TIME SERIES ANALYSIS & RECURRENT NEURAL NETWORKS

#9

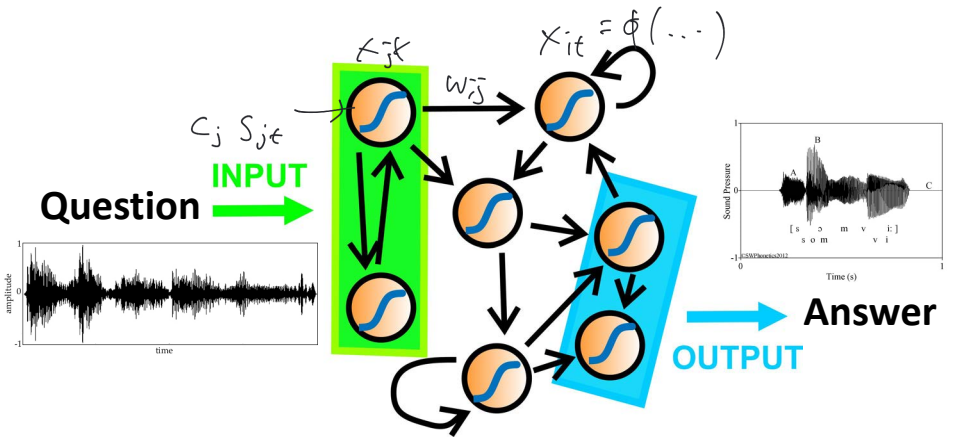
- Universal approx. theorem for RNN
- Training RNN by grad. descent, Real Time Recurrent Learning

Main lecture: Daniel Durstewitz

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

Heidelberg University

Recurrent Neural Network



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

RNN are

- computat. universal: can implement any TM (Siegelmann & Sontag 95, Koiran 94, Cu et al. 2017)
- dynamically universal: can approx. any DS to arbitrary degree

Univ. Approx. Theorem (UAT) for RNN

1) UAT for FF networks → apply to the flow $F(x)$ of DS
2) Reformulate the usu. FFN as a RNN

UAT for FFN (Cybenko 89)

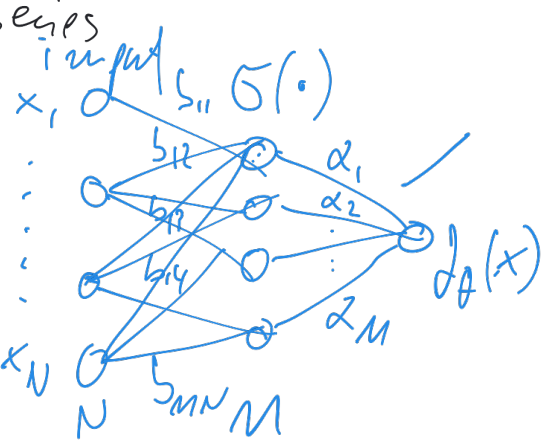
Any contin., smooth func. $F(x)$, $x \in \mathbb{R}^N$, can be approx. to precis. ε by the series

$$g_\theta(x) = \sum_{j=1}^{M} \alpha_j \sigma(b_j x + h_j)$$

(b_j: $N \times N$, h_j: $N \times 1$)

$$\sigma(\gamma) \longrightarrow \begin{cases} 1 & \text{for } \gamma \to +\infty \\ 0 & \text{for } \gamma \to -\infty \end{cases}$$

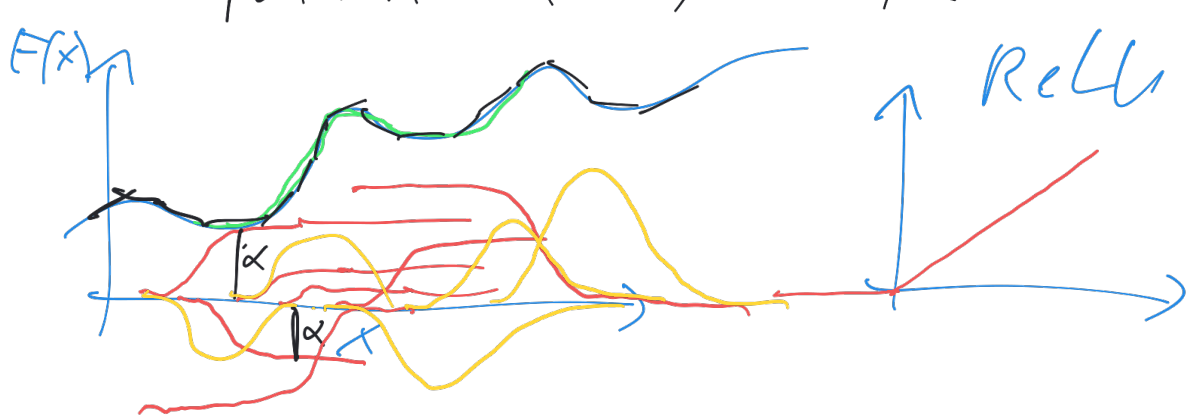
sigmoid, differentiable



such that $|F(x) - g_\theta(x)| < \varepsilon \quad \forall \varepsilon > 0$

for appropr. choices of N, M on a compact set of $\mathbb{R}^N$

- Kolmogorov, Stone-Weierstrass
- Hornik (91)
- Chatziafratis et al. (2020) ICLR, ReLU



UAT for RNN

- approx. any DS arbitr. closely
  - Funahashi & Nakamura (93) NN
  - Kimura & Nakano (98) NN
  - Hanson & Raginsky (2020) PMLR

$$x_t = \phi(\underbrace{W x_{t-1} + h}_{=: z_{t-1}})$$

$$\Rightarrow x_t = W^{-1} z_t - W^{-1} h$$

$$\Rightarrow W^{-1}(z_t - h) = \phi(z_{t-1})$$

$$\Rightarrow z_t = W \phi(z_{t-1}) + h$$

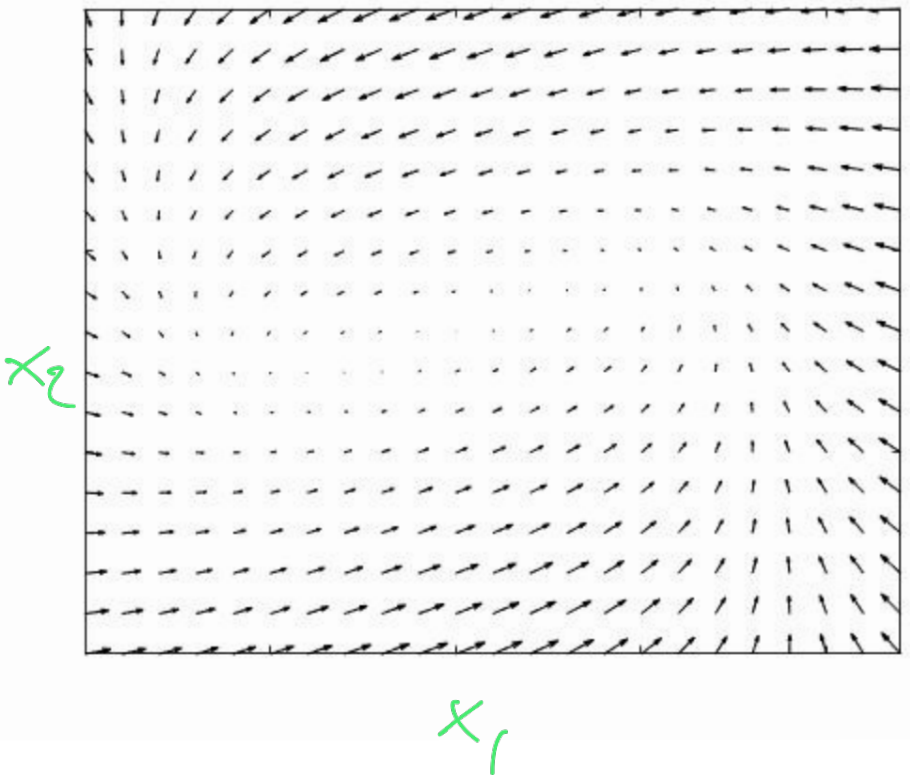
$$z_t = \underbrace{A}_{\text{diag}} z_{t-1} + W \phi(z_{t-1}) + h$$

$$z_t = (1 - \Delta t) I z_{t-1} + \Delta t W \phi(z_{t-1}) + \Delta t h$$

note $\Delta t = 1$

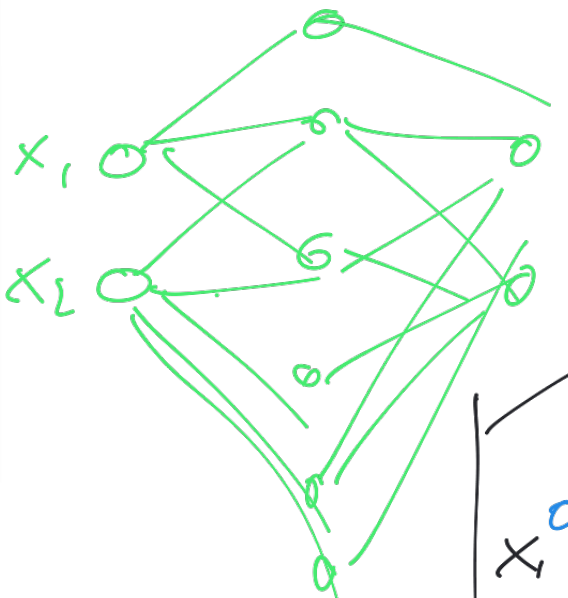
$$\Rightarrow \frac{z_t - z_{t-1}}{\Delta t} = -z_{t-1} + W \phi(z_{t-1}) + h$$

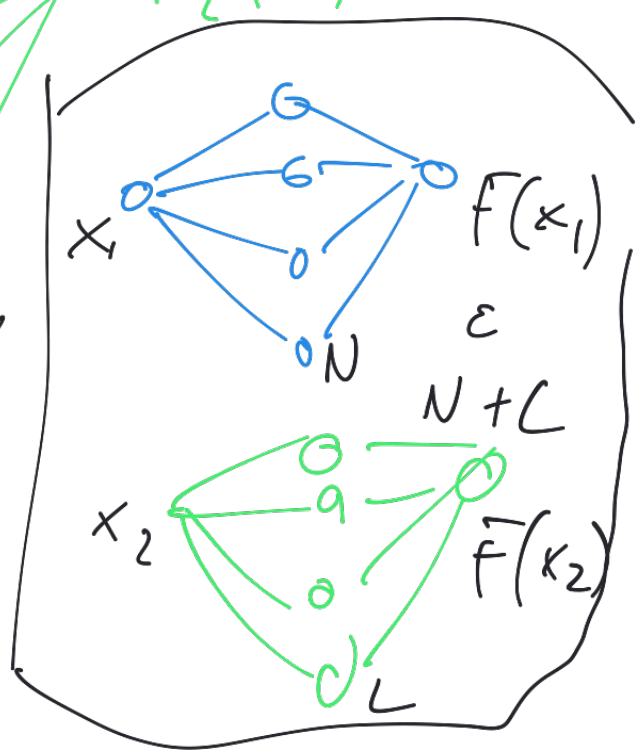
$$\lim_{\Delta t \to 0} \quad \dot{z} = -z + W \phi(z) + h$$



$$\dot{x} \equiv \frac{dx}{dt} = F(x)$$

↑ hidden layer FFN





Let $F(x)$ be the flow (vector) field of a N-dim. DS $\dot{x} = F(x)$, $x \in \mathbb{R}^N$

looking for $N \times M$ $M \times N$

$$|F(x) - \underbrace{A\phi(Bx + L)}| < \varepsilon$$

acc. UAT for FNN

everywhen, $\forall \varepsilon > 0$

Let $\dot{y} = -\frac{1}{\tau} y + A\phi(By + L)$, Σ very Carefo

Training RNN

Supervised tasks

$$\{s_\tau^{(p)}\}_{\tau \in \mathcal{T}} \longrightarrow \{\tilde{x}_t\}_{t \in T}$$

Real-Time Recurrent Learning (RTRL)

→ online learning algo.

Williams & Zipser (1990)

Pearlmutter (1989)

$$D = \{\{s^{(p)}_{\tau \in \mathcal{T}}\}, \{\tilde{x}^{(p)}_{t \in \mathcal{T}}\}\}, \quad p = 1 \dots P$$

$$x_t = \phi(W x_{t-1} + h + \cancel{(\varepsilon_t)}), \quad i = 1 \dots M \text{ units}$$

↳ M-dim. col. vec.

Loss func.

$$\text{error signal } \varepsilon_{it} = \begin{cases} \tilde{x}_{it} - x_{it} & \text{if } \tilde{x}_{it} \text{ is def.} \\ 0 & \text{otherwise} \end{cases} (i,t)$$

$$\mathcal{L}(W, h) := \frac{1}{2} \cancel{\sum_{p=1}^{P}} \sum_{t=1}^{T} \sum_{i=1}^{M} \varepsilon_{it}^2$$

$$= \frac{1}{2} \sum_{t=1}^{T} \sum_{i=1}^{M} \mathbb{I}\{\varepsilon_{it} \neq 0\} (\tilde{x}_{it} - x_{it})^2$$

$$\frac{\partial \mathcal{L}(w)}{\partial w_{ij}} = \frac{1}{2} \sum_{t=1}^{T} \frac{\partial \mathcal{L}_t(w)}{\partial w_{ij}}$$

$$\frac{\partial \mathcal{L}_t(w)}{\partial w_{ij}} = \sum_{k=1}^{M} \mathbb{I}\{\varepsilon_{kt} \neq 0\} (\tilde{x}_{kt} - x_{kt}) \frac{\partial x_{kt}}{\partial w_{ij}}$$

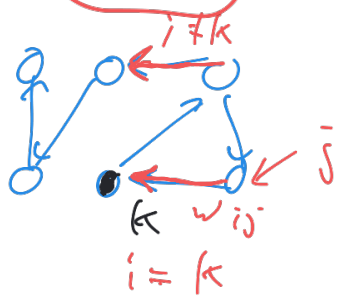
$$\frac{\partial x_{kt}}{\partial w_{ij}} = \phi'\left(\sum_{l=1}^{M} w_{kl} x_{l,t-1} + h_k\right) \cdot \left[\delta_{ki} x_{j,t-1} + \sum_{l=1}^{M} w_{kl} \frac{\partial x_{l,t-1}}{\partial w_{ij}}\right]$$

δ_{ki}: Kronecker

$\gamma^{(t)}_{kij}$

$\gamma^{(t-1)}_{lij}$

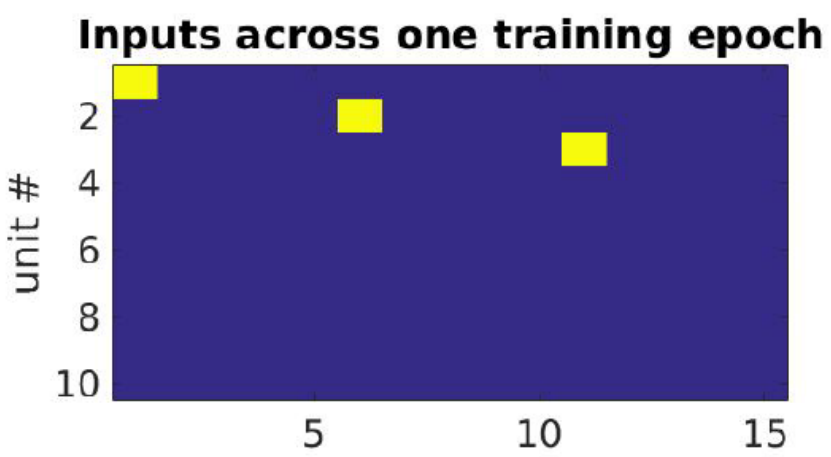
$(fg)' = f'g + fg'$ for $i = k$

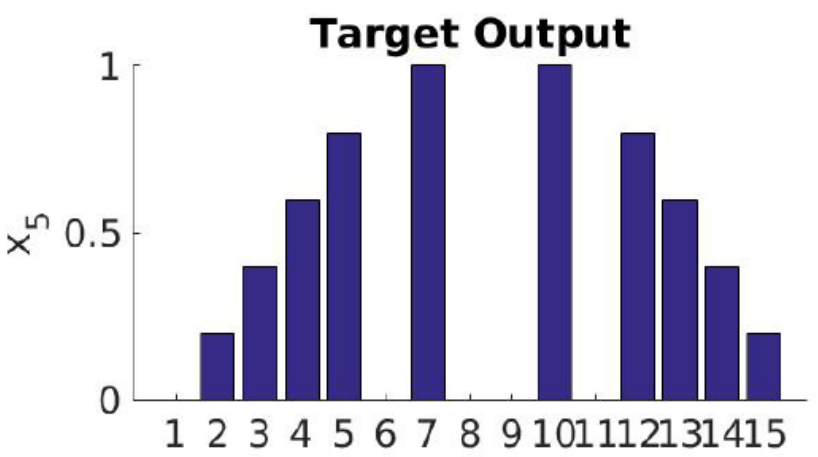


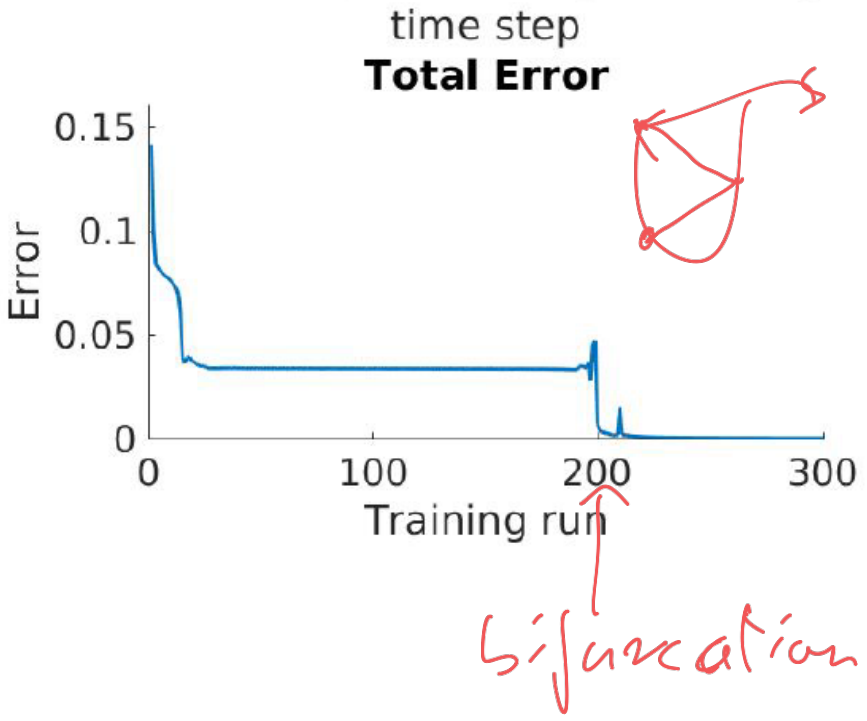
RTRL algo.

- init. cond.: $\gamma^{(0)}_{lij} = 0 \quad \forall (l, i, j)$
- update $\gamma^{(t)}_{lij}$ along with act. val. $x^{(t)}_l$
- update $w^{(t+1)}_{ij} = w^{(t)}_{ij} - \alpha \frac{\partial \mathcal{L}_t}{\partial w_{ij}}$

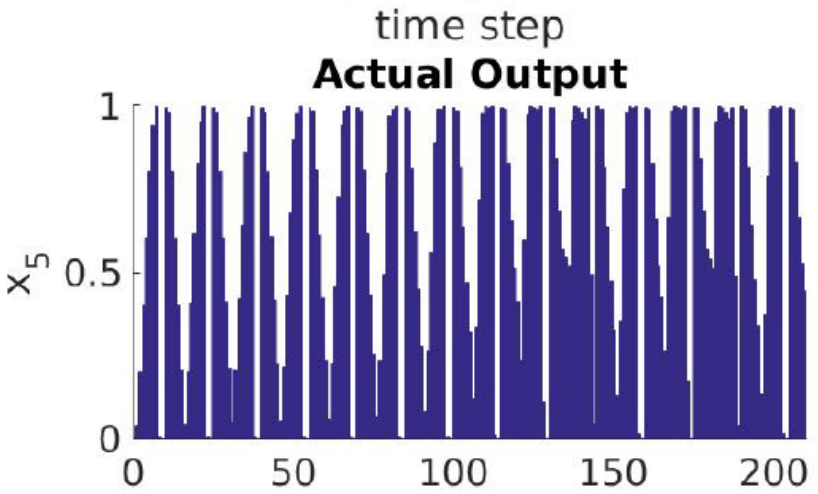
Source: Durstewitz (2017) *Advanced Data Analysis in Neuroscience.* Springer.







bifurcation



Source: Durstewitz (2017) *Advanced Data Analysis in Neuroscience.* Springer