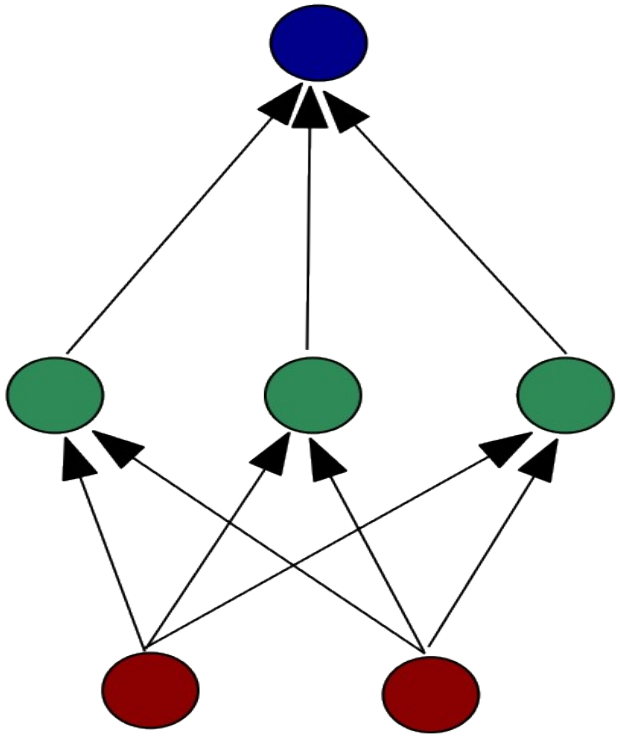
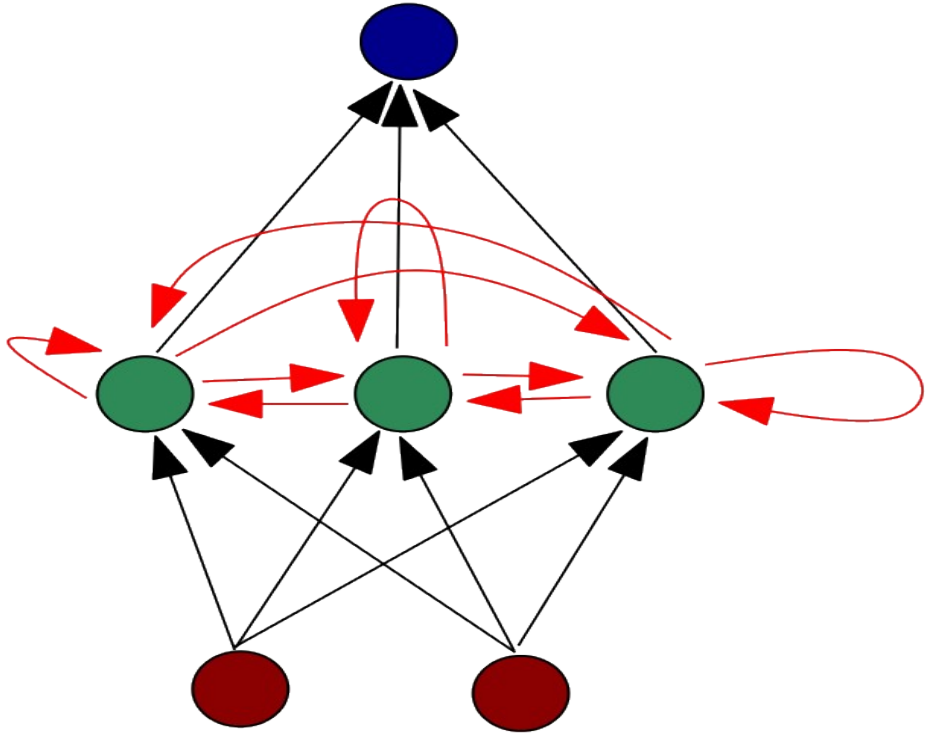
Brief Introduction to Recurrent Neural Models

Razvan Pascanu

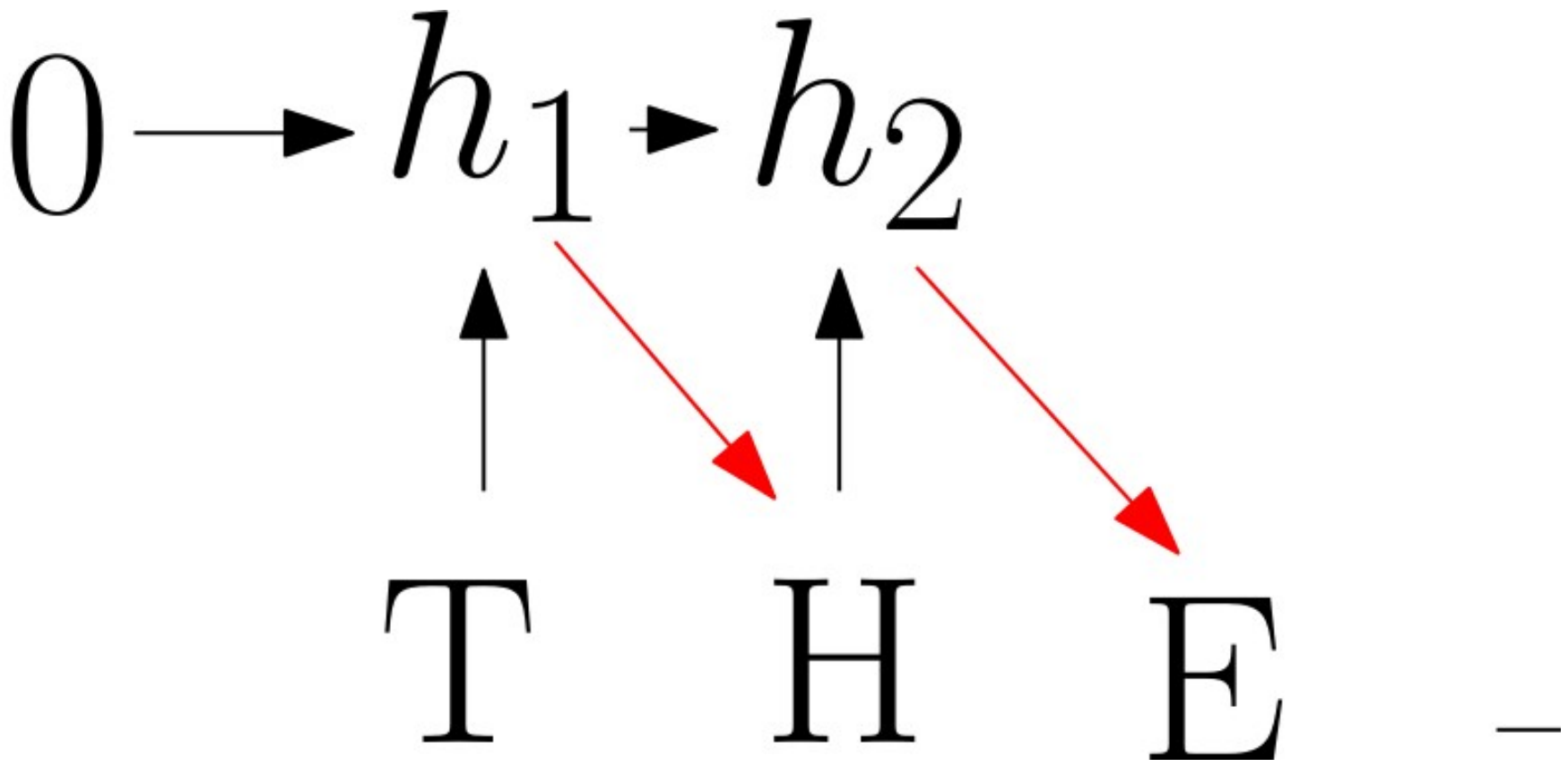
Université de Montréal

Recurrent Neural Models

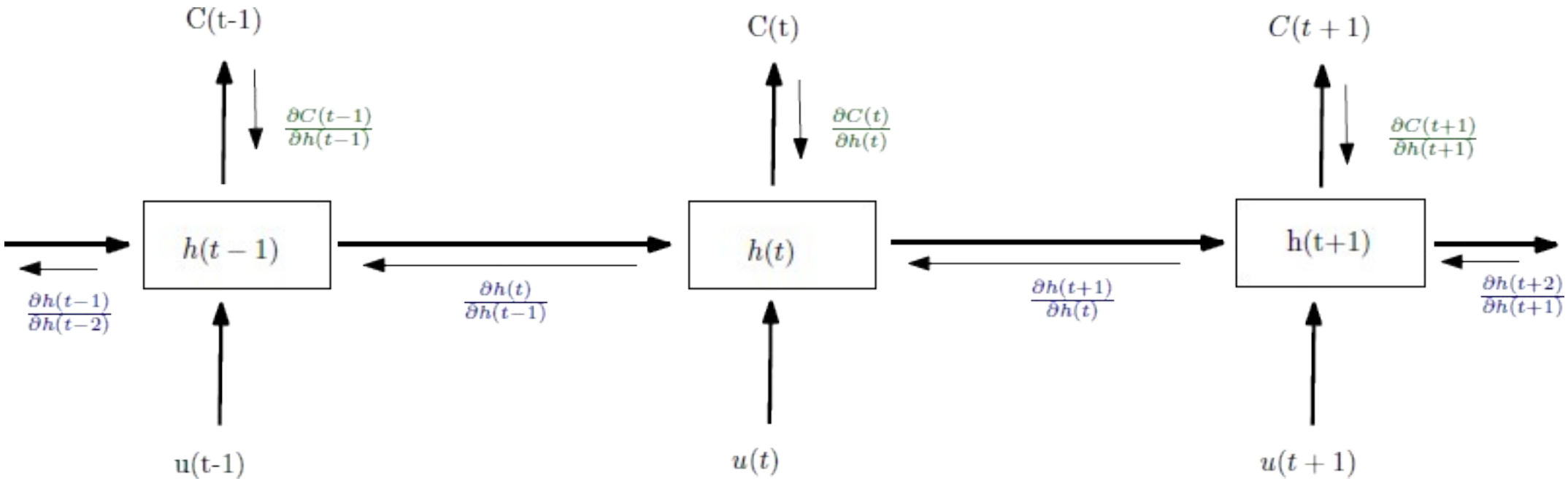




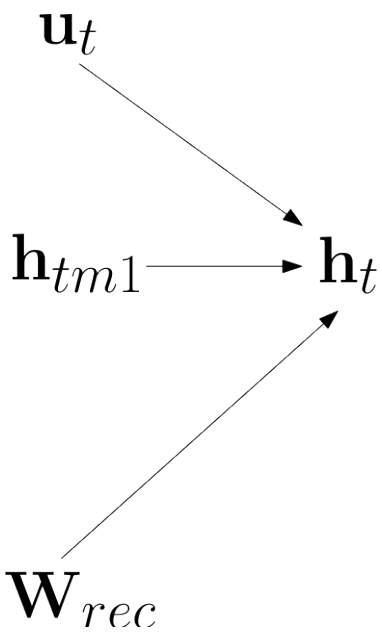
RNN for Language modelling



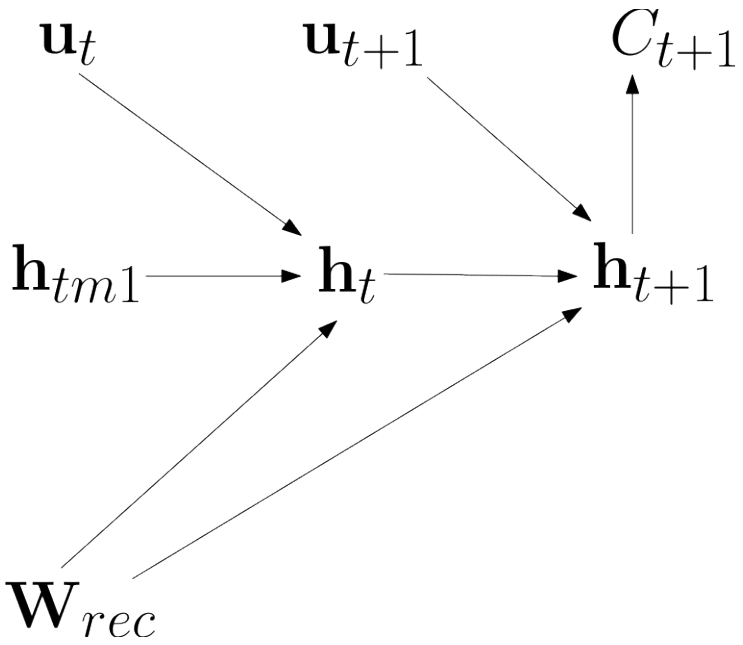
RNNs – Backpropagation Through Time



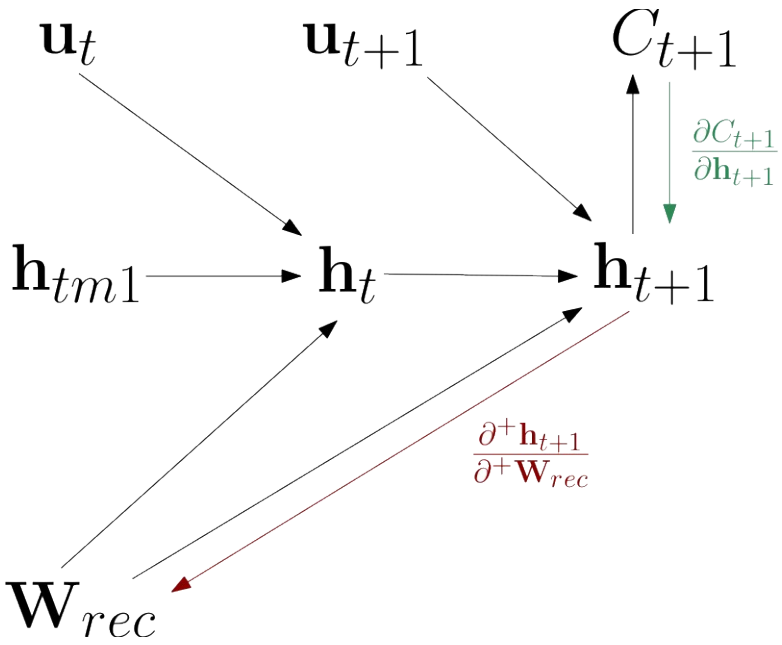
RNNs – Backpropagation Through Time



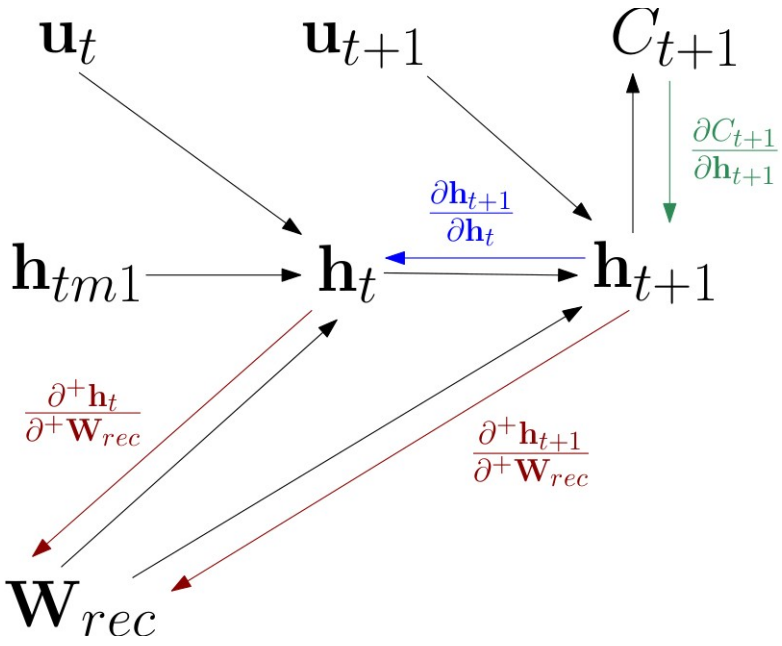
RNNs – Backpropagation Through Time



RNNs – Backpropagation Through Time



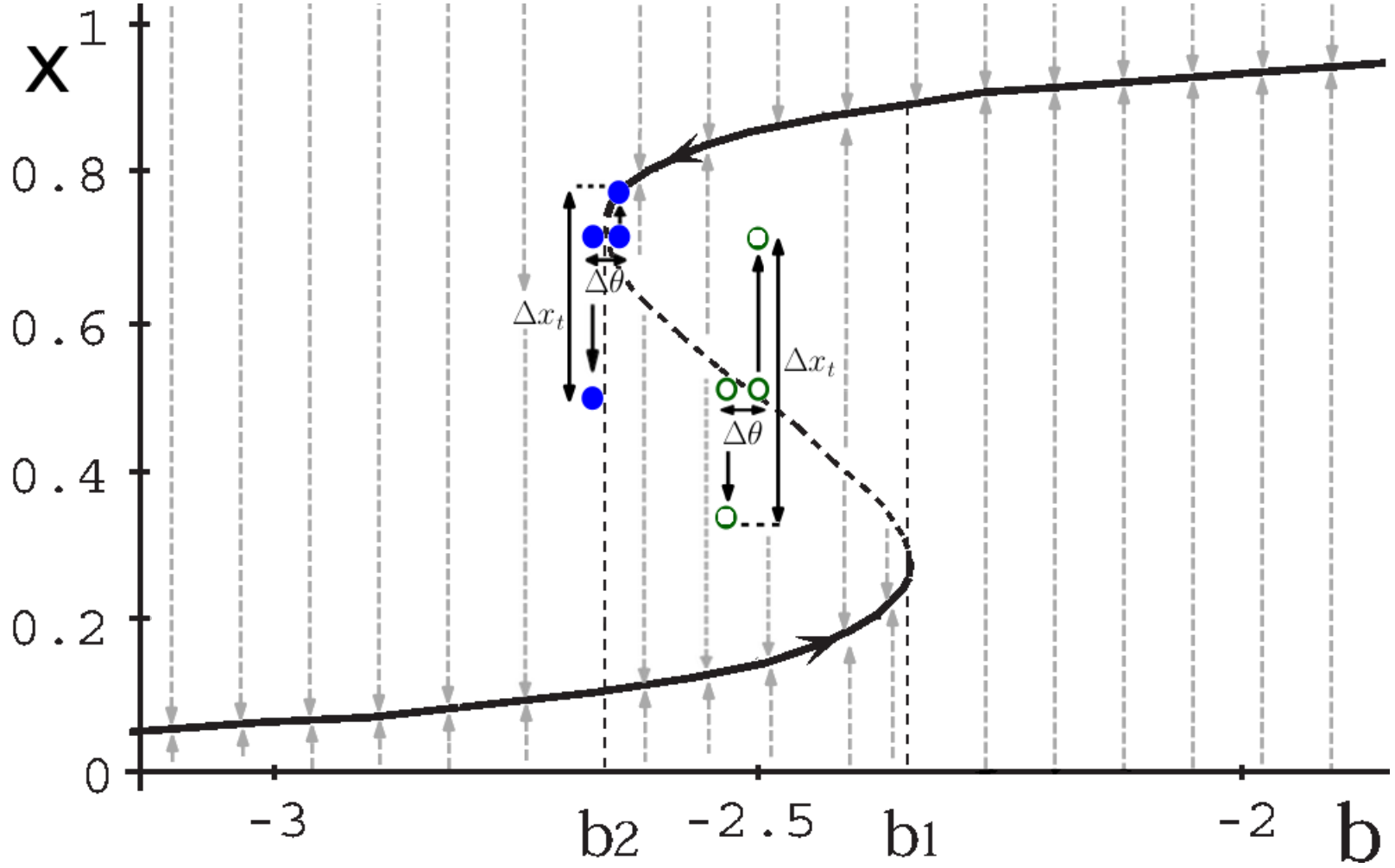
RNNs – Backpropagation Through Time



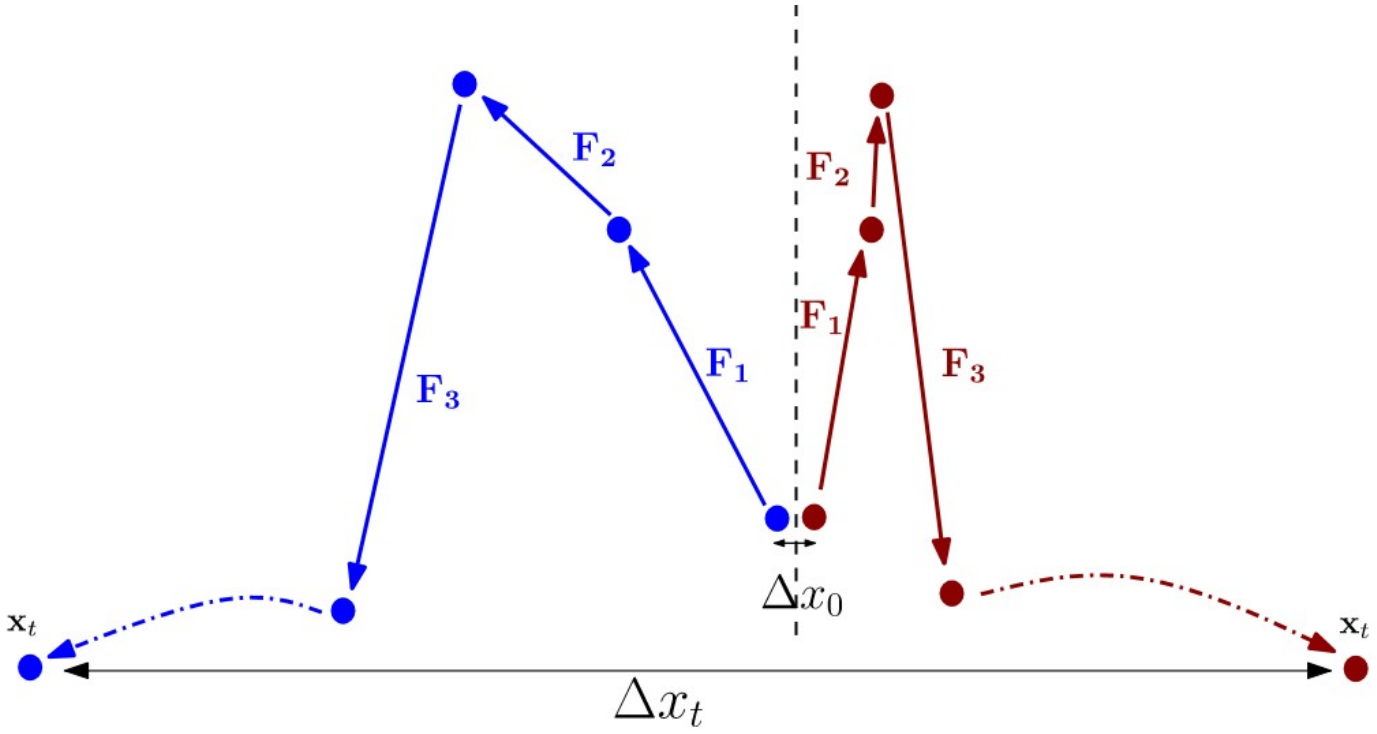
$$\frac{\partial C}{\partial \mathbf{W}} = \sum_t \frac{\partial C(t)}{\partial \mathbf{W}} = \sum_t \frac{\partial C(t)}{\partial \mathbf{h}(t)} \left(\sum_{k=0}^{t} \frac{\partial \mathbf{h}(t)}{\partial \mathbf{h}(t-k)} \frac{\partial \mathbf{h}(t-k)}{\partial \mathbf{W}} \right)$$

$$\frac{\partial \mathbf{h}(t)}{\partial \mathbf{h}(t-k)} = \prod_{j=k+1}^{t} \frac{\partial \mathbf{h}(j)}{\partial \mathbf{h}(j-1)}$$

Dynamical System Perspective



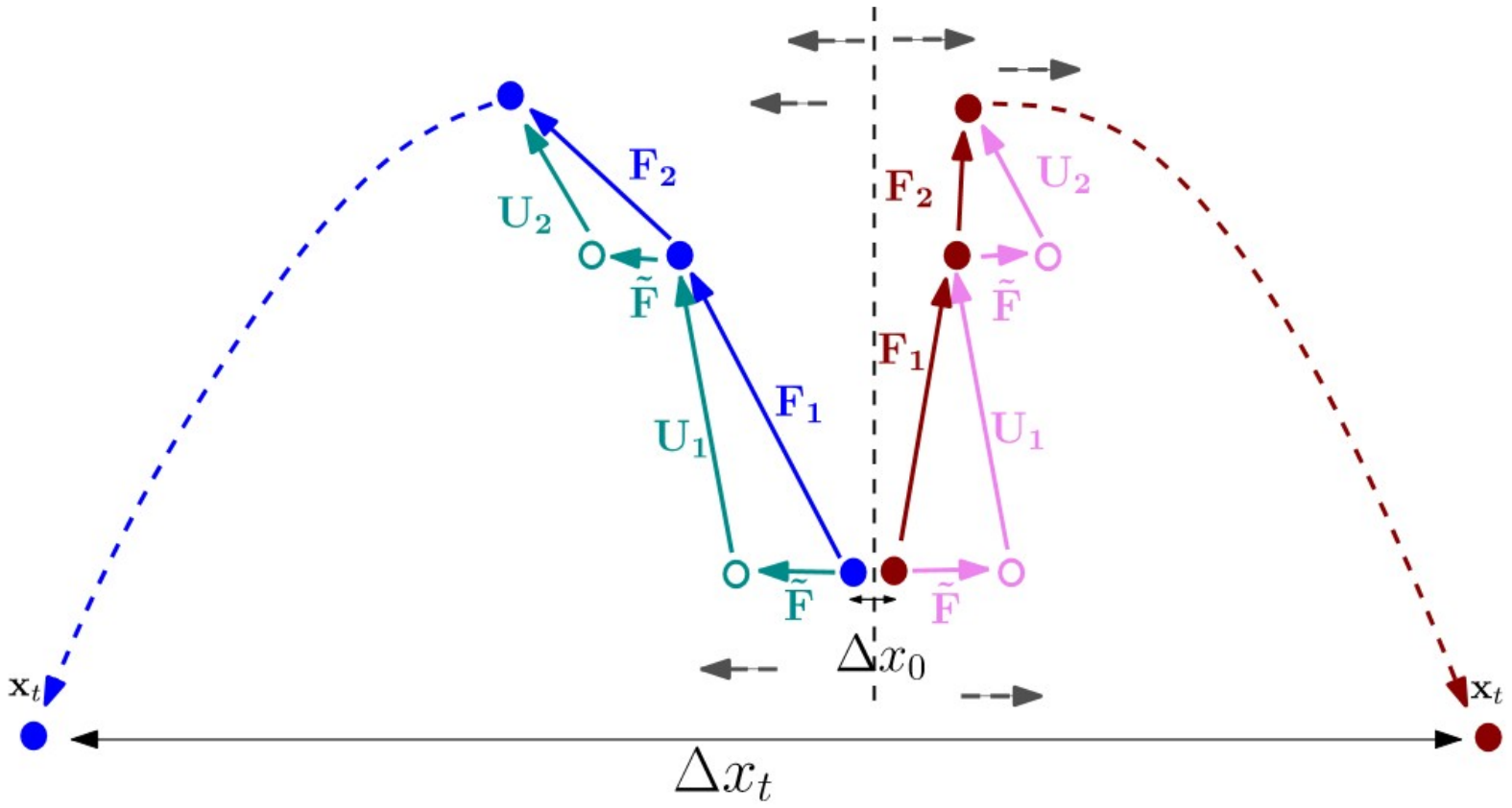
Dynamical System Perspective



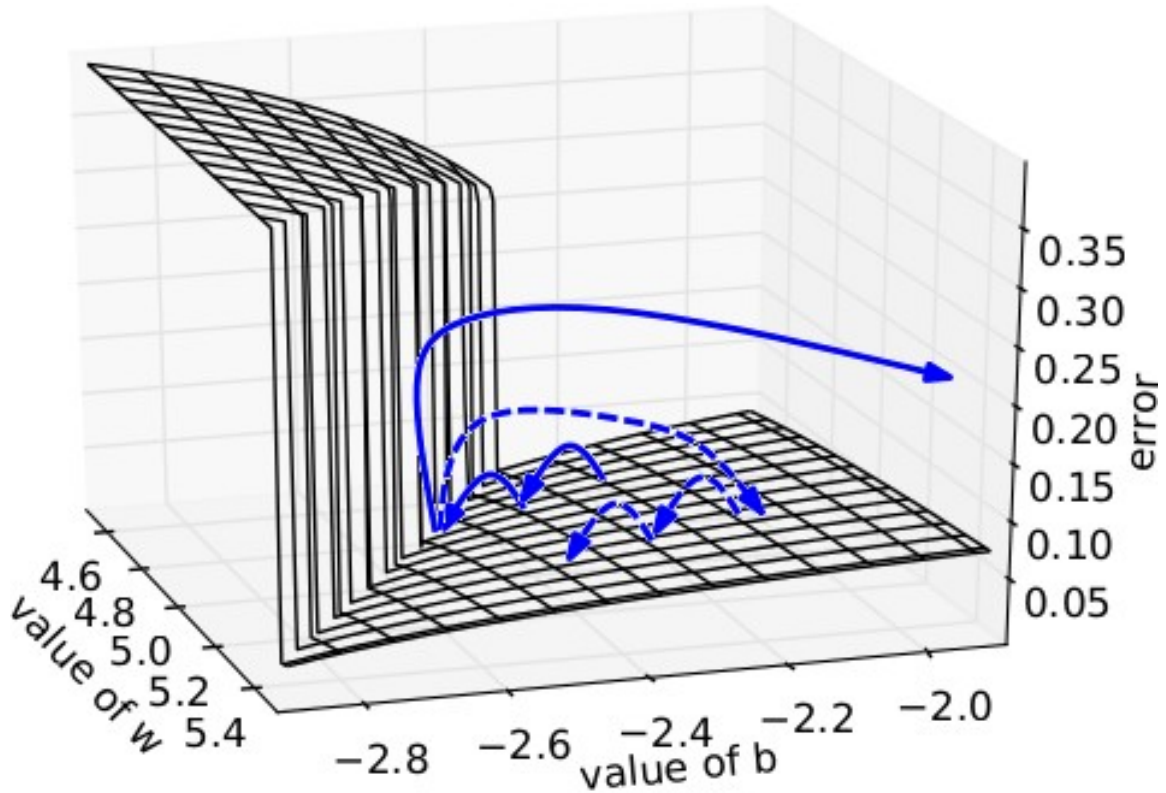
$$\mathbf{x}_t = F(\mathbf{x}_{t-1}, \mathbf{u}_t, \theta)$$

Dynamical System Perspective

$$\mathbf{x}_t = \mathbf{W}_{rec}\sigma(\mathbf{x}_{t-1}) + \mathbf{W}_{in}\mathbf{u}_t + \mathbf{b}$$



Geometrical View



The error is $(x(50)-0.7)^2$ for $x(t) = w\sigma(x(t-1)) + b$ with $x(0) = 0.5$

Norm Clipping

- Originally used by Tomas to get state of the art results in LM
- Modified here to be more theoretically justifiable

$$\hat{\mathbf{g}} \leftarrow \frac{\partial error}{\partial \theta}$$

$$\textbf{if } \|\hat{\mathbf{g}}\| \geq threshold \textbf{ then}$$

$$\hat{\mathbf{g}} \leftarrow \frac{threshold}{\|\hat{\mathbf{g}}\|}\hat{\mathbf{g}}$$

$$\textbf{end if}$$

Regularization term

$$\Omega = \sum_k \Omega_k = \sum_k \left(\frac{\left\| \frac{\partial C}{\partial \mathbf{x}_{k+1}} \frac{\partial \mathbf{x}_{k+1}}{\partial \mathbf{x}_k} \right\|}{\left\| \frac{\partial C}{\partial \mathbf{x}_{k+1}} \right\|} - 1 \right)^2$$

Regularization term

$$\frac{\partial^+ \Omega}{\partial \mathbf{W}_{rec}} = \sum_k \frac{\partial^+ \Omega_k}{\partial \mathbf{W}_{rec}}$$

$$= \sum_k \frac{\partial^+ \left(\frac{\left\| \frac{\partial C}{\partial \mathbf{x}_{k+1}} \mathbf{W}_{rec}^T diag(\sigma'(\mathbf{x}_k)) \right\|^2}{\left\| \frac{\partial C}{\partial \mathbf{x}_{k+1}} \right\|^2} - 1 \right)^2}{\partial \mathbf{W}_{rec}}$$

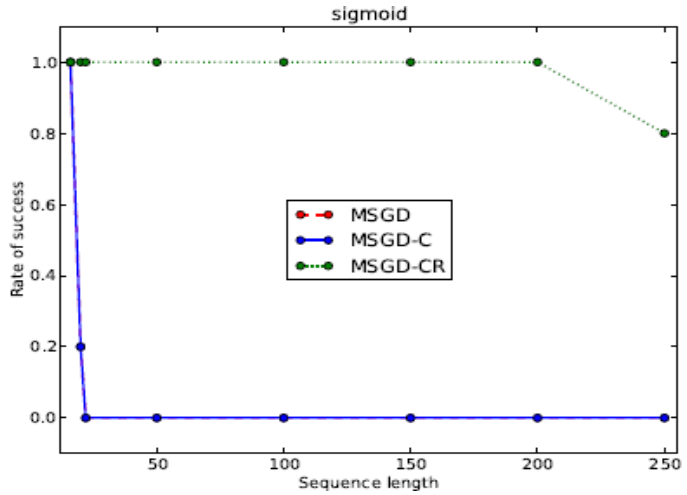
Other approaches

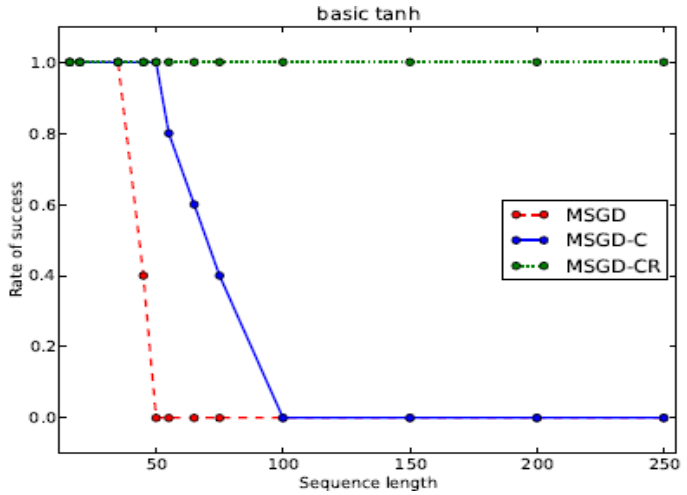
- LSTMs
- ESNs
- L1/L2 norm
- Hessian-Free
- Truncated BPTT

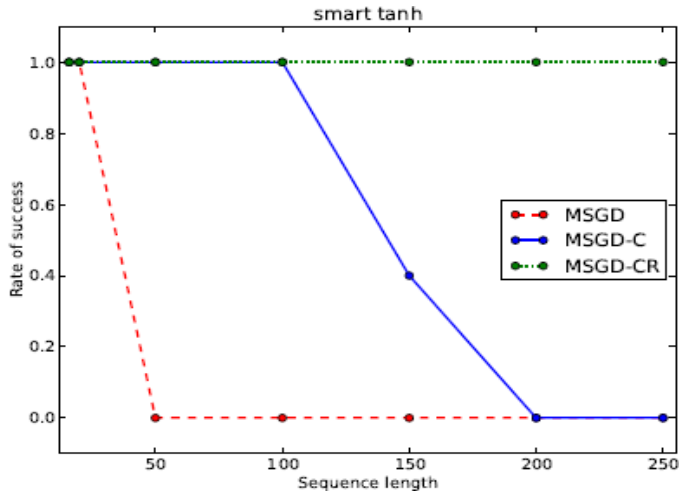
Some results

Data set	Data fold	MSGD	MSGD+C	MSGD+CR	State of the art for RNN	State of the art
Piano-midi.de (nll)	train	6.87	6.81	**7.01**	7.04	6.32
	test	7.56	7.53	**7.46**	7.57	7.05
Nottingham (nll)	train	3.67	3.21	**2.95**	3.20	1.81
	test	3.80	3.48	**3.36**	3.43	2.31
MuseData (nll)	train	8.25	6.54	**6.43**	6.47	5.20
	test	7.11	7.00	**6.97**	6.99	5.60
Penn Treebank 1 step (bits/char)	train	1.46	**1.34**	1.36	N/A	N/A
	test	1.50	1.42	**1.41**	**1.41**	1.37
Penn Treebank 5 steps (bits/char)	train	N/A	3.76	**3.70**	N/A	N/A
	test	N/A	3.89	**3.74**	N/A	N/A

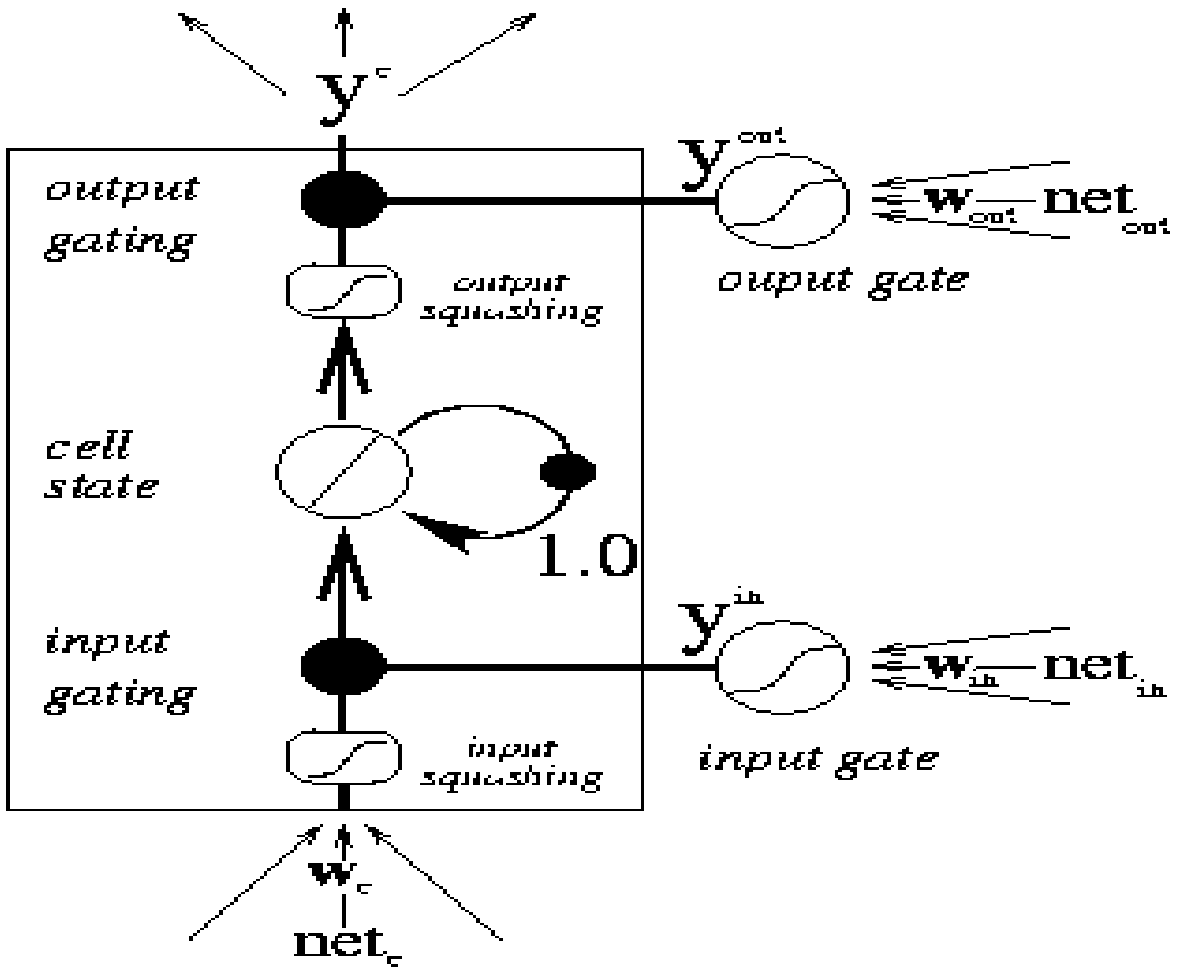
Temporal order task:



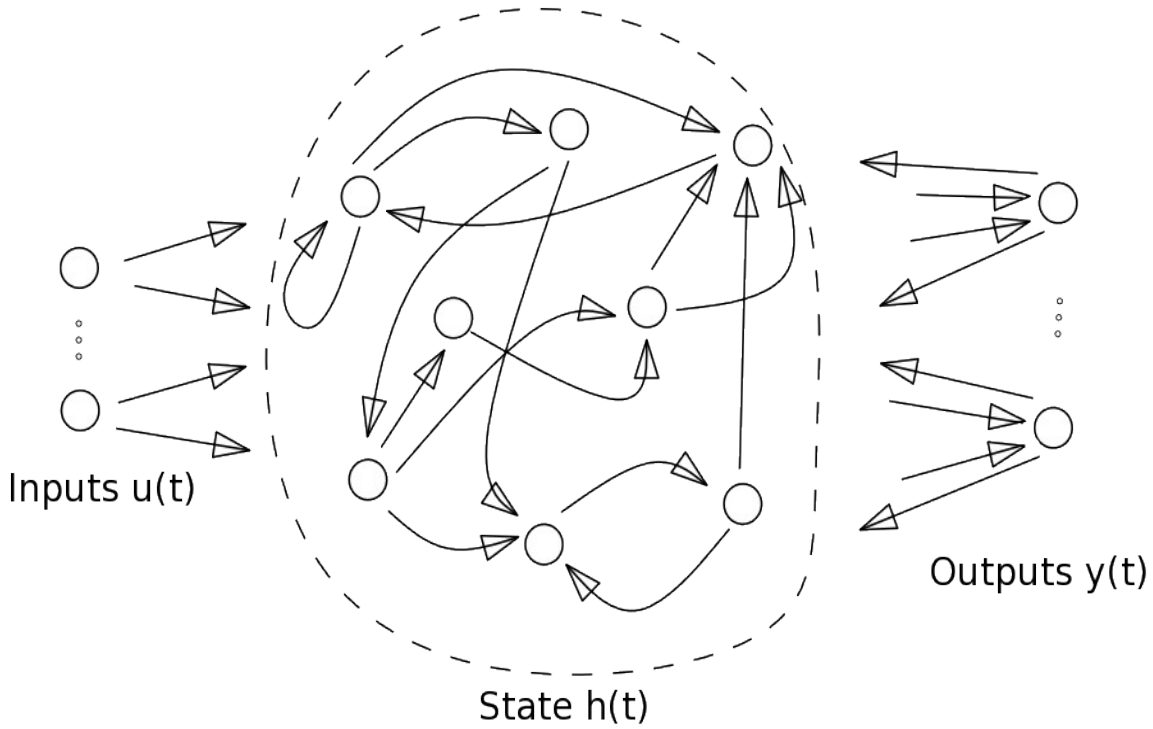




LSTM

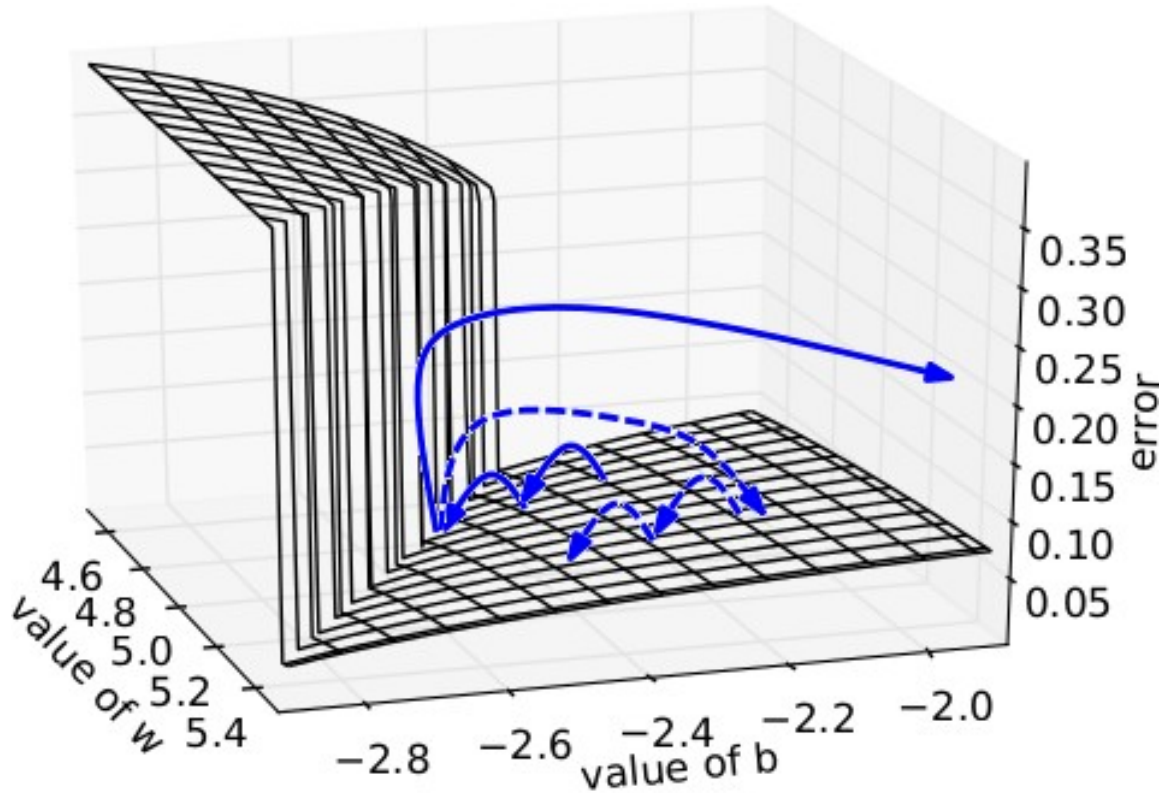

output gating
output squashing
cell state
1.0
input gating
input squashing
output gate
input gate

Echo State Property



$$\mathbf{x}_t = F(\mathbf{x}_{t-1}, \mathbf{u}_t, \theta)$$

Hessian-Free



Thank you !

- Questions ?