyay et al. [20] successfully predict the future observations of the Lorenz system, while Pathak et al. [17], Vlachas et al. [21], and Hart et al. [22] additionally recover topological invariants including Lyapunov exponents, fixed point eigenvalues and homology groups. The authors are successful in their numerical experiments because the Lorenz attractor is *mixing* which implies it is ergodic, suggesting the conditions Theorem 3.6 hold and we can $L^2(\mu)$ approximate target functions on the Lorenz attractor.

Proving that the Lorenz attractor is mixing was a tremendous achievement, built upon the works of Afraimovich et al. [31], Guckenheimer and Williams [32], Pesin [33], Williams [34], and Tucker [35] culminating with the seminal paper by Tucker [36], which resolved Smale's 14th problem *'Is the dynamics of the ordinary differential equations of Lorenz (1963) that of the geometric Lorenz attractor of Williams, Guckenheimer and Yorke?'* [37]. To formalise some of these ideas, we will begin with the definition of a mixing dynamical system.

Definition 4.1 (*Mixing*)**.** Let $\phi : M \to M$ be a measure preserving transformation on the measure space (M, Σ, μ) with $\mu(M) = 1$. Then ϕ is mixing if for any $A, B \in \Sigma$

$$\lim_{\ell\to\infty} \mu\big(A \cap \phi^{-\ell}(B)\big) = \mu(A)\mu(B).$$

Lemma 4.2 (*Mixing Implies Ergodic*)**.** *Let $\phi : M \to M$ be a measure preserving transformation on the measure space (M, Σ, μ) with $\mu(M) = 1$. Suppose ϕ is mixing, then ϕ is ergodic.*

Proof. Suppose ϕ is mixing and $A, B \in \Sigma$. Then

$$\begin{aligned} &\lim_{\ell\to\infty} \mu\big(A \cap \phi^{-\ell}(B)\big) = \mu(A)\mu(B) \\ \implies &\lim_{\ell\to\infty} \frac{1}{\ell}\sum_{k=0}^{\ell-1} \mu\big(A \cap \phi^{-k}(B)\big) = \mu(A)\mu(B) \\ \implies &\lim_{\ell\to\infty} \frac{1}{\ell}\sum_{k=0}^{\ell-1} \mu\big(A \cap \phi^{-k}(A)\big) = \mu(A)^2. \end{aligned} \tag{3}$$

Now suppose $\mu(A) = \mu\big(\phi^{-1}(A)\big)$. Then (3) reduces to $\mu(A) = \mu(A)^2$ hence $\mu(A) = 1$ or $\mu(A) = 0$, so ϕ is ergodic. □

Definition 4.3 (*Stably Mixing*)**.** Let $\phi : M \to M$ be a measure preserving transformation on the measure space (M, Σ, μ) with $\mu(M) = 1$. Then ϕ is stably mixing if sufficiently small C^1 perturbations of ϕ are mixing.

Theorem 4.4. *The Lorenz [38] system*

$$\begin{aligned} \dot{\xi} &= \sigma(\upsilon - \xi) \\ \dot{\upsilon} &= \xi(\rho - \zeta) - \upsilon \\ \dot{\zeta} &= \xi\upsilon - \beta\zeta \end{aligned} \tag{4}$$

with parameters $\sigma = 10$, $\beta = 8/3$, $\rho = 28$ admits a robust attractor that is stably mixing.

Proof. Luzzatto et al. [23]. □

Since the Lorenz attractor is stably mixing, so is any sufficiently good C^1 approximation to the evolution operator ϕ, obtained by numerical methods. Consequently, a numerically approximated Lorenz system is ergodic, by Lemma 4.2. Thus, we expect that an ESN, trained using Tikhonov least squares, on a sequence of observations of a numerically integrated trajectory of the Lorenz attractor will $L^2(\mu)$ approximate arbitrary target functions on the attractor.

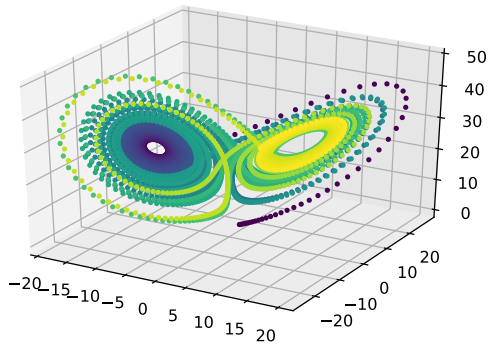


Fig. 1. A typical trajectory of the Lorenz system (4) computed for 4000 timesteps, represented by the individual dots at time intervals $\tau = 0.01$. Colour indicates the direction of travel along the trajectory: darkest colours (blue) at the earliest times and lighest colours at the most recent times (yellow). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

5. Numerical experiments

Our goal is to use an ESN to learn a mapping from the ξ component of the Lorenz attractor to the ζ component. We will sample data from a single trajectory of the Lorenz attractor. To this end, let $\phi : \mathbb{R}^3 \to \mathbb{R}^3$ denote a discretisation of the Lorenz system (4) with time step τ i.e effectively a discrete-time map of the form

$$\phi(\xi, \upsilon, \zeta) = (\xi, \upsilon, \zeta) + \int_0^\tau (\dot{\xi}, \dot{\upsilon}, \dot{\zeta})\, dt.$$

We set the timestep $\tau = 0.01$ and initial condition $(\xi_0, \upsilon_0, \zeta_0) = (0, 1.0, 1.05)$. For these initial conditions and the parameter values as in 4.4, we computed a trajectory for a 40 time units (i.e. 4000 timesteps), illustrated in Fig. 1.

We select observation and target functions to be the first and third components of the Lorenz system, i.e. we choose the function $\omega(\xi, \upsilon, \zeta) = \xi$ so that the observations z_k are the ξ components of the trajectory at the sampled time points $t = k\tau$, so that

$$z_k = \omega \circ \phi^k(\xi_0, \upsilon_0, \zeta_0).$$

We select the target function to be $\omega(\xi, \upsilon, \zeta) = \zeta$ so the targets u_k are the ζ components of the trajectory:

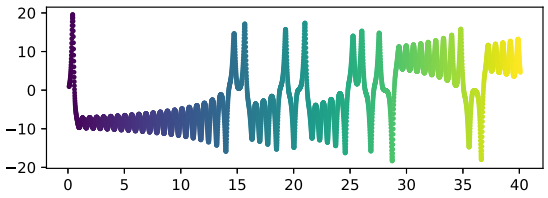
$$u_k = u \circ \phi^k(\xi_0, \upsilon_0, \zeta_0).$$

The trajectories of these two components of observations and targets are shown in Fig. 2(a) and (b), respectively.

Our goal is to use an ESN to predict the targets based on the observations. So, we set up an ESN with the following parameters:

- Reservoir size: $T = 300$,
- Activation function: $\sigma = \tanh$,
- Input matrix C and bias vector ζ: i.i.d uniform random variables $\sim U[-0.05, 0.05]$,
- Reservoir matrix A: i.i.d uniform random variables rescaled so that $\|A\|_2 = 1$,
- Regularisation parameter $\lambda = 10^{-9}$.

Iterating the ESN with observations z_k creates a discrete-time sequence of reservoir states x_k, illustrated in Fig. 3, which shows



(a) The ξ-component of the Lorenz trajectory (vertical axis) plotted against time (horizontal axis).

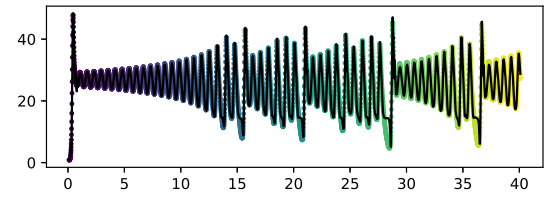


Fig. 2. Observations z_k and targets u_k drawn from the Lorenz trajectory.

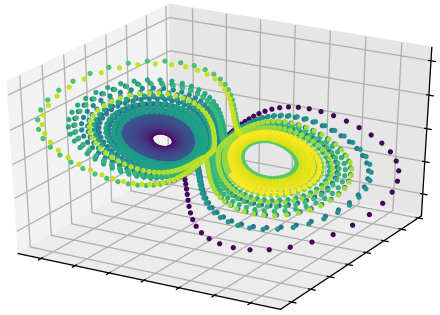


Fig. 3. Illustration of the reservoir states of the ESN driven by inputs z_k being the discrete-time samples observed from a trajectory of the Lorenz system. The figure shows the projection of the reservoir states onto their first 3 principal components.

a projection of the reservoir states onto their first the principal components. We then solved the least squares problem

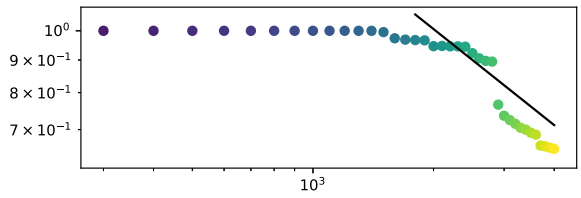
$$\min_W \sum_{k=0}^{\ell-1} \|W^\top x_k - u_k\|^2 + \lambda \|W\|^2$$

to determine the output layer W using the SVD. This offline learning method is described by Hansen et al. [39]. Our aim here is to understand how increasing the number of data points ℓ improves our approximation of the target function u. So we repeated this process with *fewer* observation-target pairs, from 300 in increments of 100 up to 4000. For each value of ℓ, we compute the best-fit readout layer W. We repeated this process once more for a 20,000 time step (i.e. 200 time unit) trajectory and computed the readout layer which for this case we denote by W_∞, assuming that it is extremely close to the readout layer we would obtain in the limit of infinitely many time steps. For each readout layer W obtained using fewer data points ($300 \leq \ell \leq 4000$) we estimated the error on the readout layer which we denote by WE:

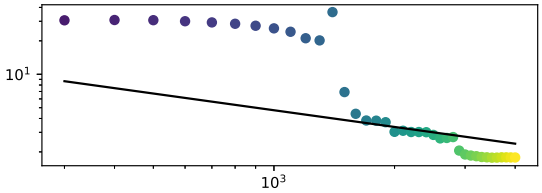
$$\text{WE} = \frac{\|W - W_\infty\|}{\|W_\infty\|},$$

and the root mean square error (RMSE) between the targets and the approximation *for the entire 20,000 point trajectory*

$$\text{RMSE} = \sqrt{\frac{1}{20000} \sum_{k=0}^{20000-1} \|W^\top x_k - u_k\|^2}.$$



(a) Log–log plot of the error on the linear readout layer W (vertical axis) against number of data points (horizontal axis) used to train the readout layer W. The line $y = 45/\sqrt{\ell}$ is plotted in black as a guide to the eye.



(b) Log–log plot of the root mean square error (RMSE) (vertical axis) against number of data points (horizontal axis) used to train the readout layer W. The line $y = 150/\sqrt{\ell}$ is plotted in black as a guide to the eye.

Fig. 4. Convergence of the error on the readout layer (WE) and convergence of the root mean square error (RMSE) displayed in log–log plots. Black lines indicate convergence with order $1/\sqrt{\ell}$ and are shown in order to compare the convergence to what might be expected if a central limit theorem applied.

We expect that as the number of data points ℓ grows the WE and RMSE will converge. The central limit theorem suggests that the matrix vector pairs $(\Sigma_\ell, v_\ell)_{\ell \in \mathbb{N}}$ (which satisfy the Gauss normal equations $\Sigma_\ell W_\ell = v_\ell$) will converge in law to a multivariate normal distribution, with standard deviation converging with order $1/\sqrt{\ell}$; as the number of data points ℓ tends to infinity. This suggests (but does not strictly imply) that the WE and RMSE might converge at a similar rate. We have been unable to derive expressions for the convergence of the RMSE and WE and remark that the need to compute W via a least-squares fit means it is not obvious that these would share the convergence rate of Σ_ℓ and v_ℓ. Typical numerical results for the convergence of the RMSE and WE are illustrated in Fig. 4.

The figures reveal that the convergence of the RMSE and WE is complicated. We observe sudden jumps which appear when the Lorenz trajectory switches to a different wing in the attractor, at such times presumably the ESN rapidly acquires new independent information which improves the fit. Furthermore, the convergence at least over this range of trajectory lengths does not (convincingly) converge with order $1/\sqrt{\ell}$. Since the sudden jumps occur on a timescale intrinsic to the dynamical system ϕ we conclude that the internal structure of the attractor and its dynamics plays an important role in the evolution of the error; appealing to the asymptotic behaviour may not always be useful.

We pushed the numerics further, hoping to detect an asymptotic regime by repeating the numerical experiments with a much longer trajectory. We computed W_∞ for a 100 000 point trajectory and compared this to the W obtained for shorter time series of lengths $\ell = 1000, 2000, \ldots, 98000$. For each ℓ we computed the WE with 10 randomly generated realisations of the ESN. The results are shown in Fig. 5 and are also (sadly) inconclusive; there is no obvious regime over which the error decreases as a power law. Sudden decreases as the trajectory switches lobes on the attractor are still visible, and the rate of convergence remains complicated.

6. Conclusions and future work

The main result of this paper (Theorem 3.6) states that an ESN trained on a sequence of observations from an ergodic dynamical

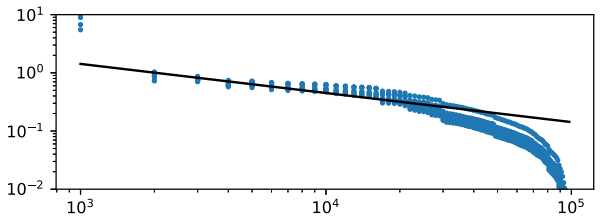


Fig. 5. The error on the readout layer (WE) (vertical axis) shown against the number of data points ℓ (horizontal axis). The black line has equation $y = 45/\sqrt{\ell}$ as a guide to the eye. Results for 10 separate realisations of an ESN are shown.

system (with invariant measure μ) using Tikhonov least squares will $L^2(\mu)$ approximate any target function u. We then summarised the result by Luzzatto et al. [23] which implies the Lorenz attractor exists and is mixing, hence ergodic. This allowed us to conclude that an ESN trained on a sequence of scalar observations taken from the Lorenz system using Tikhonov least squares should $L^2(\mu)$ approximate the dynamics in the attractor. In Section 5 we simulated the Lorenz system ourselves and designated the ξ and ζ components observations and targets respectively. We confirmed that as the number of data points (ξ_k, ζ_k) grew, the approximation of the target function improved. A good approximation was reached before the number of data points was large enough for the central limit theorem to (perhaps) become relevant. This suggests that (perhaps unfortunately) this asymptotic result may have limited practical use.

We discussed in Section 2 that the $L^2(\mu)$ norm is weaker than the C^1 norm, in the sense that convergence in C^1 implies convergence in $L^2(\mu)$, while the converse does not hold. This is somewhat unsatisfying, because (topologically conjugate) time series forecasting requires the autonomous phase of the ESN to be a C^1 approximator of the embedded (structurally stable) dynamics.

It may be a fruitful to develop a training method beyond Tikhonov least squares that guarantees a C^1 approximation. Alternatively, it may be intriguing to explore under what conditions Tikhonov least squares do provide a sufficiently good C^1 approximation, which appears to happen frequently in simulations. Authors including Pathak et al. [17], Vlachas et al. [21], and Hart et al. [22] have demonstrated that an ESNs trained with Tikhonov least squares can replicate topological invariants of dynamical systems like Lyapunov exponents, fixed point eigenvalues, and homology groups, suggesting a sufficiently good C^1 approximation was achieved.

Though the $L^2(\mu)$ approximation may not be sufficient for topological results, it may be powerful enough to prove interesting results about ESNs applied to control problems. We can view a control system as a dynamical system, for which we have at every state $x \in M$ a set of actions $a \in \mathcal{A}$ available to us. Then we seek a map $\pi : M \to \mathcal{A}$, called an optimal controller (in control theory), or an optimal policy (in reinforcement learning), which maximises some reward function. To determine the value of a policy π it suffices to determine the value function $u : M \to \mathbb{R}$ which, we can in principal approximate with an ESN from only partial observations of the control system. Developing algorithms to find the optimal controller/policy may be a rewarding direction of future work.

We also believe much of the theory presented here could be generalised or modified for other recurrent neural networks such as long short term memory networks (LSTMs). LSTMs are used extensively in industry and perform very well at *context dependent* time series problems. These are problems where events that happened a long time in the past may suddenly become important in the present. The ESN is not well suited to such problems, because the importance of events necessarily decays (at least) exponentially quickly as we move further into the past, while the structure of an LSTM sidesteps this problem. A detailed explanation of the architecture is provided by Gers [40]. Equations for a peephole LSTMS are listed below

$$
\begin{aligned}
f_k &= \varphi_g(A_f c_k + W_f^{\text{in}} u_k)\\
i_k &= \varphi_g(A_i c_k + W_i^{\text{in}} u_k)\\
o_k &= \varphi_g(A_o c_k + W_o^{\text{in}} u_k)\\
c_k &= f_k \odot c_{k-1} + i_k \odot \varphi_c(W_c^{\text{in}} u_k)\\
h_k &= \varphi_h(o_k \odot c_k)
\end{aligned}
$$

where $f_k, i_k, o_k, c_k, h_k \in \mathbb{R}^n$ are the vectors of the forget gate, input gate, output gate, cell state, and hidden state (also known as the output state) associated to the LSTM at time k. Next, $u_k \in \mathbb{R}$ is the scalar input of the LSTM at time k and $\varphi_g : \mathbb{R}^n \to \mathbb{R}^n$ is a componentwise sigmoid function, $\varphi_c : \mathbb{R}^n \to \mathbb{R}^n$ is the componentwise tanh function, and $\varphi_h : \mathbb{R}^n \to \mathbb{R}^n$ is some function that is usually the identity map. A_f, A_i, A_o, A_c are $n \times n$ matrices and $W_f^{\text{in}}, W_i^{\text{in}}, W_o^{\text{in}}, W_c^{\text{in}}$ are $1 \times n$ matrices. Finally, the symbol $\odot$ here represents the Hadamard product (taking componentwise product of 2 vectors).

We can see that LSTMs admit ESNs as a special case by fixing $A_f = 0$, $W_f^{\text{in}} = 0$, $b_f = 0$, $b_i = 0$, $b_c = \operatorname{arc\,tanh}(1/2)$, $W_c^{\text{in}} = 0$. It may therefore interest the academic community, as well as those with industrial applications in mind, to generalise the theory of ESNs presented here and elsewhere to LSTMs.

One shortcoming of Echo State Networks (that is typical for a machine learning paradigm) is that physical information about the underlying dynamical system is typically ignored. The question of how one might integrate some basic knowledge of the underlying dynamical system into the ESN architecture was recently explored numerically by Huhn and Magri [41] and Doan et al. [42]. Developing their ideas further may be an intriguing direction of future work.

CRediT authorship contribution statement

Allen G. Hart: Conceptualisation, Writing the draft, Performing the numerical experiments. **James L. Hook:** Supervision, Edited the manuscript, Improved the layout of the proofs, Performed additional numerical experiments. **Jonathan H.P. Dawes:** Supervision, Edited the manuscript, Improved the layout of the proofs, Performed additional numerical experiments.

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 thank the examiners of A.G. Hart's PhD confirmation viva, Alastair Spence and Chris Guiver, who offered helpful criticism of much of the material which formed the basis of this paper. We are grateful to an anonymous reviewer for their comments and careful reading of the manuscript which have significantly helped to improve its presentation. A.G. Hart is supported by a scholarship from the EPSRC, UK Centre for Doctoral Training in Statistical Applied Mathematics at Bath (SAMBa), under the project EP/L015684/1.

References

[1] H. Jaeger, The "echo state" approach to analysing and training recurrent neural networks, 2001.

[2] W. Maass, T. Natschläger, H. Markram, Real-time computing without stable states: A new framework for neural computation based on perturbations, Neural Comput. 14 (2002) 2531–2560, http://dx.doi.org/10.1162/089976602760407955.

[3] M.D. Skowronski, J.G. Harris, Automatic speech recognition using a predictive echo state network classifier, Neural Netw. 20 (2007) 414–423, http://dx.doi.org/10.1016/j.neunet.2007.04.006, echo State Networks and Liquid State Machines.

[4] M.H. Tong, A.D. Bickett, E.M. Christiansen, G.W. Cottrell, Learning grammatical structure with echo state networks, Neural Netw. 20 (2007) 424–432, http://dx.doi.org/10.1016/j.neunet.2007.04.013, echo State Networks and Liquid State Machines.

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

[6] X. Lin, Z. Yang, Y. Song, Short-term stock price prediction based on echo state networks, Expert Syst. Appl. 36 (2009) 7313–7317, http://dx.doi.org/10.1016/j.eswa.2008.09.049.

[7] J.D. Ser, I. Lana, E.L. Manibardo, I. Oregi, J.L.L. Eneko Osaba, M.N. Bilbao, E.I. Vlahogianni, Deep echo state networks for short-term traffic forecasting: Performance comparison and statistical assessment, 2020, arXiv:2004.08170.

[8] H. Peng, C. Chen, C.C. Lai, L.C. Wang, Z. Han, A predictive on-demand placement of uav base stations using echo state network, 2019, arXiv:1909.11598.

[9] T. Gürel, S.R.U. Egert, Functional identification of biological neural networks using reservoir adaptation for point processes, J. Comput. Neurosci. (2010) 279–299.

[10] G. Tanaka, T. Yamane, J.B. Héroux, R. Nakane, N. Kanazawa, S. Takeda, H. Numata, D. Nakano, A. Hirose, Recent advances in physical reservoir computing: A review, Neural Netw. 115 (2019) 100–123, http://dx.doi.org/10.1016/j.neunet.2019.03.005.

[11] P. Verzelli, C. Alippi, L. Livi, Learn to synchronize, synchronize to learn, 2020, arXiv:2010.02860.

[12] J. Xi, Z. Shi, M. Han, Analyzing the state space property of echo state networks for chaotic system prediction, in: Proceedings. 2005 IEEE International Joint Conference on Neural Networks, Vol. 3, 2005, pp. 1412–1417, http://dx.doi.org/10.1109/IJCNN.2005.1556081.

[13] B. Schrauwen, D. Verstraeten, J. Van Campenhout, An overview of reservoir computing: theory, applications and implementations, in: Proceedings of the 15th European Symposium on Artificial Neural Networks, 2007, pp. 471–482.

[14] Z. Shi, M. Han, Support vector echo-state machine for chaotic time-series prediction, IEEE Trans. Neural Netw. 18 (2007) 359–372.

[15] Yong Song, Yibin Li, Qun Wang, Caihong Li, 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), 2010, pp. 669–672.

[16] I.B. Yildiz, H. Jaeger, S.J. Kiebel, Re-visiting the echo state property, Neural Netw. 35 (2012) 1–9, http://dx.doi.org/10.1016/j.neunet.2012.07.005.

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

[18] S. Løkse, F.M. Bianchi, R. Jenssen, Training echo state networks with regularization through dimensionality reduction, Cogn. Comput. 9 (2017) 364–378, http://dx.doi.org/10.1007/s12559-017-9450-z.

[19] K. Yeo, Data-driven reconstruction of nonlinear dynamics from sparse observation, J. Comput. Phys. 395 (2019) 671–689, http://dx.doi.org/10.1016/j.jcp.2019.06.039.

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

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

[22] A.G. Hart, J.L. Hook, J.H.P. Dawes, Embedding and approximation theorems for echo state networks, 2019, arXiv:1908.05202.

[23] S. Luzzatto, I. Melbourne, F. Paccaut, The Lorenz attractor is mixing, Comm. Math. Phys. 260 (2005) 393–401, http://dx.doi.org/10.1007/s00220-005-1411-9.

[24] G.D. Birkhoff, Proof of the ergodic theorem, Proc. Natl. Acad. Sci. 17 (1931) 656–660, http://dx.doi.org/10.1073/pnas.17.2.656, arXiv:https://www.pnas.org/content/17/12/656.full.pdf.

[25] L. Kocarev, U. Parlitz, Generalized synchronization, predictability, and equivalence of unidirectionally coupled dynamical systems, Phys. Rev. Lett. 76 (1996) 1816–1819.

[26] L. Grigoryeva, A. Hart, J.P. Ortega, Chaos on compact manifolds: Differentiable synchronizations beyond takens, 2020, arXiv:2010.03218.

[27] K. Hornik, M. Stinchcombe, H. White, Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks, Neural Netw. 3 (1990) 551–560, http://dx.doi.org/10.1016/0893-6080(90)90005-6.

[28] V. Bogachev, Measure Theory, Vol. 1, Springer-Verlag, 2007.

[29] A.G. Kachurovskii, The rate of convergence in ergodic theorems, Russian Math. Surveys 51 (1996) 653–703.

[30] J.A.L. Camí, Ergodic theory, 2010.

[31] V.S. Afraimovich, V.V. Bykov, L.P. Shilnikov, On the origin and structure of the Lorenz attractor, Dokl. Akad. Nauk SSSR 234 (1977) 336–339.

[32] J. Guckenheimer, R.F. Williams, Structural stability of Lorenz attractors, Publ. Math. Inst. Hautes Eétudes Sci. 50 (1979) 59–72.

[33] Y.B. Pesin, Dynamical systems with generalized hyperbolic attractors: hyperbolic, ergodic and topological properties, Ergodic Theory Dynam. Systems 12 (1992) 123–151, http://dx.doi.org/10.1017/S0143385700006635.

[34] R.F. Williams, The structure of Lorenz attractors, Publ. Math. Inst. Hautes Eétudes Sci. 50 (1979) 73–99.

[35] W. Tucker, The Lorenz attractor exists, C. R. Acad. Sci., Paris I 328 (1999) 1197–1202.

[36] W. Tucker, A rigorous ode solver and smale's 14th problem, Found. Comput. Math. 2 (2002) 53–117.

[37] S. Smale, Mathematical problems for the next century, Math. Intelligencer 20 (1998) 7–15.

[38] E.N. Lorenz, Deterministic nonperiodic flow, J. Atmos. Sci. 20 (1963) 130–141.

[39] P.C. Hansen, J.G. Nagy, D.P. O'leary, Deblurring Images: Matrices, Spectra, and Filtering, SIAM, 2006.

[40] F. Gers, Learning to forget: continual prediction with lstm, IET Conf. Proc. 850–855 (5) (1999).

[41] F. Huhn, L. Magri, Learning ergodic averages in chaotic systems, 2020, arXiv:2001.04027.

[42] N.A.K. Doan, W. Polifke, L. Magri, Learning hidden states in a chaotic system: A physics-informed echo state network approach, 2020, arXiv:2001.02982.