Embedding and Approximation Theorems for Echo State Networks

Allen Hart, James Hook*, and Jonathan Dawes

Department of Mathematical Sciences
University of Bath

J.H.P.Dawes@bath.ac.uk

people.bath.ac.uk/jhpd20/

* Now at Jump Trading LLC

Outline

- Introduction
  - Setup - notation
  - Definition of the Echo State Map f: from observations to reservoir
  - Driven phase
  - Autonomous phase
- Results: *(various reasonable assumptions required)*
  - (I) Driven phase: (Takens' Theorem)
    - The Echo State Map f exists – a synchronisation result
    - The Echo State Map f is (with positive probability) an embedding
  - (II) Autonomous phase: (Randomness)
    - For a **suitably extended reservoir**, there exists an autonomous phase that has dynamics which are conjugate to the dynamics of the observations.
- Summary and future work

Introduction

Challenge: time series prediction

- Given a sequence of states

$$\dots, \quad \phi^{-3}(x), \quad \phi^{-2}(x), \quad \phi^{-1}(x), \quad x$$

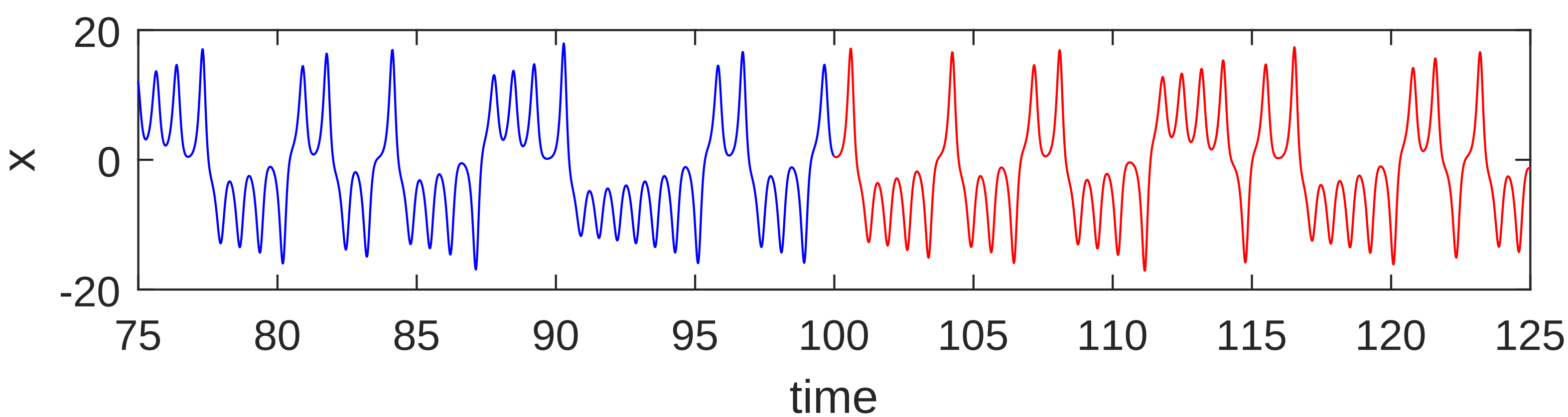
of a discrete-time dynamical system $\phi : M \to M$, **we wish to estimate $\phi(x)$**.

- Usually we have only a (scalar) observation u_k of the state:

$$\dots, \quad u_{-3} = \omega \circ \phi^{-3}(x), \quad u_{-2} = \omega \circ \phi^{-2}(x), \quad u_{-1} = \omega \circ \phi^{-1}(x), \quad u_0 = \omega(x)$$

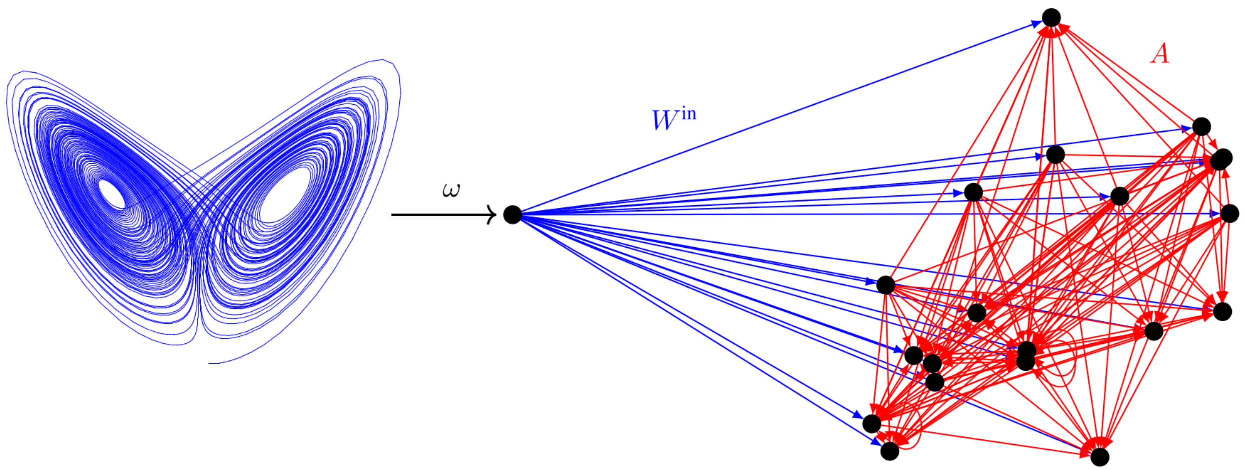
- General aim is to learn about features of ϕ, especially if ϕ samples points on trajectories of nonlinear ODEs.

E.g. for the Lorenz equations with standard parameter values:

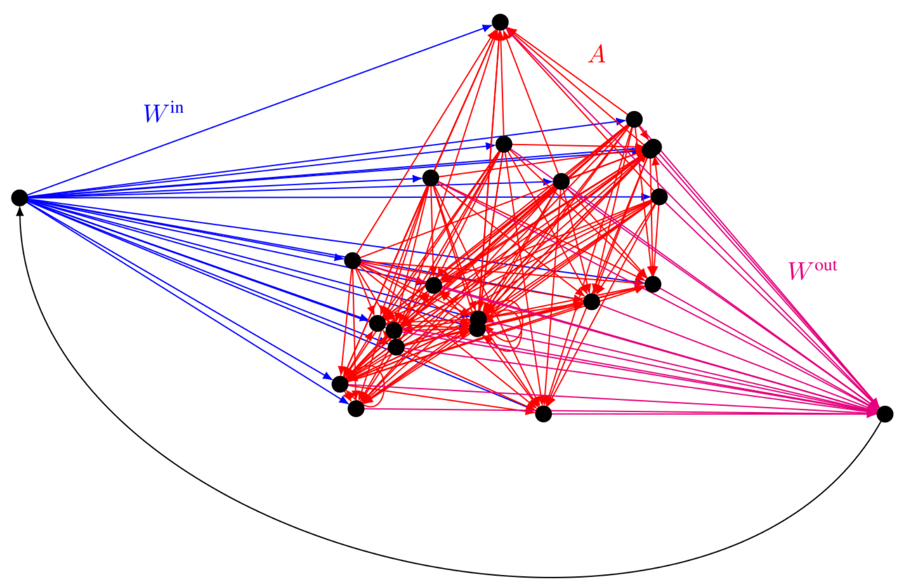


Typical Echo State Network operation

Stage 1. Driving / Synchronisation - ‘embed’ the input data into the reservoir:



Stage 2. ‘Train’ a readout layer W^{out} to mimic the inputs $\Rightarrow$ autonomous dynamics:

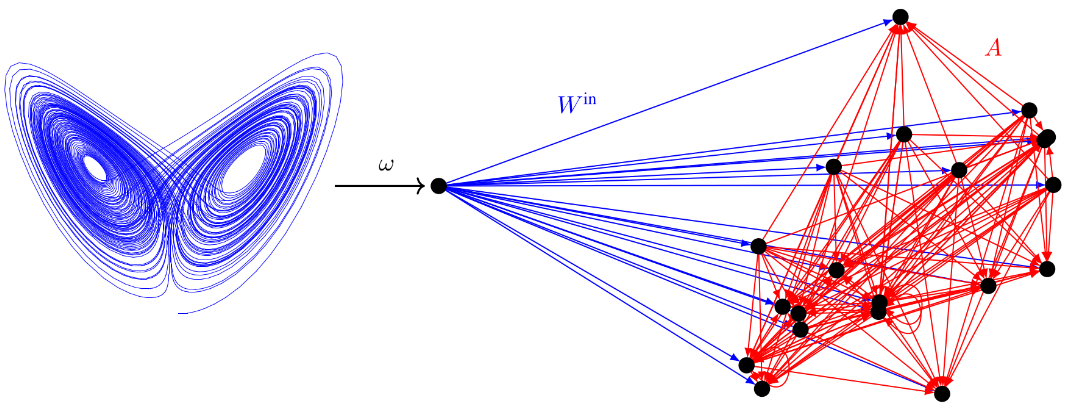


Echo State Network construction

- Generalised synchronisation: use the observations $\{u_k\}$ to drive a dynamical system (the 'reservoir') with phase space $I_n = [-1, 1]^n$, $n \gg 1$:

$$r_{k+1} \quad = \quad \varphi(A r_k + W^{\text{in}} u_k)$$

- $r_k \in I_n$ is the reservoir state
- $[\varphi(r)]_i = \sigma(r_i + b_i)$ for $i = 1, \ldots, n$; $\sigma()$ is the activation function, e.g. tanh(), defined pointwise, and b_i is a constant – a bias
- A is the adjacency matrix for the reservoir (usually generated randomly)
- W^{in} is an $n \times 1$ vector that projects u_k into the reservoir:



Results - driven phase

Construct a family of maps $M \to I_n$

- An ESN is a triple $(\varphi, A, W^{\rm in})$.
- For a fixed ESN, and a fixed initial reservoir state r_0, define a family of maps iteratively as follows:

$$
\begin{aligned}
f_0^{r_0}(x) &= r_0 && \text{constant map} \\
f_1^{r_0}(x) &= \varphi(Ar_0 + W^{\rm in}\omega(x)) && \text{1 observation used} \\
f_2^{r_0}(x) &= \varphi(Af_1^{r_0}((\phi^{-1}(x)) + W^{\rm in}\omega(x)) && \text{2 observations used} \\
&\vdots \\
f_{k+1}^{r_0}(x) &= \varphi(Af_k^{r_0}(\phi^{-1}(x)) + W^{\rm in}\omega(x)) && k+1 \text{ observations used}
\end{aligned}
$$

- The map $f_k^{r_0}$ computes a new reservoir state based on the k observations $u_{-k+1},\ \ldots,\ u_0$, where $u_k \equiv \omega \circ \phi^k(x)$.
- Question: does the sequence $\{f_k^{r_0}\}$, as $k \to \infty$, converge?
  If so, a limit map f would satisfy

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

Echo State Mapping Theorem

Theorem. If $\|A\|_2 <$ min $\left(1, 1/\|D\phi^{-1}\|_\infty\right)$ then there exists a unique solution $f \in C^1(M, \mathbb{R}^n)$ to

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

such that for all $r_0 \in I_n$ the sequence $\{f_k^{r_0}\}$ converges to f as $k \to \infty$.

We call f the **Echo State Map** (ESM)

Idea of the Proof. Define $\Psi : C^1(M, \mathbb{R}^n) \to C^1(M, \mathbb{R}^n)$ by

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

Note that $f_{k+1}^{r_0} = \Psi(f_k^{r_0})$. Then we can show that (under the condition on $\|A\|_2$ above), Ψ is a contraction map in the C^1 norm $\|f\|_{C^1} := \|f\|_\infty + \|Df\|_\infty$.

$\square$

Echo State Mapping Theorem

Remarks:

- In the case that ϕ corresponds to the evolution operator corresponding to the integration of nonlinear ODEs over a short time Δt, ϕ will be close to the identity and so we would expect $\|D\phi^{-1}\|_\infty \gtrsim 1$.

  So the constraint $\|A\|_2 <$ min $\left(1, 1/\|D\phi^{-1}\|_\infty\right)$ is mild.

- The fixed point f is unique and independent of r_0, which corresponds to the ESN having the *Echo State Property* defined by Jaeger.

Embeddings

What properties does the Echo State Map have ?

Definition: a C^1 *embedding* is an injective immersion whose domain and image are diffeomorphic.

i.e. let $M \subseteq \mathbb{R}^m$ and $N \subseteq \mathbb{R}^n$ be differentiable submanifolds and $F : M \to N$ a C^1 map. Then for F to be an embedding of M we require

- $F : M \to F(M) \subseteq N$ is injective
- $DF_x : T_x M \to T_{F(x)} N$ is injective $\forall\, x \in M$

Theorem (Whitney's Weak Embedding Theorem, 1944). If $n > 2m$ then the set of C^1 embeddings is open and dense in $C^1(M, \mathbb{R}^n)$ with respect to the C^1 topology.

$\Rightarrow$ the Echo State Map f may not be an embedding, but it is definitely arbitrarily close to an embedding.

Takens' Theorem

Theorem (Takens' Theorem as formulated by Jeremy Huke, 2006). Suppose that $\phi : M \to M$ is a diffeomorphism having the properties:

(1) ϕ has only finitely many periodic points with periods $k \leq 2m$.

(2) If $x \in M$ is periodic with period $k < 2m$ then evalues of $D\phi^k\big|_x$ are distinct.

Then for a generic (i.e. an open and dense subset) C^2 observation function ω the delay observation map $\Phi_{(\phi,\omega)} : M \to \mathbb{R}^{2m+1}$ defined by

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

is a C^1 embedding.

Structure of Huke's Proof of Takens' Theorem.

- Note that here we fix on one ϕ and consider generic observation functions ω.
- Step 1: show that $\Phi_{(\phi,\omega)}$ is a C^1 embedding for an open subset of C^2 observation functions.
- Step 2 (harder): show that $\Phi_{(\phi,\omega)}$ is a C^1 embedding for a dense subset of all C^2 observation functions.

J.P. Huke *Embedding nonlinear dynamical systems: a guide to Takens' theorem*. Manchester Institute for Mathematical Sciences MIMS EPrint 2006.26 ISSN 1749-9097. `http://eprints.maths.manchester.ac.uk/` (2006)

Takens' delay map is nearly an ESN 1/2

For a generic observation function ω, consider the ESN corresponding to the choices

$$A = \begin{pmatrix} 0 & 0 & 0 & \cdots 0 \\ 1 & 0 & 0 & \cdots 0 \\ 0 & 1 & 0 & \cdots 0 \\ 0 & 0 & 1 & \cdots 0 \\ \vdots & & & \\ 0 & 0 & 0 & \ddots 0 \end{pmatrix} \quad \text{and} \quad W^{\text{in}} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

We find that the ESM f that solves the equation $f = \varphi(Af \circ \phi^{-1} + W^{\text{in}}\omega)$ is

$$f = \begin{pmatrix} \varphi_1 \circ \omega \\ \varphi_2 \circ \varphi_1 \circ \omega \circ \phi^{-1} \\ \varphi_3 \circ \varphi_2 \circ \varphi_1 \circ \omega \circ \phi^{-2} \\ \vdots \\ \varphi_n \circ \cdots \circ \varphi_1 \circ \omega \circ \phi^{-n+1} \end{pmatrix} =: g \circ \Phi_{(\phi,\omega)}$$

Takens' delay map is nearly an ESN 2/2

$$f = \begin{pmatrix} \varphi_1 \circ \omega \\ \varphi_2 \circ \varphi_1 \circ \omega \circ \phi^{-1} \\ \varphi_3 \circ \varphi_2 \circ \varphi_1 \circ \omega \circ \phi^{-2} \\ \vdots \\ \varphi_n \circ \cdots \circ \varphi_1 \circ \omega \circ \phi^{-n+1} \end{pmatrix} =: g \circ \Phi_{(\phi,\omega)} \qquad \text{where} \qquad g = \begin{pmatrix} \varphi_1 \\ \varphi_2 \circ \varphi_1 \\ \varphi_3 \circ \varphi_2 \circ \varphi_1 \\ \vdots \\ \varphi_n \circ \cdots \circ \varphi_1 \end{pmatrix}$$

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

$$\Phi_{(\phi,\omega)} := \big(\omega(x),\ \omega \circ \phi^{-1}(x),\ \omega \circ \phi^{-2}(x),\ \ldots,\ \omega \circ \phi^{-n+1}(x)\big)$$

- Since (by Takens' theorem) $\Phi_{(\phi,\omega)}$ is a C^1 embedding, and g is a diffeomorphism, $\Rightarrow f$ is a C^1 embedding, so the set of ESNs for which the ESM is a C^1 embedding is non-empty.
- We could insert a scale factor $q < 1$ in the definition of A and it would affect g but not $\Phi_{(\phi,\omega)}$.

Embeddings and ESMs

- The set of ESNs for which f is an embedding is open. More precisely:
  Lemma. Let Ω be the subset of $(A, W^{\mathrm{in}}, \omega)$ for which the resulting ESM f is a C^1 embedding. Then Ω is open.

  Idea of Proof. Follow Huke's approach for Takens' theorem, and use Whitney's result that C^1 embeddings form an open subset of $C^1(M, \mathbb{R}^n)$.

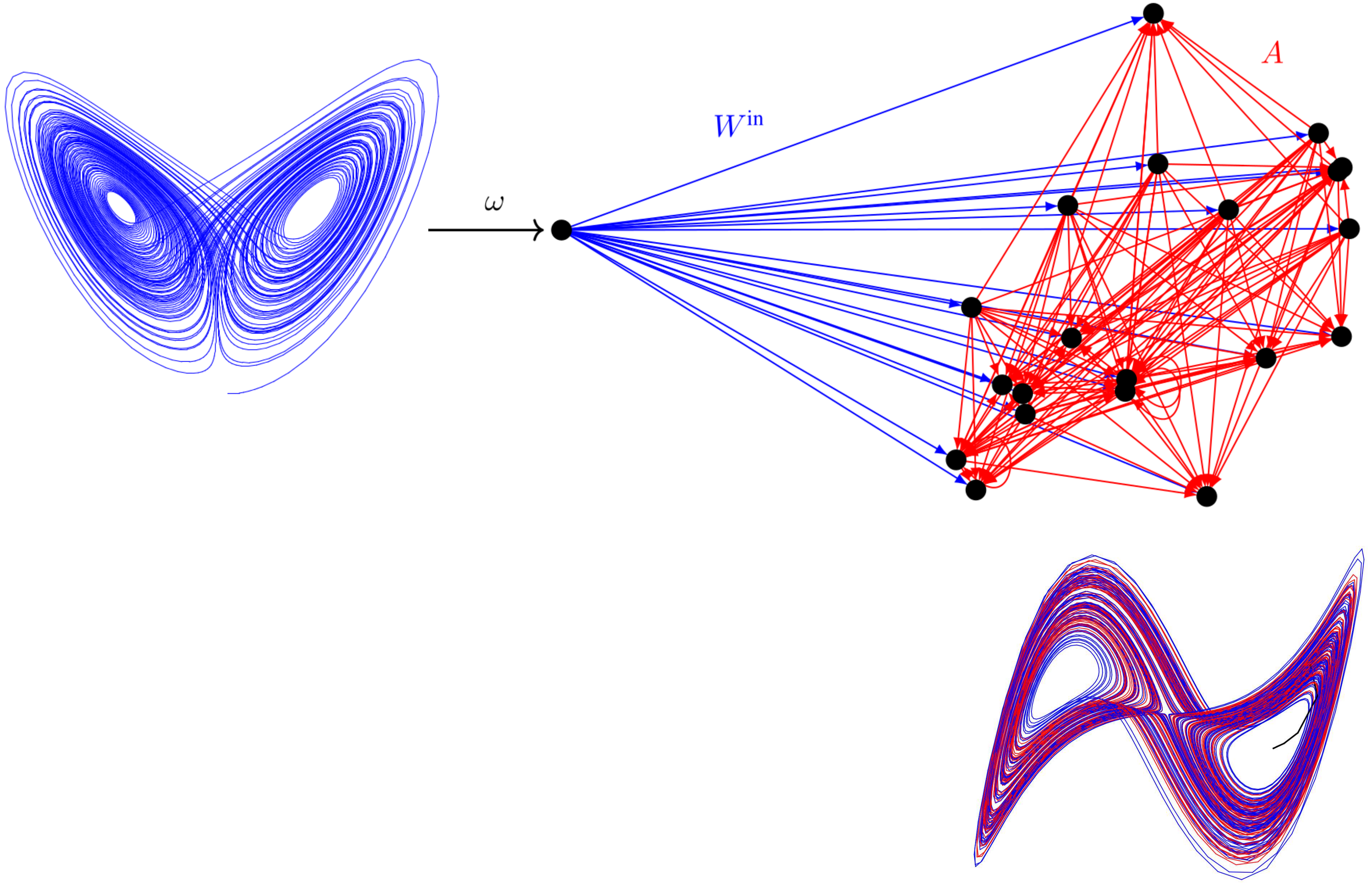
- The set of ESNs for which f is an embedding is non-empty. We just built an example where Takens' theorem implied f was an embedding.

- The set of ESNs for which f is an embedding is dense.
  We haven't been able to prove this (yet).

- Putting the first two results together we can state a
  Theorem. (Weak ESN Embedding Theorem). For randomly chosen A and W^{in}, with suitable assumptions, and a generic C^2 observation function ω, the ESM f is a C^1 embedding with positive probability.

- Good news: by open-ness, there exists some ℓ such that the finite-data approximation $f_k^{r_0}$ will also be an embedding for $k > \ell$.

Summary



Results - autonomous phase

Autonomous phase

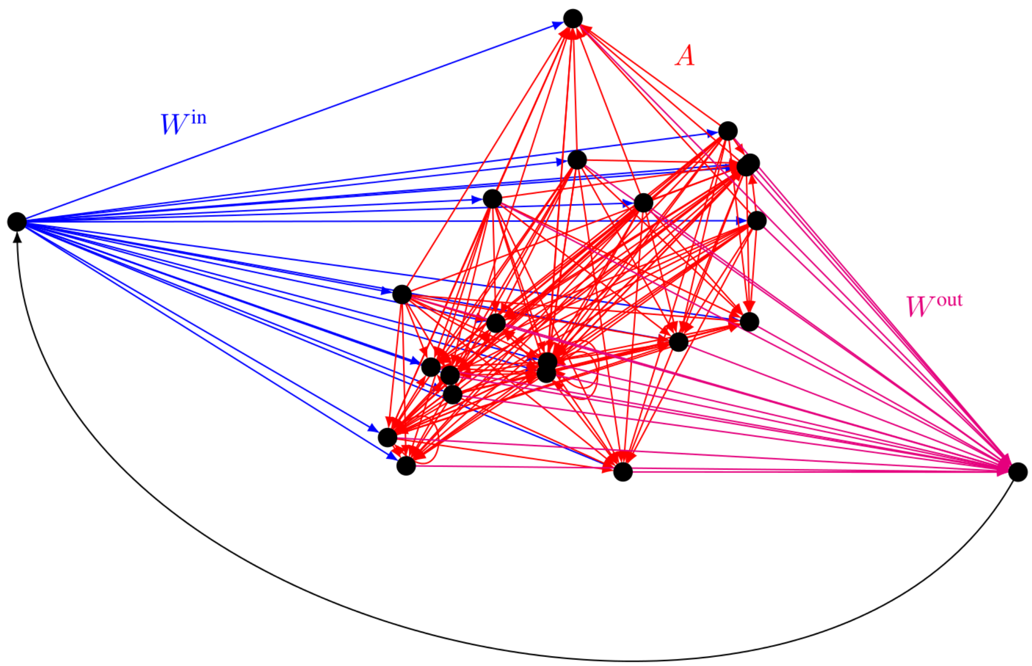
After iterating the reservoir using $r_{k+1} = \varphi(Ar_k + W^{\text{in}}u_k)$, one tries to find a $1 \times n$ readout layer (output matrix) W^{out} so that the output of the reservoir is close to a target sequence a_k. For example by taking W^{out} to be

$$W_*^{\text{out}} \quad = \quad \arg\min_{W^{\text{out}}} \sum_{k=1}^{K} \|W^{\text{out}} r_k - a_k\|^2 + \lambda \|W^{\text{out}}\|_2^2.$$

If we choose the target sequence $a_k = u_k$ then we can use this trained readout layer to run the ESN further into the future:

$$s_{k+1} \quad = \quad \varphi(As_k + W^{\text{in}} W_*^{\text{out}} s_k)$$

setting $s_0 = r_K$.



Approximating the dynamics

Central aim:

Given a diffeomorphism $\phi : M \to M$, we want to show that there exists an ESN which can approximate the dynamics of ϕ arbitrarily closely.

- Key ingredient:
  Theorem (Random Universal Approximation Theorem). For any map $G \in C^1(I_n, \mathbb{R})$, random variables $b_j \in \mathbb{R}$ and $v_j \in \mathbb{R}^n$ with full support, and for any $\alpha \in (0,1)$ and $\varepsilon > 0$, there exists an N such that with probability greater than α there exist (scalar) weights w_j such that a realisation of the random neural network

$$g(x) := \sum_{j=1}^{N} w_j \sigma(v_j^T x + b_j) \quad \text{satisfies } \|G - g\|_{C^1} < \varepsilon.$$

Huang, Zhu & Siew (2006), *Neurocomputing* **70**, 489.

Gonon, Grigoryeva & Ortega (2020), arXiv: 2002.05933.

Random UAT

- The RUAT builds on the Universal Approximation Theorem:
  Theorem (UAT; Hornik et al, 1990). Functions $\hat{g} : I_n \to \mathbb{R}$ of the form

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

  are dense in $C^1(I_n, \mathbb{R})$, as long as $\sigma(x)$ is bounded.

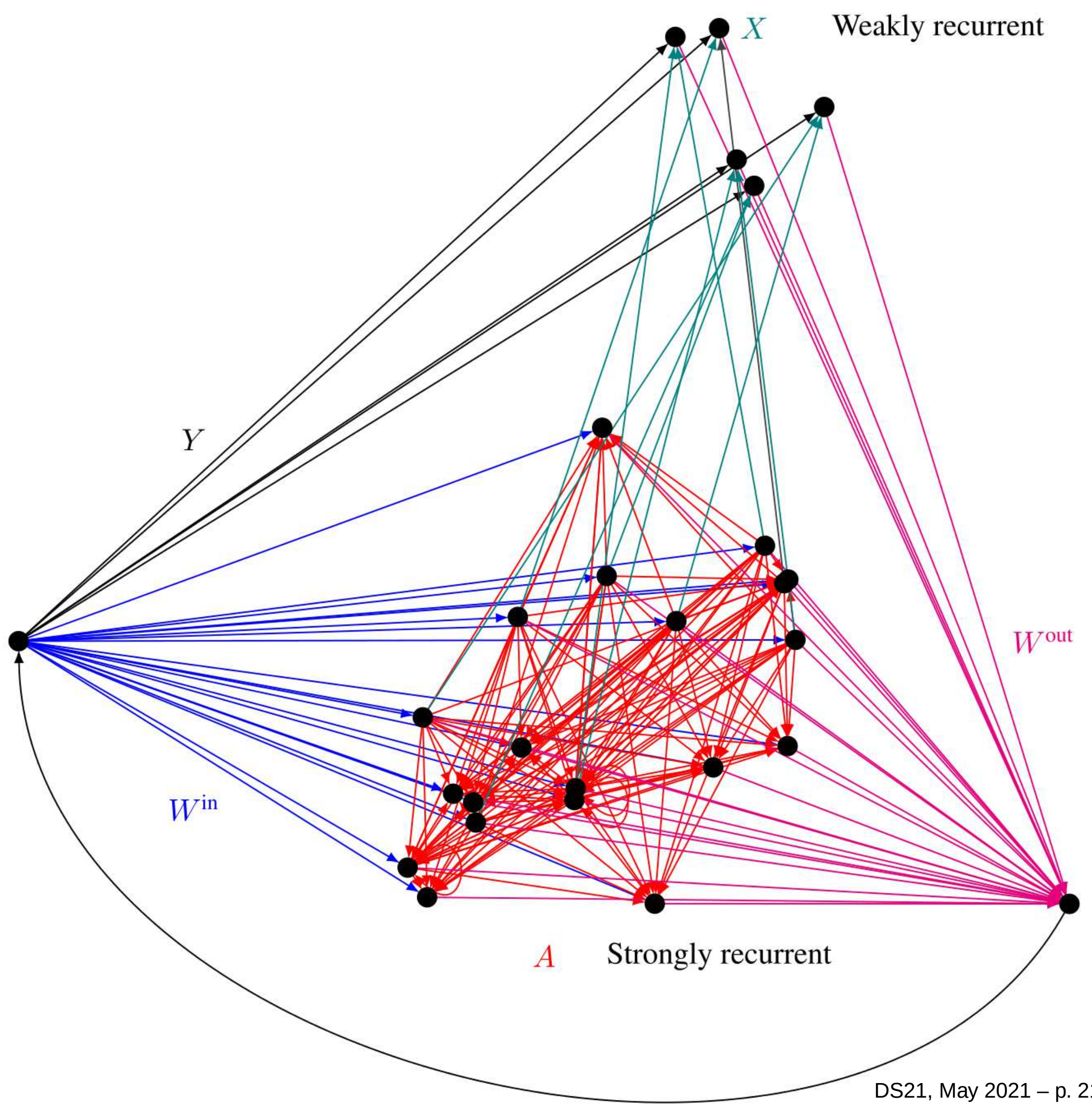
  Idea of the proof of the Random UAT.

- First, using the UAT we can approximate $G(x)$ by $\hat{g}(x) = \sum_{i=1}^{\ell} \hat{w}_i \sigma(\hat{v}_i^T x + \hat{b}_i)$
- Second, with arbitrarily high probability α we can approximate $\hat{g}$ by $g(x) = \sum_{j=1}^{N} w_j \sigma(v_j^T x + b_j)$ by taking N large enough and selecting the weight w_j to match $\hat{w}_i$ when the pair (b_j, v_j) is close enough to $(\hat{b}_i, \hat{v}_i)$, or to be zero otherwise.

$\square$

Approximating the dynamics

- We want to show that iterating the autonomous ESN can approximate the dynamics ϕ.
- Slightly more precisely, we can show that, assuming knowledge of the ESM f and the current point $x \in M$, we can with high probability build a reservoir that can approximate the next value $\omega \circ \phi(x)$ of the input sequence.
- The RUAT shows that this is possible by suitably extending the reservoir.

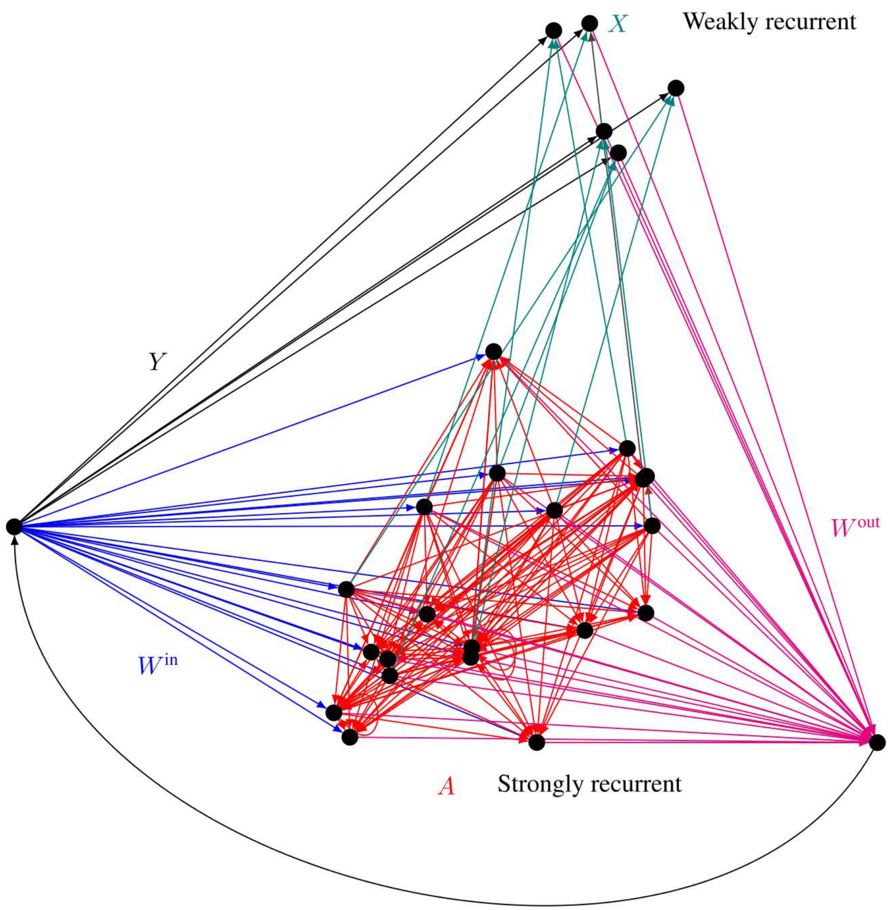


ESN Approximation Theorem

Theorem. Let $(\varphi, A, W^{\text{in}})$ be an ESN with an ESM $f \in C^1(M, \mathbb{R}^n)$. For any given probability $\alpha \in (0,1)$ there exists an (expanded) ESN $(\tilde{\varphi}, \tilde{A}, \tilde{W}^{\text{in}})$ and a readout layer W^{out} such that the autonomous ESN defined by

$$\begin{aligned} s_{k+1} &:= \psi(s) \\ &= \tilde{\varphi}\left(\tilde{A}s_k + \tilde{W}^{\text{in}} W^{\text{out}} s_k\right) \end{aligned}$$

is C^1–conjugate to the original input dynamics ϕ.



where

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

ESN Approximation: details

- (I) We construct a map $\omega \circ \phi \circ y^{-1} : \mathbb{R}^{n+1} \to \mathbb{R}$ that takes the previous reservoir state plus the observation of the current state $\omega(x)$ and produces the observation of the next state of the original dynamical system.
- By the RUAT the map $\omega \circ \phi \circ y^{-1}$ can be approximated by a sum of the form

$$g(z) \quad = \quad \sum_{i=1}^{d} W_i^{\mathrm{out}} \sigma \left(\left[\tilde{W}^{\mathrm{in}} \;\; \tilde{A} \right]_i z + b_i \right).$$

- (II) Separately, since f is a C^1 embedding, there exists a diffeomorphism η defined on an open subset Ω of the image $f(M)$. And $f(M)$ is a normally hyperbolic attracting submanifold (NHASM), on which $\eta|_{f(M)} = f \circ \phi \circ f^{-1}$.
- Since NHASMs 'persist under approximation' and we can then assert the existence of a nearby NHASM on which ψ is defined and is conjugate to η, and hence is conjugate to ϕ. $\square$

Summary and future work

- The Echo State Map f exists and, for randomly chosen A and W^{in}, is an embedding with positive probability.
  This doesn't demand a very large reservoir (cf Taken's theorem).

- There exists an extension of the ESN and a choice of W^{out} for which the dynamics ψ of the autonomous phase of the ESN is C^1-conjugate to the original input dynamics.
  This may require a much larger reservoir (i.e. $d \gg n$).

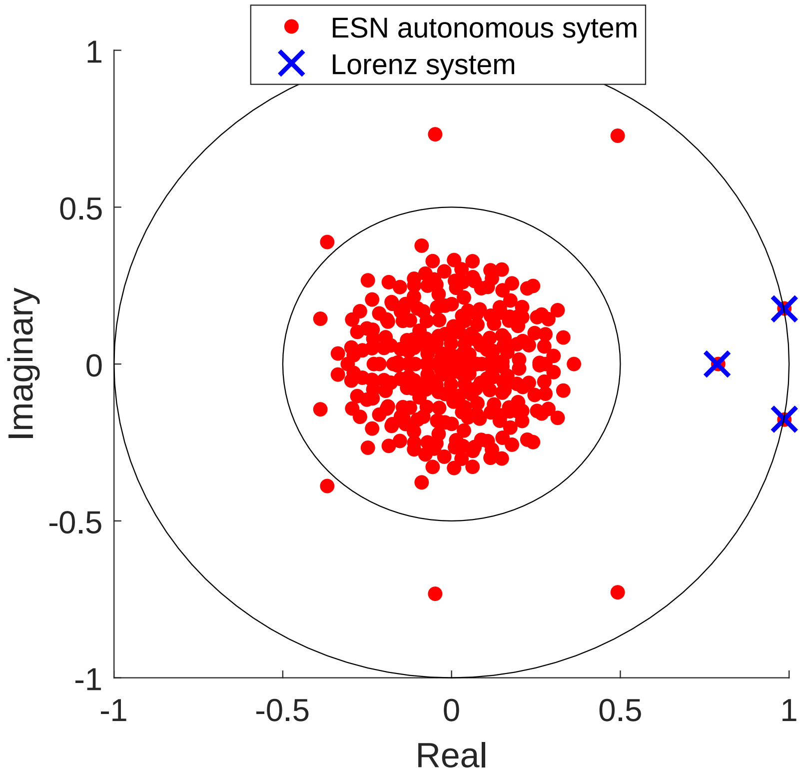
Future work / open questions:

- How best actually to choose W^{out} ?
- Find sufficient conditions for f to actually be an embedding (our 'ESN Embedding Conjecture').
- Generalisations (e.g. Grigoryeva, Hart & Ortega, arxiv: 2010.03218).

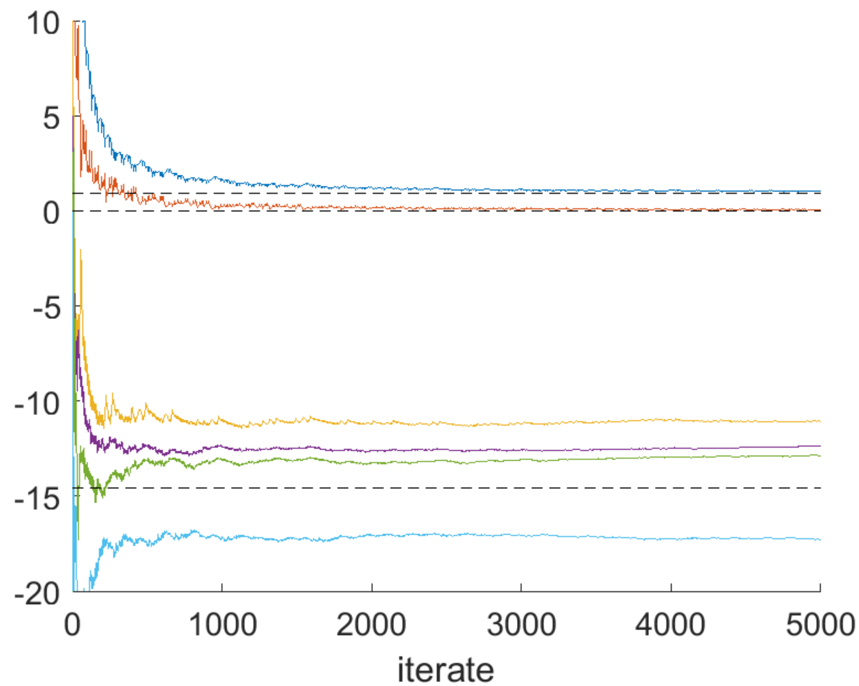
A.G. Hart, J.L. Hook and J.H.P. Dawes, Embedding and approximation theorems for echo state networks. *Neural Networks* **128**, 234–247 (2020)

Why does this matter?

- Helps to explain preservation of metric features of the whole attractor:



Eigenvalues of $\exp\left(\Delta t\, Df|_{x^*}\right)$ at a non-trivial eqm pt x^* for the Lorenz equations.



Convergence of Lyapunov exponents for the autonomous phase, with the values $0.9056, 0, -14.5723$ (4 d.p.) shown for comparison.