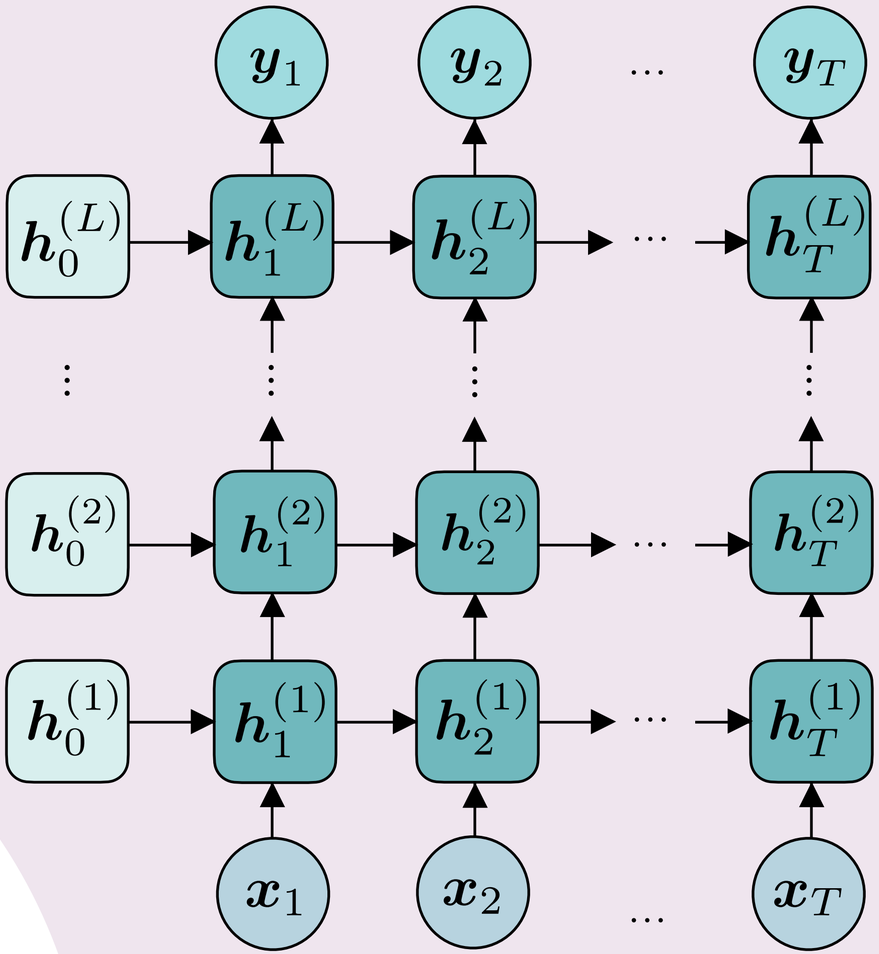
On the Role of Depth in the Expressivity of RNNs

Maude Lizaire,
Michael Rizvi-Martel,
Éric Dupuis &
Guillaume Rabusseau

Poster **#171**
Come see us
:-)

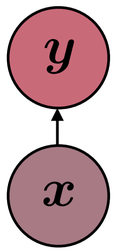


Mila

Université de Montréal

Depth in Feedforward Neural Network

Shallow

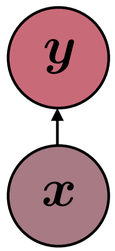


Essentially
linear model

$$\mathbf{y} = \phi(\mathbf{V}\mathbf{x} + \mathbf{c})$$

Depth in Feedforward Neural Network

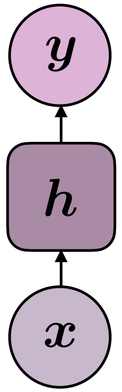
Shallow



Essentially
linear model

$$\mathbf{y} = \phi(\mathbf{V}\mathbf{x} + \mathbf{c})$$

Just a little bit of depth



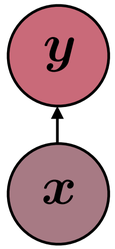
Universal
approximation
Theorem

$$\mathbf{y} = \phi(\mathbf{V}\mathbf{x} + \mathbf{c})$$

$$\mathbf{h} = \sigma(\mathbf{W}\mathbf{x} + \mathbf{b})$$

Depth in Feedforward Neural Network

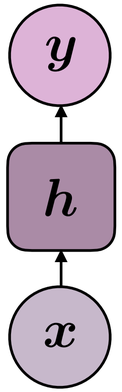
Shallow



Essentially linear model

$$\mathbf{y} = \phi(\mathbf{V}\mathbf{x} + \mathbf{c})$$

Just a little bit of depth

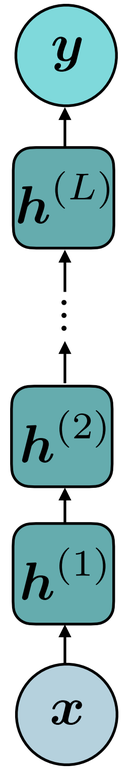


Universal approximation Theorem

$$\mathbf{y} = \phi(\mathbf{V}\mathbf{x} + \mathbf{c})$$

$$\mathbf{h} = \sigma(\mathbf{W}\mathbf{x} + \mathbf{b})$$

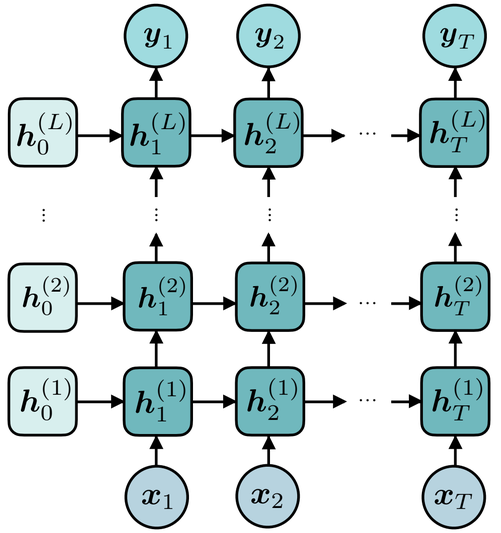
Deep FNN



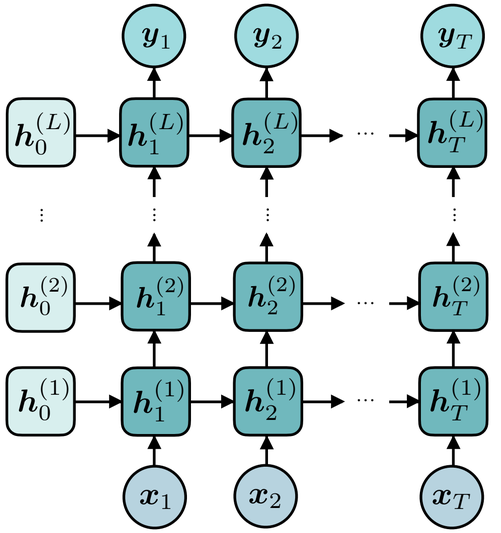
More layers
=
Exponentially less parameters required

But Linear FNN collapses ! —> Shallow

Our work : What about RNNs?



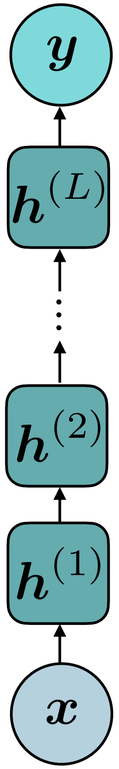
Our work : What about RNNs?



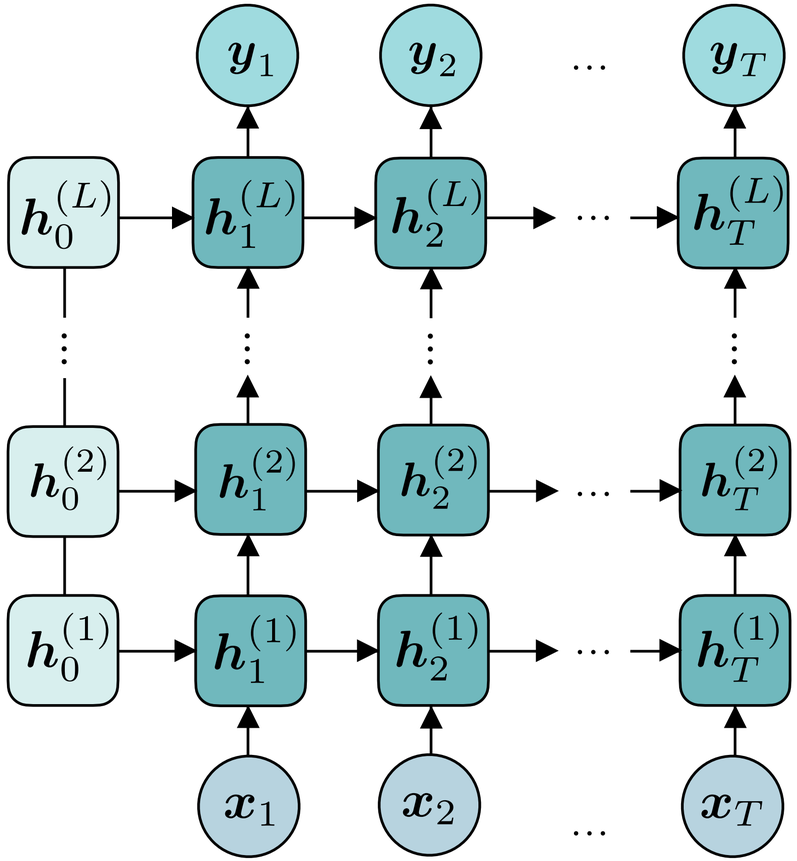
- Are there **benefits to depth** in RNNs?
- What about **linear** RNNs? Do they **collapse** like FNNs?
- Is it the same for 2RNNs? $\mathbf{h}_t^{(l)} = \underbrace{\boldsymbol{\mathcal{A}}^{(l)} \times_1 \mathbf{h}_{t-1}^{(l)} \times_2 \mathbf{h}_t^{(l-1)}}_{\text{Multiplicative interactions}} + \mathbf{U}^{(l)}\mathbf{h}_t^{(l-1)} + \mathbf{V}^{(l)}\mathbf{h}_{t-1}^{(l)} + \mathbf{b}^{(l)}$
- Depth vs. Multiplicative interactions?

Are there benefits to depth in RNNs? Yes, of course.

They inherit the same benefits as FNNs.



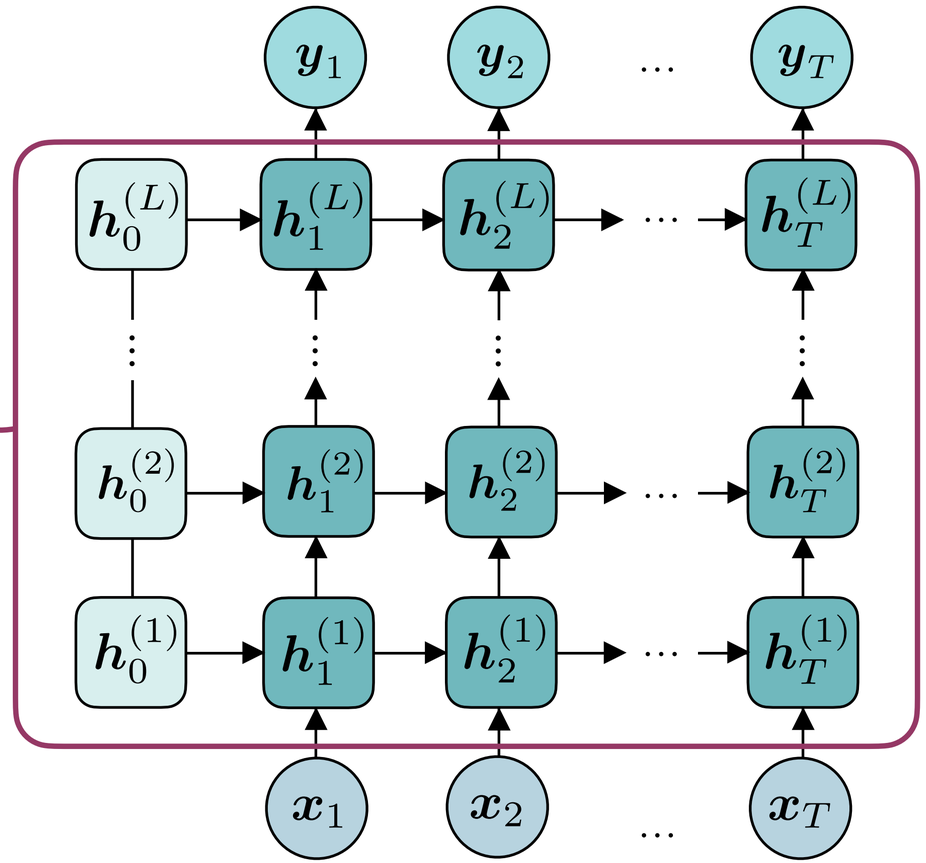
Composition of nonlinear activations



What about linear RNNs, do they collapse?

Linear **RNN**

Linear map

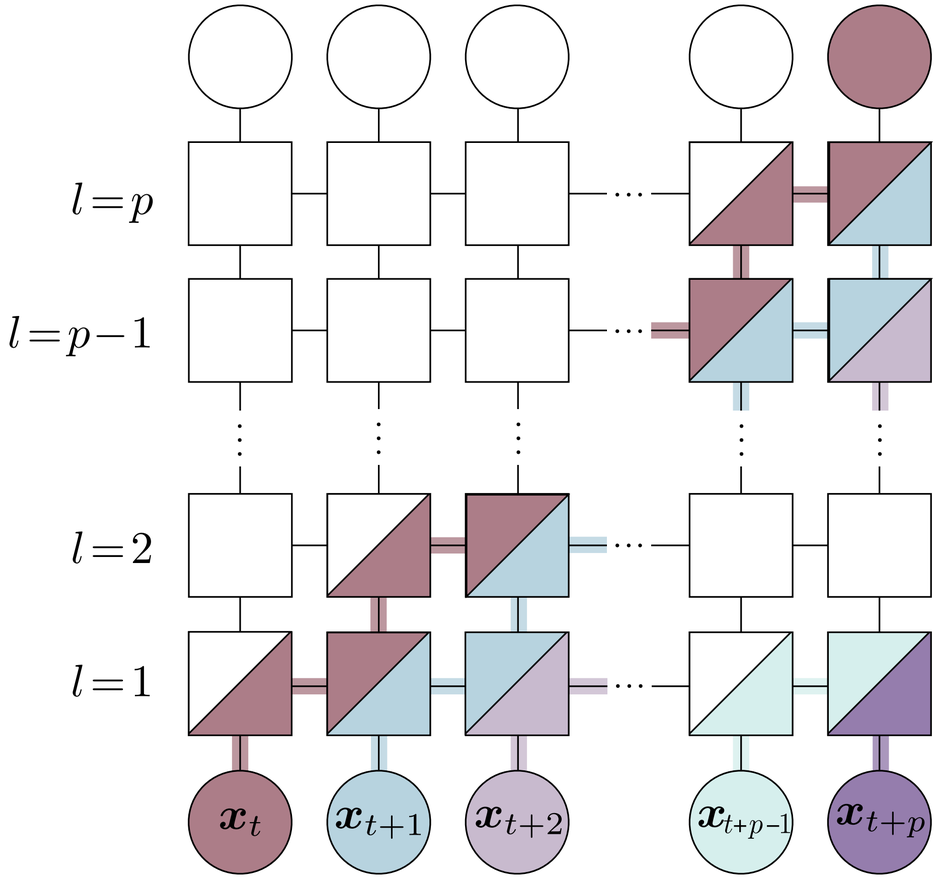


What about linear RNNs, do they collapse? No!

Deep linear RNNs are **strictly more expressive** than shallow ones.

Depth increases **memory** capacity

And it does so **efficiently**!

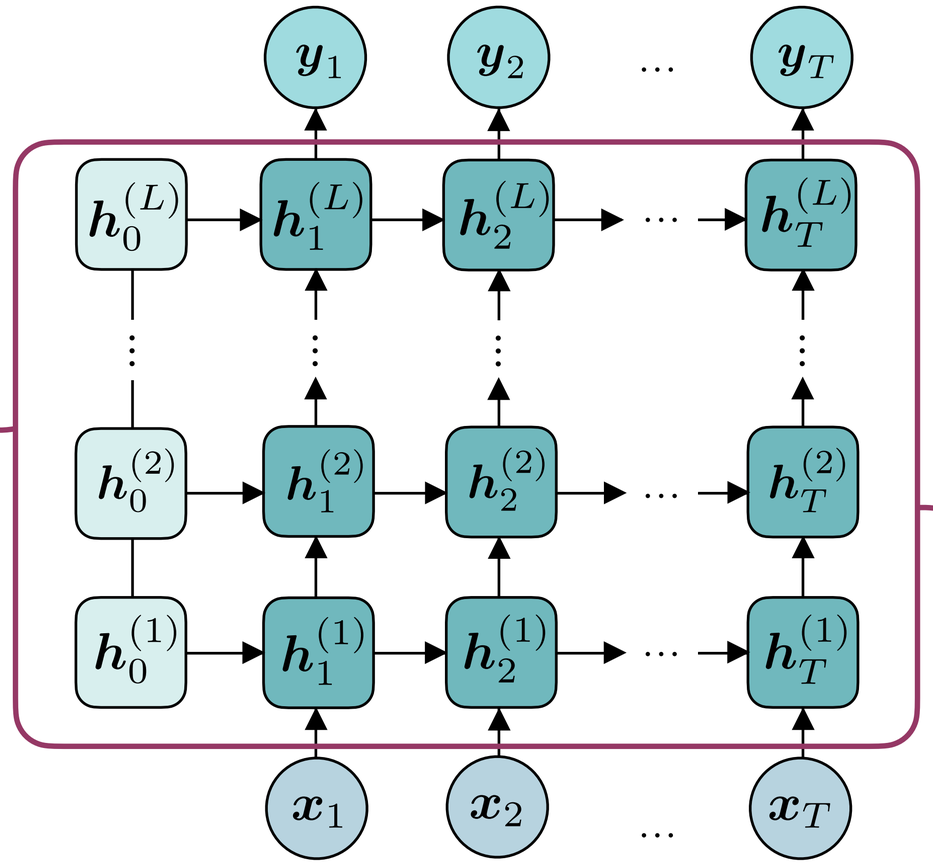


Is it the same for 2RNNs? No, there is more!

Depth broadens the class of functions 2RNNs can represent.

Linear **RNN**

Linear map



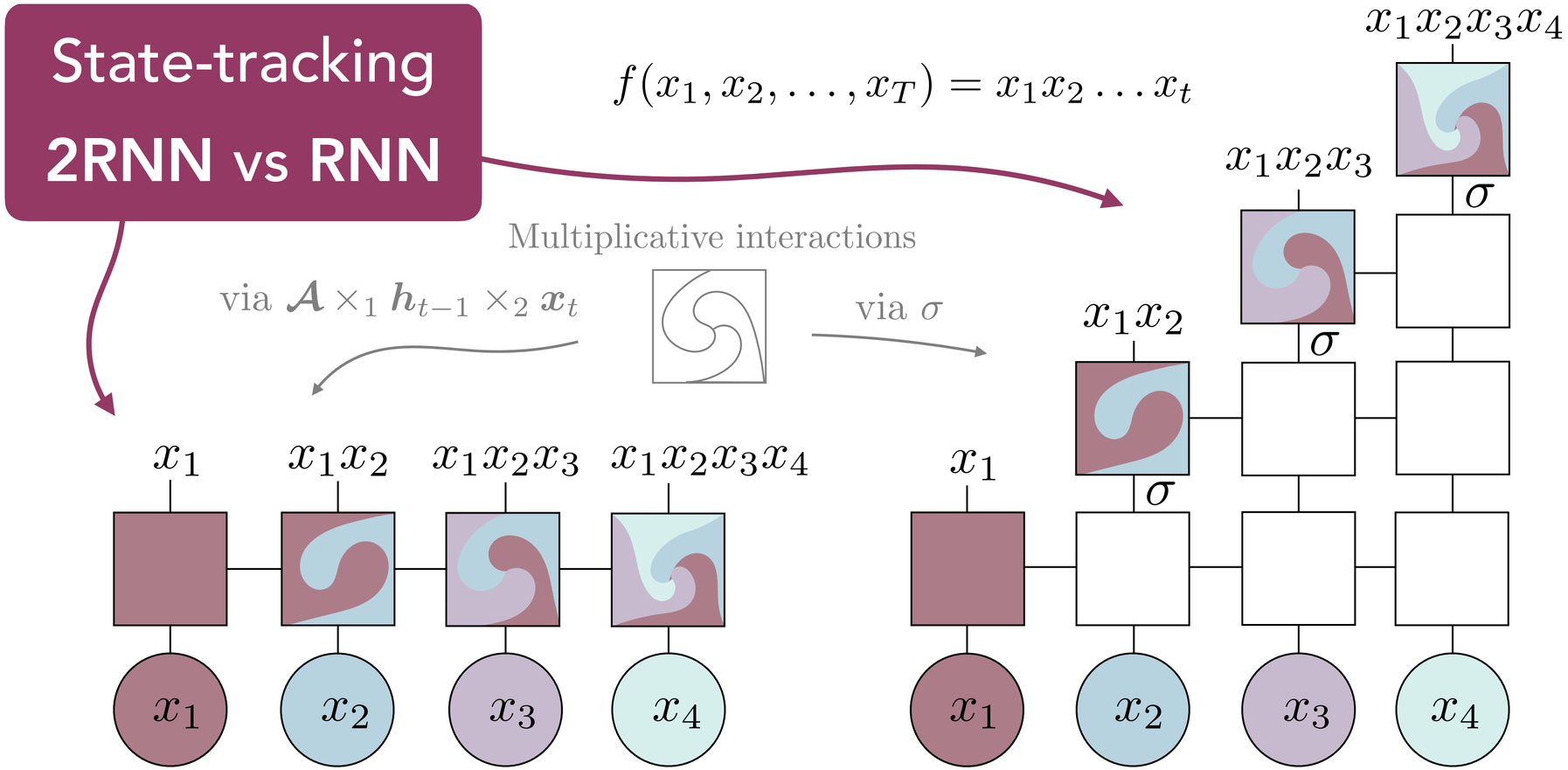
Linear **2RNN**

Polynomial of degree T^L

We are far from the shallow now …

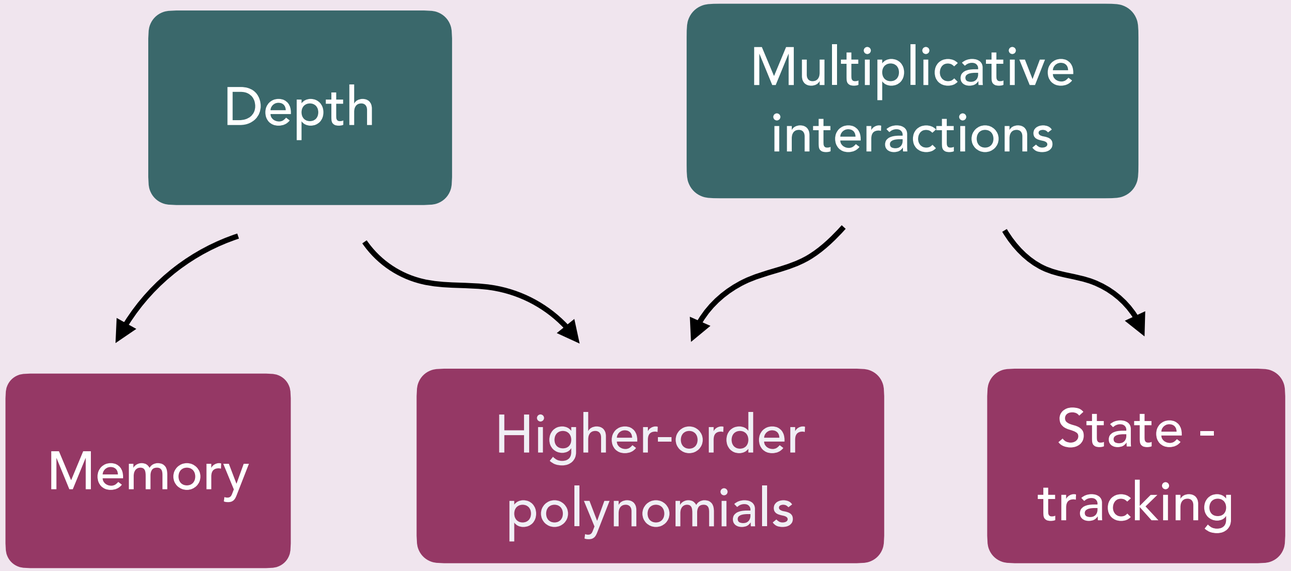
Depth vs multiplicative interactions? Not the same!

Depth-wise nonlinearities cannot replace multiplicative interactions



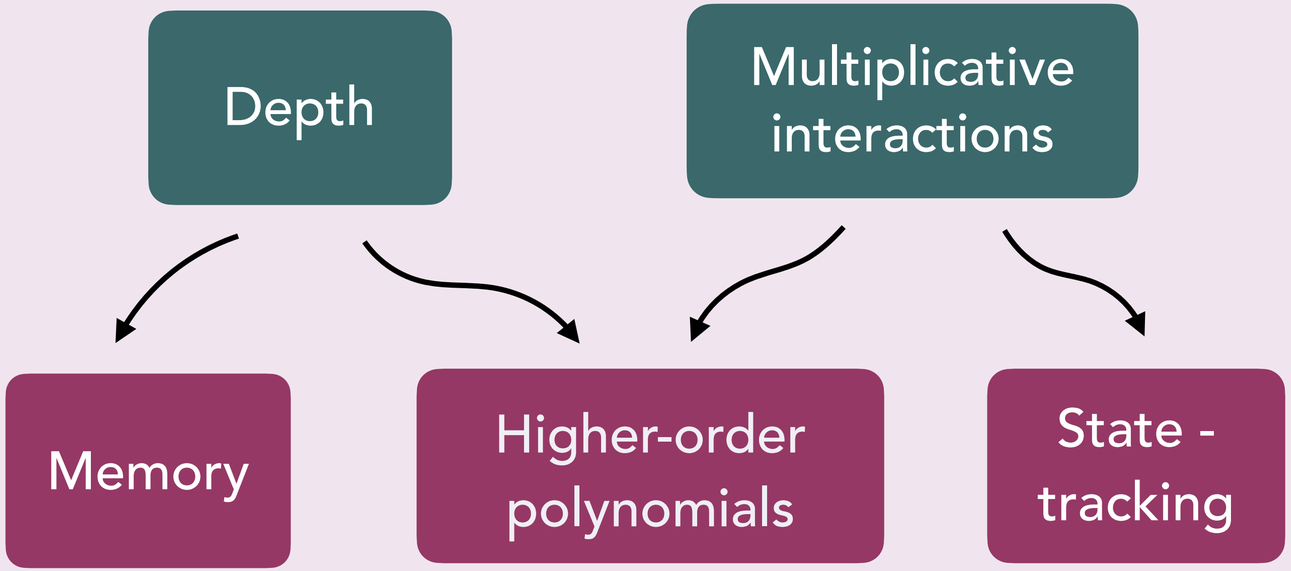
Take aways

How architectural features influence the expressivity of recurrent networks



Take aways

How architectural features influence the expressivity of recurrent networks



Poster #171
Come see us
:-)