Predictive State Recurrent Neural Networks

Carlton Downey, Ahmed Hefny, Byron Boots, Geoffrey Gordon

{cmdowney,ahefny,ggordon}@cs.cmu.edu, bboots@cc.gatech.edu

Abstract

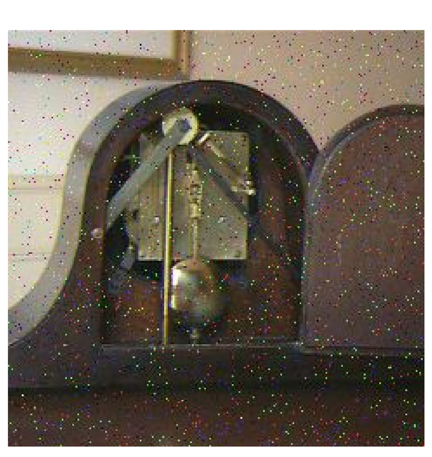
- Recurrent Neural Networks (RNNs) have seen a massive surge in popularity in recent years, particularly with the advent of modern architectures such as LSTMs.
- Despite their success, we still struggle to provide a rigorous theoretical analysis of these models, or to truly understand the mechanism behind their success.
- Prior to the success of RNNs, time series data modelling was dominated by Bayes Filters in their many forms. In contrast to RNNs Bayes Filters are grounded in axiomatic probability theory, resulting in a class of models which can be easily analyzed and whose action is well understood.
- In this work we propose a new class of models called ***Predictive State Recurrent Neural Networks, which combine the axiomatic probability theory of Bayes Filters with the rich functional forms and practical success of RNNs.***
- We show that PSRNNs can be learned effectively by combining Backpropogation Through Time (BPTT) with a method-of-moments initialization called Two-Stage Regression.
- Furthermore we show PSRNNs ***reveal interesting connections between Kernel Bayes Rule and conventional RNN architectures such as LSTMs and GRUs.***
- Finally we show ***PSRNNs outperform conventional RNN architectures, including LSTMs, on a range of datasets including both text and robotics data.***

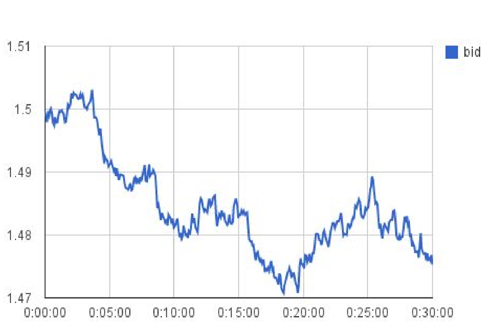
Introduction

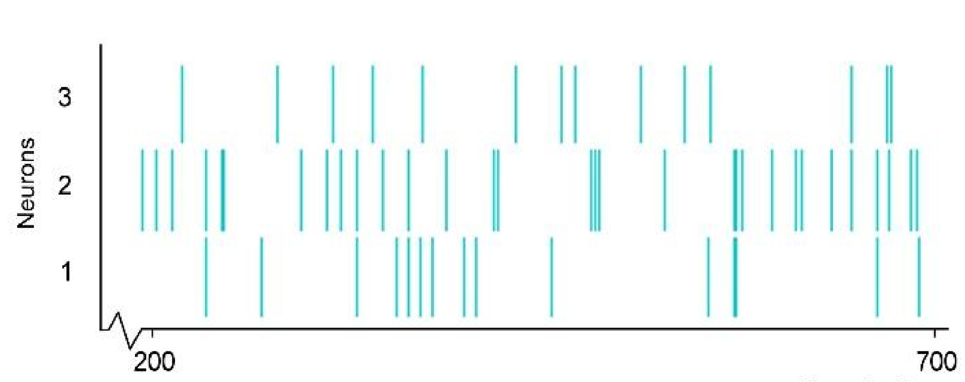
Dynamical System: A system whose state changes over time

Examples:

- Fluid Dynamics
- Test
- Speech
- GPS Tracking
- Lidar
- Video
- etc.







Modelling Dynamical Systems:

We present a new model, ***Predictive State Recurrent Neural Networks (PSRNNs)*** which draw insights and advantages from both RNNs and PSRs (A type of Bayes Filter)

- Strong statistical theory leading to consistent learning algorithms
- Possess a rich functional form

Bayes Filters	Recurrent Neural Networks
Probabilistic Belief State	Arbitrary State
Update using Bayes Rule	Arbitrary (smooth) Update
Strong Axiomatic Statistical Theory	No Axiomatic Statistical Theory
Simple Functional Forms	Rich Functional Forms
Globally Optimal Solutions	Local Optima

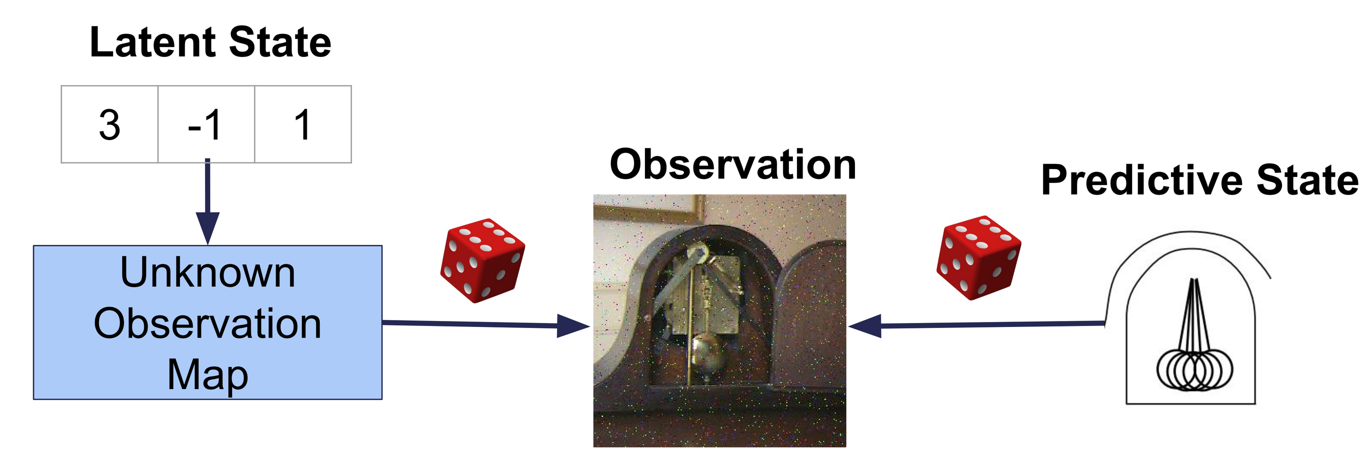
Predictive State Representations

Type of Bayes Filter (State update follows Bayes Rule)

- State is the expected value of future observation features conditioned on past observations

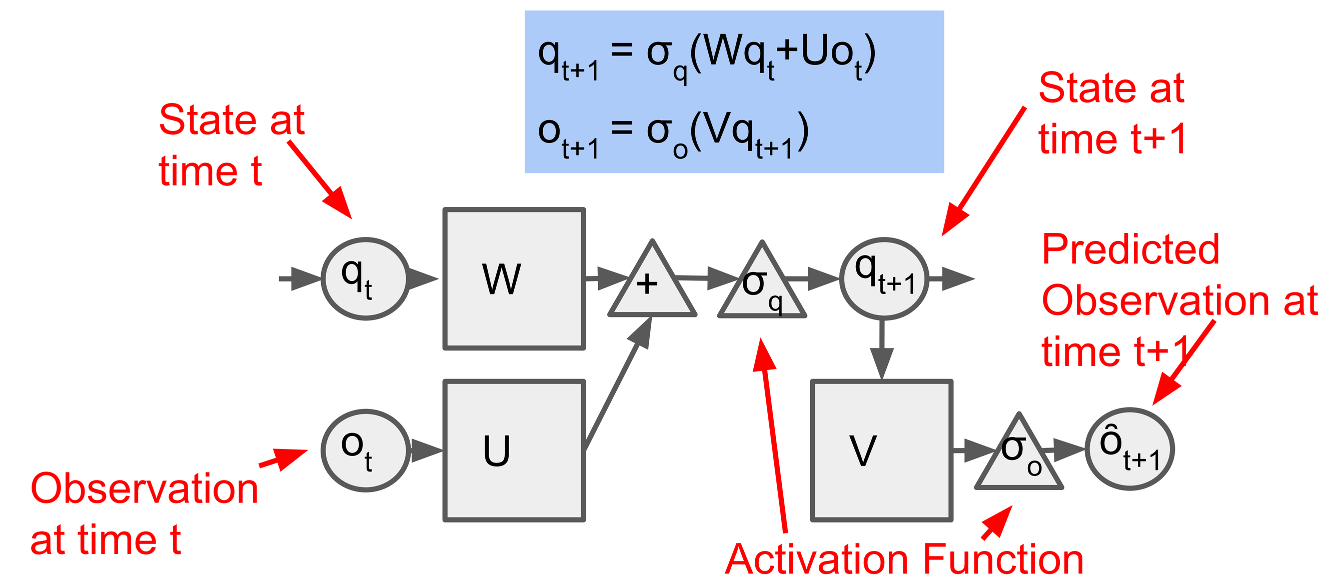
Key Advantage: PSR state is **observable**, not **latent**.

- Observable state leads to globally consistent method-of-moments learning algorithm.



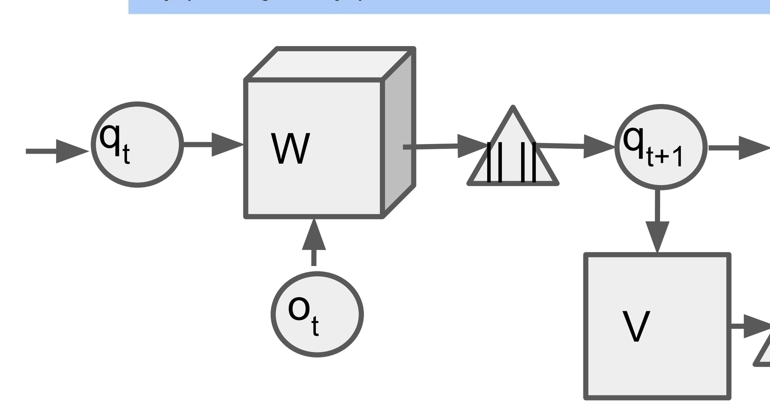
RNNs

$$q_{t+1} = \sigma_q(Wq_t+Uo_t)$$
$$o_{t+1} = \sigma_o(Vq_{t+1})$$



PSRNNs

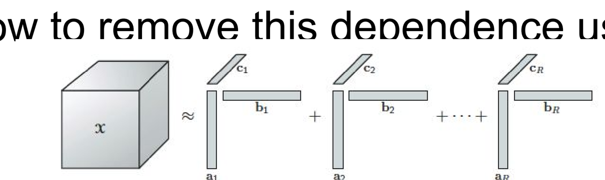
$$\alpha_{t+1} = A^T(Bo_t \odot Cq_t) \qquad q_{t+1} = \alpha_{t+1}/||\alpha_{t+1}||$$
$$o_{t+1} = \sigma_o(Vq_{t+1})$$



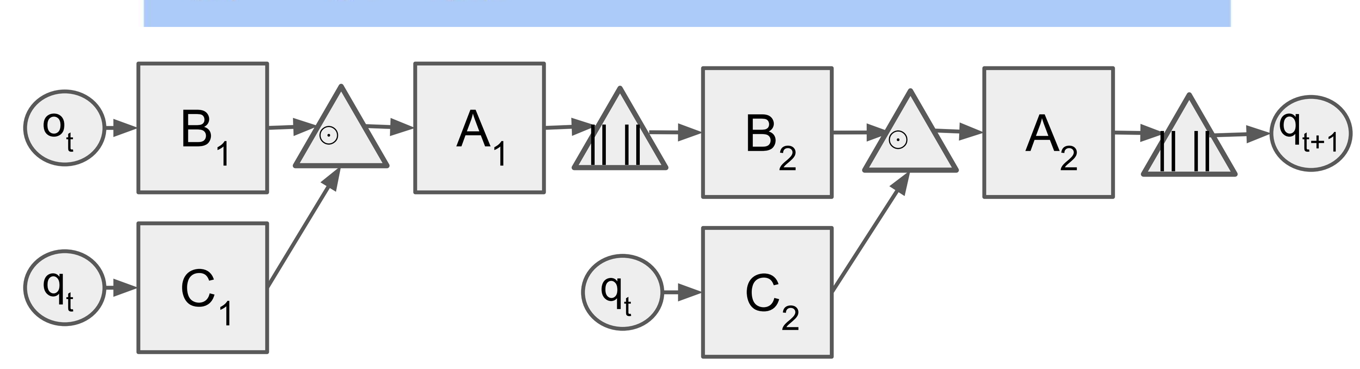
Factorized PSRNNS

Naive formulation of PSRNNs has nice gradients but is still cubic in size

- We now show how to remove this dependence using tensor CP Decomposition



$$\alpha_{t+1} = W \times_3 q_t \times_2 o_t \qquad q_{t+1} = \alpha_{t+1}/||\alpha_{t+1}||$$
$$o_{t+1} = \sigma_o(Vq_{t+1})$$



PSRNNS and LSTMs/GRUs

There is a deep connection between PSRNNs and LSTMs/GRUs

- Kernel bayes rule can be expressed as $A^T(Bo_t \odot Cq_t)$
- A very similar update is also at the heart of popular RNN architectures such as GRUs and LSTMS
- In other words GRUs/LSTMs have a version of kernel bayes rule as part of their update

GRU Update

$$z_{t+1} = \sigma_g(W_z o_t + U_z q_t)$$
$$r_{t+1} = \sigma_g(W_r o_t + U_r q_t)$$
$$q_{t+1} = z_t \odot q_t + (1 - z_t) \odot \sigma_q(W_q o_t + \boxed{U_q(r_t \odot q_t)})$$

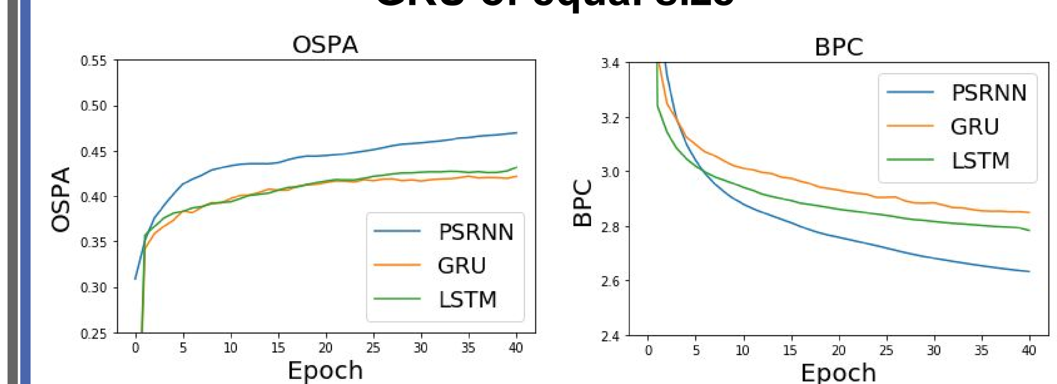
Experiments

Text Data

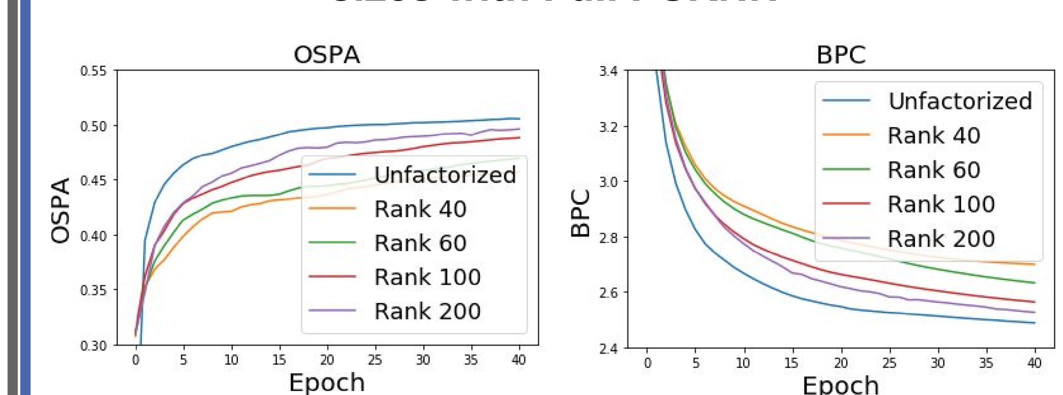
Setup

- Penn Tree Bank
- Character Model
- Metrics:
  - One Step Prediction Accuracy (OSPA)
  - Bits Per Character (BPC)

Comparison of PSRNN, LSTM, and GRU of equal size



Comparison of factorized PSRNNs of different sizes with Full PSRNN

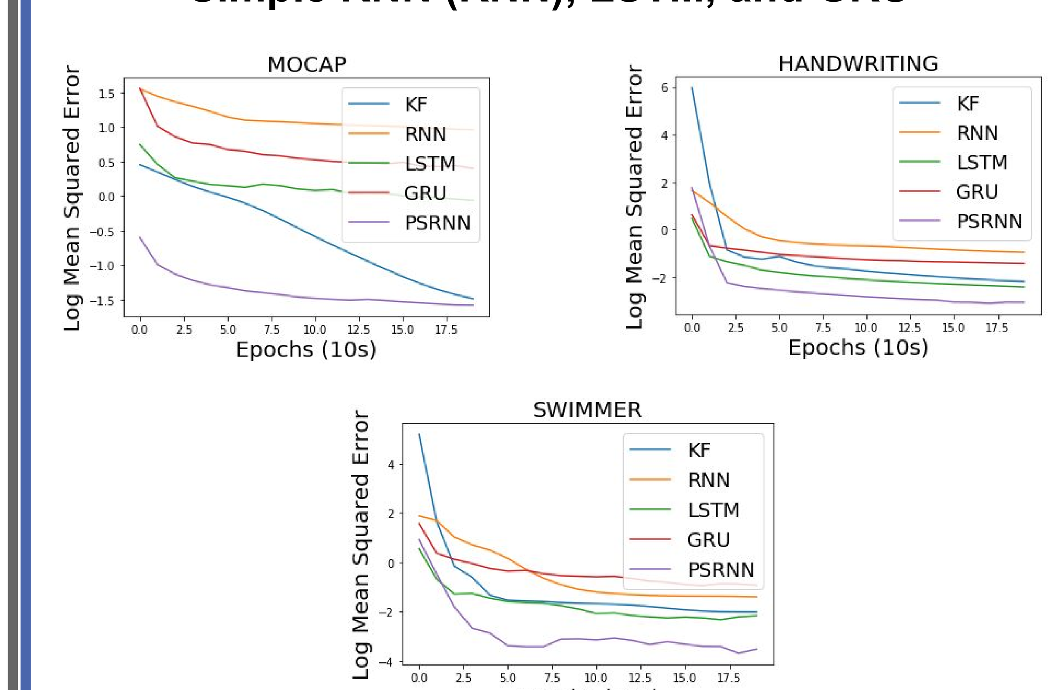


Continuous Data

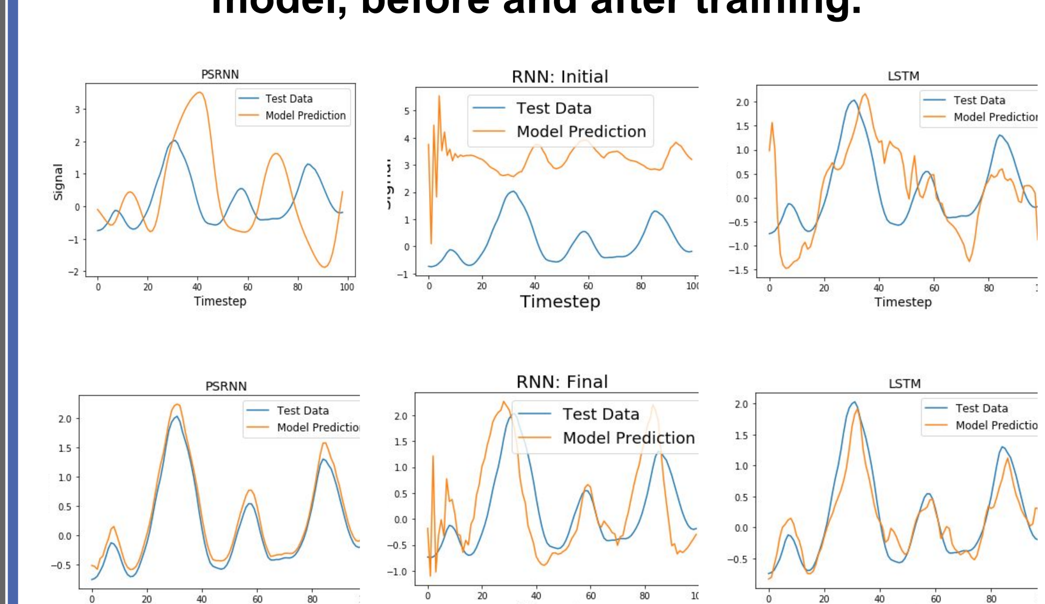
Setup

- Data
  - Swimmer (OpenAI Gym)
  - Motion Capture
  - Handwriting (UCI Repository)
- Mean Squared Error (MSE)

Comparison of PSRNN, Kalman Filter (KF), Simple RNN (RNN), LSTM, and GRU



Visualization of a single feature for each model, before and after training.



Future Work

Quantum PSRNNs!!!