Efficiently Modelling Sparse Dynamical Systems with Compressed Predictive State Representations

William Hamilton, Mahdi Fard, Joelle Pineau

Department of Computer Science
McGill University

June 8, 2013

Modelling a Dynamical System Using Time-Series Data

Problem: Efficiently learning a model of a dynamical system using time-series data.

- Focus on systems with the following properties:
  - Large discrete observation spaces.
  - Partially observability.
  - Sparsity.
- Example: Robot navigation without GPS

Latent-State Approaches to Learning.

Some popular examples:

- Expectation maximization learning with Hidden Markov Models (uncontrolled) [Rabiner, 1990] and POMDPs (controlled) [Kaelbling et al., 1998].
- Kalman Filtering [Kalman, 1960].

Limitations of these approaches:

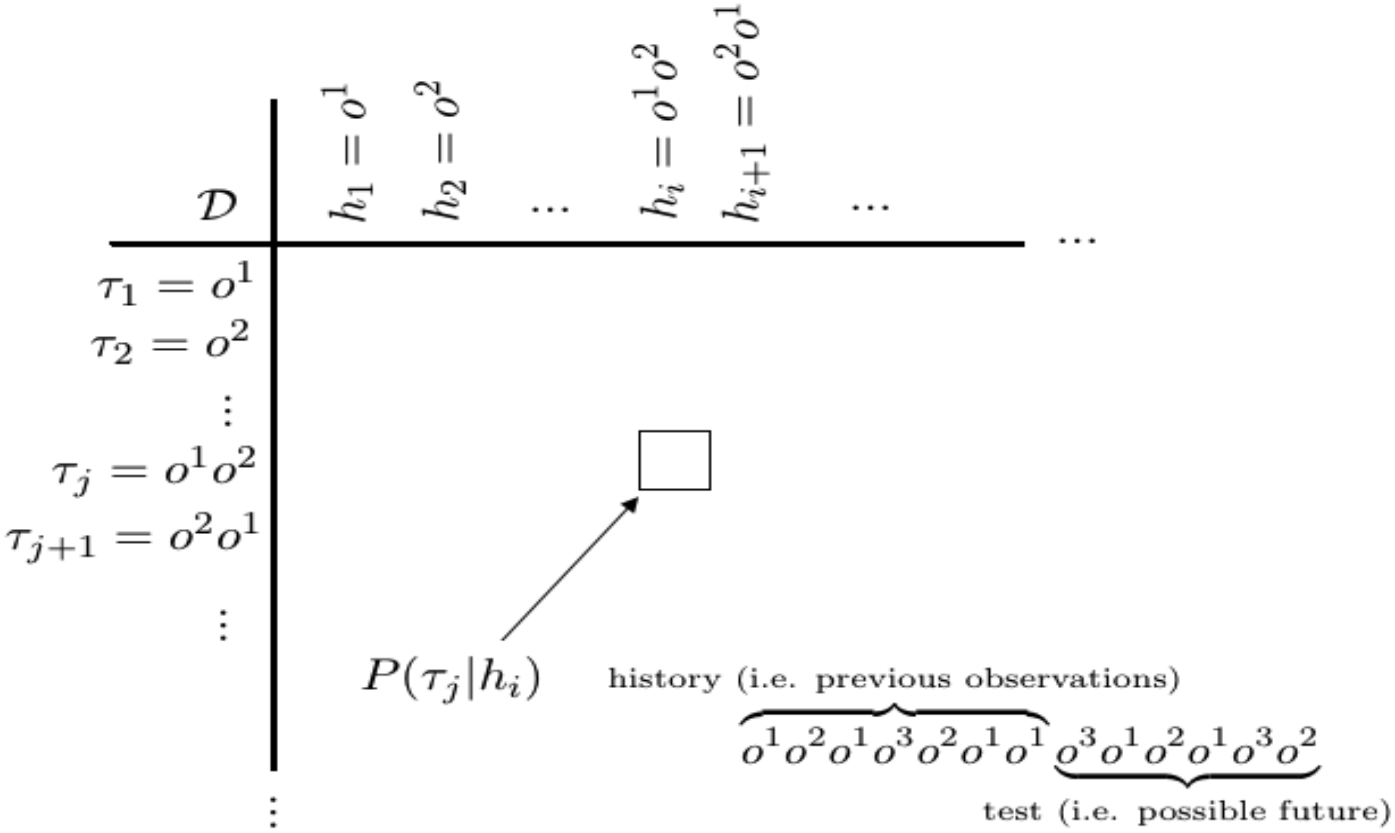
- Assumptions.
- Local minima.
- Scalability.

Event-Based Approaches to Learning.

- To avoid these limitations: work directly with observable events.
  - Build model by determining probabilities of the form: $P(o^i_{t+2} o^j_{t+1} | o^k_t)$
  - Learn how to compactly represent these probabilities as *predictive states*.
- Allows for model learning algorithms that are:
  - More general [Singh et al., 2004].
  - Immune to local minima [Rosencrantz et al., 2004].
- Examples:
  - Spectral learning methods [Hsu et al., 2008], Observable Operator Models, [Jaeger, 2000], Predictive State Representations [Littman et al., 2002].

The System Dynamics Matrix, $\mathcal{D}$

- We want $P(\tau_i|h_i)\ \forall i\ \forall j$
- Rank finite and bounded [Littman et al., 2002].
- Tests corresponding to row basis called *core tests*.



Learning Compact Approximations of Predictive States (Previous Approaches)

- Predictive State Representations
  - Discover core tests through combinatorial search [Littman et al., 2002].
  - Exponential complexity and not very useful in practice.
- Spectral Methods and Transformed Predictive State Representations (TPSRs) [Rosencrantz et al., 2004, Boots et al., 2009].
  - Estimate large sub-matrices of $\mathcal{D}$.
  - Project to low-dimensional subspace using SVD.
  - Computationally expensive, $O(|\mathcal{T}|^2|H|)$.
  - Consistency requires knowledge of $rank(\mathcal{D})$ [Boots et al., 2009].

A new approach: Compressed Predictive State Representations (CPSRs)

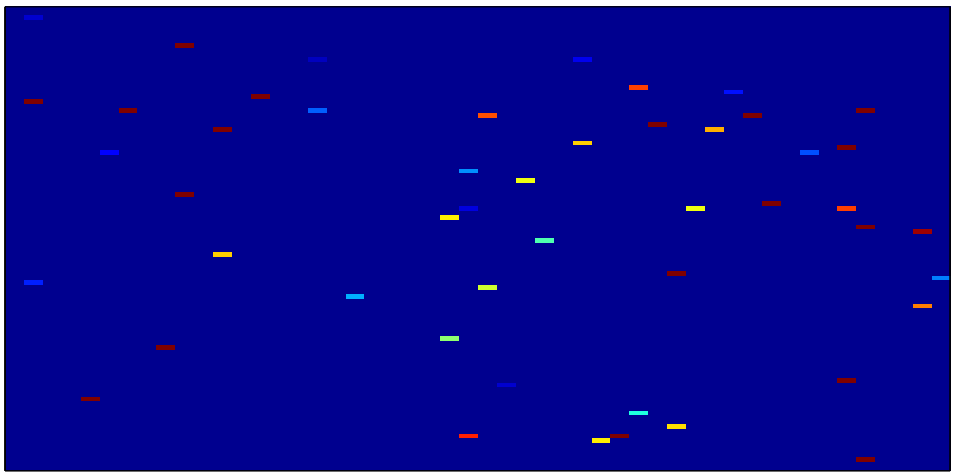
- As with spectral methods, core test discovery avoided.
- Large sub-matrices of $\mathcal{D}$ are estimated in a *compressed space* using random projections.
  - Computationally efficient (projection has no cost).
  - Regularizes the solution (i.e. the learned model parameters).
  - Relies on the sparsity of the system.
- Compressed estimates and regression are used to learn compact model.

Key Assumption: Sparsity in $\mathcal{D}$

- We say $\mathcal{D}$ is k sparse if

$$k \geq ||\mathbf{c}_i||_0 \; \forall \mathbf{c_i} \in \mathcal{D}$$

- I.e. only k tests possible given any history h_i.
- Can we assume that many systems are sparse?
  - In Poc-Man domain, large sub-matrix estimates of $\mathcal{D}$ empirically observed to have an average 99.902% column sparsity.



Compressing a Matrix using Random Projections

- Exploit sparsity using random projections.
  - Compress $m \times n$ matrix Y to a $d \times n$ matrix $\mathbf{X}$, where $d << m$.
  - Use

$$\mathbf{X} = \Phi \mathbf{Y}.$$

    where Φ is a $d \times m$ projection matrix with entries from $\mathcal{N} - (0, 1/d)$.

- In our case:
  - Projection via standard matrix multiplication unnecessary. Multiplication done "online" and Y matrix never held in memory.
  - Theoretical guarantees on compression fidelity.

Compressing a Matrix using Random Projections

- Exploit sparsity using random projections.
  - Compress $m \times n$ matrix Y to a $d \times n$ matrix $\mathbf{X}$, where $d << m$.
  - Use

    $$\mathbf{X} = \Phi\mathbf{Y}.$$

    where Φ is a $d \times m$ projection matrix with entries from $\mathcal{N} - (0, 1/d)$.
- In our case:
  - Projection via standard matrix multiplication unnecessary. Multiplication done "online" and Y matrix never held in memory.
  - Theoretical guarantees on compression fidelity.

The CPSR Algorithm

Algorithm

- Obtain compressed estimates for sub-matrices of $\mathcal{D}$, $\Phi\mathcal{P}_{\mathcal{T},\mathcal{H}}$, $\Phi\mathcal{P}_{\mathcal{T},o',\mathcal{H}}$s, and $\mathcal{P}_{\mathcal{H}}$ by sampling time series data.
  - Estimate $\Phi\mathcal{P}_{\mathcal{T},\mathcal{H}}$ in compressed space by adding ϕ_i to column j each time t_i observed after h_i (Likewise for $\Phi\mathcal{P}_{\mathcal{T},o',\mathcal{H}}$s).
- Compute CPSR model:
  - $\mathbf{c}_0 = \Phi\hat{\mathcal{P}}(\tau|\emptyset)$
  - $\mathbf{C}_o = \Phi\mathcal{P}_{\mathcal{T},o',\mathcal{H}}(\Phi\mathcal{P}_{\mathcal{T},\mathcal{H}})^+$
  - $\mathbf{c}_\infty = (\Phi\mathcal{P}_{\mathcal{T},\mathcal{H}})^+\hat{\mathcal{P}}_H$

Using the compact representation.

State definition and necessary equations

- $\mathbf{c}_0$ serves as initial prediction vector (i.e. state vector).
- Update state vector after seeing observation with
  - $\mathbf{c}_{t+1} = \frac{\mathbf{C}_o \mathbf{c}_t}{\mathbf{C}_\infty \mathbf{C}_o \mathbf{c}_t}$
- Predict k-steps into the future using
  - $P(o^j_{t+k}|h_t) = \mathbf{b}_\infty \mathbf{C}_{o^j} (\mathbf{C}_\star)^{k-1} \mathbf{C}_t$ where $\mathbf{C}_\star = \sum_{o^i \in \mathcal{O}} \mathbf{C}_{o_i}$.

Theory: Overview

- Unlike spectral methods, we allow projection to subspaces of dimension $d < rank(\mathcal{D})$.
  - If $d \geq rank(\mathcal{D})$ then model trivially consistent [Boots et al., 2009].
- Results build upon work on compressed regression [Fard et al., 2012].
  - Analyze how compression provides regularization.
  - Provide error bounds and necessary projection size.
- We had to analyze how noisy targets affect these results.

Theory: Preliminaries

- Results from the compressed regression literature [Maillard et al., 2012, Fard et al., 2012]:
  - For random projection of size d, there exists a generic upper bound function ϵ, such that with probability no less than $1 - \delta$

$$\|f(\mathbf{x}) - \hat{f}_d(\mathbf{x})\|_{\rho(\mathbf{x})} \leq \epsilon(n, D, d, \|\mathbf{w}\|\|\mathbf{x}\|_{\rho(\mathbf{x})}, \sigma^2, \delta)$$

- Our sparsity assumptions:
  - For all h, $\mathcal{P}_{\mathcal{Q},h}$ and $\mathcal{P}_{\mathcal{Q},o,h}$ are k-sparse.

Theory: Main Results

Error of the CPSR parameters

With probability no less than $1 - \delta$ we have:

$$\left\| \mathbf{C}_o(\Phi\mathcal{P}_{\mathcal{Q},h}) - \Phi\mathcal{P}_{\mathcal{Q},o,h} \right\|_{\rho(\mathbf{x})} \leq \sqrt{d}\epsilon(|\mathcal{H}|, |\mathcal{Q}|, d, L_o, \sigma_o^2, \delta/d)$$

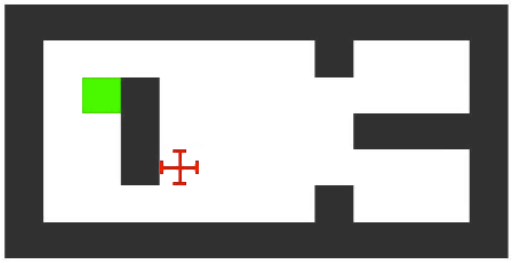
Error propagation

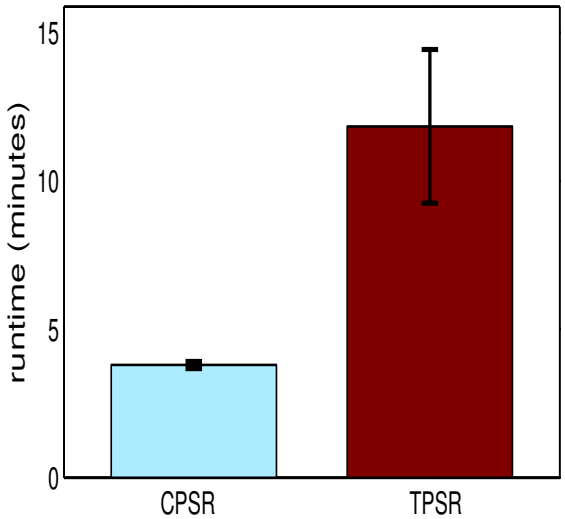
The total propagated error for T steps is bounded by $\epsilon(c^T - 1)/(c - 1)$.

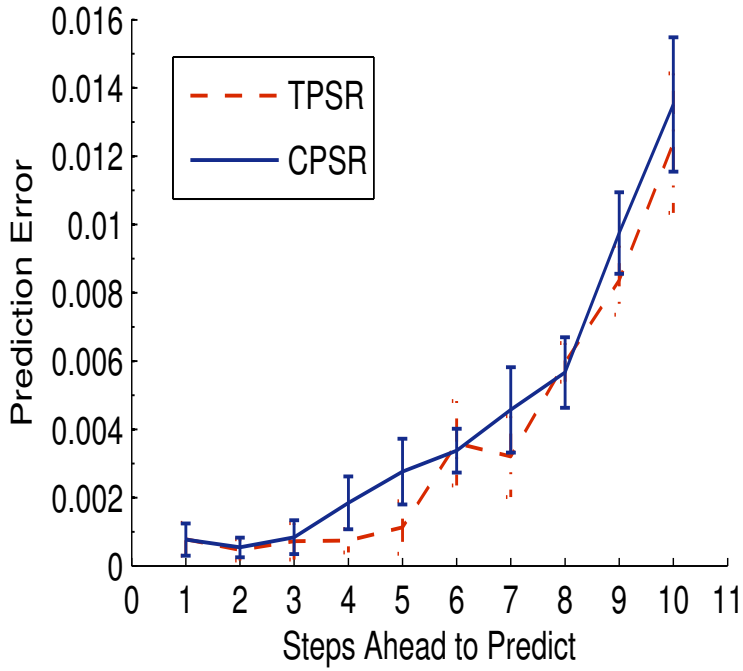
Projection size

A projection size of $d = O(k \log |Q|)$ suffices in a majority of systems.

GridWorld: Increased time-efficiency in small simple systems

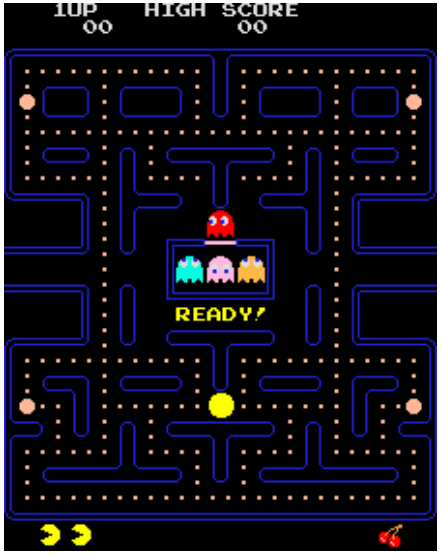


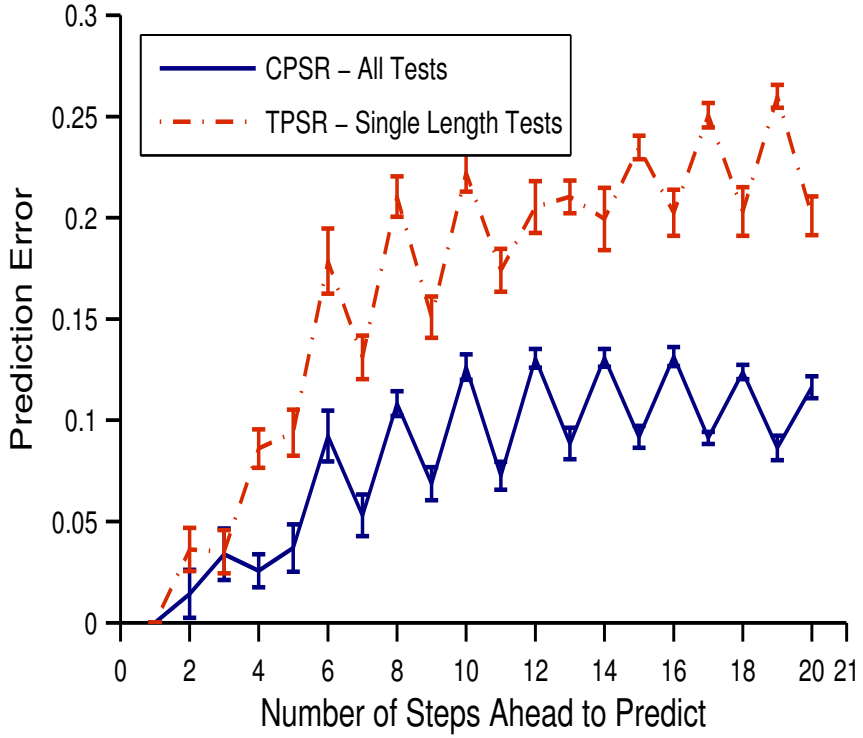




Poc-Man: Better model quality in large difficult systems

Partially observable variant of Pac-Man video-game with $|\mathcal{S}| = 10^{56}$ and $|\mathcal{O}| = 2^{10}$ [Silver and Veness, 2010].





Summary

- Developed an efficient algorithm for modelling dynamical systems.
- The compression technique has extremely low computational cost.
- Model can be used for planning.

- Directions for further work:
  - Determine how feature mapping affects sparsity.
  - Examine different model averaging techniques.
  - Formally analyze value-function based planning approach.

References I


Boots, B., Siddiqi, S., and Gordon, G. (2009).
Closing the learning-planning loop with predictive state representations.
In *Proceedings of Robotics: Science and Systems VI*.


Fard, M., Grinberg, Y., Pineau, J., and Precup, D. (2012).
Compressed least-squares regression on sparse spaces.
AAAI.


Hsu, D., Kakade, S., and Zhang, T. (2008).
A spectral algorithm for learning hidden markov models.

In *COLT*.

References II


Jaeger, H. (2000).
Observable operator models for discrete stochastic time series.
Neural Computation, 12(6):1371–1398.


Kaelbling, L., Littman, M., and Cassandra, A. (1998).
Planning and acting in partially observable stochastic domains.
Artificial Intelligence, 101:99–134.


Kalman, R. (1960).
A new approach to linear filtering and prediction problems.
In *Transactions of the ASME, Journal of Basic Engineering*, volume 82, pages 35–45.

References III


Littman, M., Sutton, R. S., and Singh, S. (2002).
Predictive representations of state.
In *In Advances In Neural Information Processing Systems*.


Maillard, O., Munos, R., et al. (2012).
Linear regression with random projections.
Journal of Machine Learning Research.


Rabiner, L. R. (1990).
Readings in speech recognition.
chapter A tutorial on hidden Markov models and selected applications in speech recognition, pages 267–296.

References IV


Rosencrantz, M., Gordon, G., and Thrun, S. (2004).
Learning low dimensional predictive representations.
In *Proceedings of the twenty-first international conference on Machine learning*.


Silver, D. and Veness, J. (2010).
Monte-carlo planning in large pomdps.
In Advances In Neural Information Processing Systems, 47:1–9.


Singh, S., James, M., and Rudary, M. (2004).
Predictive state representations: a new theory for modeling dynamical systems.
In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*.