Chaos & RNNs

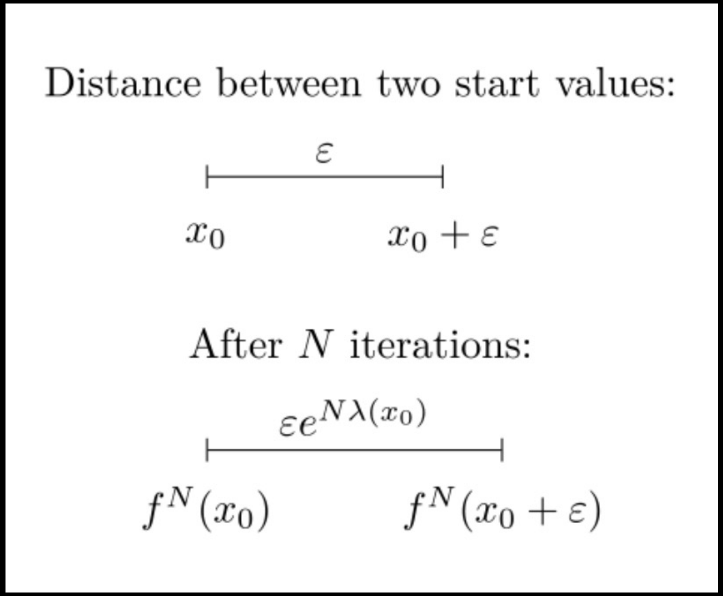
December 29th, 2022

Yejun Jang, Club Gauss
Department of Electrical Engineering

Contents

Chaos : A Brief Explanation

- Usually occurs in Nonlinear Dynamical Systems.
- Sensitivity to Initial Conditions – Butterfly effects (Lorenz)



Source: Wikipedia, Lyapunov Exponent

More specifically, we can measure the (Maximum) Lyapunov Exponent by taking the log of the geometric-mean-like value of the Jacobian spectral norms.

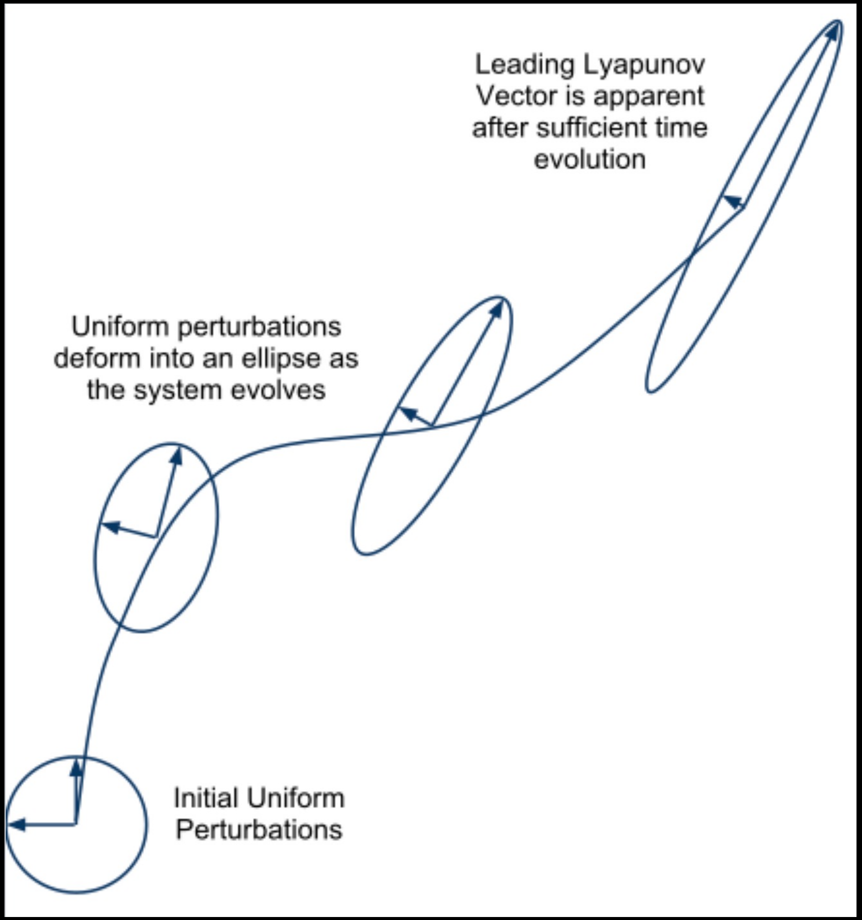
$$\lambda_{max} := \lim_{T \to \infty} \frac{1}{T} \log \left\| \prod_{r=0}^{T-2} \boldsymbol{J}_{T-r} \right\|,$$

$$||A||_2 = \bar{\sigma}(A)$$

$$\bar{\sigma}(A \cdot B) \leq \bar{\sigma}(A) \cdot \bar{\sigma}(B)$$

$\| A \|$ denotes the spectral norm, defined $\rho(A^T A)$,
where $\rho(\cdot)$ is the maximum eigenvalue.
For square matrices, it is the maximum absolute eigenvalue.

Chaos : A Brief Explanation



Source: Wikipedia, Lyapunov Exponent

- Depending of the orientation of the initial perturbation, the degree of separation may differ.
- Eventually, The eigenvector corresponding to the "maximum" Lyapunov exponent, or the leading Lyapunov vector becomes apparent.

Chaos：A Brief Explanation

- There are a several more properties regarding the topology on the state space for us to fully define chaos.
- However, the sensitivity to initial condition is the only property of chaos that we will be using for today, hence the full definition is omitted.
- Besides, scientists do not yet agree on a specific definition of chaos.

Examples of Chaos

$$f\colon \mathbb{R} \mapsto \mathbb{R}$$
$$f\colon x \mapsto rx(1-x)$$

$$x_{n+1} = f(x_n)$$

Logistic Map

Lorenz Attractor

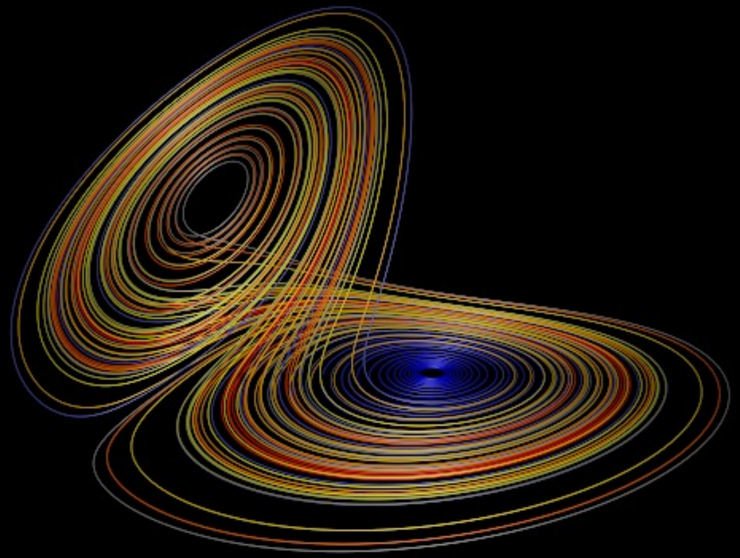
$$\dot{x} = \sigma(y-x)$$
$$\dot{y} = x(\rho - z) - y$$
$$\dot{z} = xy - \beta z$$

The Double Pendulum

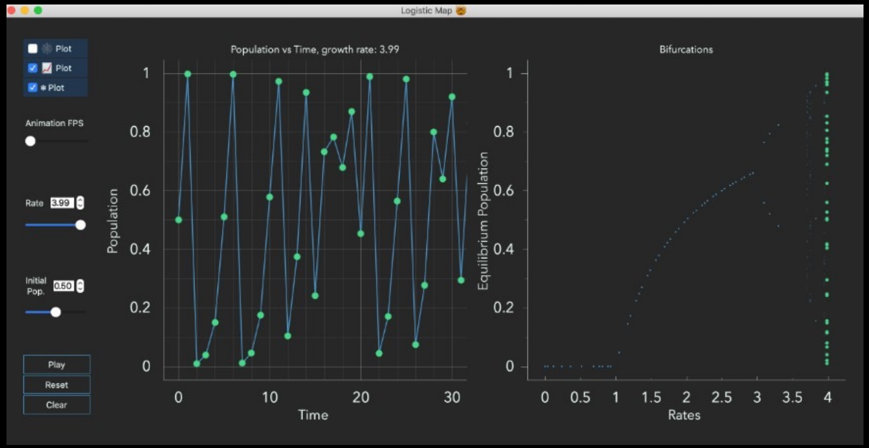
$$x_1 = \frac{l}{2}\sin\theta_1$$
$$y_1 = -\frac{l}{2}\cos\theta_1$$
$$x_2 = l\left(\sin\theta_1 + \tfrac{1}{2}\sin\theta_2\right)$$
$$y_2 = -l\left(\cos\theta_1 + \tfrac{1}{2}\cos\theta_2\right)$$

Source: Wikipedia, Chaos Theory

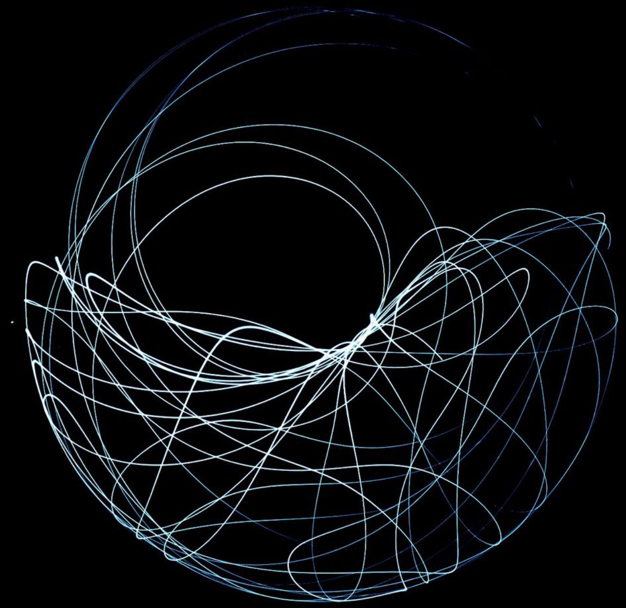
Examples of Chaos



Lorenz
Attractor



Logistic Map



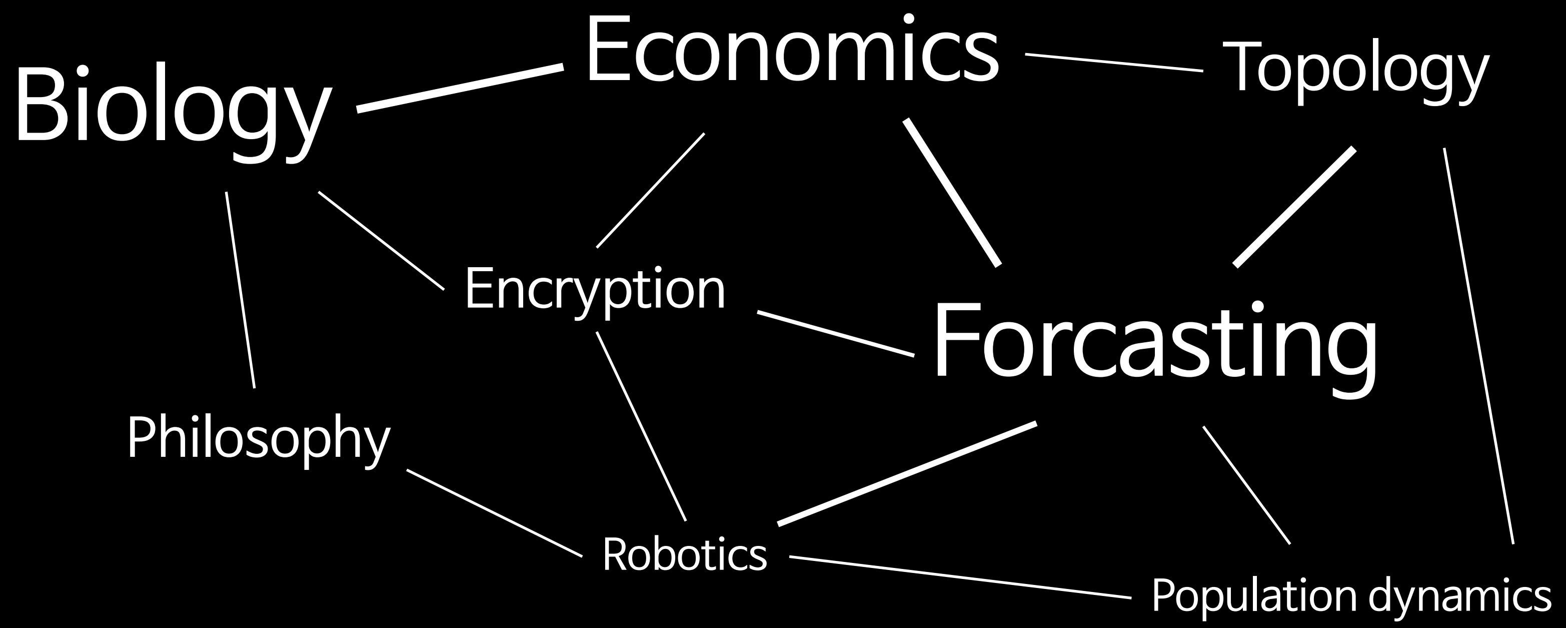
The Double
Pendulum

Source:

Double Pendulum, Wikipedia
The Lorenz Attractor in 3D, paulbourke.net
Visualizations of the connections between chaos theory and fractals through the logistic map, Python Awesome

Examples of Chaos

- Lesson: Deterministic Dynamical Systems can be impossible to predict, if we only have access to finitely accurate real-world data.



... Chaotic systems appear basically everywhere

Today's paper:

On the difficulty of learning chaotic dynamics with RNNs

Jonas M. Mikhaeil[1,2,*], **Zahra Monfared**[1,4*], and **Daniel Durstewitz**[1,2,3]

j.mikhaeil@columbia.edu, {zahra.monfared, daniel.durstewitz}@zi-mannheim.de

[1]Department of Theoretical Neuroscience, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany

[2]Faculty of Physics and Astronomy, Heidelberg University, Heidelberg, Germany

[3]Interdisciplinary Center for Scientific Computing, Heidelberg University

[4]Department of Mathematics & Informatics and Cluster of Excellence STRUCTURES, Heidelberg University, Heidelberg, Germany

[*]These authors contributed equally

Key points

- EVGP (Exploding & Vanishing Gradient Problem)

- RNNs producing stable equilibrium or cyclic behavior have bounded gradients.
- gradients of RNNs with chaotic dynamics **always diverge**.

- Applying theory to practice：Sparse Teacher Forcing

Some Background Knowledge...

RNNs are discrete time DS

$$\boldsymbol{z}_t = F_{\boldsymbol{\theta}}(\boldsymbol{z}_{t-1}, \boldsymbol{s}_t),$$

$$\boldsymbol{J}_t := \frac{\partial F_{\boldsymbol{\theta}}(\boldsymbol{z}_{t-1}, \boldsymbol{s}_t)}{\partial \boldsymbol{z}_{t-1}} = \frac{\partial \boldsymbol{z}_t}{\partial \boldsymbol{z}_{t-1}}.$$

$$\lambda_{max} := \lim_{T \to \infty} \frac{1}{T} \log \left\| \prod_{r=0}^{T-2} \boldsymbol{J}_{T-r} \right\|,$$

(Max. Lyapunov Exponent)

(Spectral Norm of the product of J_t's)

Suppose the loss function decomposes through time

$$\mathcal{L} = \sum_{t=1}^{T} \mathcal{L}_t$$

$$\frac{\partial \mathcal{L}}{\partial \theta} = \sum_{t=1}^{T} \frac{\partial \mathcal{L}_t}{\partial \theta} \quad \text{with} \quad \frac{\partial \mathcal{L}_t}{\partial \theta} = \sum_{r=1}^{t} \frac{\partial \mathcal{L}_t}{\partial \boldsymbol{z}_t} \frac{\partial \boldsymbol{z}_t}{\partial \boldsymbol{z}_r} \frac{\partial^+ \boldsymbol{z}_r}{\partial \theta},$$

$$\begin{aligned} \frac{\partial \boldsymbol{z}_t}{\partial \boldsymbol{z}_r} &= \frac{\partial \boldsymbol{z}_t}{\partial \boldsymbol{z}_{t-1}} \frac{\partial \boldsymbol{z}_{t-1}}{\partial \boldsymbol{z}_{t-2}} \cdots \frac{\partial \boldsymbol{z}_{r+1}}{\partial \boldsymbol{z}_r} \\ &= \prod_{k=0}^{t-r-1} \frac{\partial \boldsymbol{z}_{t-k}}{\partial \boldsymbol{z}_{t-k-1}} = \prod_{k=0}^{t-r-1} \boldsymbol{J}_{t-k}, \end{aligned}$$

Attractor & Basin of attraction

- <u>Attractor</u>
  - A set in state space, s.t., it is the limit set of orbits originating from a set of initial conditions of positive Lebesgue measure.
- <u>Basin of Attraction</u>
  - The set of initial conditions leading to long-time behavior that approaches that attractor.

Now, we will cover a few theorems.

Theorem 1 | The Asymptotically Periodic Case.

- Convergence to a stable fixed point or k−cycle : Gradient does not diverge.

Theorem 1. *Consider an RNN $F_{\boldsymbol{\theta}} \in \mathcal{R}$ parameterized by $\boldsymbol{\theta}$, and assume that it converges to a stable fixed point or k-cycle Γ_k $(k \geq 1)$ with $\mathcal{B}_{\Gamma_k}$ as its basin of attraction. Then for every $\boldsymbol{z}_1 \in \mathcal{B}_{\Gamma_k}$ (i) the Jacobian $\frac{\partial \boldsymbol{z}_T}{\partial \boldsymbol{z}_1}$ exponentially vanishes as $T \to \infty$; (ii) for Γ_k the tangent vectors $\frac{\partial \boldsymbol{z}_T}{\partial \theta}$ and thus the gradient of the loss function, $\frac{\partial \mathcal{L}_T}{\partial \boldsymbol{\theta}}$, will be bounded from above, i.e. will not diverge for $T \to \infty$; and (iii) for the PLRNN (27) both $\left\|\frac{\partial \boldsymbol{z}_T}{\partial \theta}\right\|$ and $\left\|\frac{\partial \mathcal{L}_T}{\partial \boldsymbol{\theta}}\right\|$ will remain bounded for every $\boldsymbol{z}_1 \in \mathcal{B}_{\Gamma_k}$ as $T \to \infty$.*

Theorem 1-(i) (Skip)

Proof. (i) Assume that Γ_k is a stable k-cycle $(k \geq 1)$ denoted by

$$\Gamma_k = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \cdots, \boldsymbol{z}_T, \cdots\} = \{\boldsymbol{z}_{t^*k}, \boldsymbol{z}_{t^*k-1}, \cdots, \boldsymbol{z}_{t^*k-(k-1)}, \boldsymbol{z}_{t^*k}, \boldsymbol{z}_{t^*k-1}, \cdots, \boldsymbol{z}_{t^*k-(k-1)}, \cdots\}. \tag{7}$$

Then, the largest Lyapunov exponent of Γ_k is given by

$$\begin{aligned} \lambda_{\Gamma_k} &= \lim_{t\to\infty} \frac{1}{t} \ln \left\| J_t^* \, J_{t-1}^* \cdots J_2^* \right\| \\ &= \lim_{j\to\infty} \frac{1}{jk} \ln \left\| \left(\prod_{s=0}^{k-1} J_{t^*k-s} \right)^j \right\|. \end{aligned} \tag{8}$$

Theorem 1-(i) (Skip)

By assumption of stability of Γ_k we have $\lambda_{\Gamma_k} < 0$ and also $\rho\big(\prod_{s=0}^{k-1} J_{t^{*k}-s}\big) < 1$ (the spectral radius), which implies

$$\lim_{t\to\infty} J_t^* \, J_{t-1}^* \cdots J_2^* = \lim_{j\to\infty} \left(\prod_{s=0}^{k-1} J_{t^{*k}-s}\right)^j = 0. \tag{9}$$

Now suppose that $\mathcal{O}_{\boldsymbol{z}_1}$ is an orbit of the map eqn. (1) converging to Γ_k, i.e. $\boldsymbol{z}_1 \in \mathcal{B}_{\Gamma_k}$. Since $\mathcal{O}_{\boldsymbol{z}_1}$ and Γ_k have the same largest Lyapunov exponent, we have

$$\lambda_{\mathcal{O}_{\boldsymbol{z}_1}} = \lim_{T\to\infty} \frac{1}{T} \ln \|J_T \, J_{T-1} \cdots J_2\| = \lambda_{\Gamma_k} < 0, \tag{10}$$

and hence for $\boldsymbol{z}_1 \in \mathcal{B}_{\Gamma_k}$

$$\lim_{T\to\infty} \left\| \frac{\partial \boldsymbol{z}_T}{\partial \boldsymbol{z}_1} \right\| = \lim_{T\to\infty} \|J_T \, J_{T-1} \cdots J_2\| = 0. \tag{11}$$

(ii) & (iii) See Appx. A.2.1. $\square$

Theorem 2 | The Chaotic Case. ⭐⭐⭐⭐⭐

- RNN going through a chaotic orbit – Gradient will diverge.

Theorem 2. *Suppose that an RNN $F_{\boldsymbol{\theta}} \in \mathcal{R}$ (parameterized by $\boldsymbol{\theta}$) has a chaotic attractor Γ^* with $\mathcal{B}_{\Gamma^*}$ as its basin of attraction. Then, for almost every orbit with $\boldsymbol{z}_1 \in \mathcal{B}_{\Gamma^*}$, (i) the Jacobians connecting temporally distal states $\boldsymbol{z}_T$ and $\boldsymbol{z}_t$ $(T \gg t)$, $\frac{\partial \boldsymbol{z}_T}{\partial \boldsymbol{z}_t}$, will exponentially explode for $T \to \infty$, and (ii) the tangent vector $\frac{\partial \boldsymbol{z}_T}{\partial \theta}$ and so the gradients of the loss function, $\frac{\partial \mathcal{L}_T}{\partial \theta}$, will diverge as $T \to \infty$.*

Theorem 2-(i)

Proof. Let the RNN $F_{\boldsymbol{\theta}} \in \mathcal{R}$ have a chaotic orbit denoted by $\Gamma^* = \{\boldsymbol{z}_1^*, \boldsymbol{z}_2^*, \cdots, \boldsymbol{z}_T^*, \cdots\}$. Then, denoting by J_T^* the Jacobian of (1) at $\boldsymbol{z}_T^* \in \Gamma^*$, the largest Lyapunov exponent of Γ^* is given by

$$\lambda = \lim_{T \to \infty} \frac{1}{T} \ln \left\| J_T^* \, J_{T-1}^* \cdots J_2^* \right\|. \tag{12}$$

Since Γ^* is chaotic, so $\lambda > 0$. Hence, from (12), it is concluded that

$$\lim_{T \to \infty} \left\| J_T^* \, J_{T-1}^* \cdots J_2^* \right\| = \lim_{T \to \infty} \left\| \frac{\partial \boldsymbol{z}_T^*}{\partial \boldsymbol{z}_t^*} \right\| = \infty, \quad T \gg t. \tag{13}$$

Now, according to Oseledec's multiplicative ergodic Theorem, almost all the points in the basin of attraction of Γ^* have the same largest Lyapunov exponent λ. Thus, (13) holds for almost every $\boldsymbol{z}_1 \in \mathcal{B}_{\Gamma^*}$.

(ii) See Appx. A.2.2. $\square$

Theorem 3 | The Quasi–periodic Case.

Theorem 3. *Assume that an RNN $F_{\boldsymbol{\theta}} \in \mathcal{R}$ (parameterized by $\boldsymbol{\theta}$) has a quasi-periodic attractor Γ with $\mathcal{B}_\Gamma$ as its basin of attraction. Then, for every $\boldsymbol{z}_1 \in \mathcal{B}_\Gamma$*

$$\forall\, 0 < \epsilon < 1 \;\; \exists\, T_0 > 1 \;\; s.t. \;\; \forall\, T \geq T_0 \implies$$

$$(1-\epsilon)^{T-1} < \left\| \frac{\partial \boldsymbol{z}_T}{\partial \boldsymbol{z}_1} \right\| < (1+\epsilon)^{T-1}. \tag{14}$$

Proof. See Appx. A.2.3. $\square$

Solution : Sparse Teacher Forcing.

(Incomplete)

linear output layer $\hat{\boldsymbol{x}}_t = \boldsymbol{B}\boldsymbol{z}_t,$

(control signal) → $\tilde{\boldsymbol{z}}_t = (\boldsymbol{B}^\top \boldsymbol{B})^{-1} \boldsymbol{B}^\top \boldsymbol{x}_t.$

(The Moore-Penrose Pseudoinverse)

$$\boldsymbol{z}_{t+1} = \begin{cases} RNN(\tilde{\boldsymbol{z}}_t) & \text{if } t \in \{n\tau^{?} + 1\}_{n \in \mathbb{N}_0} \\ RNN(\boldsymbol{z}_t) & \text{else} \end{cases}$$

$$\tau_{\text{pred}} = \frac{\ln 2}{\lambda_{\max}}.$$

Predictability Time
is given by ln2 / (Max. Lyapunov Exponent)

The predictability time is calculated only once, using the observation data.

Solution : Sparse Teacher Forcing.

(Incomplete)

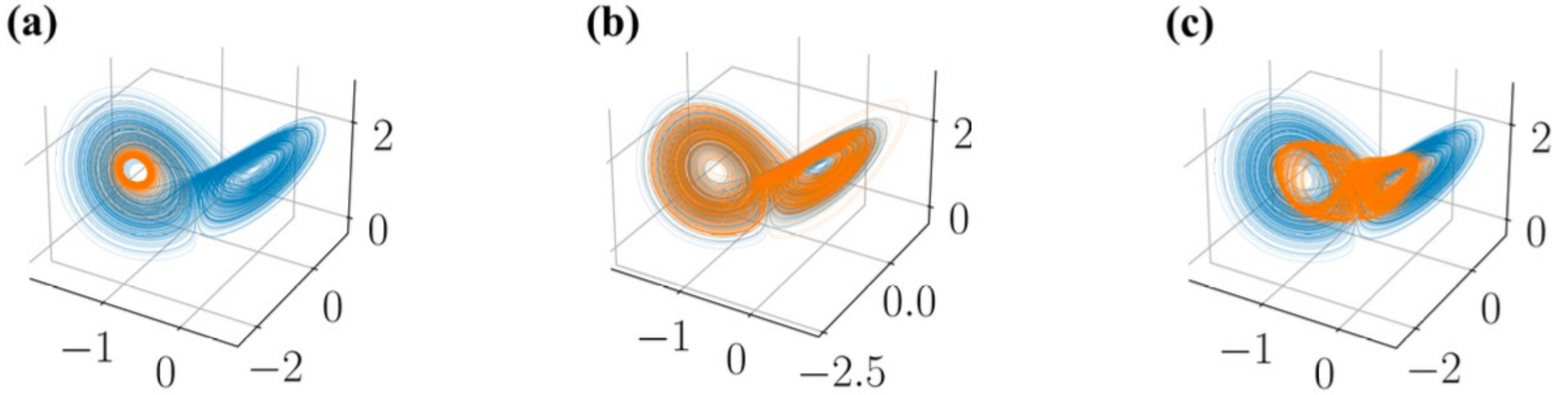


Figure 3: Lorenz attractor (blue) and example reconstructions by an LSTM (orange) trained with a learning interval (a) chosen too small ($\tau = 5$), (b) chosen optimally ($\tau = 30$), and (c) chosen too large ($\tau = 200$). See Fig. 14 for a vanilla RNN example.

Discussion and Conclusions

5 Discussion and conclusions

In this paper we proved that RNN dynamics and loss gradients are intimately related for all major types of RNNs and activation functions. If the RNN is "well behaved" in the sense that its dynamics converges to a fixed point or cycle, loss gradients will remain bounded, and established remedies [35, 80] can be used to refrain them from vanishing. However, if the dynamics are chaotic, gradients will always explode. This constitutes a *principle* problem in RNN training that cannot easily be mastered through architectural design or gradient clipping. This is because to avoid exploding gradients while training on time series from chaotic systems, one either needs to constrain the RNN so much that chaotic behavior is completely disabled to begin with (i.e., ultimately by forcing all Lyapunov exponents to be smaller or equal to zero), implying a very poor fit to such data. Or one needs to be a bit more lenient and thereby allow for the possibility of exploding gradients (as LSTMs or PLRNNs in fact do). This problem is furthermore practically highly relevant, as most time series we encounter in nature, and many from man-made systems as well, are inherently chaotic.