Recurrent neural networks: vanishing and exploding gradients are not the end of the story

Nicolas Zucchet
Antonio Orvieto

ETH zürich
ellis INSTITUTE TÜBINGEN

Motivation

Deep state-space models / RNNs models in practice (Mamba, Griffin, xLSTMs, S4…)

Deep state-space models / recurrent networks in theory

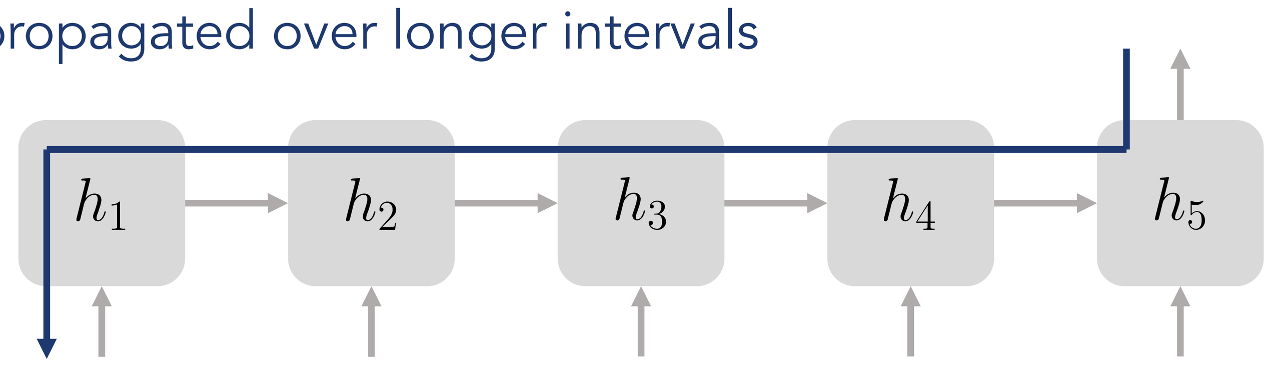
Gu and Dao, *COLM* 2023; De et al. 2024; Beck et al. 2024; Gu et al. *ICLR* 2023

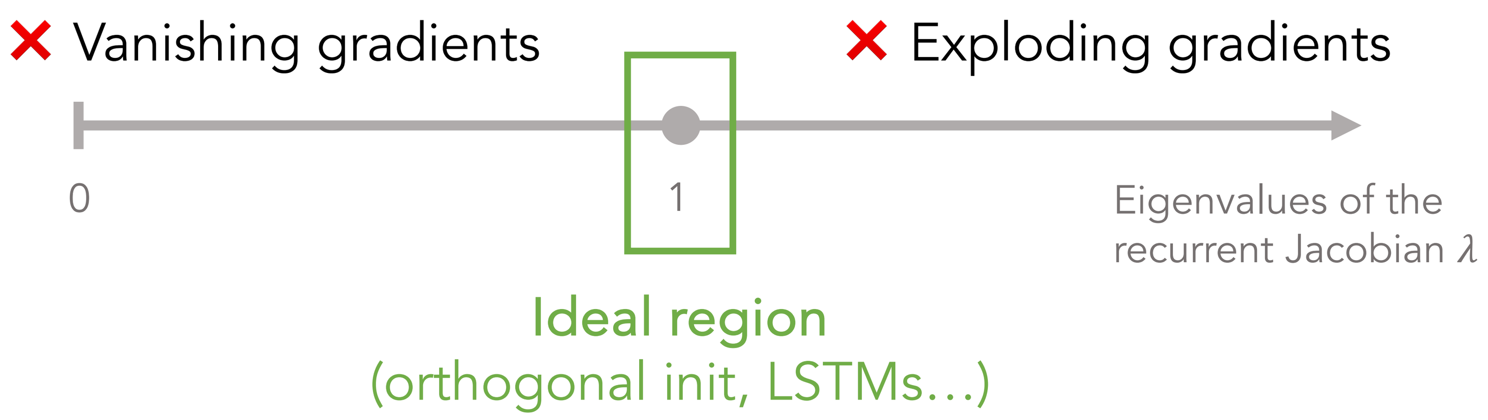
Our contributions

- We identify, for the first time, a **critical issue** arising in the training of RNNs. We analyze it theoretically in great details.
- We show that **successful RNNs mitigate this issue**.
- Our theory **precisely captures learning dynamics** in a controlled teacher-student task and is a good approximation for more realistic scenarios.

Vanishing and exploding gradients

Backpropagated errors tend to **vanish** or **explode** as they are propagated over longer intervals

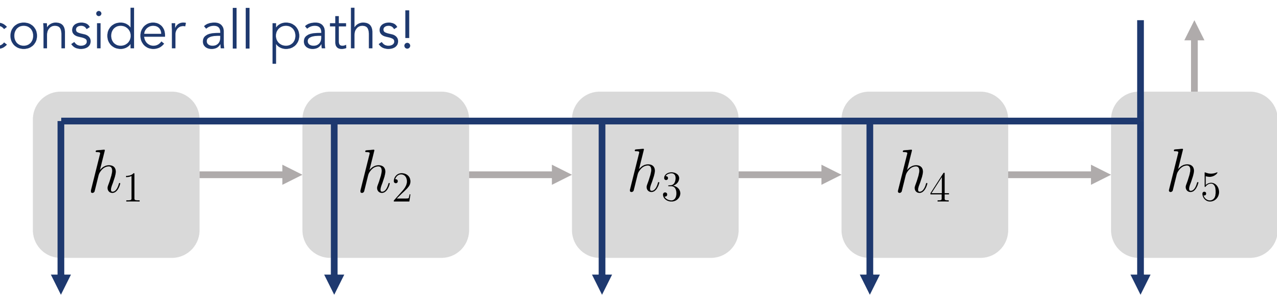




Hochreiter, *Master's thesis* 1991; Bengio et al., *IEEE trans. neural net.* 1994; Hochreiter et al., *IEEE* 2001; Pascanu et al., *ICML* 2013.

The curse of memory: intuition

As the **same function** is repeated **multiple times**, we need to consider all paths!



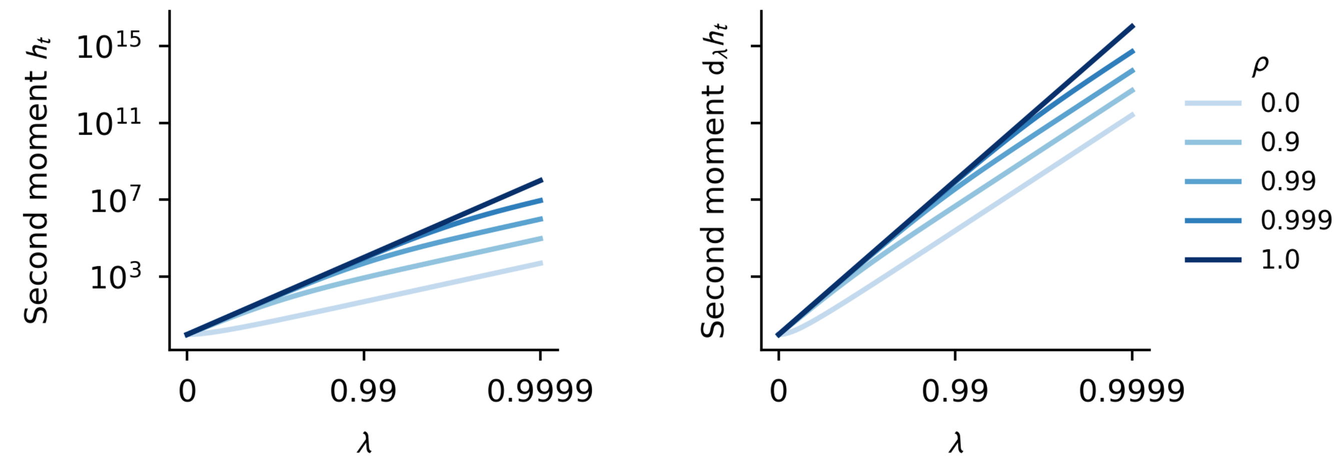
Even if each term decays exponentially fast, the sum (i.e. the gradient) **diverges as λ goes to 1**.

The curse of memory: theory

Deep SSMs (e.g. S4, S5, LRU) are great for mathematical analysis because their **recurrent update is very simple**.

$$\underbrace{h_{t+1}}_{\text{hidden state}} = \underbrace{\lambda}_{\text{recurrent parameter}} h_t + \underbrace{x_{t+1}}_{\text{input}}$$

We can calculate how signals propagate when the **time horizon is infinite**, and the **input distribution is wide-sense stationary**.

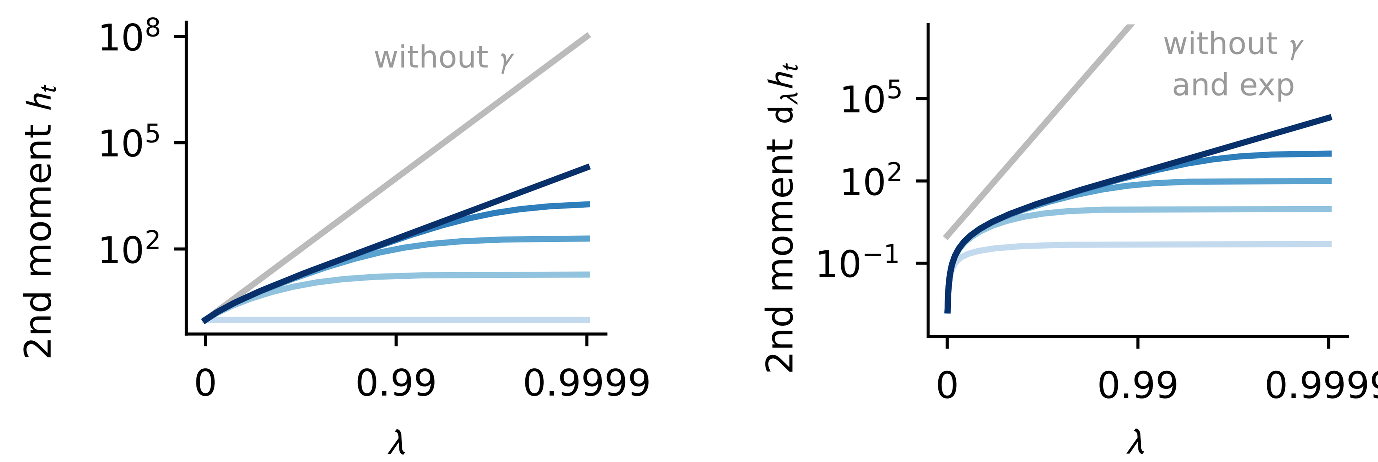


Deep SSMs (and LSTMs) mitigate the curse of memory

These architectures all features some form of **normalization** and **reparameterization**.

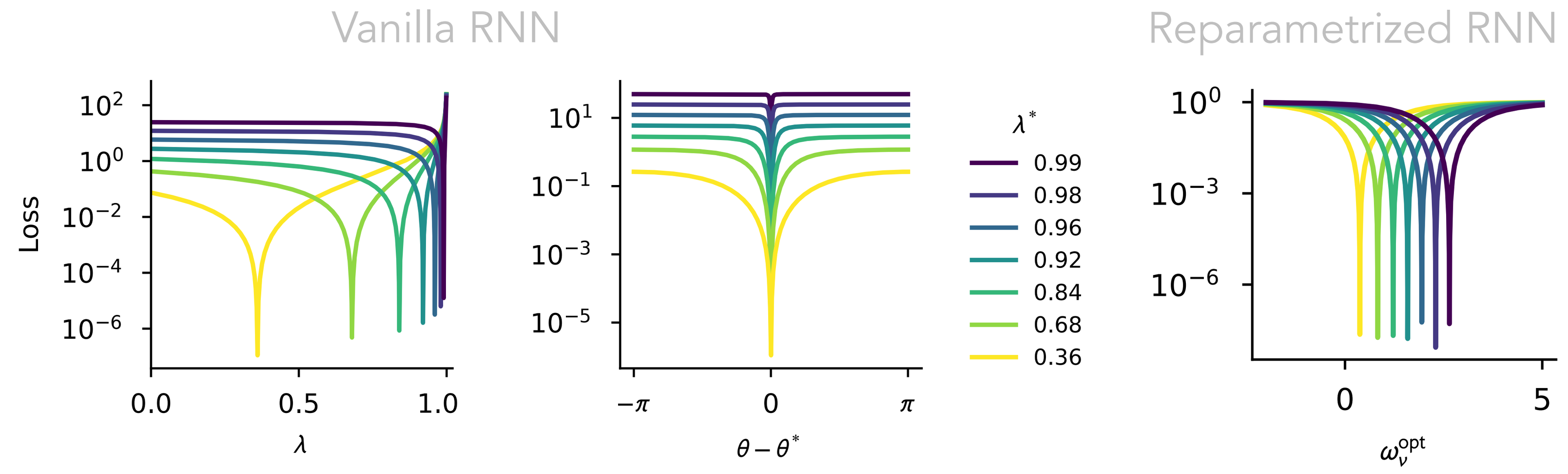
$$h_{t+1} = \lambda(\omega) h_t + \gamma(\lambda) x_{t+1}$$

	Normalization	Reparameterization
Deep SSMs	Discretization ODE	Powerful discretization scheme
LSTM	Input gate	Forget gate

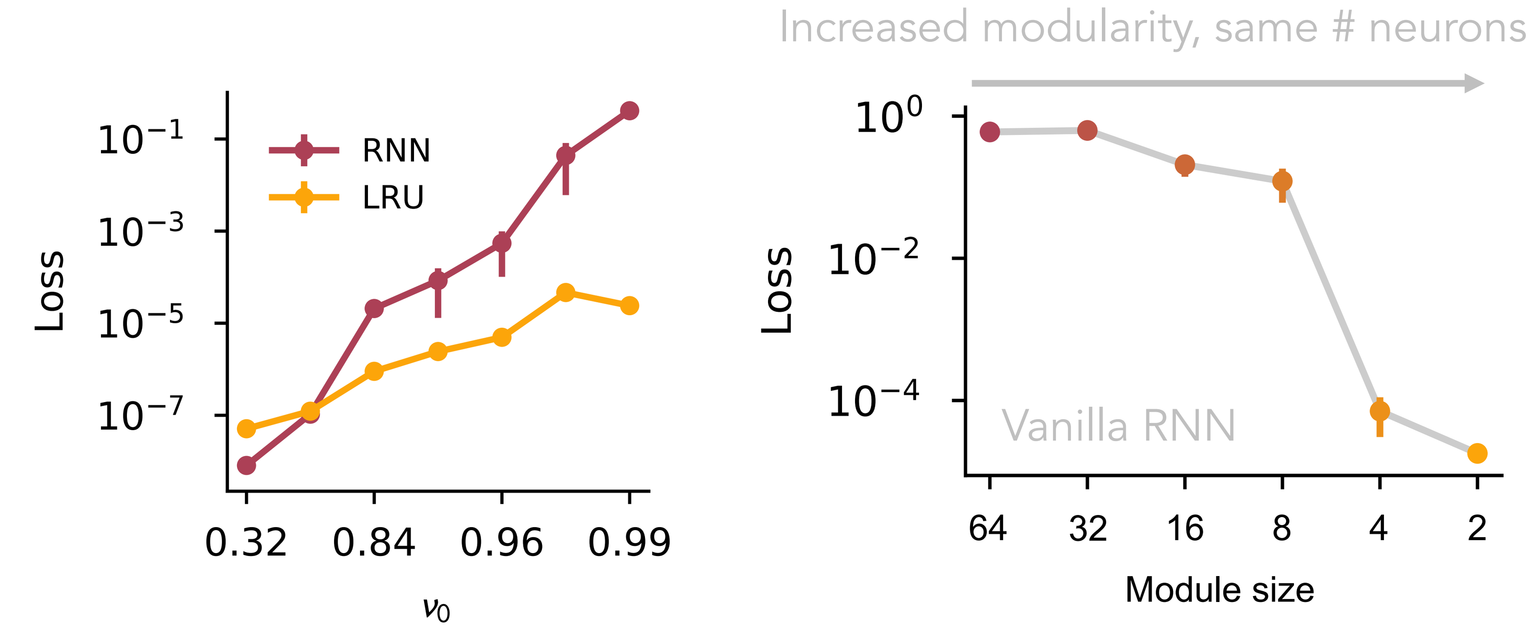


Linear teacher – linear student task

Goal: find the simplest setup in which **(deep) SSMs outperform vanilla RNNs**. A simple teacher-student task is enough!



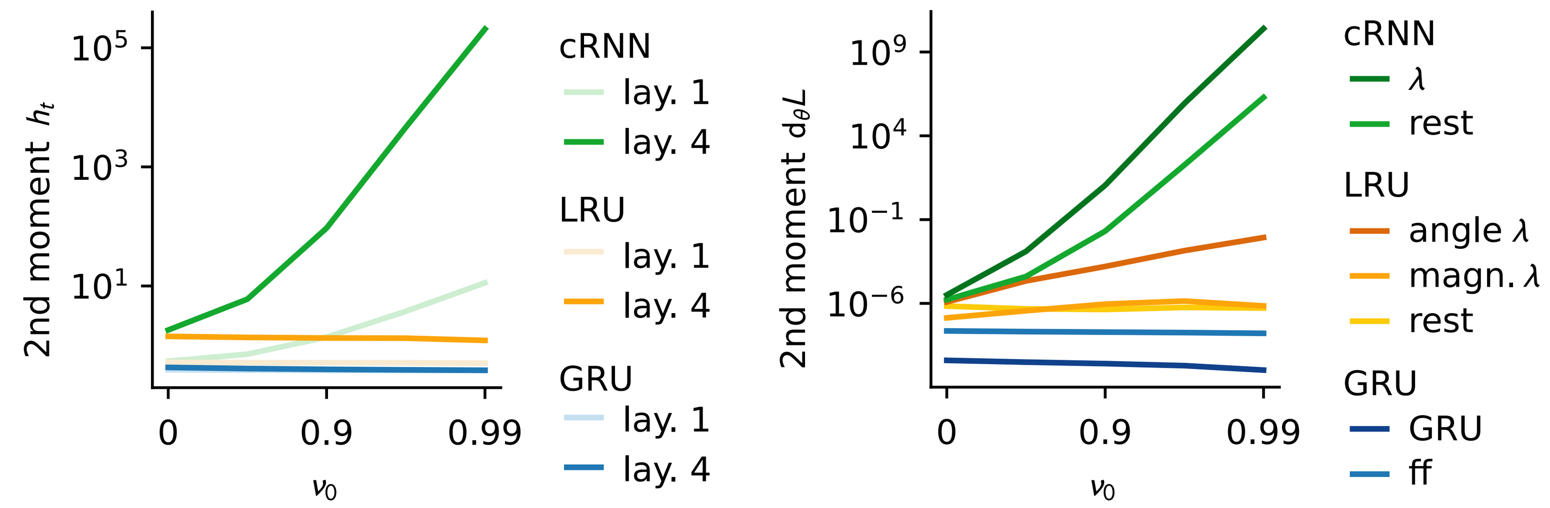
Key results. 1. The reparametrization strategy **greatly simplifies optimization**. 2. **Modularity** makes it possible for **Adam** to **compensate** for the imperfections of reparametrization.



In the paper: many **more analyses** + **theory explaining this**!

More realistic scenarios

Next token prediction task on BERT embeddings (Wiki data).



How to improve RNN training?

Reparameterization cannot work for strongly connected RNNs, let's **(re)investigate more elaborate optimizers**.

Modularity as a promising direction to **improve optimization** + nice connections with the brain (**cortical columns**),