TIME SERIES ANALYSIS & RECURRENT NEURAL NETWORKS #10

- Training RNN (cont.): BPTT
- Vanishing/ exploding grad. problem
- LSTM, GRU, rPLRNN

Main lecture: Daniel Durstewitz

Exercises: Leonard Bereska, Manuel Brenner, Daniel Kramer, Georgia Koppe

Heidelberg University

Training RNN by grad. descent

$$x_t = \phi\left(W x_{t-1} + h + C s_t\right)$$

$$D = \left\{ \{s_t^{(p)}\}, \{\tilde{x}_t^{(p)}\} \right\}, \quad p = 1 \ldots P$$

$$L(W, h, C) = \frac{1}{2} \sum_{p=1}^{P} \sum_{t=1}^{T} \sum_{i=1}^{M} I\{\varepsilon_{it} = 1\} \left(\tilde{x}_{it} - x_{it}\right)^2 \frac{1}{P} \frac{1}{T}$$

$$\nabla_W L := \left(\frac{\partial L}{\partial w_{ij}}\right), \quad \gamma_{kij} := \frac{\partial x_k}{\partial w_{ij}}, \quad O(M^3)$$

Back-Prop. through time (BPTT)

Werbos (88), Willia

Making RNN deep in time: "Vanishing/ exploding gradient problem"

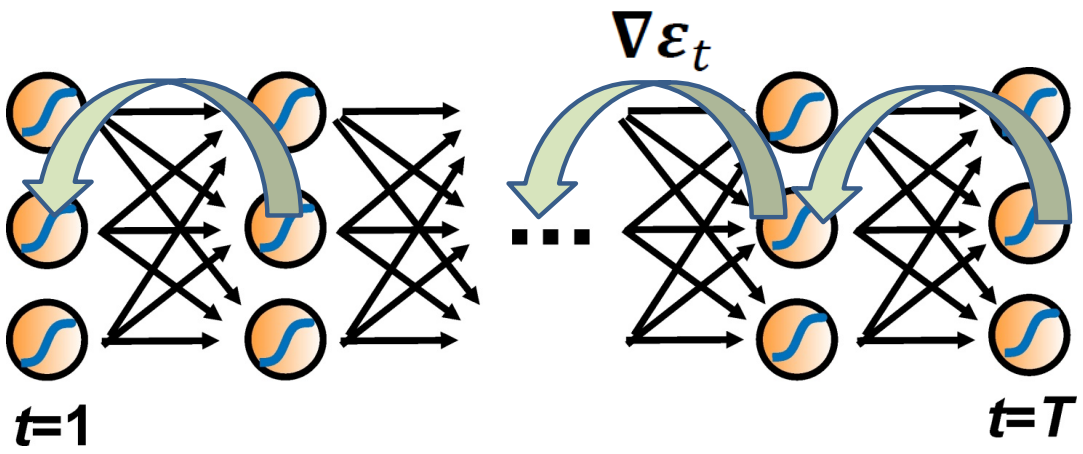
"Inputs": $\boldsymbol{S} = \{\boldsymbol{s}_t\}$ "Outputs": $\tilde{\boldsymbol{Z}} = \{\tilde{\boldsymbol{z}}_t\}, t = 1 \dots T$

RNN $\boldsymbol{z}_t \in \mathbb{R}^M$

$$\boldsymbol{z}_t = \phi(\boldsymbol{W}\boldsymbol{z}_{t-1} + \boldsymbol{h} + \boldsymbol{C}\boldsymbol{s}_t)$$

Loss function

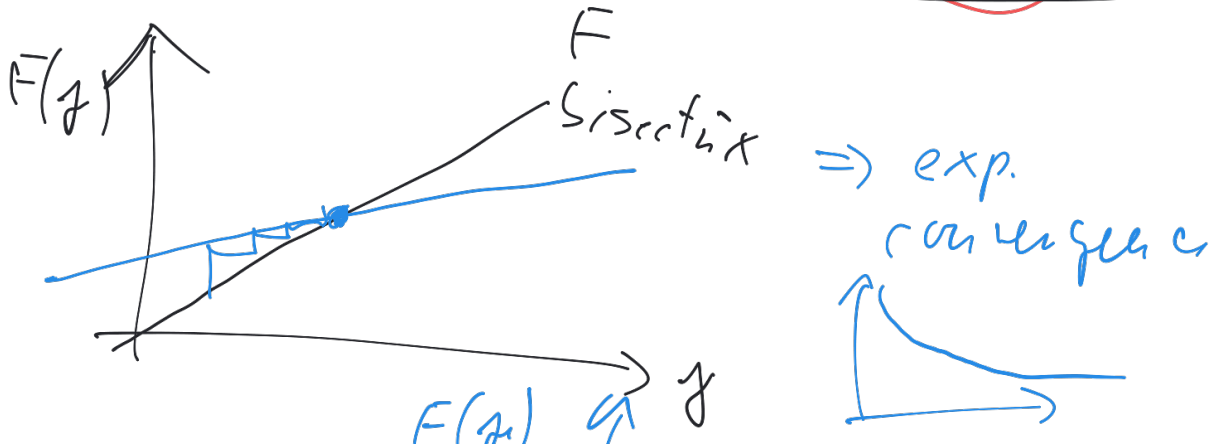
$$\varepsilon = \sum_{r=1}^{R} \sum_{t=1}^{T} \|\tilde{\boldsymbol{z}}_t - \boldsymbol{z}_t\|_2^2$$

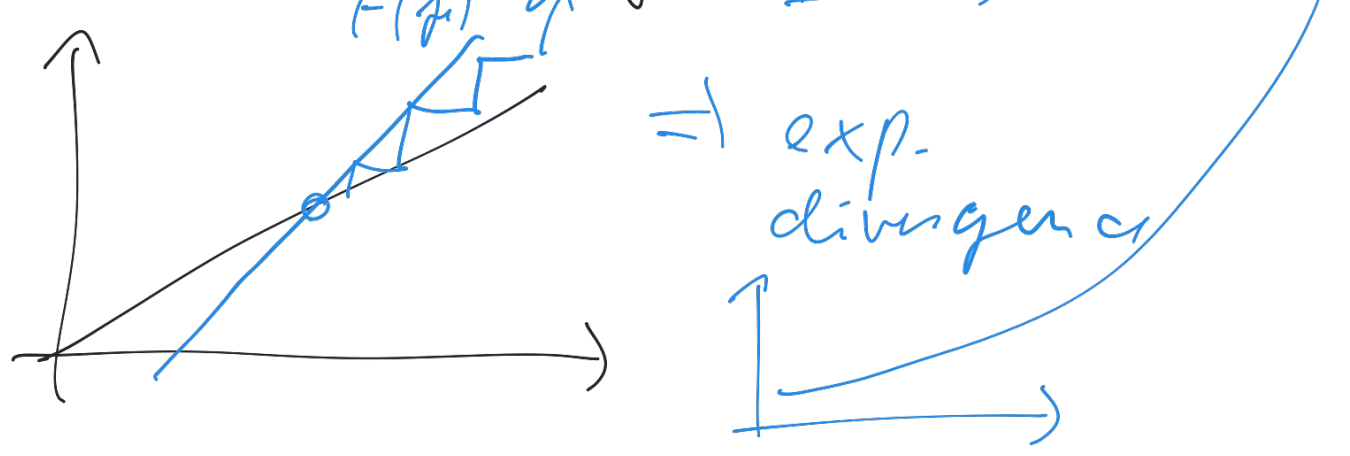


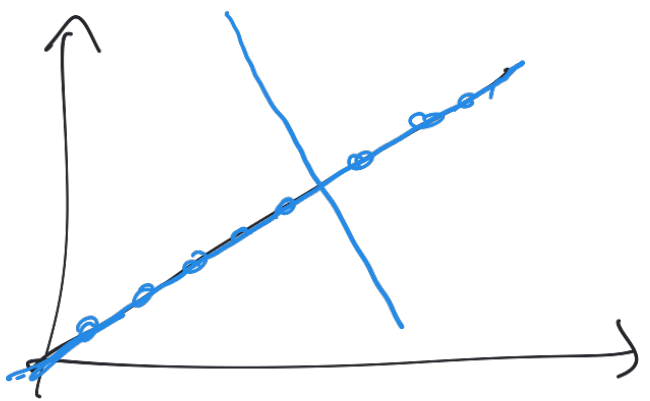
RTRL

$$\mathcal{L} = \frac{1}{2} \sum_t \sum_k \sum_i I\{\varepsilon_{it} = 1\} \left(\tilde{x}_{it} - x_{it}\right)^2$$

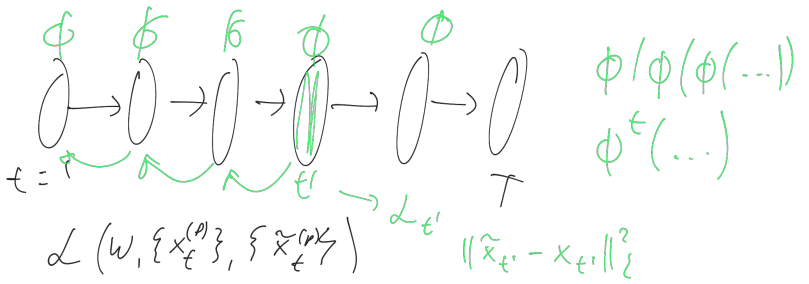
$$\gamma_{kij}^{(t)} := \frac{\partial x_{kt}}{\partial w_{ij}} = \underbrace{\phi' \cdot \left[\delta_{ki}\, x_{j,t-1} + \sum_{l=1}^{n} w_{kl} \frac{\partial x_{l,t-1}}{\partial w_{ij}} \right]}_{F}$$







⇒ neither vanish nor explode



$\phi / \phi(\phi(\dots))$

$\phi^t(\dots)$

$\mathcal{L}\left(w, \{x_t^{(n)}\}, \{\tilde{x}_t^{(n)}\}\right) \quad \mathcal{L}_{t'} \quad \|\tilde{x}_{t'} - x_{t'}\|_2^2$

$$\frac{\partial \mathcal{L}}{\partial w_{ij}} = \sum_{t=1}^{T} \frac{\partial \mathcal{L}_t}{\partial w_{ij}}$$

$$\frac{\partial \mathcal{L}_t}{\partial w_{ij}} = \sum_{\tau=1}^{t} \frac{\partial \mathcal{L}_t}{\partial x_t} \frac{\partial x_t}{\partial x_\tau} \frac{\partial x_\tau}{\partial w_{ij}}$$

$1 \times M$; $M \times 1$; $M \times M$

$$\frac{\partial x_t}{\partial x_\tau} = \begin{pmatrix} \frac{\partial x_{1t}}{\partial x_{1\tau}} & \frac{\partial x_{1t}}{\partial x_{2\tau}} & \cdots & \frac{\partial x_{1t}}{\partial x_{M\tau}} \\ \vdots & \ddots & & \vdots \\ \frac{\partial x_{Mt}}{\partial x_{1\tau}} & \cdots & & \frac{\partial x_{Mt}}{\partial x_{M\tau}} \end{pmatrix}$$

$$\frac{\partial x_t}{\partial x_\tau} = \frac{\partial x_t}{\partial x_{t-1}} \frac{\partial x_{t-1}}{\partial x_{t-2}} \cdots \frac{\partial x_{\tau+1}}{\partial x_\tau} = \prod_{n=\tau+1}^{t} \frac{\partial x_n}{\partial x_{n-1}}$$

$x_n = \phi(W x_{n-1} + b)$

$$\prod \left(W \, \mathrm{diag}\left[\left(\frac{\partial \phi}{\partial x_{n-1}} \right) \right] \right)$$

$\Rightarrow \max |\mathrm{eig}(W \phi')| > 1$: divergence! explode!

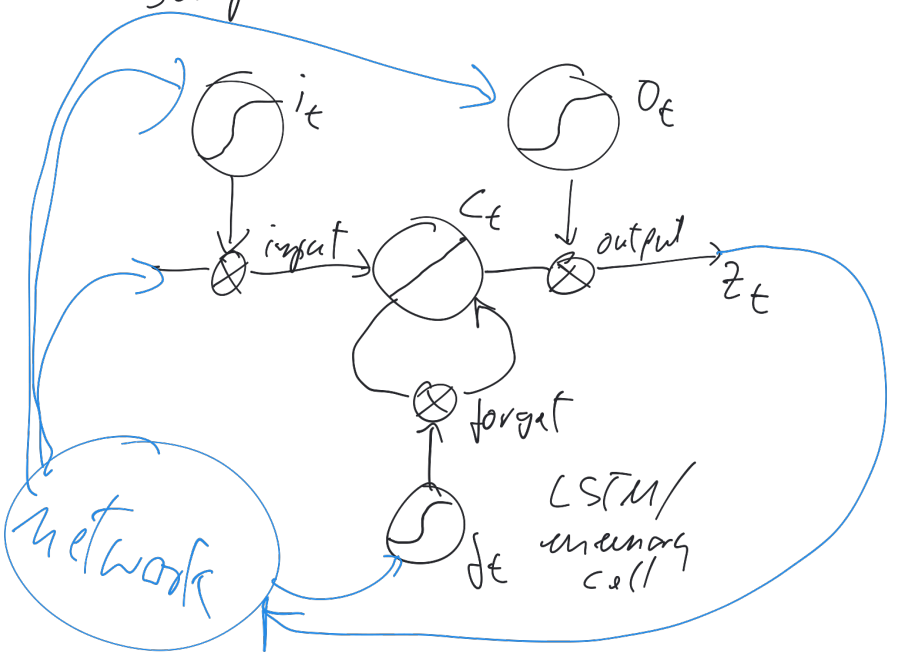
$\max |\mathrm{eig}(W \phi')| < 1$: convergence! vanish!

solutions

Long-Short Term Memory (LSTM)

Hochreiter & Schmidhuber (97) NC

→ Bengio



$$c_t = f_t \circ c_{t-1} + i_t \circ \tanh\left[W_c z_{t-1} + b_c\right]$$

(elem-wise multiplic.)

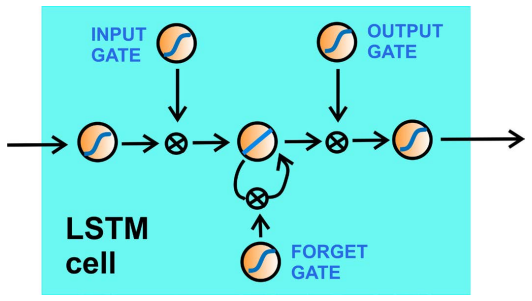
$$z_t = o_t \circ \tanh(c_t)$$

$$i_t = \sigma(W_i z_{t-1} + b_i)$$

$$o_t = \sigma(W_o z_{t-1} + b_o) \in [0, 1],$$

$$f_t = \sigma(W_f z_{t-1} + b_f) \qquad \sigma(\eta) = \frac{1}{1 + e^{-\eta}}$$

Long Short-Term Memory



Hochreiter & Schmidhuber (1997)

$$
\begin{aligned}
\boldsymbol{i}_t^l &= \sigma(W_i^l[\boldsymbol{\chi}_t; \boldsymbol{h}_{t-1}^l; \boldsymbol{h}_t^{l-1}] + \boldsymbol{b}_i^l) \\
\boldsymbol{f}_t^l &= \sigma(W_f^l[\boldsymbol{\chi}_t; \boldsymbol{h}_{t-1}^l; \boldsymbol{h}_t^{l-1}] + \boldsymbol{b}_f^l) \\
\boldsymbol{s}_t^l &= \boldsymbol{f}_t^l \boldsymbol{s}_{t-1}^l + \boldsymbol{i}_t^l \tanh(W_s^l[\boldsymbol{\chi}_t; \boldsymbol{h}_{t-1}^l; \boldsymbol{h}_t^{l-1}] + \boldsymbol{b}_s^l) \\
\boldsymbol{o}_t^l &= \sigma(W_o^l[\boldsymbol{\chi}_t; \boldsymbol{h}_{t-1}^l; \boldsymbol{h}_t^{l-1}] + \boldsymbol{b}_o^l) \\
\boldsymbol{h}_t^l &= \boldsymbol{o}_t^l \tanh(\boldsymbol{s}_t^l)
\end{aligned}
$$

Gated Recurrent Unit (GRU)

Cho ... Bengio (2014)

update gate: $u_t = \sigma(W_u z_{t-1} + b_u) \in [0,1]$

reset gate: $r_t = \sigma(W_r z_{t-1} + b_r)$

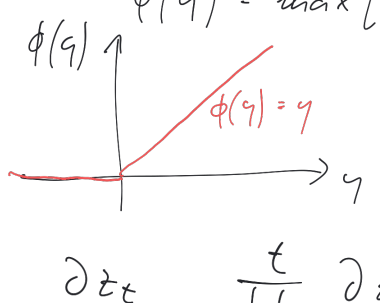
hidden state: $z_t = u_t \circ z_{t-1} + (1-u_t) \cdot \tanh\left[W_z(r_t \circ z_{t-1}) + b_z + C s_t\right]$

Initialization / Regularization

Le ... Hinton (2015) arXiv

$$z_t = \phi(W z_{t-1} + b + C s_t)$$

$$\phi(y) = \max[y, 0], \quad \text{ReLU}$$



$\in \{0, 1\}$

$$\frac{\partial z_t}{\partial z_\tau} = \prod_{u=\tau+1}^{t} \frac{\partial z_u}{\partial z_{u-1}} = \prod_{u=\tau+1}^{t} W \operatorname{diag}[\phi']$$

$W = I$

$$z_t = A z_{t-1} + W \underbrace{\phi(z_{t-1})}_{\max[0, z_{t-1}]} + b + C s_t$$

$$A = \begin{pmatrix} 1 & & & 0 \\ & 1 & & \\ & & \ddots & \\ 0 & & & 1 \end{pmatrix}, \quad W = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ * & * & * & * \end{pmatrix}$$

$$b = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ * \end{pmatrix}$$

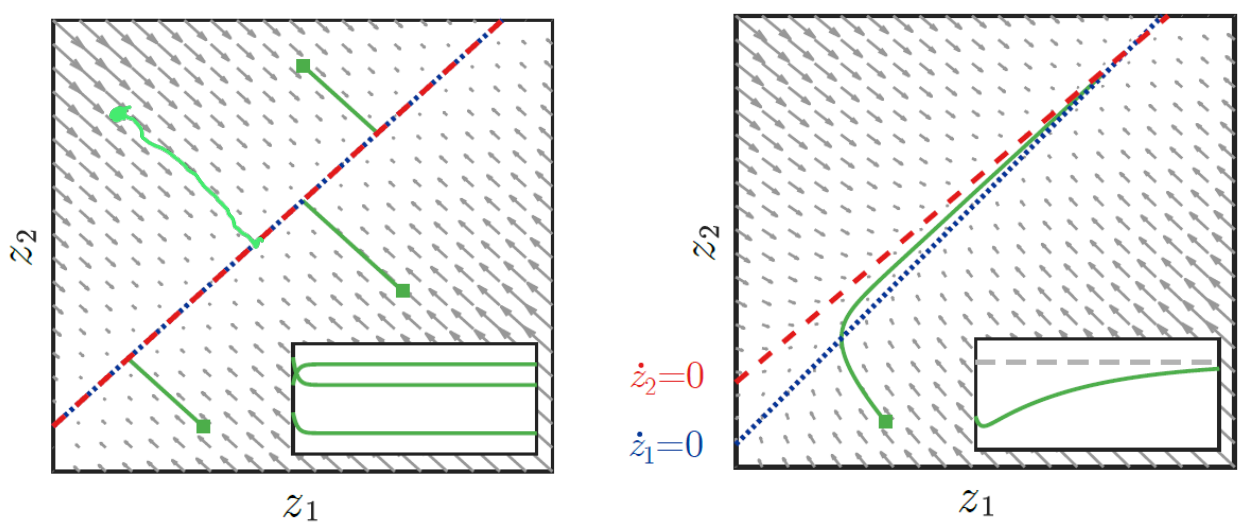
$z_t \in \mathbb{R}^M$, $A \in \mathbb{R}^{M \times M}$ diag. mtx., $W \in \mathbb{R}^{M \times M}$

$M_{reg} < M$

$$C = \mathcal{L}_{MSE} + \lambda\left[\sum_{i=1}^{M_{reg}} (1 - a_{ii})^2 + \sum_{i=1}^{M_{reg}} \sum_{j=1}^{M} w_{ij}^2 + \sum_{i=1}^{M_{reg}} b_i^2\right]$$

Schmidt, Kappel, ..., D (2021) ICLR

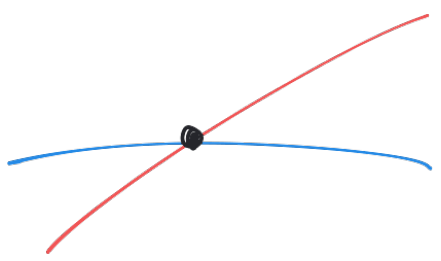
Schmidt et al. (2021) ICLR



nullclines

$\dot{z}_1 = 0$

$\dot{z}_2 = 0$



$z_t = (1-\varepsilon) z_{t-1}$

ε very small

Manifold-attractor regularization

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

$$\boldsymbol{A} = \mathrm{diag}([a_{11} \ldots a_{MM}]) \in \mathbb{R}^{M\times M}$$

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

penalty term $M_{\mathrm{reg}} \leq M$

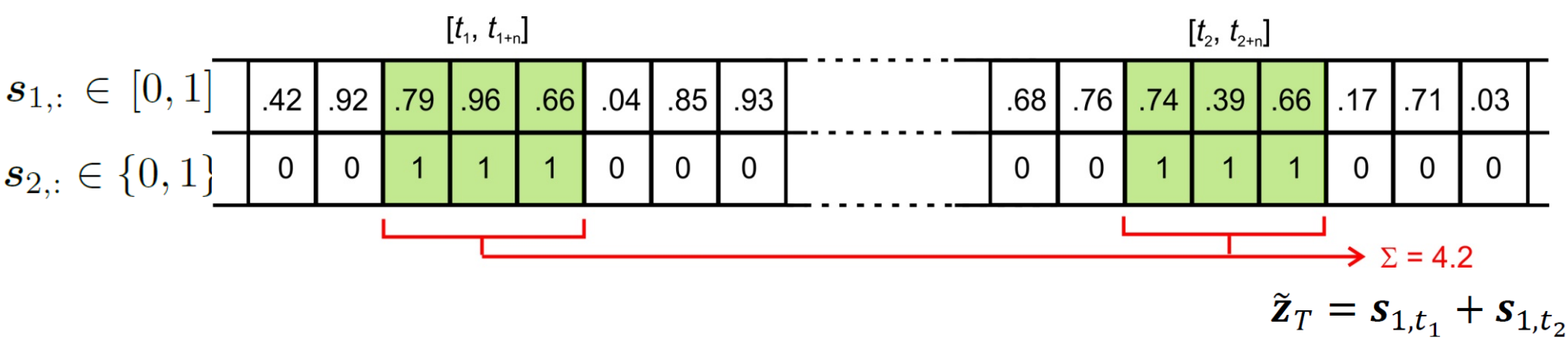
$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & \times & 0 & 0 \\ 0 & 0 & 0 & 0 & \times & 0 \\ 0 & 0 & 0 & 0 & 0 & \times \end{pmatrix} \quad \boldsymbol{A}$$

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \times & \times & \times & 0 & \times & \times \\ \times & \times & \times & \times & 0 & \times \\ \times & \times & \times & \times & \times & 0 \end{pmatrix} \quad \boldsymbol{W}$$

$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ \times \\ \times \\ \times \end{pmatrix} \quad \boldsymbol{h}$$

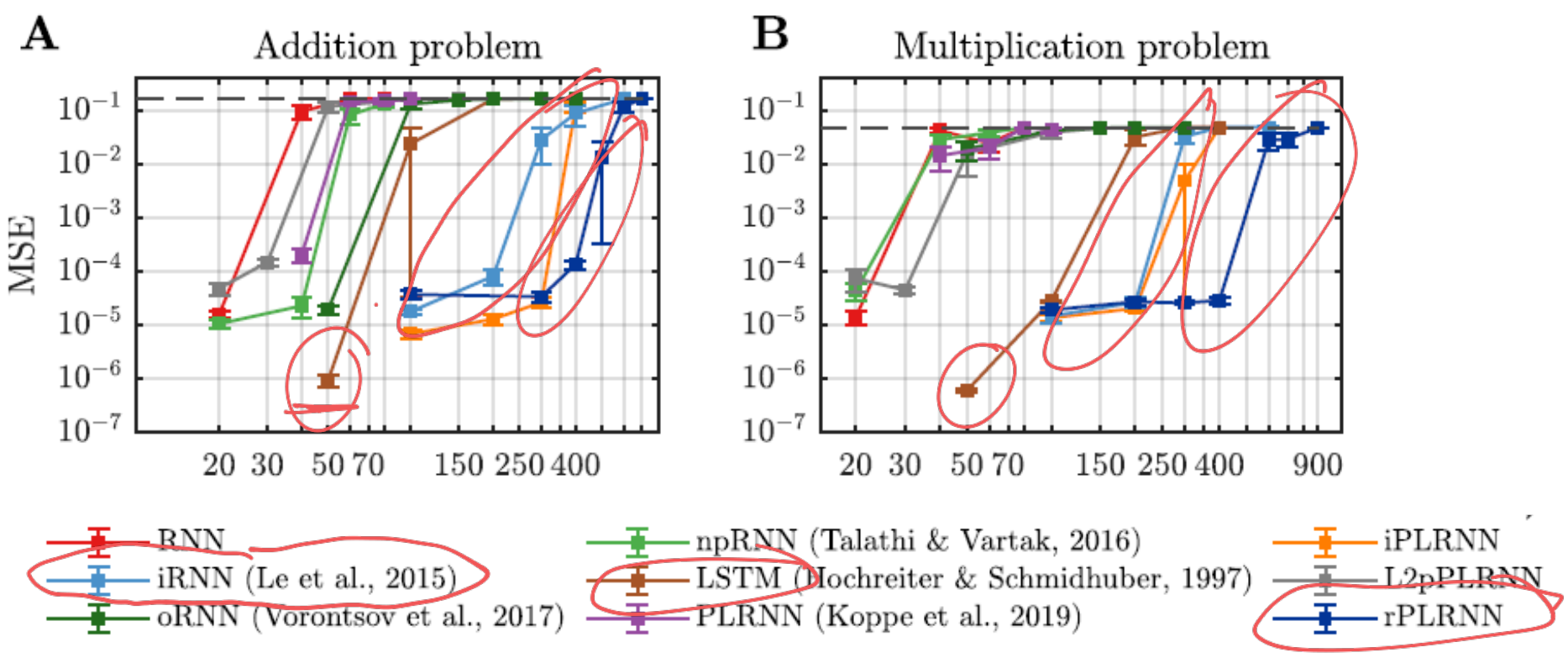
Schmidt et al. (2021) ICLR

Machine Learning benchmark tasks for RNN

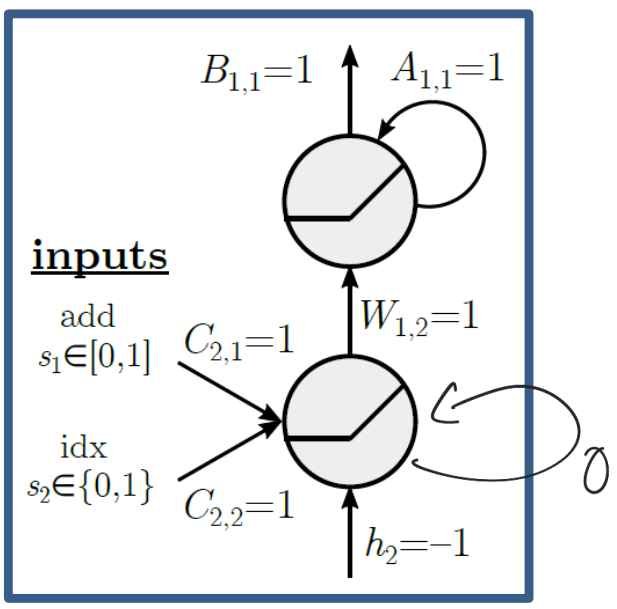


$$\tilde{\boldsymbol{z}}_T = \boldsymbol{s}_{1,t_1} + \boldsymbol{s}_{1,t_2}$$

from Monfared & D (2020) ICML; after Hochreiter & Schmidhuber (1997) NC



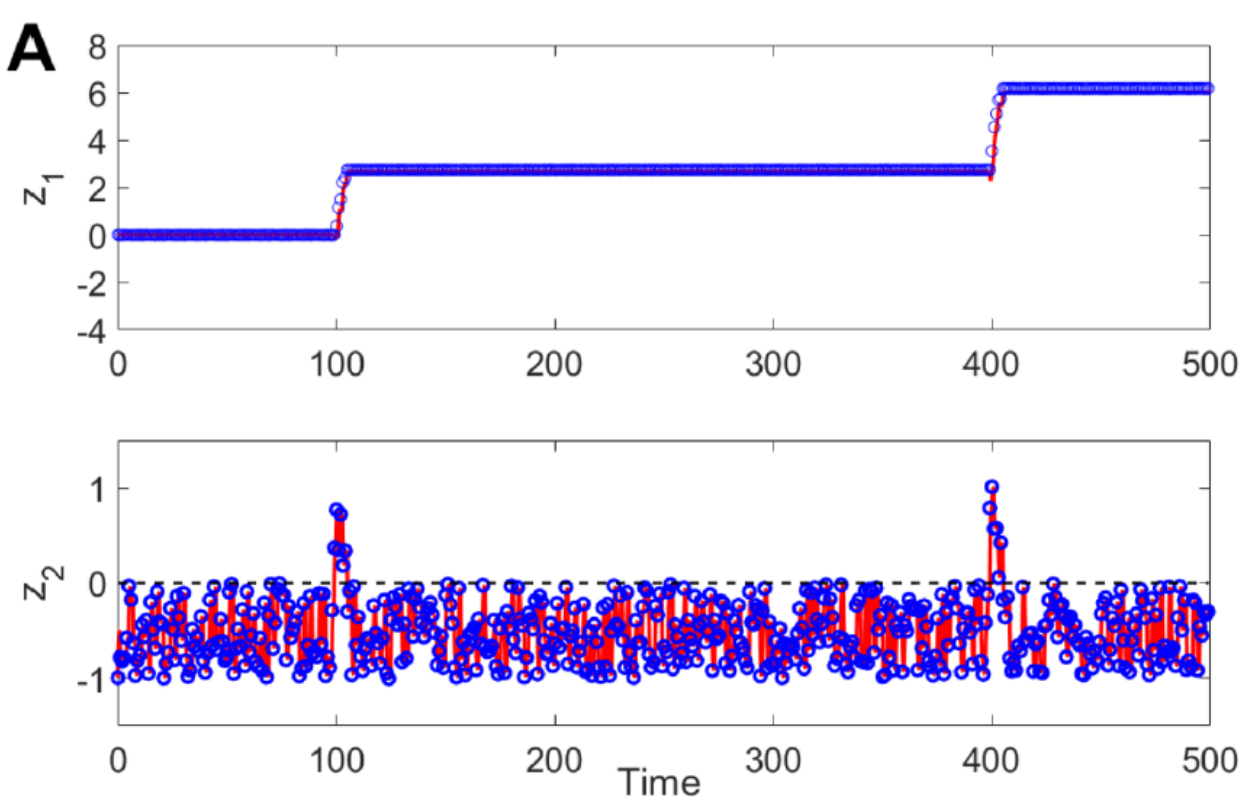
Schmidt et al. (2021) ICLR

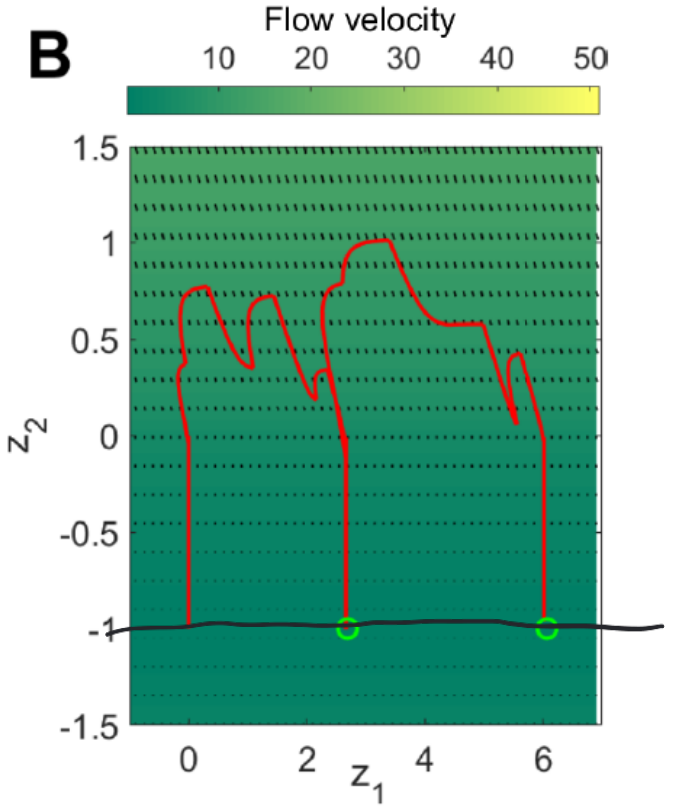


Max

Beutelsparchen

$z_2 = Cs_t - 1$

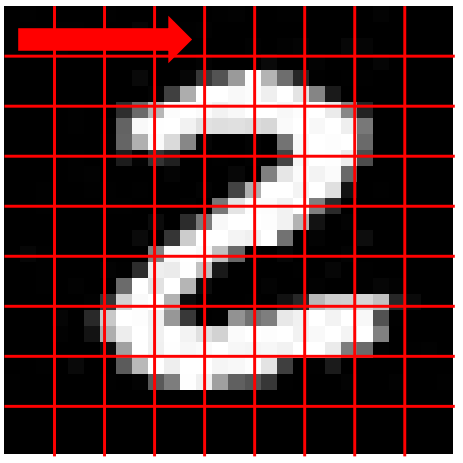




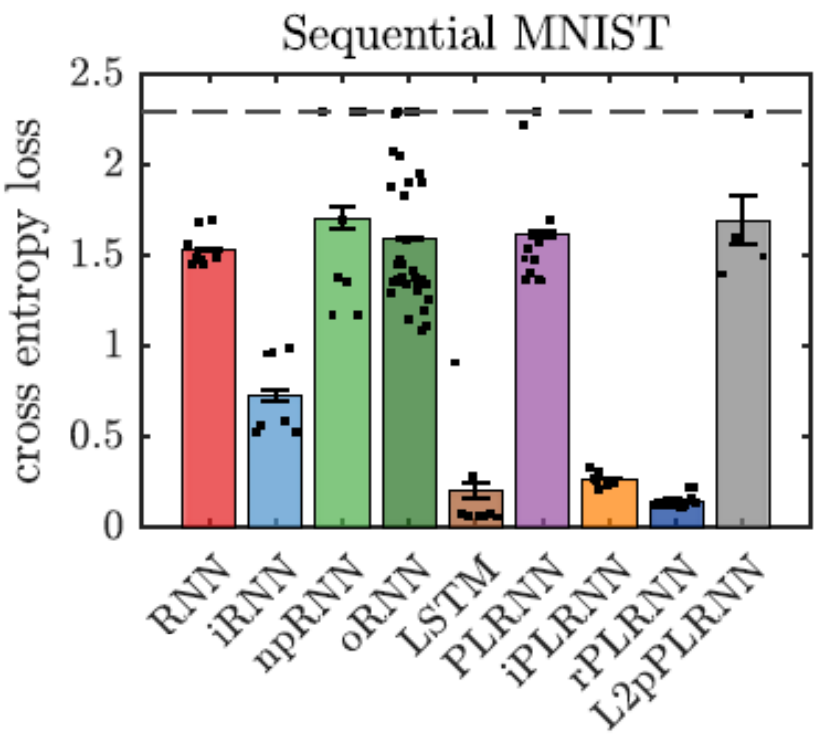
Schmidt et al. (2021) ICLR, Monfared & Durstewitz (2020) ICML

Sequential MNIST benchmark

By Josef Steppan - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=64810040



"2"



W orthog.
orthogonal

Schmidt et al. (2021) ICLR