Generalized Teacher Forcing for Learning Chaotic Dynamics

Florian Hess[1 2 3] Zahra Monfared[1 3] Manuel Brenner[1 2] Daniel Durstewitz[1 2 4]

[1]Department of Theoretical Neuroscience, Central Institute of Mental Health, Mannheim

[2]Faculty of Physics and Astronomy, Heidelberg University

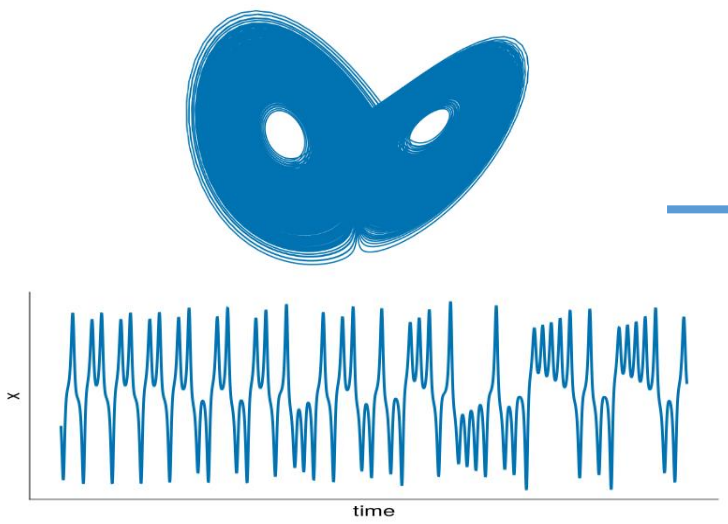
[3]Cluster of Excellence STRUCTURES, Heidelberg University

[4] Interdisciplinary Center for Scientific Computing, Heidelberg University

ICML
International Conference On Machine Learning

Dynamical Systems Reconstruction (DSR)

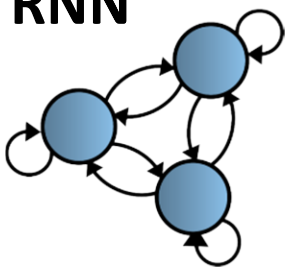
Observe system at discrete times $\{t_n\}$



Infer

$\boldsymbol{\theta}, \boldsymbol{\phi}$

RNN



$$\begin{pmatrix} \boldsymbol{z}_t = \boldsymbol{F}_{\boldsymbol{\theta}}(\boldsymbol{z}_{t-1}) \\ \boldsymbol{x}_t = \boldsymbol{G}_{\boldsymbol{\phi}}(\boldsymbol{z}_t) \end{pmatrix}$$

Generate

Same geometrical and temporal properties

Empirical data often come with …

- Observational and dynamical **noise**
- **Multiple** temporal and spatial time **scales**
- **Non-stationarity**

⋮

… and almost always **chaotic**!

Chaotic Dynamics and Loss Gradients

[1] showed: Training RNNs via BPTT on chaotic data is ill-posed:

Generic RNN:
$$\boldsymbol{z}_t = \boldsymbol{F}_{\boldsymbol{\theta}}(\boldsymbol{z}_{t-1}, \boldsymbol{s}_t)$$

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

Maximum Lyapunov exponent of an RNN orbit $\boldsymbol{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T, \ldots\}$:

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

$\lambda_{max} > 0$ necessary condition for chaos!

Backpropagation through time (BPTT) with loss $L = \sum_{t=1}^{T} L_t$

$$\frac{\partial L_t}{\partial \theta_i} = \sum_{r=1}^{t} \frac{\partial L_t}{\partial \boldsymbol{z}_t} \frac{\partial \boldsymbol{z}_t}{\partial \boldsymbol{z}_r} \frac{\partial^+ \boldsymbol{z}_r}{\partial \theta_i}$$

with

$$\frac{\partial \boldsymbol{z}_t}{\partial \boldsymbol{z}_r} = \prod_{k=0}^{t-r-1} \boldsymbol{J}_{t-k}$$

Loss gradients during training on chaotic data will **inevitably explode** for $T \to \infty$.

[1] Mikhaeil et al. (NIPS, 2022). On the difficulty of learning chaotic dynamics with RNNs.

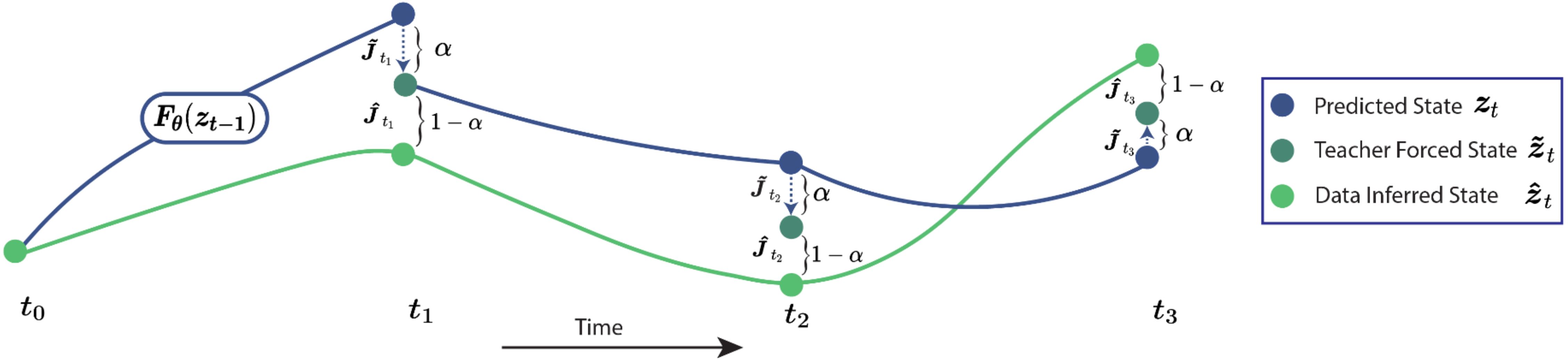
Generalized Teacher Forcing (GTF)

During training, GTF [2] linearly interpolates between RNN state $\boldsymbol{z}_t$ and data-inferred state $\hat{\boldsymbol{z}}_t$ with parameter $0 \leq \alpha \leq 1$

$$\boxed{\begin{aligned} \boldsymbol{z}_t &= \boldsymbol{F}_{\boldsymbol{\theta}}(\tilde{\boldsymbol{z}}_{t-1}) \\ \tilde{\boldsymbol{z}}_{t-1} &= (1-\alpha)\boldsymbol{z}_{t-1} + \alpha\hat{\boldsymbol{z}}_{t-1} \end{aligned}}$$

Jacobian factorizes

$$\boxed{\boldsymbol{J}_t = \frac{\partial \boldsymbol{z}_t}{\partial \boldsymbol{z}_{t-1}} = \frac{\partial \boldsymbol{F}_{\boldsymbol{\theta}}(\tilde{\boldsymbol{z}}_{t-1})}{\partial \tilde{\boldsymbol{z}}_{t-1}} \frac{\partial \tilde{\boldsymbol{z}}_{t-1}}{\partial \boldsymbol{z}_{t-1}} = \tilde{\boldsymbol{J}}_t(1-\alpha)}$$



[2] Doya (1992, IEEE). Bifurcations in the learning of recurrent neural networks.

Generalized Teacher Forcing (GTF)

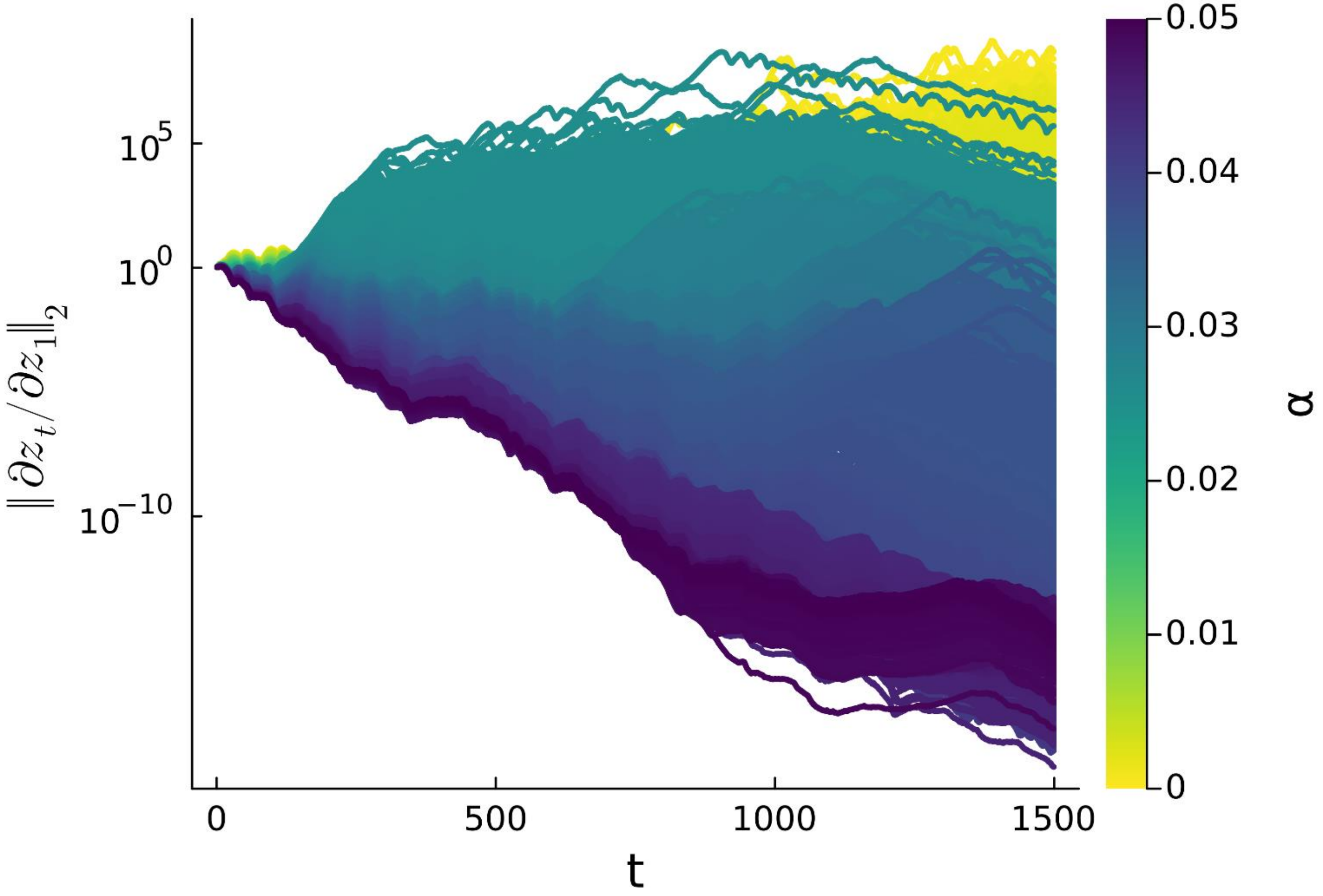
This allows α to control the Jacobian product norm during training.

$$\boxed{\frac{\partial \boldsymbol{z}_t}{\partial \boldsymbol{z}_r} = (1-\alpha)^{t-r} \prod_{k=0}^{t-r-1} \tilde{J}_{t-r}}$$

Choosing

$$\boxed{\alpha = \alpha^* := 1 - \frac{1}{\tilde{\sigma}_{max}}}$$

leads to **strictly all-time bounded gradients during BPTT training.**

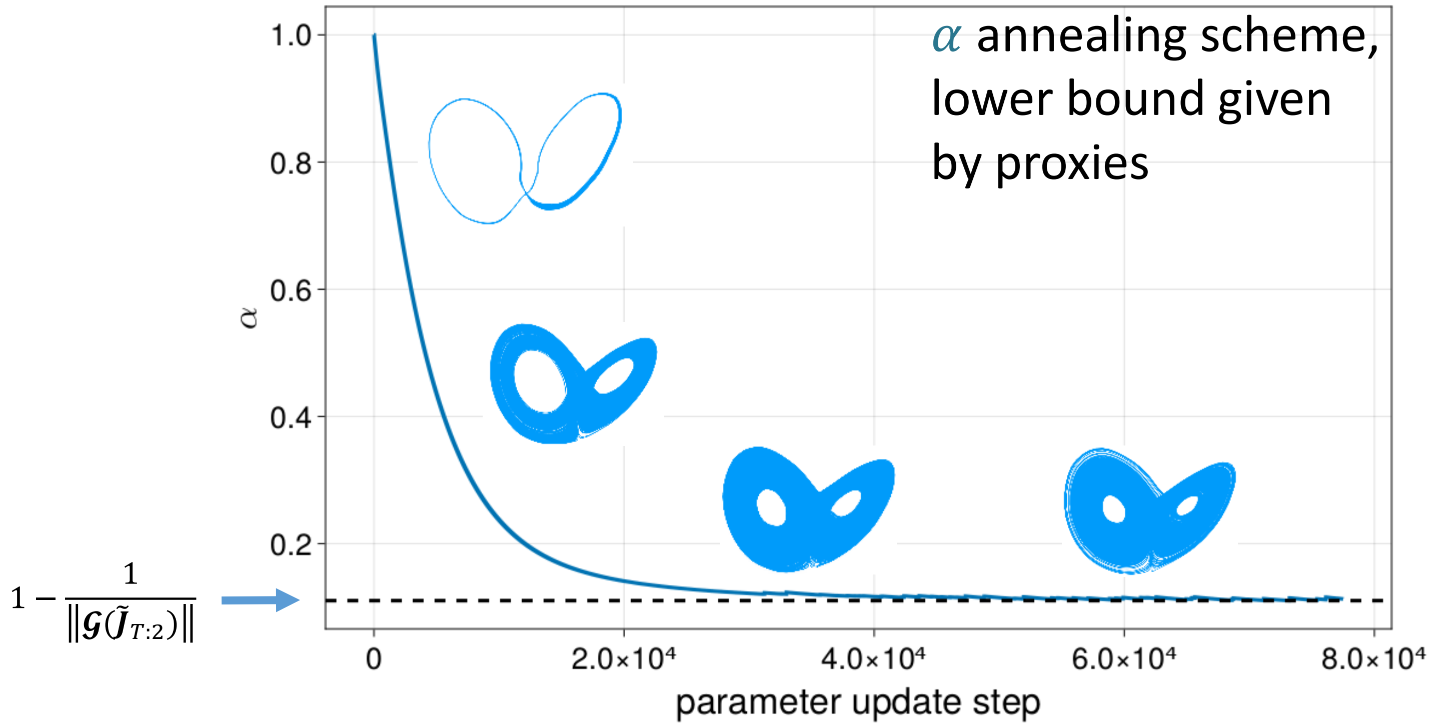


Adaptive GTF (aGTF)

Computing α^* in practice is too costly, use data-based proxies instead

$$\boxed{\frac{\partial \boldsymbol{z}_T}{\partial \boldsymbol{z}_1} = (1-\alpha)^{\mathrm{T}-1} \prod_{k=0}^{T-2} \tilde{\boldsymbol{J}}_{T-k} \overset{!}{=} \boldsymbol{I}} \longrightarrow \boxed{\alpha = 1 - \frac{1}{\left\| \mathcal{G}(\tilde{\boldsymbol{J}}_{T:2}) \right\|}}$$

with $\mathcal{G}(\tilde{\boldsymbol{J}}_{T:2}) := \left(\prod_{k=0}^{T-2} \tilde{\boldsymbol{J}}_{T-k}\right)^{\frac{1}{T-1}} \approx \frac{1}{T-1} \sum_{t=2}^{T} \tilde{\boldsymbol{J}}_t$



GTF is only used in training, **not** testing!

Model Reformulation: shallow PLRNN

- Reformulation of the dendritic piecewise linear RNN (dendPLRNN, [3]) into a 1-hidden-layer design

$$\boxed{\boldsymbol{z}_t = \boldsymbol{F}_{\boldsymbol{\theta}}(\boldsymbol{z}_{t-1}) = \boldsymbol{A}\boldsymbol{z}_{t-1} + \boldsymbol{W}_1 ReLU(\boldsymbol{W}_2 \boldsymbol{z}_{t-1} + \mathbf{h}_2) + \mathbf{h}_1}$$

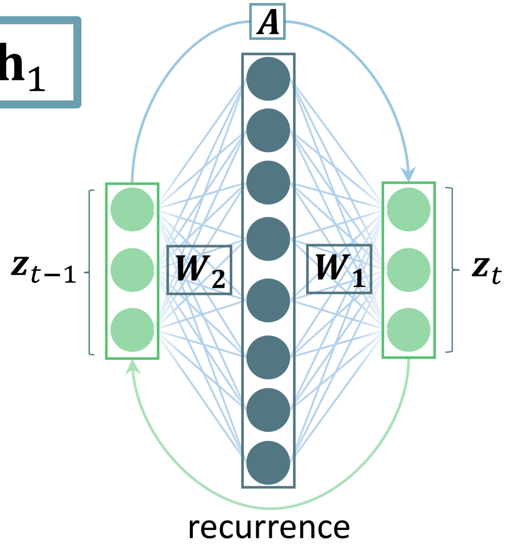
$\boldsymbol{A} \in \mathbb{R}^{M \times M}$ diagonal

$\boldsymbol{W}_1 \in \mathbb{R}^{M \times L}, \boldsymbol{W}_2 \in \mathbb{R}^{L \times M}$

$\boldsymbol{h}_1 \in \mathbb{R}^{M}, \boldsymbol{h}_2 \in \mathbb{R}^{L}$

M: model's state space dimensionality

L: hidden layer size



- Can learn DS in **very low-dimensional state spaces**
- Retains **semi-analytic access to fixed points and k-cycles**

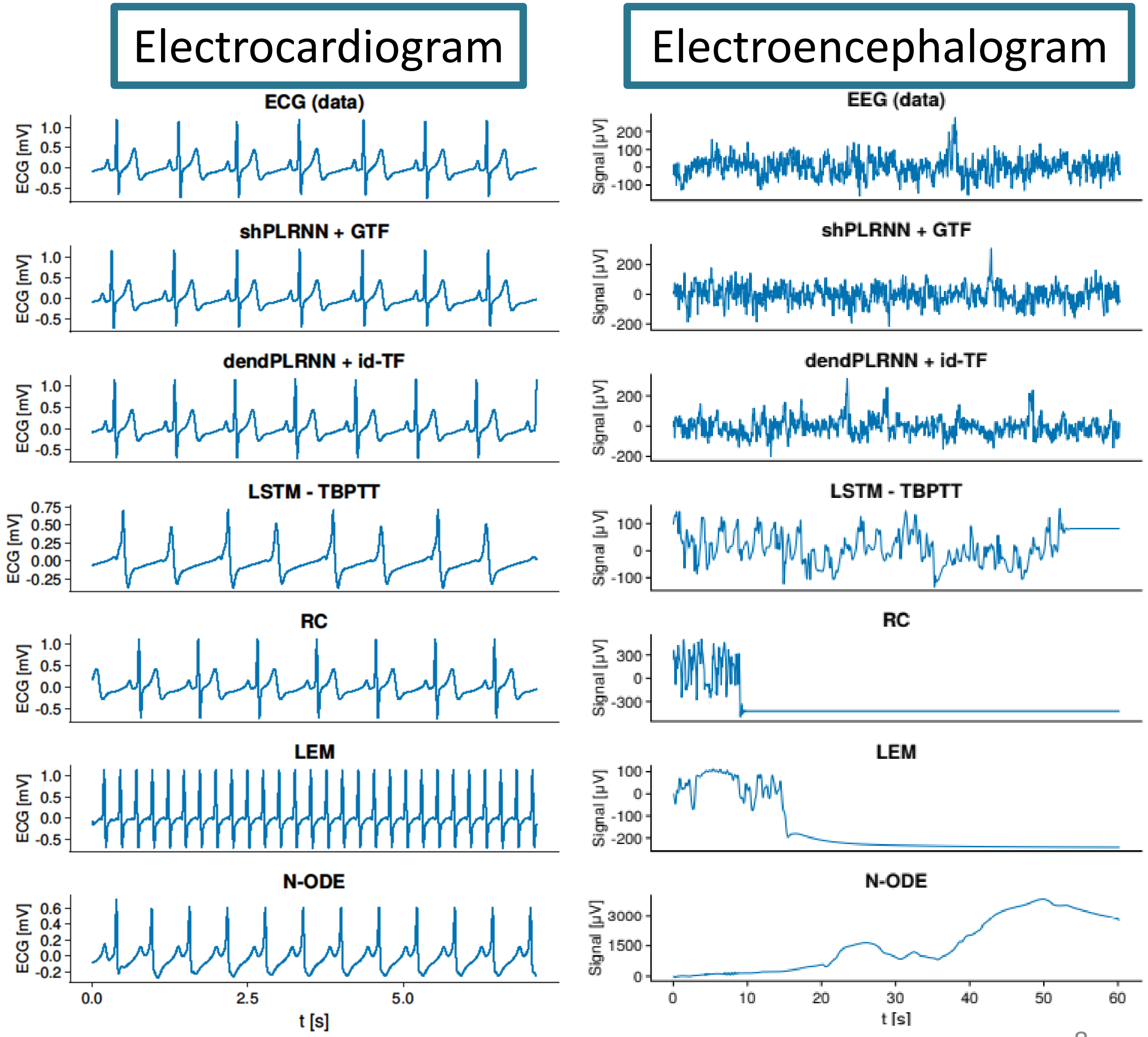
[3] Brenner & Hess et al. (2022, ICML). Tractable Dendritic RNNs for Reconstructing Nonlinear Dynamical Systems.

Reconstruction of DS from empirical data

- Competitive performance of shPLRNN + GTF compared to 4 major classes of DSR algorithms

  - Gated RNN architectures (LSTMs)
  - Reservoir Computers (RC)
  - Library-based methods (SINDy)
  - ODE-based RNNs like Long Expressive Memory (LEM) and Neural ODEs (N-ODE)

Freely generated orbits **after** training!



Reconstruction of DS from empirical data

Agreement in attractor geometry:

$$D_{stsp} = \int_{\boldsymbol{x}\in\mathbb{R}^N} p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}$$

Hellinger distance between power spectra:

$$D_H = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \int_{-\infty}^{\infty}\sqrt{f_i(\omega)g_i(\omega)}\,d\omega\right)^{\frac{1}{2}}$$

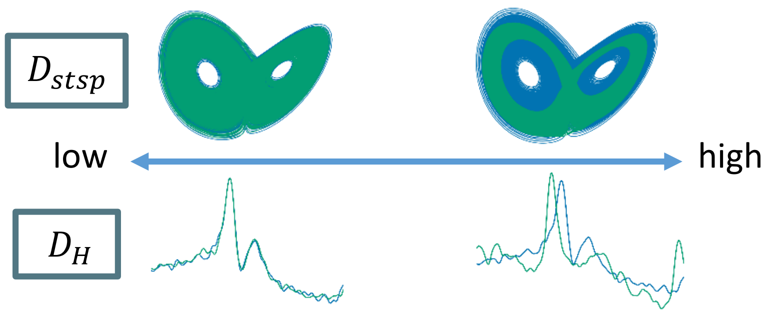
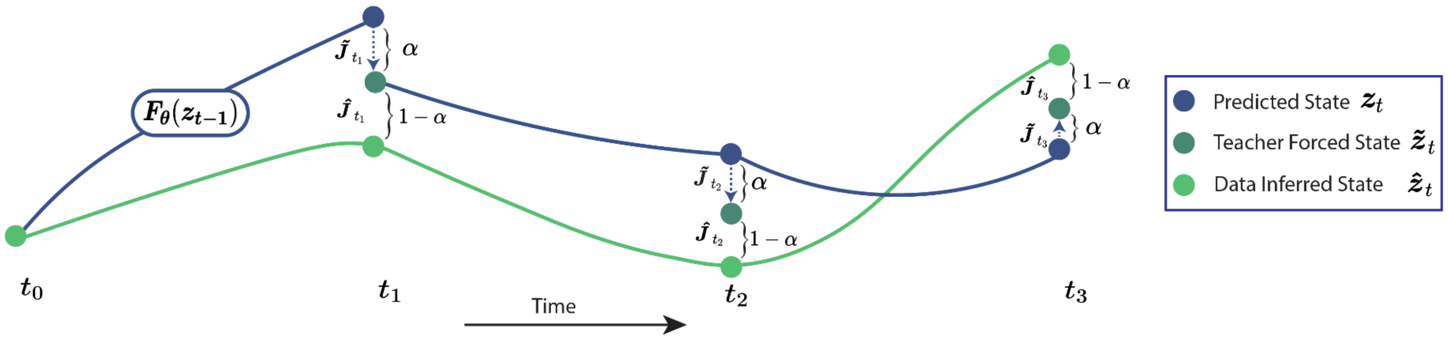


Table 1. SOTA comparisons. Reported values are median $\pm$ median absolute deviation over 20 independent training runs. 'dim' refers to the model's state space dimensionality (number of dynamical variables). $|\theta|$ denotes the total number of *trainable* parameters.

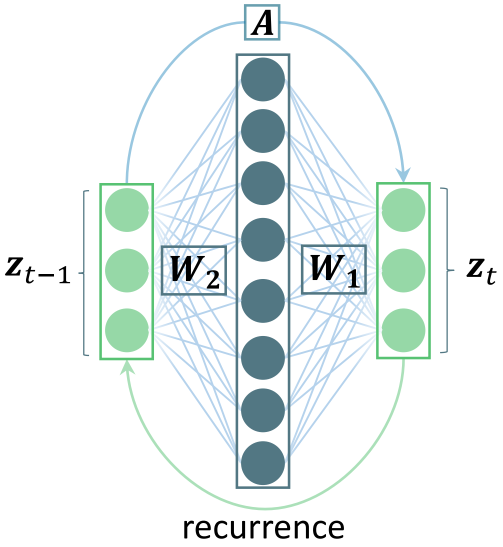
Dataset	Method	$D_{stsp}\downarrow$	$D_H\downarrow$	PE(20) $\downarrow$	dim	$\lvert\theta\rvert$
ECG (5d)	shPLRNN + GTF	$\mathbf{4.3 \pm 0.6}$	$\mathbf{0.34 \pm 0.02}$	$\mathbf{(2.4 \pm 0.1) \cdot 10^{-3}}$	5	2785
	shPLRNN + aGTF	$\mathbf{4.5 \pm 0.4}$	$\mathbf{0.34 \pm 0.02}$	$\mathbf{(2.4 \pm 0.2) \cdot 10^{-3}}$	5	2785
	shPLRNN + STF	7.1 ± 1.8	0.38 ± 0.03	$(5 \pm 2) \cdot 10^{-3}$	5	2785
	dendPLRNN + id-TF	5.8 ± 0.6	0.37 ± 0.06	$(4.0 \pm 0.4) \cdot 10^{-3}$	35	3245
	RC	5.3 ± 1.7	0.39 ± 0.05	$(4 \pm 1) \cdot 10^{-3}$	1000	5000
	LSTM-TBPTT	15.2 ± 0.5	0.73 ± 0.02	$(2.5 \pm 0.5) \cdot 10^{-2}$	70	5920
	SINDy	diverging	diverging	diverging	5	3960
	N-ODE	12.2 ± 0.7	0.7 ± 0.03	$(4.1 \pm 0.1) \cdot 10^{-1}$	5	4955
	LEM	16.3 ± 0.2	0.56 ± 0.04	$(7.4 \pm 0.1) \cdot 10^{-1}$	62	4872
EEG (64d)	shPLRNN + GTF	$\mathbf{2.1 \pm 0.2}$	$\mathbf{0.11 \pm 0.01}$	$(5.5 \pm 0.1) \cdot 10^{-1}$	16	17952
	shPLRNN + aGTF	$\mathbf{2.4 \pm 0.2}$	$\mathbf{0.13 \pm 0.01}$	$(5.4 \pm 0.6) \cdot 10^{-1}$	16	17952
	shPLRNN + STF	14 ± 7	0.50 ± 0.16	$\mathbf{(2.5 \pm 0.3) \cdot 10^{-1}}$	16	17952
	dendPLRNN + id-TF	3 ± 1	0.13 ± 0.04	$(3.4 \pm 0.1) \cdot 10^{-1}$	105	18099
	RC	14 ± 7	0.54 ± 0.15	$(5.9 \pm 0.3) \cdot 10^{-1}$	448	28672
	LSTM-TBPTT	30 ± 21	0.2 ± 0.1	$(9.2 \pm 2.3) \cdot 10^{-1}$	160	51584
	SINDy	diverging	diverging	diverging	64	133120
	N-ODE	20 ± 0.5	0.47 ± 0.01	$(5.5 \pm 0.2) \cdot 10^{-1}$	64	17995
	LEM	10.2 ± 1.5	0.38 ± 0.06	$(8.2 \pm 0.6) \cdot 10^{-1}$	76	18304

Conclusion

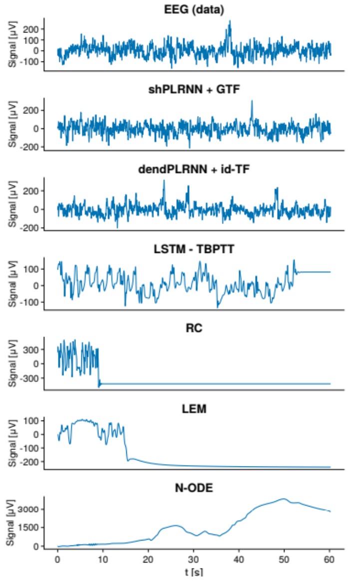
Generalized Teacher Forcing:
Training algorithm to solve the exploding gradient problem in BPTT training on chaotic systems.



Shallow PLRNN:
Model reformulation that allows for low-dimensional state spaces while providing analytic access to FPs and k-cycles.



Reconstructing DS from empirical data:
Competitive algorithm + model for dynamical systems reconstruction compared to other SOTA methods in the field.



Thanks for your attention!

Deutsche Forschungsgemeinschaft
German Research Foundation

This work was funded by the German Research Foundation (DFG) within Germany's Excellence Strategy EXC 2181/1 – 390900948 (STRUCTURES), by DFG grants Du354/10-1 & Du354/15-1 to DD, and by the European Union Horizon-2020 consortium SC1-DTH-13-2020 (IMMERSE).