

# Identifying nonlinear dynamical systems with multiple time scales and long-range dependencies

**Dominik Schmidt, Georgia Koppe, Zahra Monfared, Max Beutelspacher,  
Daniel Durstewitz**

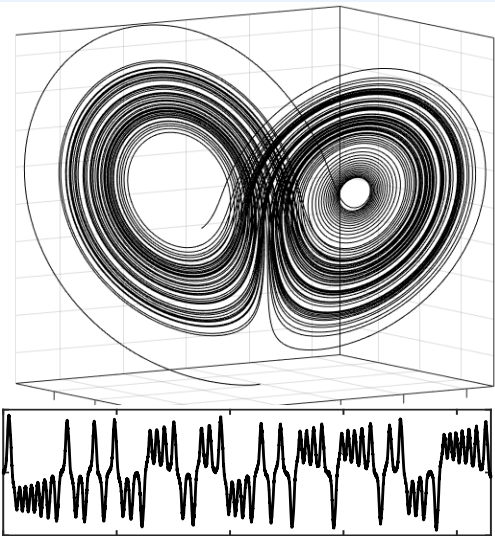
Department of Theoretical Neuroscience &  
Bernstein Center for Computational Neuroscience  
Central Institute of Mental Health/Medical Faculty Mannheim, Heidelberg University

Ninth International Conference on Learning Representations (ICLR 2021)

# Nonlinear dynamical systems reconstruction

## dynamical system

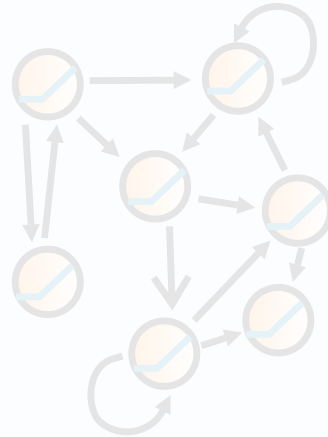
$$\begin{aligned}\dot{x}_1 &= a(x_2 - x_1) \\ \dot{x}_2 &= x_1(b - x_3) - x_2 \\ \dot{x}_3 &= x_1x_2 - cx_3\end{aligned}$$



train

## RNN

$$z_t = F_\theta(z_{t-1}, s_t)$$

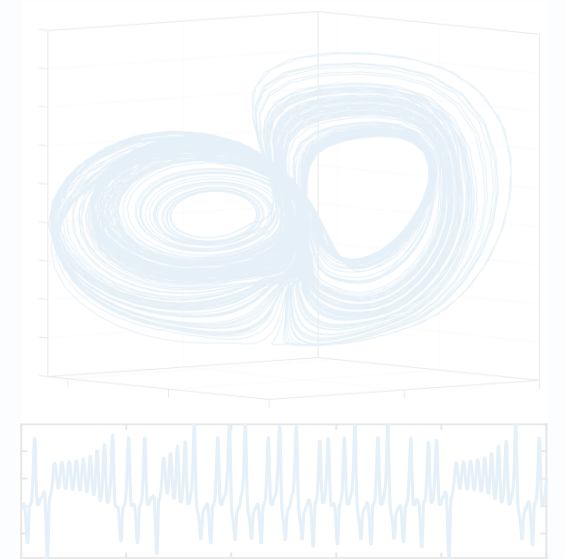


universal approximator  
of dynamical systems  
(Hanson & Raginsky, 2020)

generate

## simulated trajectory

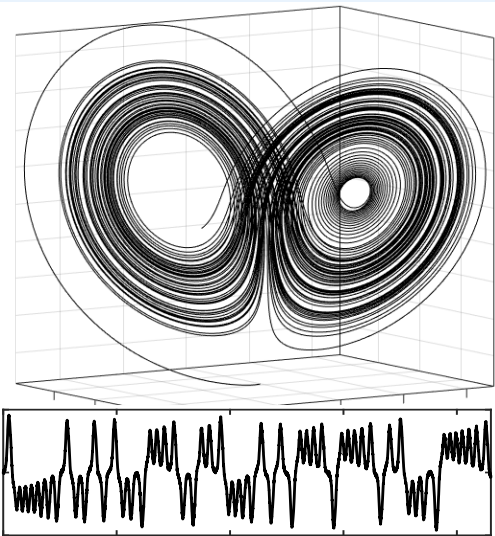
$$\begin{aligned}z_t &= Az_{t-1} \\ &\quad + W \max(z_{t-1}, 0) + h\end{aligned}$$



# Nonlinear dynamical systems reconstruction

## dynamical system

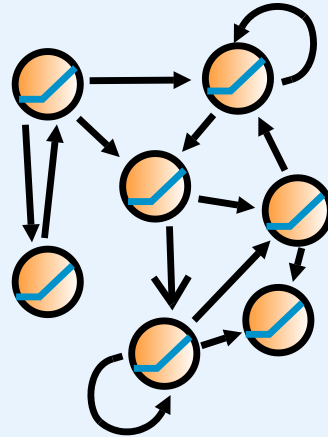
$$\begin{aligned}\dot{x}_1 &= a(x_2 - x_1) \\ \dot{x}_2 &= x_1(b - x_3) - x_2 \\ \dot{x}_3 &= x_1x_2 - cx_3\end{aligned}$$



train

## RNN

$$z_t = F_\theta(z_{t-1}, s_t)$$

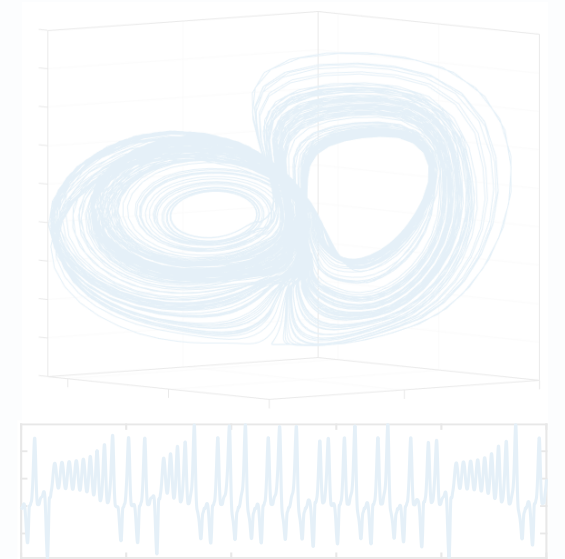


**universal approximator  
of dynamical systems**  
(Hanson & Raginsky, 2020)

generate

## simulated trajectory

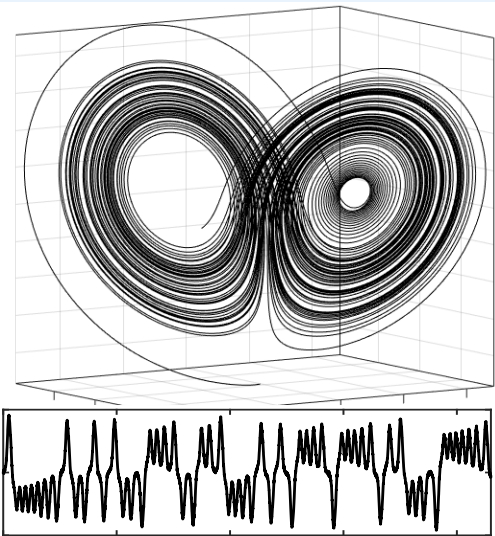
$$\begin{aligned}z_t &= Az_{t-1} \\ &\quad + W \max(z_{t-1}, 0) + h\end{aligned}$$



# Nonlinear dynamical systems reconstruction

## dynamical system

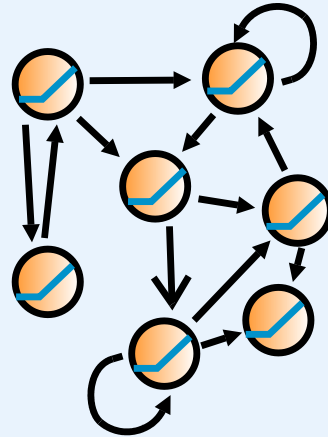
$$\begin{aligned}\dot{x}_1 &= a(x_2 - x_1) \\ \dot{x}_2 &= x_1(b - x_3) - x_2 \\ \dot{x}_3 &= x_1x_2 - cx_3\end{aligned}$$



train

## RNN

$$z_t = F_\theta(z_{t-1}, s_t)$$

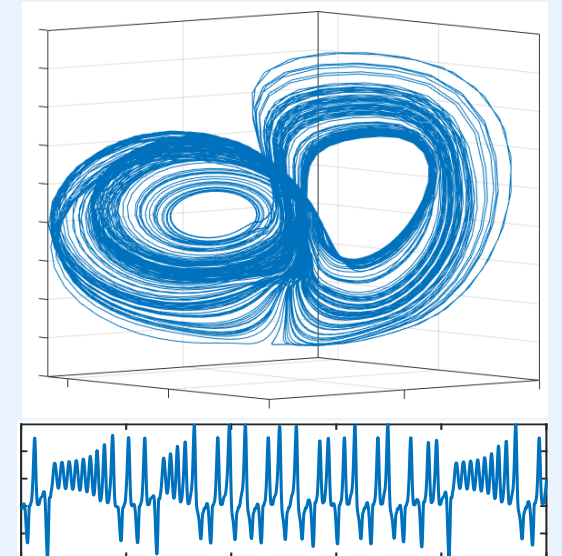


universal approximator  
of dynamical systems  
(Hanson & Raginsky, 2020)

generate

## simulated trajectory

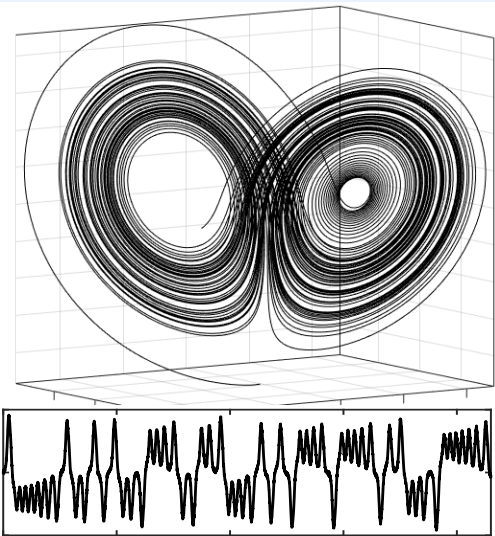
$$\begin{aligned}z_t &= Az_{t-1} \\ &\quad + W \max(z_{t-1}, 0) + h\end{aligned}$$



# Nonlinear dynamical systems reconstruction

## dynamical system

$$\begin{aligned}\dot{x}_1 &= a(x_2 - x_1) \\ \dot{x}_2 &= x_1(b - x_3) - x_2 \\ \dot{x}_3 &= x_1x_2 - cx_3\end{aligned}$$

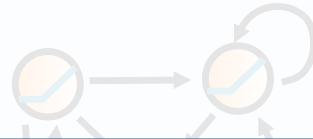


Same geometrical and temporal structure

universal approximator  
of dynamical systems  
(Hanson & Raginsky, 2020)

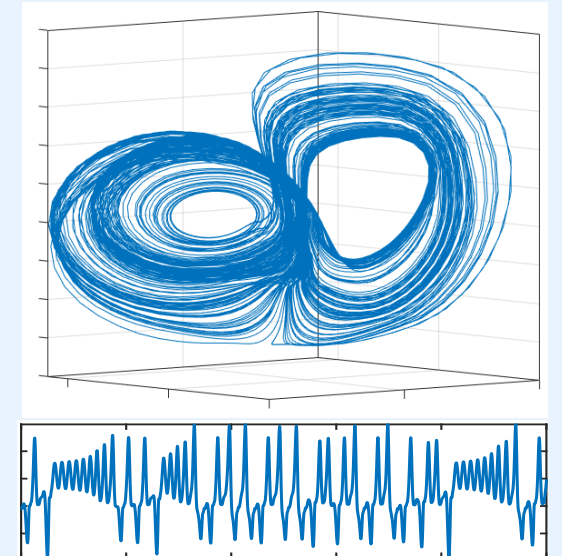
## RNN

$$z_t = F_\theta(z_{t-1}, s_t)$$



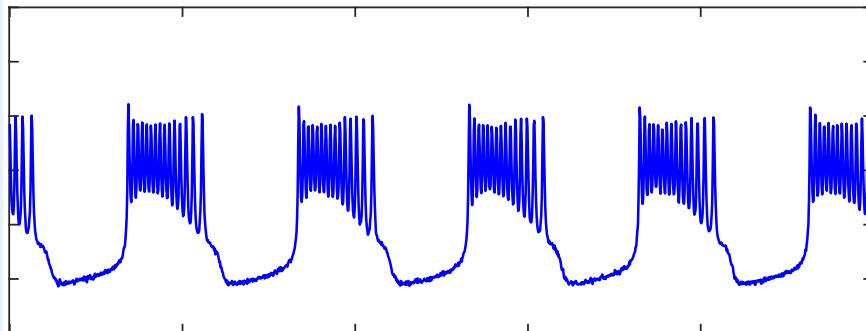
## simulated trajectory

$$\begin{aligned}z_t &= Az_{t-1} \\ &\quad + W \max(z_{t-1}, 0) + h\end{aligned}$$



# Challenges in training

multiple time  
scales & long-  
range  
dependencies



vanishing/  
exploding  
gradients

$$\frac{\partial z_t}{\partial z_k} = \prod_{t \geq i > k} \frac{\partial z_i}{\partial z_{i-1}} = \prod_{t \geq i > k} W \text{diag}(\phi'(z_{i-1}))$$



Previous solutions

Special architectures:  
e.g. LSTM, GRU

Hochreiter & Schmidhuber (1997)  
Cho et al. (2004)

Initialization of W

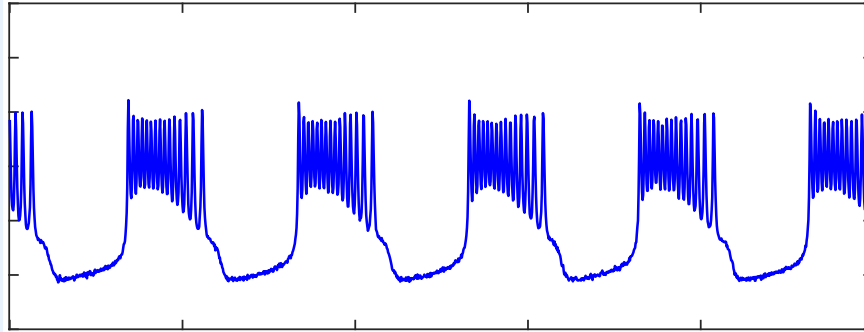
Le et al. (2015)  
Talathi & Vartak (2016)

Unitary/orthogonal RNN

Arjovsky et al. (2016)  
Helfrich et al. (2018)

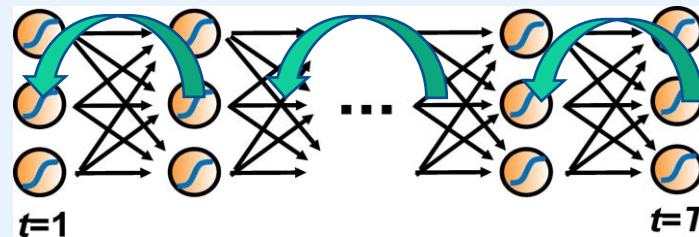
# Challenges in training

multiple time  
scales & long-  
range  
dependencies



vanishing/  
exploding  
gradients

$$\frac{\partial z_t}{\partial z_k} = \prod_{t \geq i > k} \frac{\partial z_i}{\partial z_{i-1}} = \prod_{t \geq i > k} W \text{diag}(\phi'(z_{i-1}))$$



Previous solutions

Special architectures:  
e.g. LSTM, GRU

Hochreiter & Schmidhuber (1997)  
Cho et al. (2004)

Initialization of W

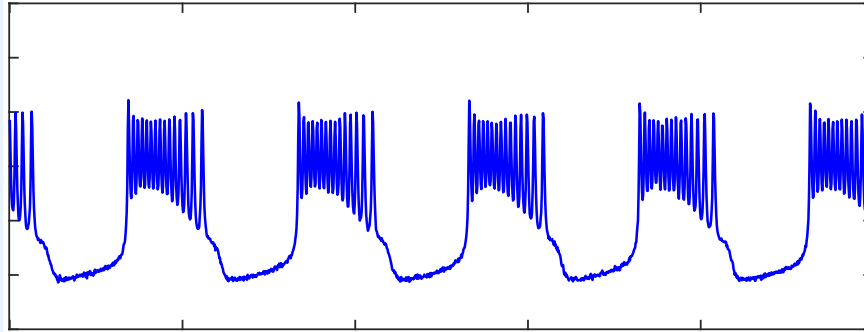
Le et al. (2015)  
Talathi & Vartak (2016)

Unitary/orthogonal RNN

Arjovsky et al. (2016)  
Helfrich et al. (2018)

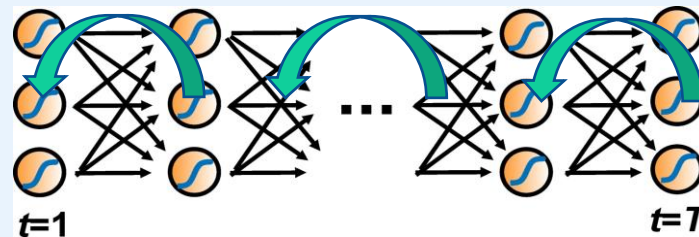
# Challenges in training

multiple time  
scales & long-  
range  
dependencies



vanishing/  
exploding  
gradients

$$\frac{\partial z_t}{\partial z_k} = \prod_{t \geq i > k} \frac{\partial z_i}{\partial z_{i-1}} = \prod_{t \geq i > k} W \text{diag}(\phi'(z_{i-1}))$$



Previous solutions

Special architectures:  
e.g. LSTM, GRU

Hochreiter & Schmidhuber (1997)  
Cho et al. (2004)

Initialization of W

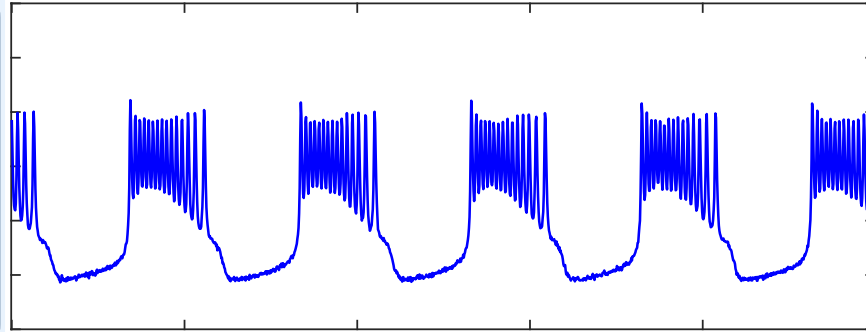
Le et al. (2015)  
Talathi & Vartak (2016)

Unitary/orthogonal RNN

Arjovsky et al. (2016)  
Helfrich et al. (2018)

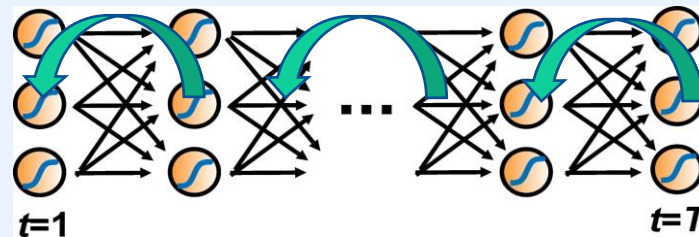
# Challenges in training

multiple time  
scales & long-  
range  
dependencies



vanishing/  
exploding  
gradients

$$\frac{\partial z_t}{\partial z_k} = \prod_{t \geq i > k} \frac{\partial z_i}{\partial z_{i-1}} = \prod_{t \geq i > k} W \text{diag}(\phi'(z_{i-1}))$$



**Previous solutions**

**Special architectures:  
e.g. LSTM, GRU**

Hochreiter & Schmidhuber (1997)  
Cho et al. (2004)

**Initialization of W**

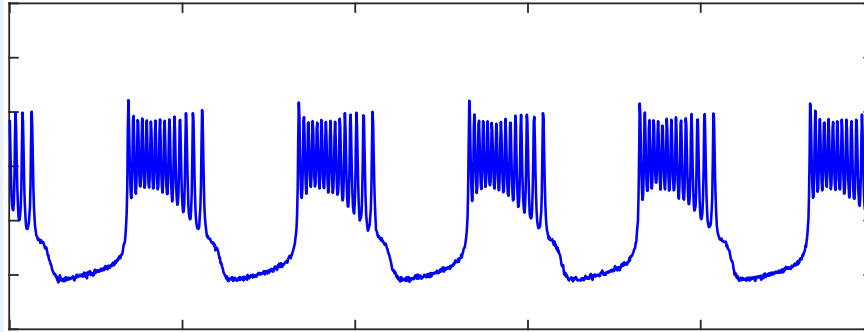
Le et al. (2015)  
Talathi & Vartak (2016)

**Unitary/orthogonal RNN**

Arjovsky et al. (2016)  
Helfrich et al. (2018)

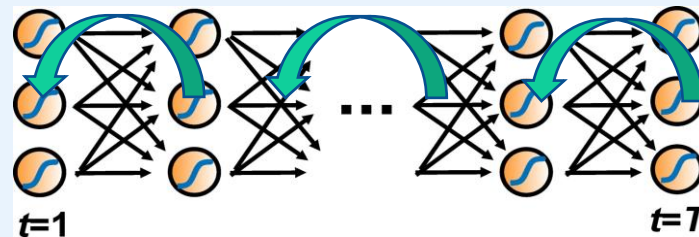
# Challenges in training

multiple time  
scales & long-  
range  
dependencies



vanishing/  
exploding  
gradients

$$\frac{\partial z_t}{\partial z_k} = \prod_{t \geq i > k} \frac{\partial z_i}{\partial z_{i-1}} = \prod_{t \geq i > k} W \text{diag}(\phi'(z_{i-1}))$$



Previous solutions

Special architectures:  
e.g. LSTM, GRU

Mathematically complex

Initialization of W

Effect fades during training

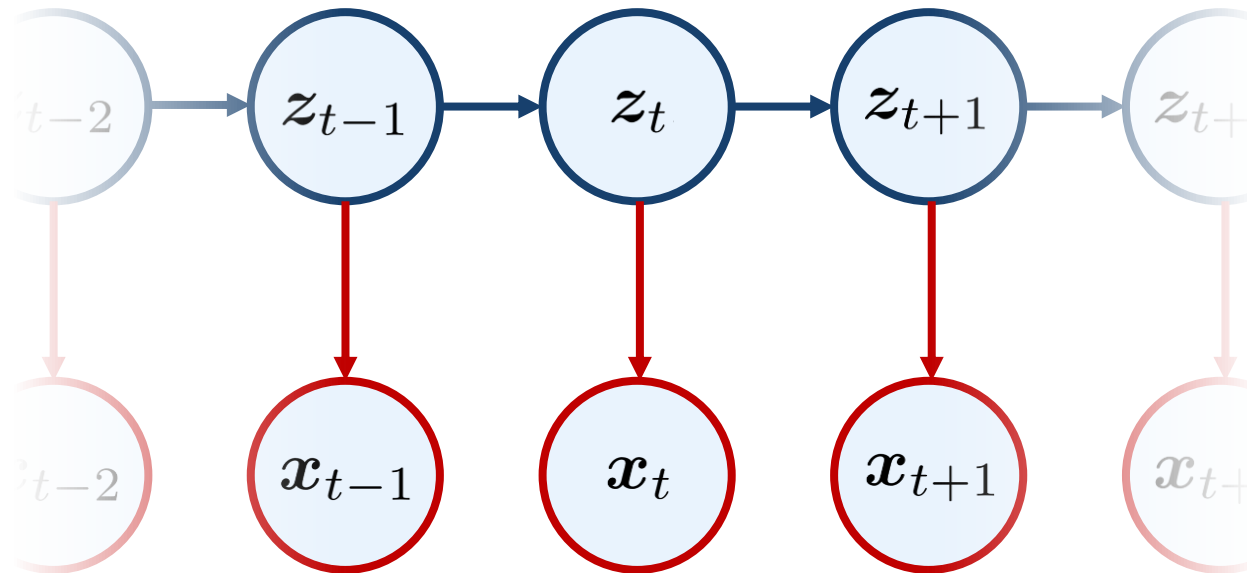
Unitary/orthogonal RNN

Too restrictive

# Piecewise Linear RNN (PLRNN)

$$z_t = Az_{t-1} + W \max(z_{t-1}, 0) + Cs_t + h + \varepsilon_t$$

$$x_t = Bg(z_t) + \eta_t$$



# Piecewise Linear RNN (PLRNN)

$$\mathbf{z}_t = \mathbf{A}\mathbf{z}_{t-1} + \mathbf{W} \max(\mathbf{z}_{t-1}, \mathbf{0}) + \mathbf{C}\mathbf{s}_t + \mathbf{h} + \boldsymbol{\varepsilon}_t$$

**DS properties analytically accessible**

**For example ...**

**Fixed points**  $\mathbf{z}^{*1} = (\mathbf{I} - \mathbf{W}_{\Omega(t^{*1})})^{-1} \mathbf{h}$

**k-cycles**  $\mathbf{z}^{*k} = \left( \mathbf{I} - \prod_{i=1}^k \mathbf{W}_{\Omega(t^{*k}-i)} \right)^{-1} \times \left[ \sum_{j=2}^k \prod_{i=1}^{k-j+1} \mathbf{W}_{\Omega(t^{*k}-i)} + \mathbf{I} \right] \mathbf{h}$

Can be transformed into equivalent continuous-time ODE system



(Monfared & Durstewitz, ICML, 2020)

## Piecewise Linear RNN (PLRNN)

$$\mathbf{z}_t = \mathbf{A}\mathbf{z}_{t-1} + \mathbf{W} \max(\mathbf{z}_{t-1}, \mathbf{0}) + \mathbf{C}\mathbf{s}_t + \mathbf{h} + \varepsilon_t$$

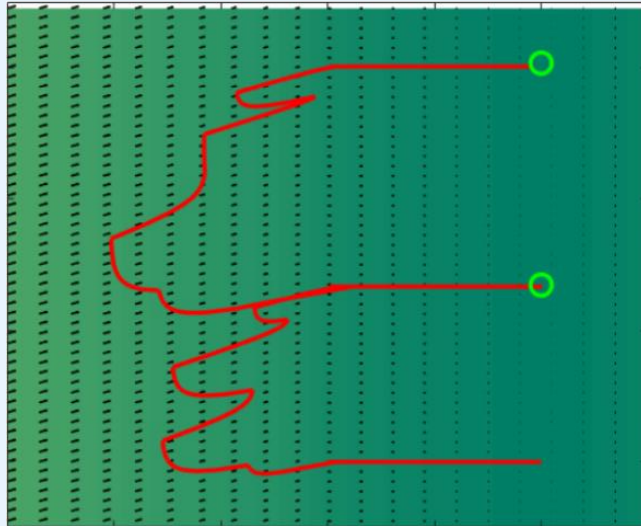
**DS properties analytically accessible**

## For example ...

**Fixed points**  $\mathbf{z}^{*1} = (\mathbf{I} - \mathbf{W}_{\Omega(t^{*1})})^{-1} \mathbf{h}$

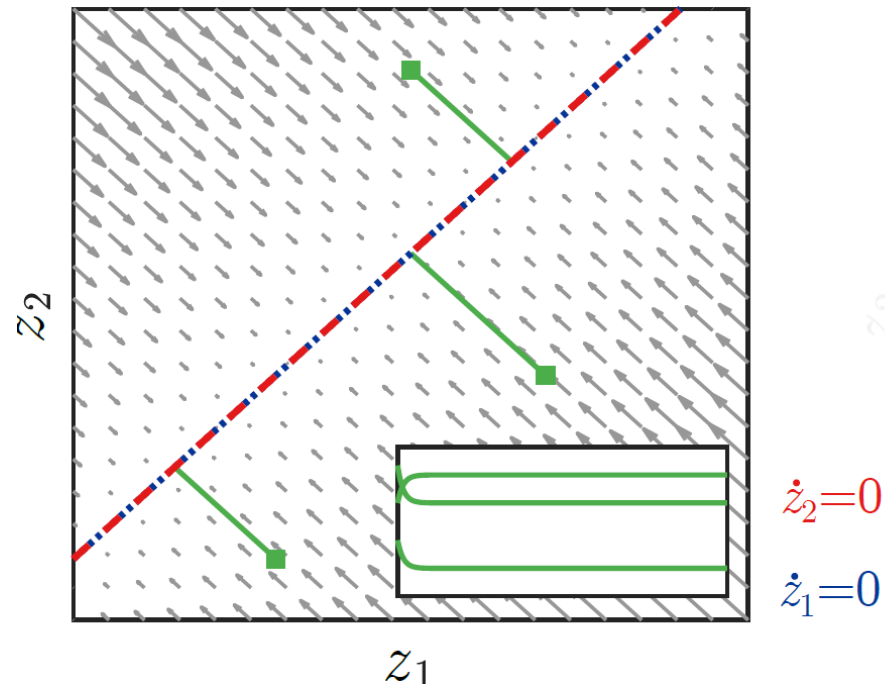
**k-cycles**  $z^{*k} = \left( I - \prod_{i=1}^k W_{\Omega(t^{*k}-i)} \right)^{-1} \times \left[ \sum_{j=2}^k \prod_{i=1}^{k-j+1} W_{\Omega(t^{*k}-i)} + I \right] h$

**Can be transformed into equivalent continuous-time ODE system**

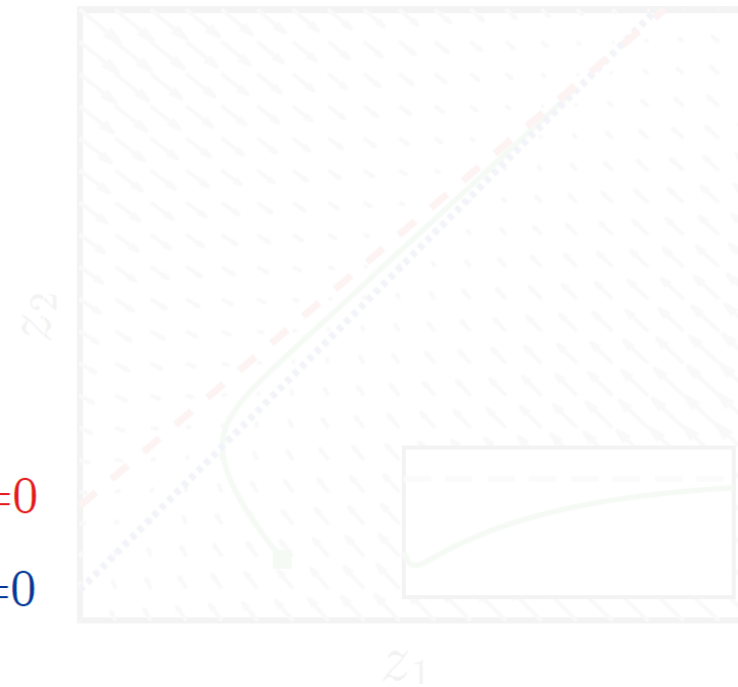


(Monfared & Durstewitz, ICML, 2020)

# Manifold-attractors for long memory & slow time scales

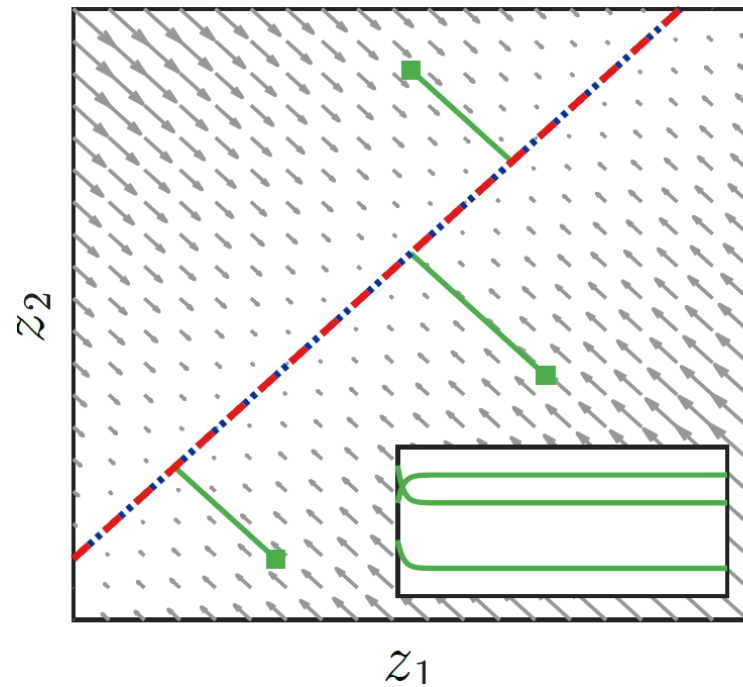


Perfect line attractor retains states indefinitely

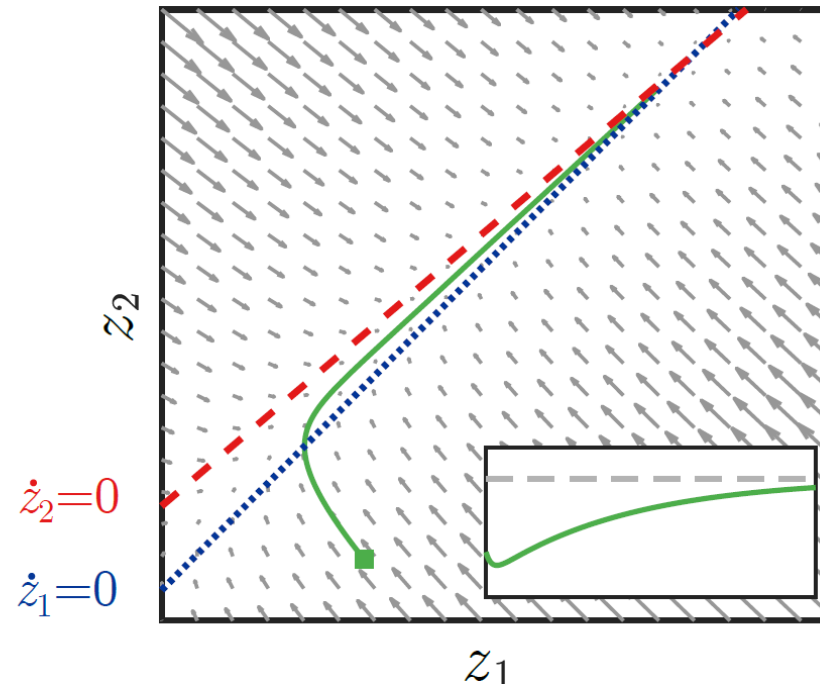


Slightly 'detuned' line attractor enables arbitrary time constants

# Manifold-attractors for long memory & slow time scales



Perfect line attractor retains states indefinitely



Slightly 'detuned' line attractor enables arbitrary time constants

# Manifold-Attractor regularization

$$\mathbf{z}_t = \mathbf{A}\mathbf{z}_{t-1} + \mathbf{W} \max(\mathbf{z}_{t-1}, \mathbf{0}) + \mathbf{C}\mathbf{s}_t + \mathbf{h} + \boldsymbol{\varepsilon}_t$$

$$\mathcal{L}_{\text{reg}} = \tau_A \sum_{i=1}^{M_{\text{reg}}} (A_{i,i} - 1)^2 + \tau_W \sum_{i=1}^{M_{\text{reg}}} \sum_{\substack{j=1 \\ j \neq i}}^M W_{i,j}^2 + \tau_h \sum_{i=1}^{M_{\text{reg}}} h_i^2$$

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & \times & 0 & 0 \\ 0 & 0 & 0 & 0 & \times & 0 \\ 0 & 0 & 0 & 0 & 0 & \times \end{pmatrix}$$

$\mathbf{A}$

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \times & \times & \times & 0 & \times & \times \\ \times & \times & \times & \times & 0 & \times \\ \times & \times & \times & \times & \times & 0 \end{pmatrix}$$

$\mathbf{W}$

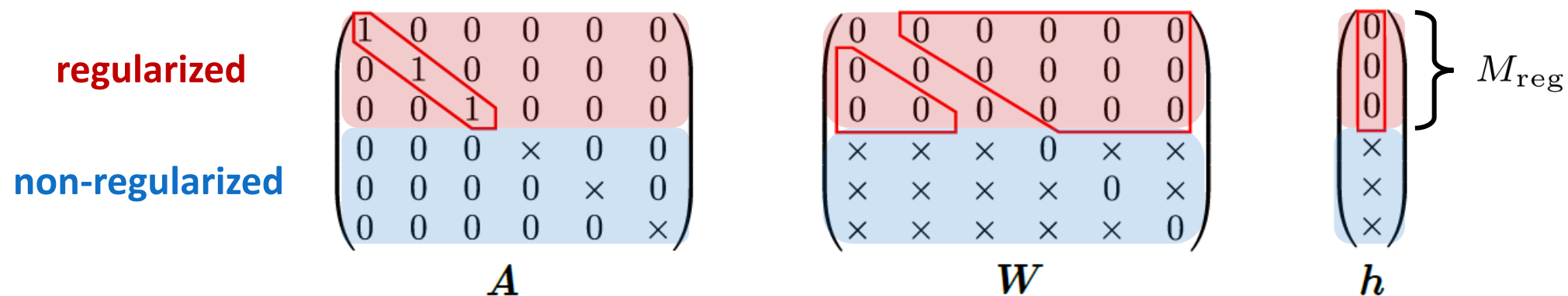
$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ \times \\ \times \\ \times \end{pmatrix}$$

$\mathbf{h}$

# Manifold-Attractor regularization

$$\mathbf{z}_t = \mathbf{A}\mathbf{z}_{t-1} + \mathbf{W} \max(\mathbf{z}_{t-1}, \mathbf{0}) + \mathbf{C}\mathbf{s}_t + \mathbf{h} + \boldsymbol{\varepsilon}_t$$

$$\mathcal{L}_{\text{reg}} = \tau_A \sum_{i=1}^{M_{\text{reg}}} (A_{i,i} - 1)^2 + \tau_W \sum_{i=1}^{M_{\text{reg}}} \sum_{\substack{j=1 \\ j \neq i}}^M W_{i,j}^2 + \tau_h \sum_{i=1}^{M_{\text{reg}}} h_i^2$$



# Theoretical analysis

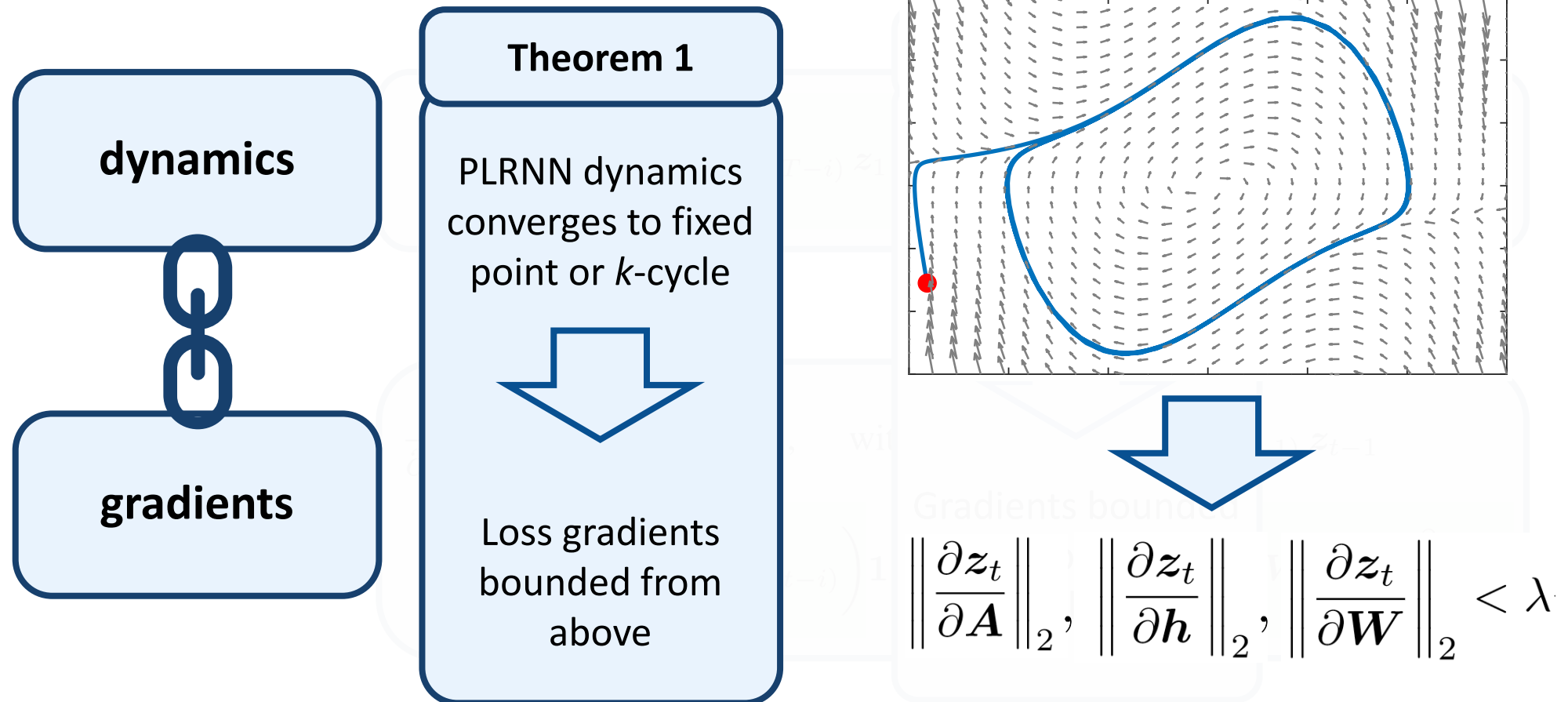
**dynamics**

$$\mathbf{z}_T = \prod_{i=1}^{T-1} \mathbf{W}_{\Omega(T-i)} \mathbf{z}_1 + \left[ \sum_{j=2}^{T-1} \prod_{i=1}^{j-1} \mathbf{W}_{\Omega(T-i)} + \mathbf{I} \right] \mathbf{h}$$

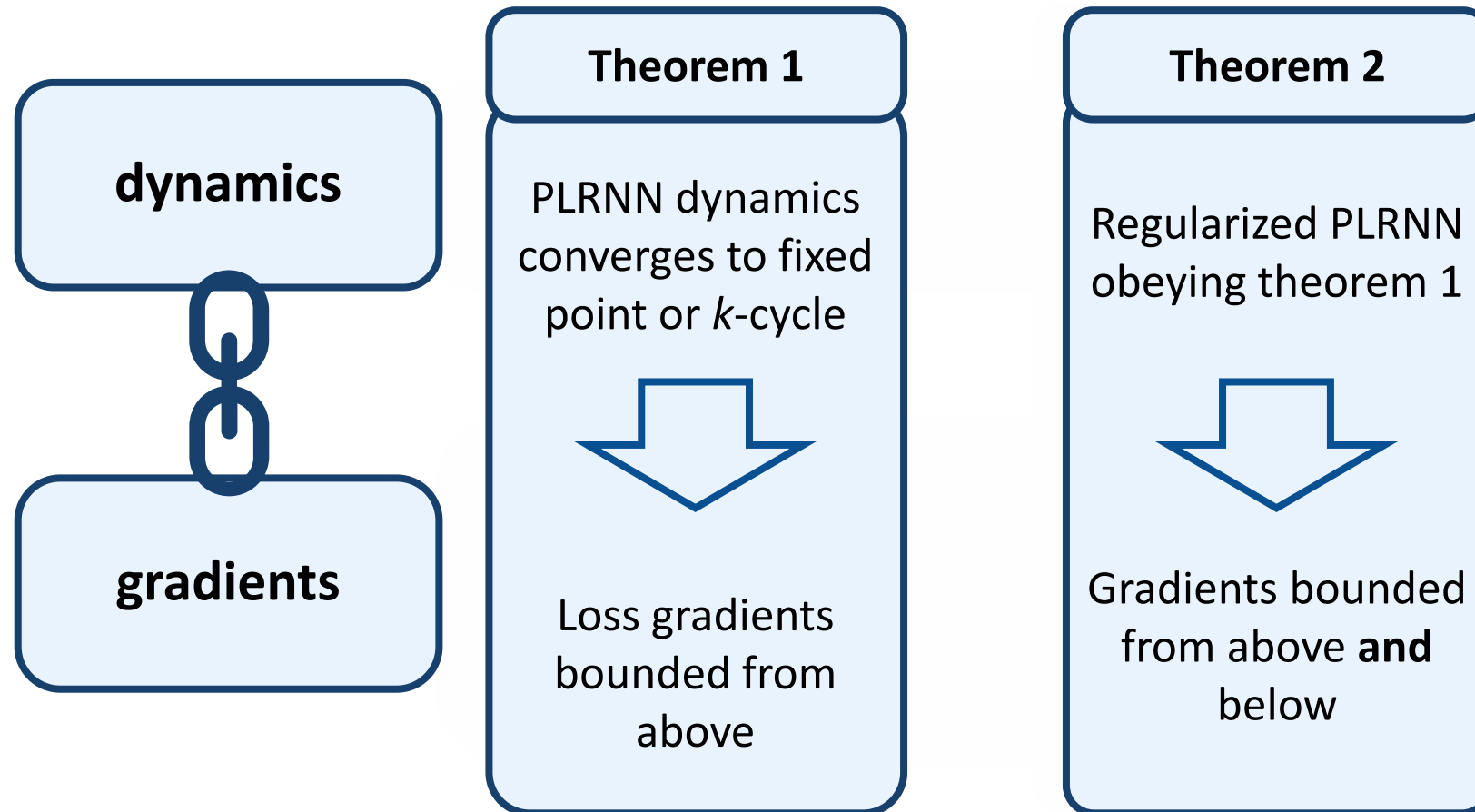
**gradients**

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial w_{mk}} &= \sum_{t=2}^T \frac{\partial \mathcal{L}_t}{\partial \mathbf{z}_t} \frac{\partial \mathbf{z}_t}{\partial w_{mk}}, \quad \text{with } \frac{\partial \mathbf{z}_t}{\partial w_{mk}} = \mathbf{1}_{(m,k)} \mathbf{D}_{\Omega(t-1)} \mathbf{z}_{t-1} \\ &+ \sum_{j=2}^{t-2} \left( \prod_{i=1}^{j-1} \mathbf{W}_{\Omega(t-i)} \right) \mathbf{1}_{(m,k)} \mathbf{D}_{\Omega(t-j)} \mathbf{z}_{t-j} + \prod_{i=1}^{t-2} \mathbf{W}_{\Omega(t-i)} \frac{\partial \mathbf{z}_2}{\partial w_{mk}} \end{aligned}$$

# Theoretical analysis

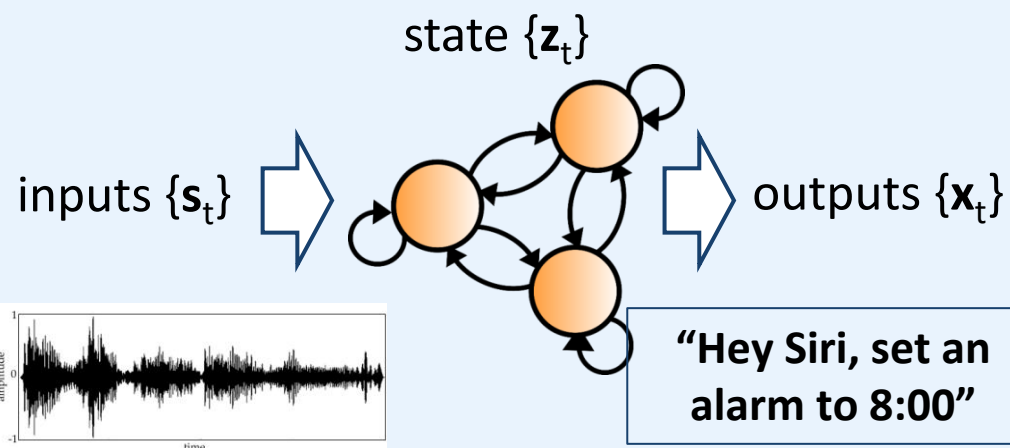


# Theoretical analysis



# Experimental setups

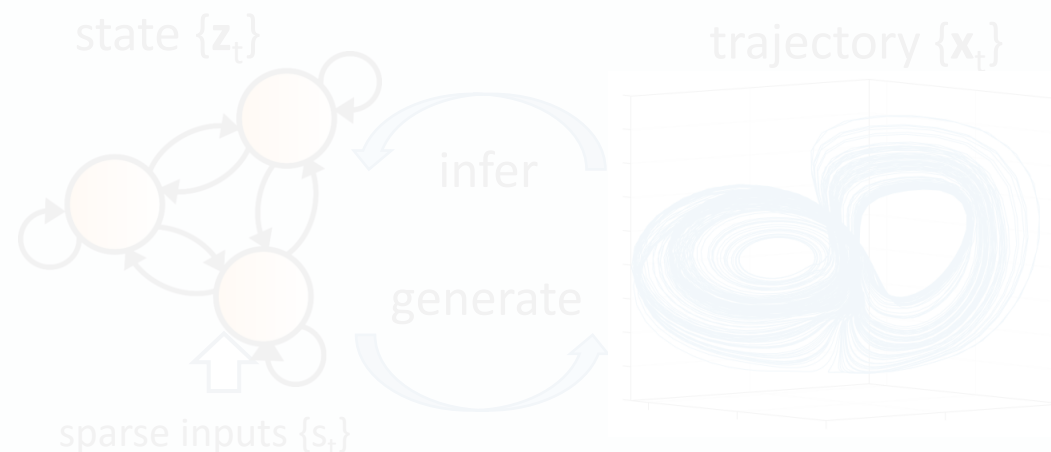
## Supervised



**Goal:**  
Good ahead  
predictions

**Training:**  
Stochastic  
gradient descent

## Unsupervised

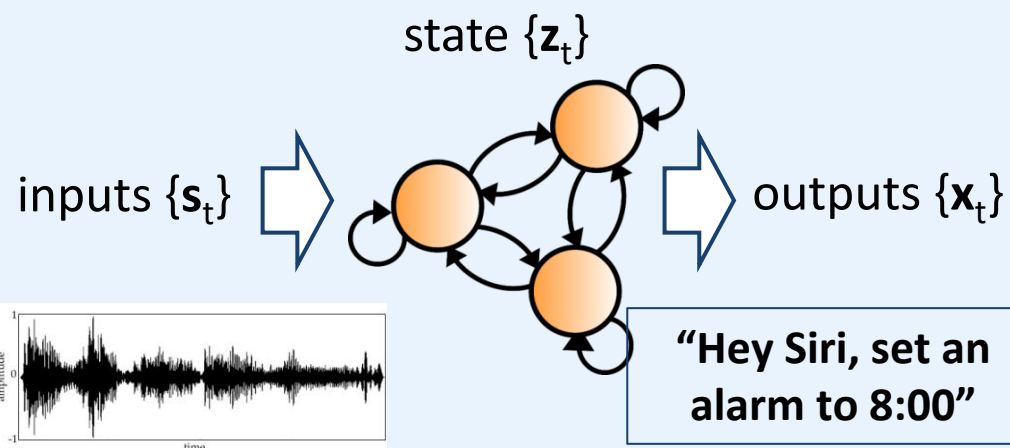


**Goal:**  
Reconstruct temporal  
& geometrical  
structure

**Training:**  
Expectation maximization  
Laplace approximation  
FP-iterations

# Experimental setups

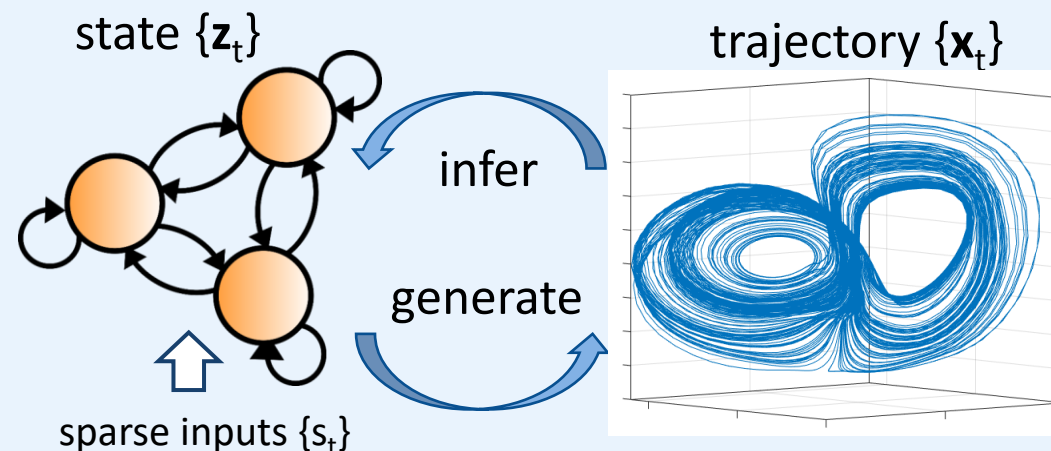
## Supervised



**Goal:**  
Good ahead  
predictions

**Training:**  
Stochastic  
gradient descent

## Unsupervised



**Goal:**  
Reconstruct temporal  
& geometrical  
structure

**Training:**  
Expectation maximization  
Laplace approximation  
FP-iterations

# Supervised: Classical ML benchmarks

Addition task

Multiplication task

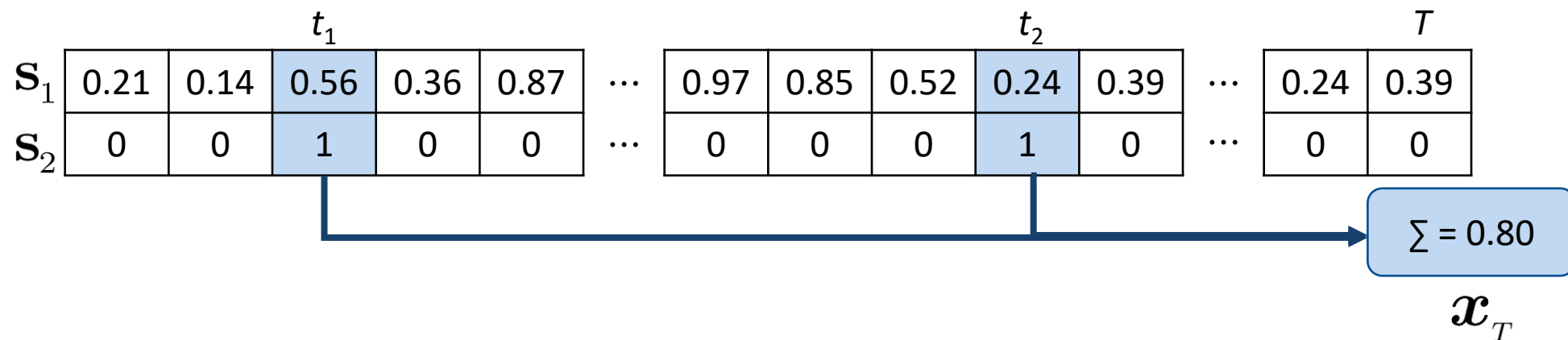
Sequential MNIST task

# Supervised: Classical ML benchmarks

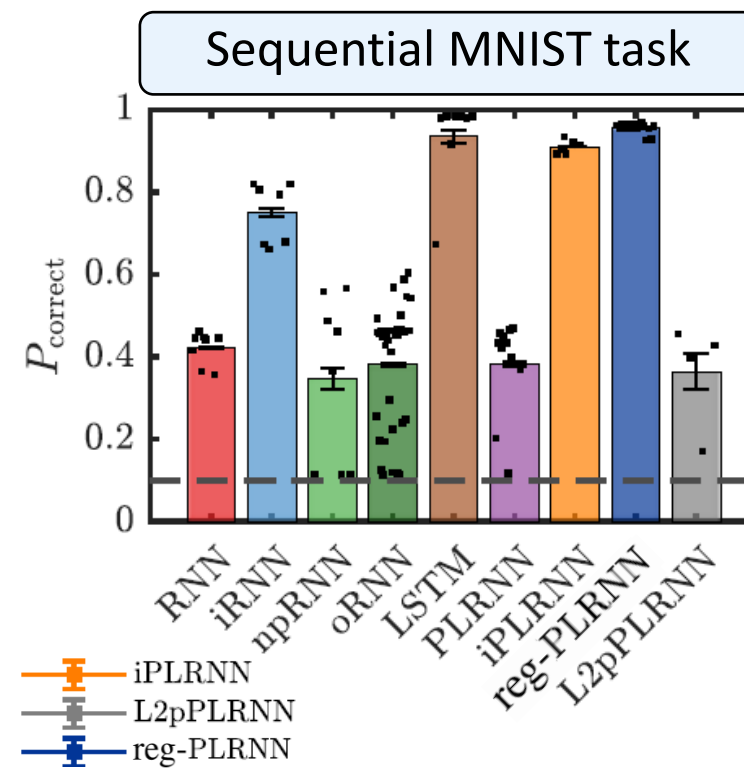
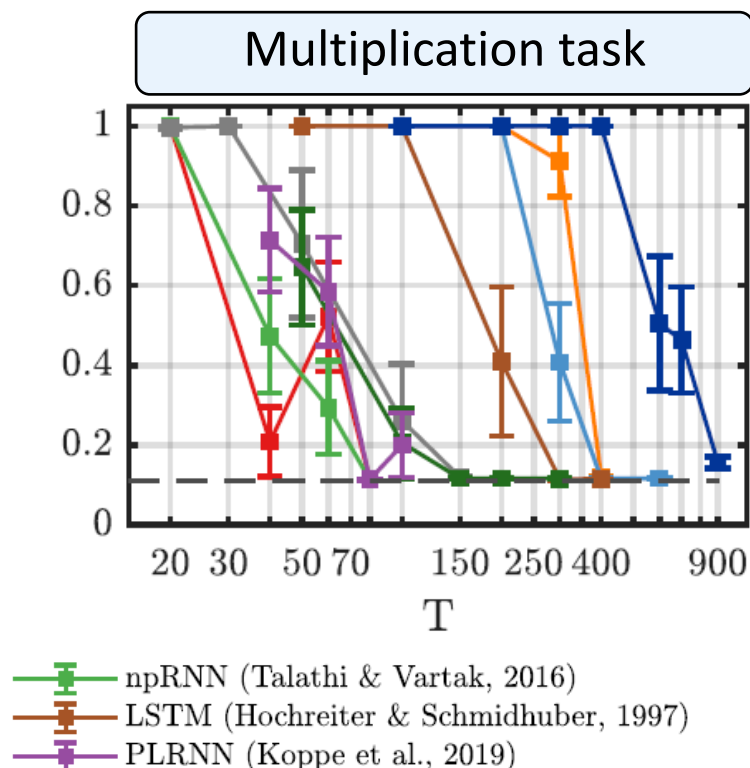
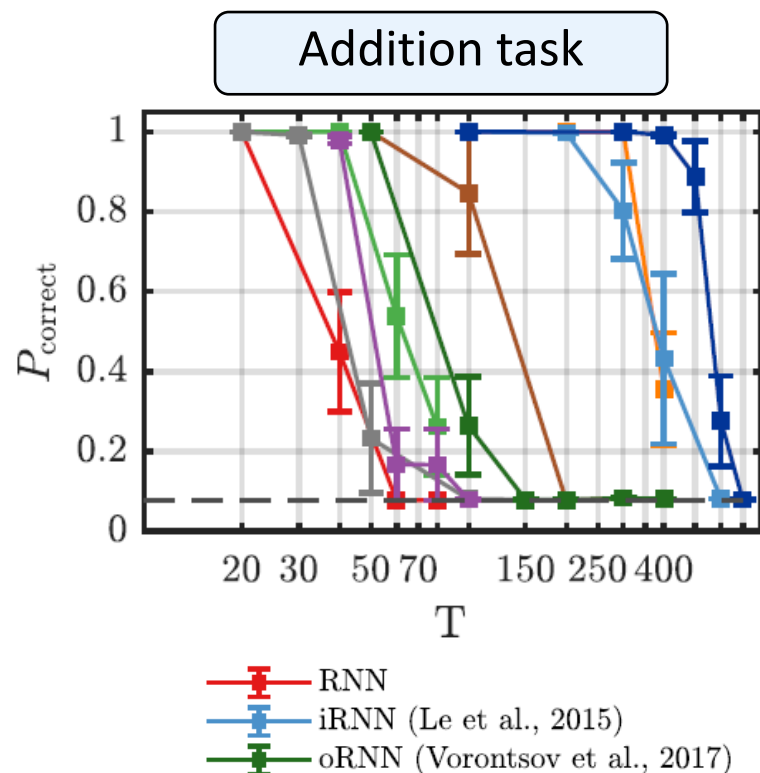
Addition task

Multiplication task

Sequential MNIST task

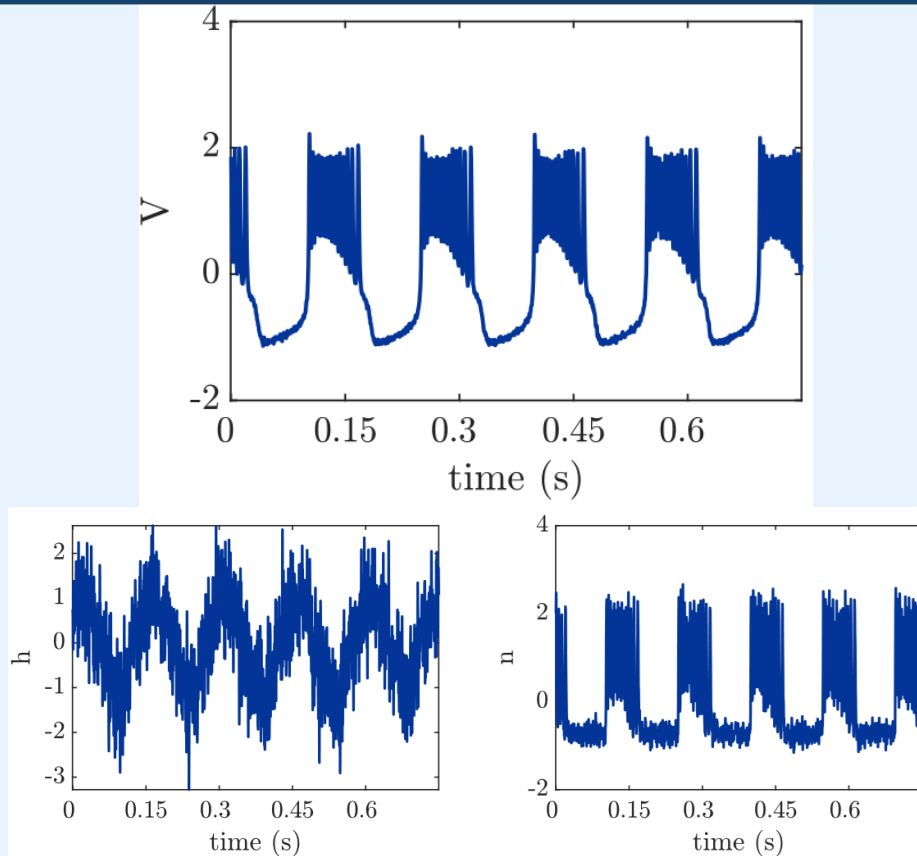


# Supervised: Classical ML benchmarks

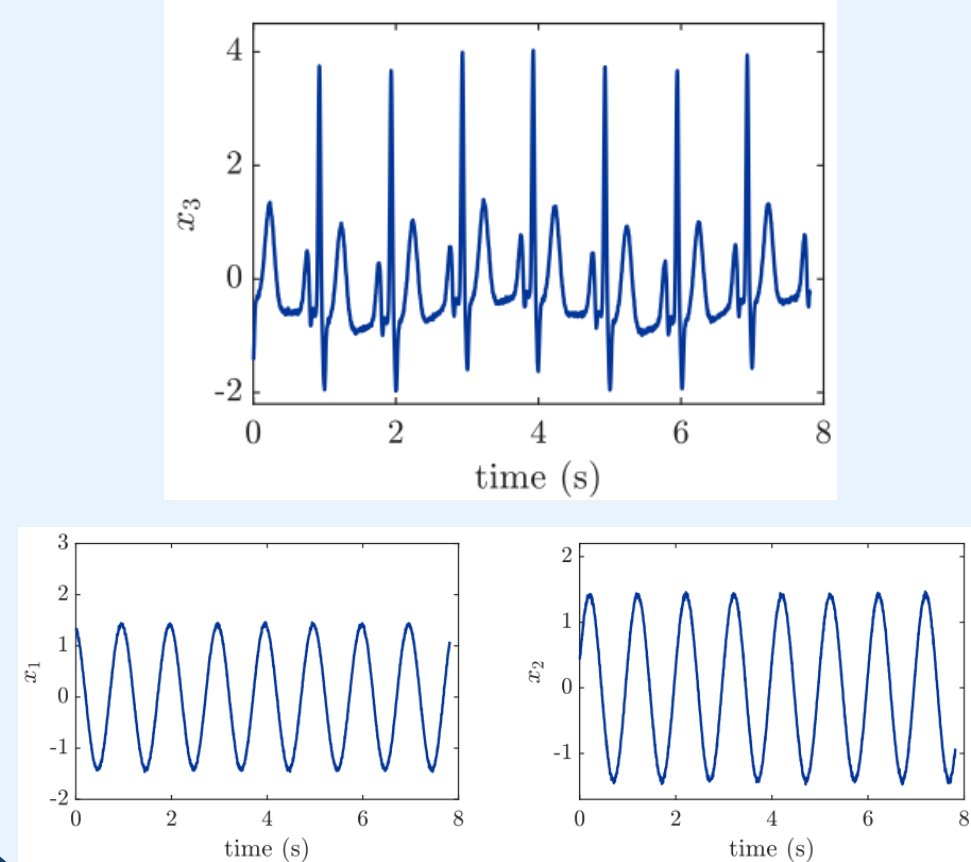


# Unsupervised: Dynamical systems reconstruction

## Bursting neuron model

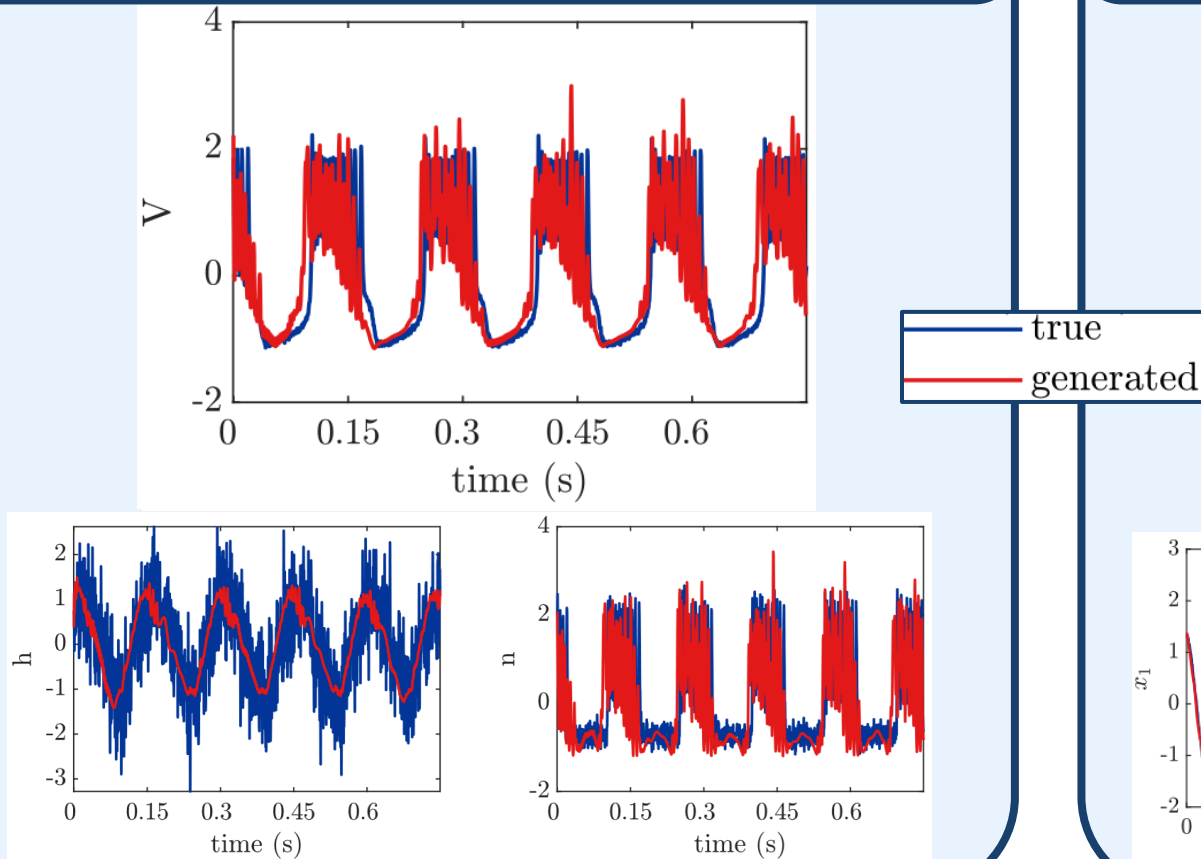


## Simulated electrocardiogram (ECG)

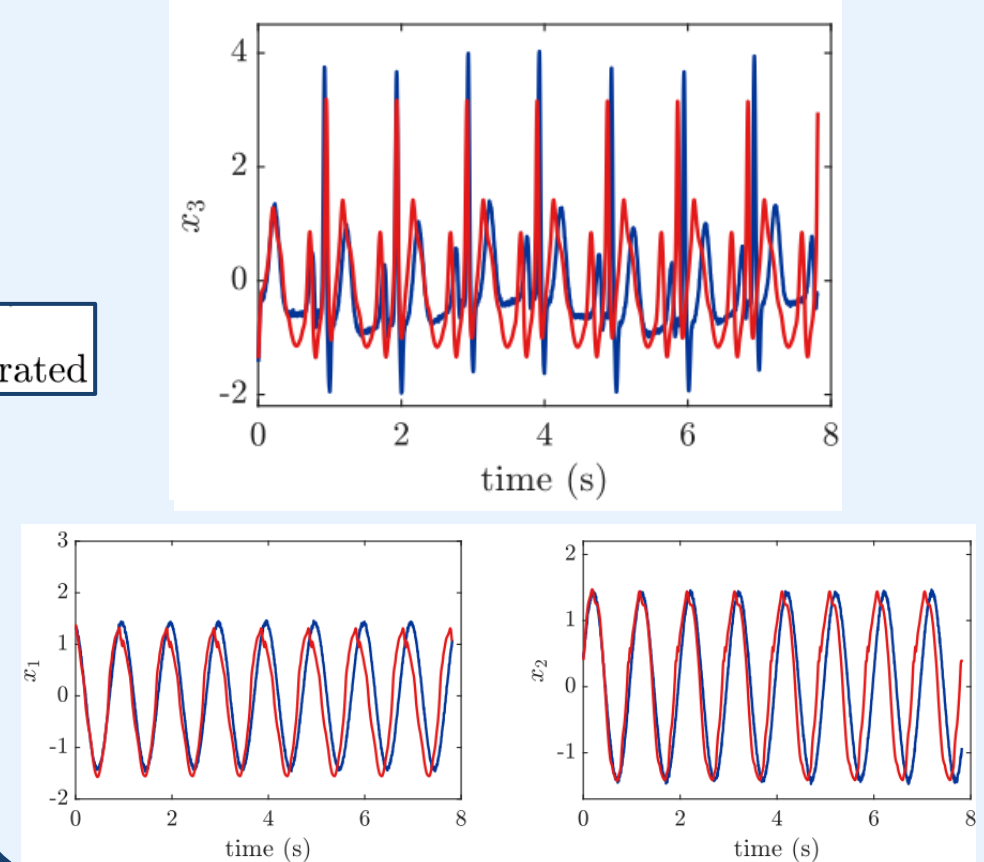


# Unsupervised: Dynamical systems reconstruction

## Bursting neuron model

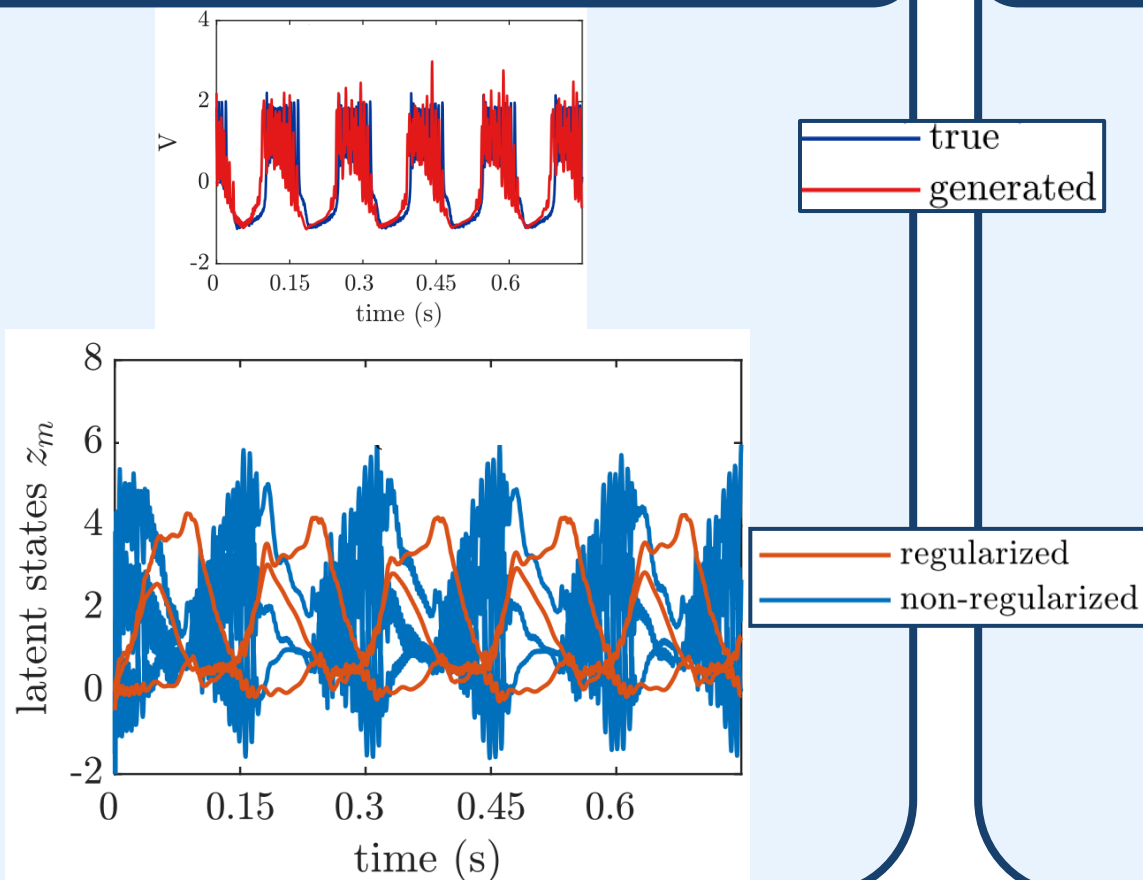


## Simulated electrocardiogram (ECG)

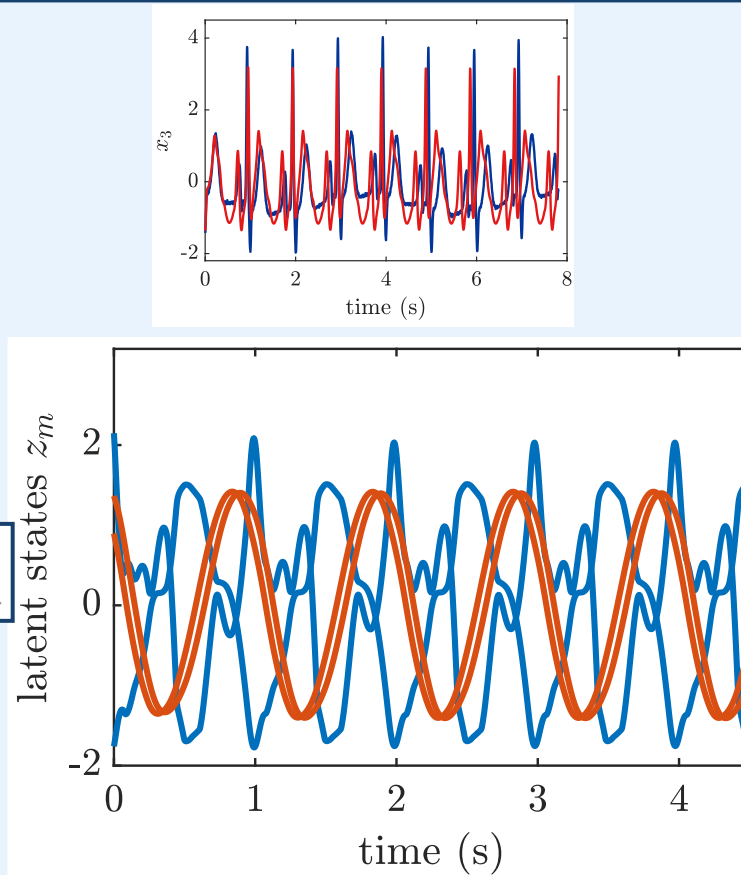


# Unsupervised: Dynamical systems reconstruction

## Bursting neuron model

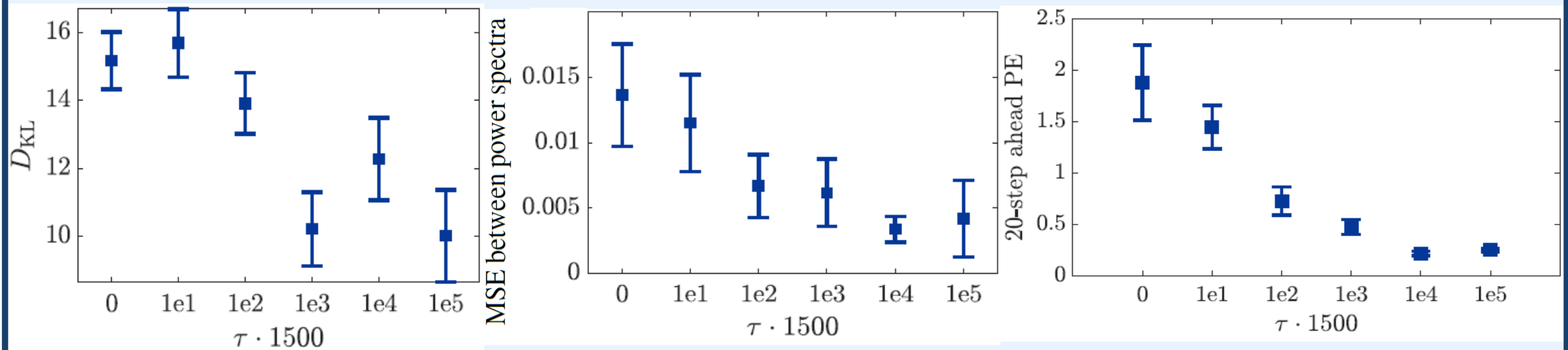


## Simulated electrocardiogram (ECG)



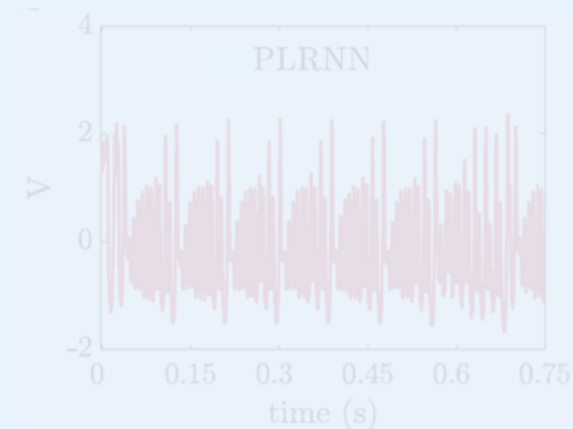
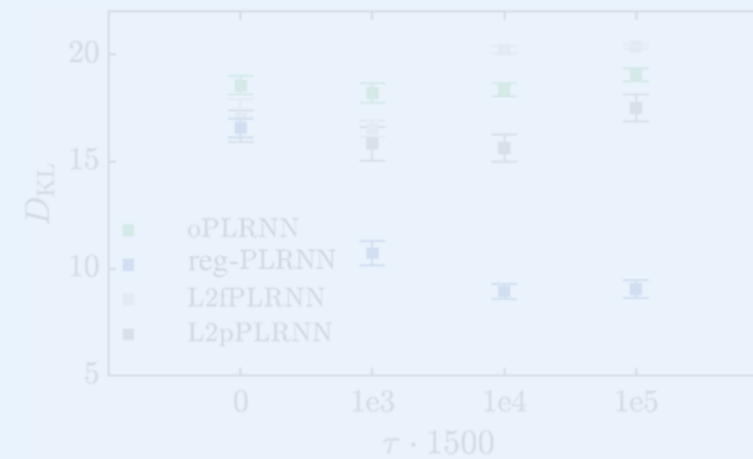
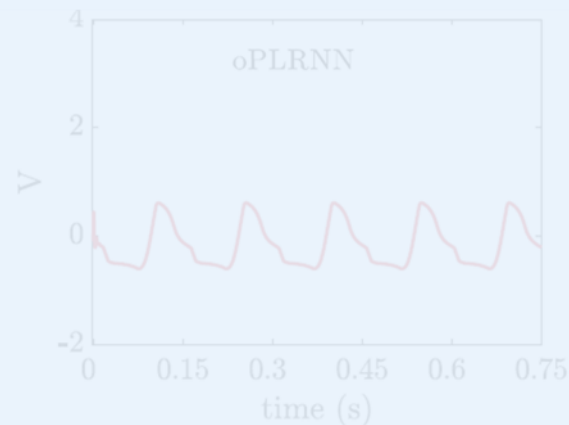
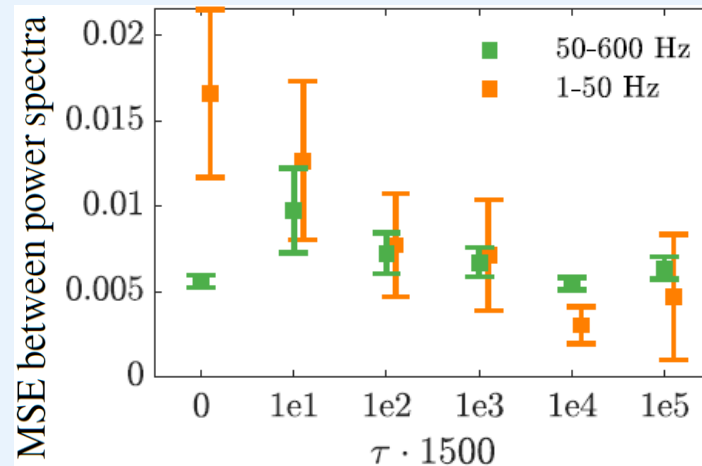
# Unsupervised: Dynamical systems reconstruction

## Dependence of performance on regularization factor



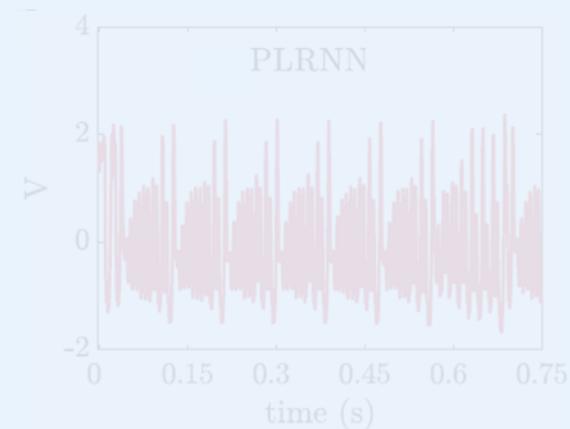
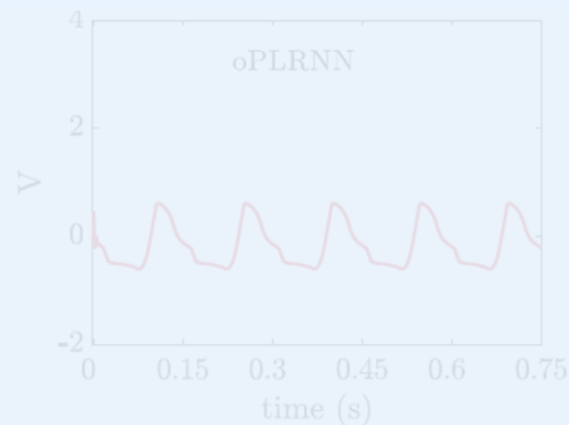
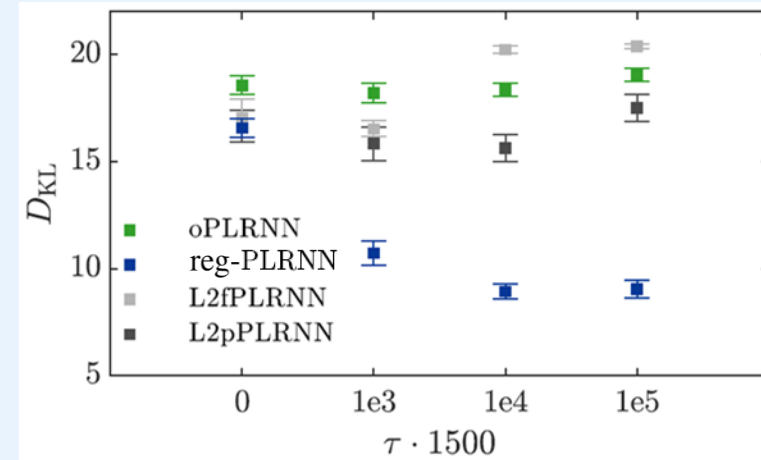
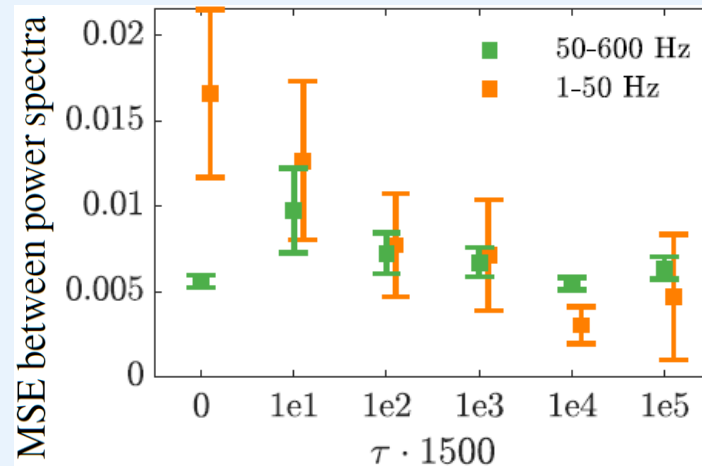
# Unsupervised: Dynamical systems reconstruction

Mechanisms: Capturing both slow and fast time scales



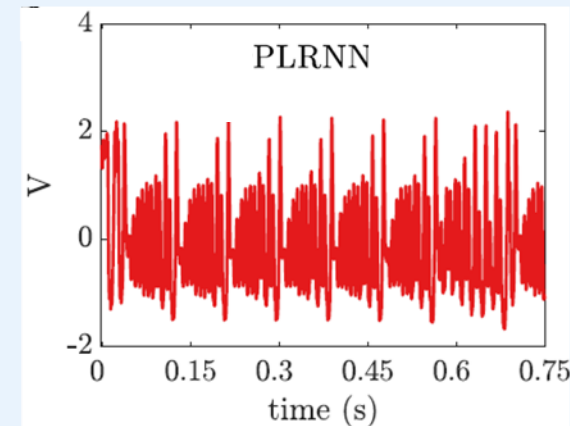
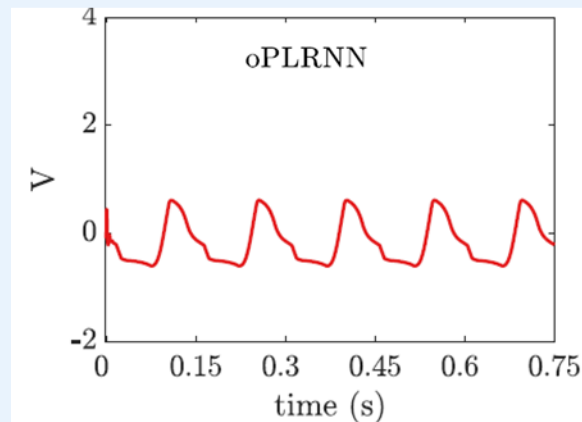
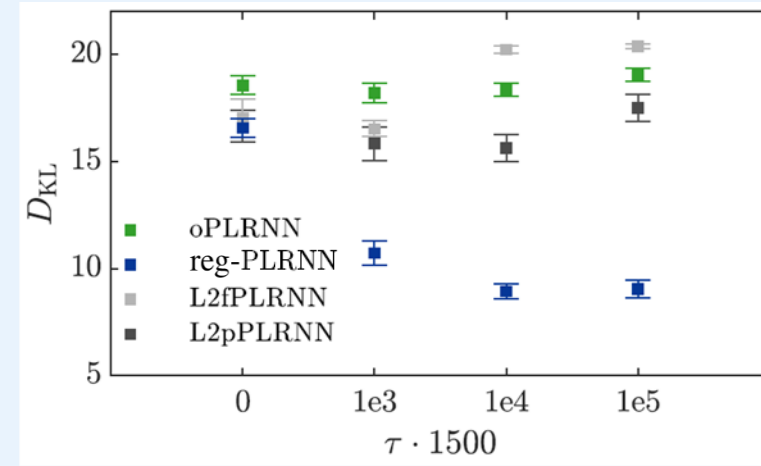
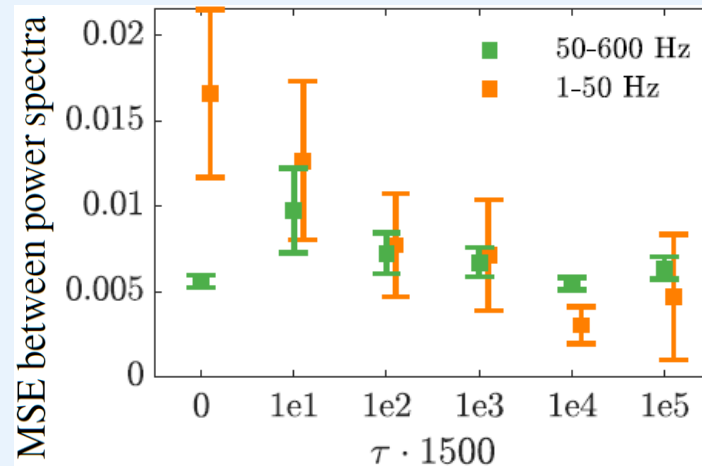
# Unsupervised: Dynamical systems reconstruction

Mechanisms: Capturing both slow and fast time scales



# Unsupervised: Dynamical systems reconstruction

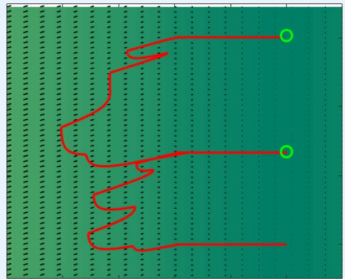
## Mechanisms: Capturing both slow and fast time scales



# Conclusions

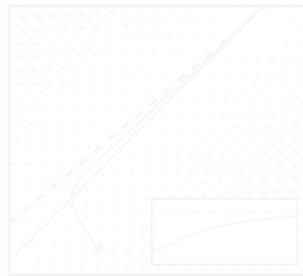
## Tractable model

Piecewise-linear RNN has tractable mathematical structure



## Regularization

Manifold-attractor encourages 'memory subspace' and slow flow



## Theorems

PLRNN-dynamics

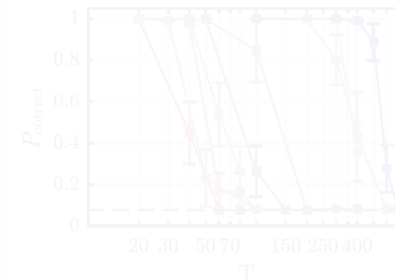


Error gradients

Bounded gradients in reg-PLRNN

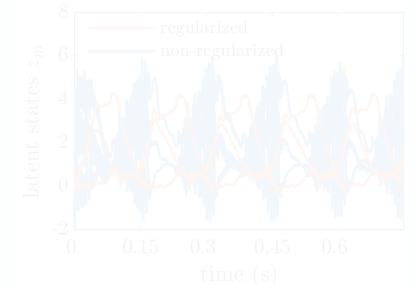
## ML benchmarks

Outperforms other RNN on ML benchmarks



## DS reconstruction

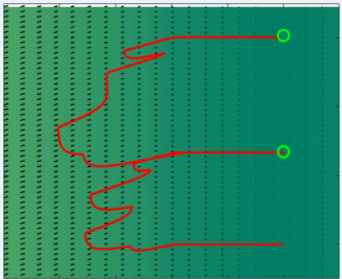
Enables multiple time scales and long-range dependencies



# Conclusions

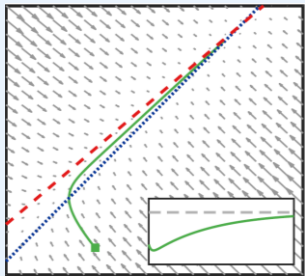
## Tractable model

Piecewise-linear RNN has tractable mathematical structure



## Regularization

Manifold-attractor encourages 'memory subspace' and slow flow



## Theorems

PLRNN-dynamics

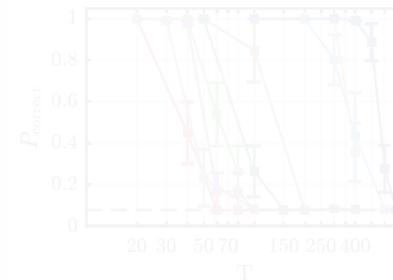


Error gradients

Bounded gradients  
in reg-PLRNN

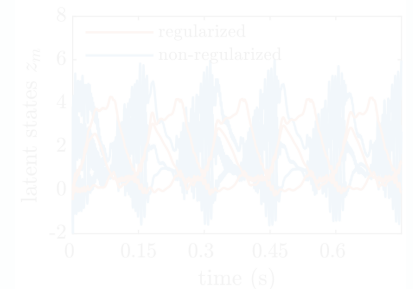
## ML benchmarks

Outperforms other RNN on ML benchmarks



## DS reconstruction

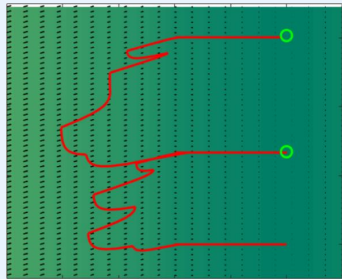
Enables multiple time scales and long-range dependencies



# Conclusions

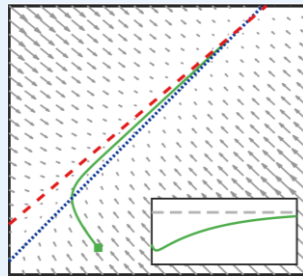
## Tractable model

Piecewise-linear RNN has tractable mathematical structure



## Regularization

Manifold-attractor encourages 'memory subspace' and slow flow



## Theorems

PLRNN-dynamics

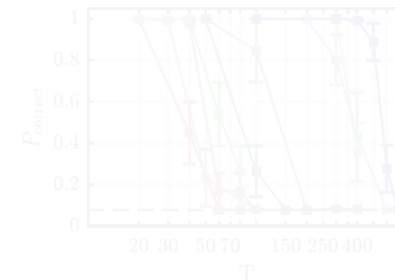


Error gradients

Bounded gradients in reg-PLRNN

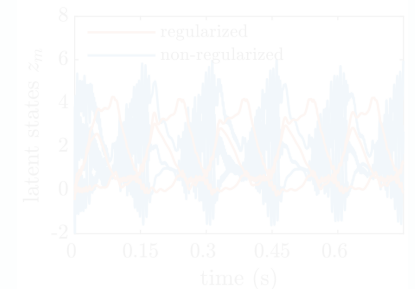
## ML benchmarks

Outperforms other RNN on ML benchmarks



## DS reconstruction

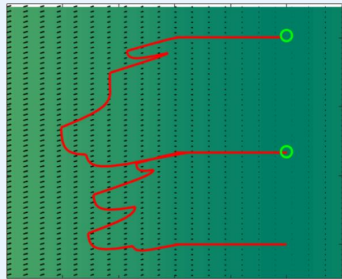
Enables multiple time scales and long-range dependencies



# Conclusions

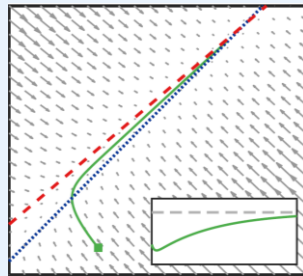
## Tractable model

Piecewise-linear RNN has tractable mathematical structure



## Regularization

Manifold-attractor encourages 'memory subspace' and slow flow



## Theorems

PLRNN-dynamics

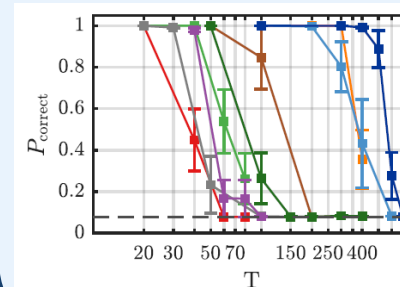


Error gradients

Bounded gradients in reg-PLRNN

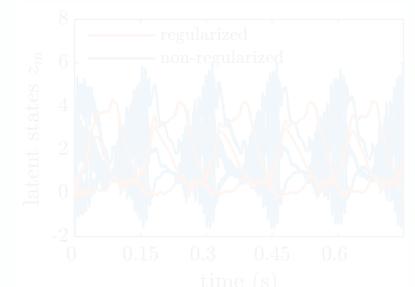
## ML benchmarks

Outperforms other RNN on ML benchmarks



## DS reconstruction

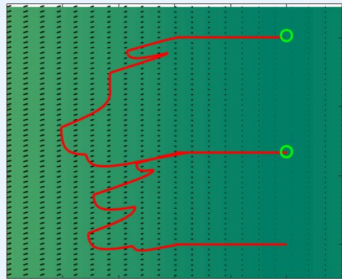
Enables multiple time scales and long-range dependencies



# Conclusions

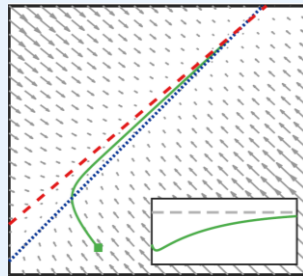
## Tractable model

Piecewise-linear RNN has tractable mathematical structure



## Regularization

Manifold-attractor encourages 'memory subspace' and slow flow



## Theorems

PLRNN-dynamics

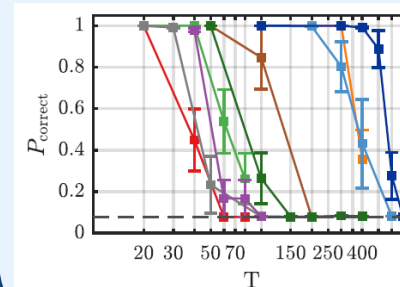


Error gradients

Bounded gradients in reg-PLRNN

## ML benchmarks

Outperforms other RNN on ML benchmarks



## DS reconstruction

Enables multiple time scales and long-range dependencies

