Closing the Learning-Planning Loop with Predictive State Representations

Byron Boots
Machine Learning Department
Carnegie Mellon University

Sajid M. Siddiqi
Robotics Institute
Carnegie Mellon University

Geoffrey J. Gordon
Machine Learning Department
Carnegie Mellon University

Motivation

A Central Problem in AI: Planning to maximize future reward under uncertainty in a partially observable environment.

Goal: Learn model of the environment directly from actions and observations, close the loop by planning in the learned model.

Significance: no other learning algorithm has botha proof of statistical consistency and a demonstration of closing the loop in an environment of this size.

Predictive State Representations

PSRs represent state as probabilities of future observations given actions (called tests) conditioned on history:

$$p(Q^O|h||Q^A) = [p(q_1^O|h||q_1^A), ..., p(q_{|Q|}^O|h||q_{|Q|}^A)]^\mathsf{T}$$

Linear PSRs update state by applying a linear **observable operator** to the current state and renormalizing:

$$p(Q^O|ho||a, Q^A) = \frac{M_{ao}p(Q^O|h||Q^A)}{m_\infty^\mathsf{T} M_{ao}p(Q^O|h||Q^A)}$$

Learning is Hard!

How do we choose a sufficient set of tests and linear model parameters?

We develop a **novel**, **consistent**, subspace identification algorithm that relies on the fact that expectations of **features** of large sets of tests and histories factor into linear functions of **true** PSR parameters:

$$[P_\mathcal{H}]_i \equiv \mathbb{E}(\phi_i^\mathcal{H}(h)) \Rightarrow P_\mathcal{H} = \Phi^{\mathcal{H}}\pi$$
$$[P_{\mathcal{T},\mathcal{H}}]_{i,j} \equiv \mathbb{E}(\phi_i^\mathcal{T}(\tau^O) \cdot \phi_j^\mathcal{H}(h)||\tau^A)$$
$$\Rightarrow P_{\mathcal{T},\mathcal{H}} = \Phi^{\mathcal{T}} RS\mathrm{diag}(\pi)\Phi^{\mathcal{H}^\mathsf{T}}$$
$$[P_{\mathcal{T},ao,\mathcal{H}}]_{i,j} \equiv \mathbb{E}(\phi_i^\mathcal{T}(\tau^O) \cdot \phi_j^\mathcal{H}(h) \cdot \delta(O = o)||\tau^A A = a)$$
$$\Rightarrow P_{\mathcal{T},ao,\mathcal{H}} = \Phi^{\mathcal{T}} RM_{ao}S\mathrm{diag}(\pi)\Phi^{\mathcal{H}^\mathsf{T}}$$

Expectations are used to recover transformed PSR parameters (see below).

Why do we prefer to use features instead of actual tests and histories?

In practice it is hard to come up with complete sets of tests and histories, but easier to come up with features that are correlated with state.

How do we deal with continuous observations?

We use kernel density estimation and prove that the resulting model remains **consistent**.

Transformed PSR Parameters

U is the matrix of left singular vectors of $P_{\mathcal{T},\mathcal{H}}$

$$b_1 \equiv U^\mathsf{T} P_{\mathcal{T},\mathcal{H}} e^\mathsf{T} = (U^\mathsf{T}\Phi^\mathcal{T} R)m_1$$
$$b_\infty^\mathsf{T} \equiv P_\mathcal{H}^\mathsf{T}(U^\mathsf{T} P_{\mathcal{T},\mathcal{H}})^\dagger = m_\infty^\mathsf{T}(U^\mathsf{T}\Phi^\mathcal{T} R)^{-1}$$
$$B_{ao} \equiv U^\mathsf{T} P_{\mathcal{T},ao,\mathcal{H}}(U^\mathsf{T} P_{\mathcal{T},\mathcal{H}})^\dagger = (U^\mathsf{T}\Phi^\mathcal{T} R)M_{ao}(U^\mathsf{T}\Phi^\mathcal{T} R)^{-1}$$

TPSR State Update

$$b_{t+1} \equiv \frac{B_{ao_t} b_t}{b_\infty^\mathsf{T} B_{ao_t} b_t}$$

TPSR Prediction

$$Ub_t$$

Learning Transformed PSRs

1. Compute empirical estimates $\widehat{P}_\mathcal{H}$, $\widehat{P}_{\mathcal{T},\mathcal{H}}$,$\widehat{P}_{\mathcal{T},ao,\mathcal{H}}$

2. Use SVD on $\widehat{P}_{\mathcal{T},\mathcal{H}}$ to compute $\widehat{U}$, the matrix of left singular vectors corresponding to the n largest singular values.

3. Compute model parameter estimates:

   (a) $\widehat{b}_1 = \widehat{U}^\mathsf{T}\widehat{P}_\mathcal{H}$,

   (b) $\widehat{b}_\infty = (\widehat{P}_{\mathcal{T},\mathcal{H}}^\mathsf{T}\widehat{U})^\dagger \widehat{P}_\mathcal{H}$,

   (c) $\widehat{B}_{ao} = \widehat{U}^\mathsf{T}\widehat{P}_{\mathcal{T},ao,\mathcal{H}}(\widehat{U}^\mathsf{T}\widehat{P}_{\mathcal{T},\mathcal{H}})^\dagger$

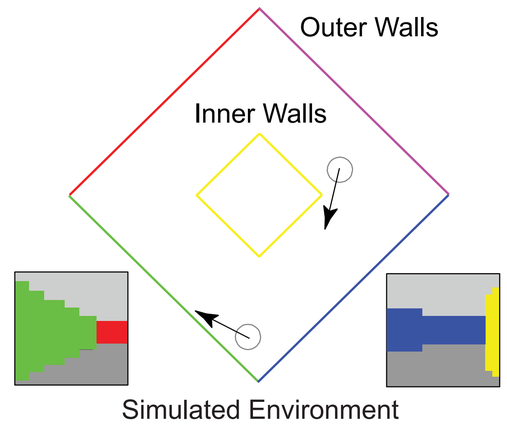
Theoretical Guarantees

Learning algorithm is statistically consistent.

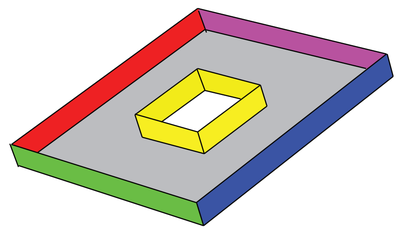
We can prove sample complexity bounds for Reduced-Rank Hidden Markov Models, an important subset of PSRs [1].

[1] S. M. Siddiqi, B. Boots, and G. J. Gordon. Reduced-Rank Hidden Markov Models. *http://arxiv.org/abs/0910.0902*, 2009.

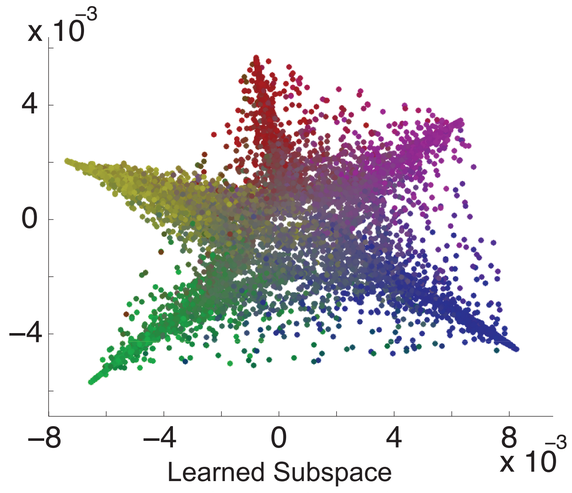
Experimental Results



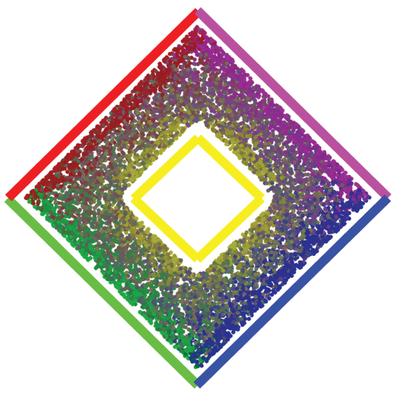
Simulated Environment



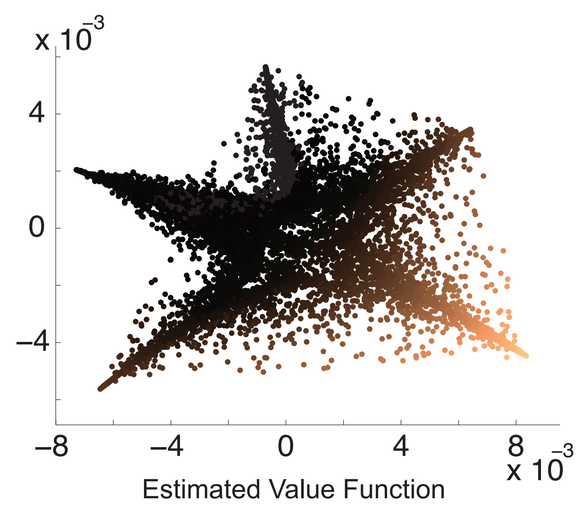
Simulated Environment
3-d View (to scale)



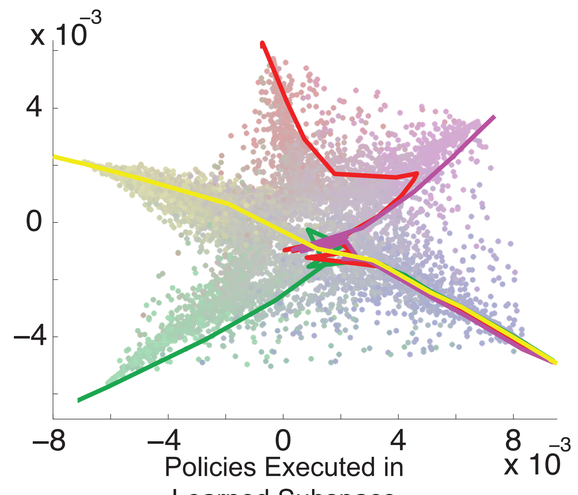
Learned Subspace
color = predominant color in image



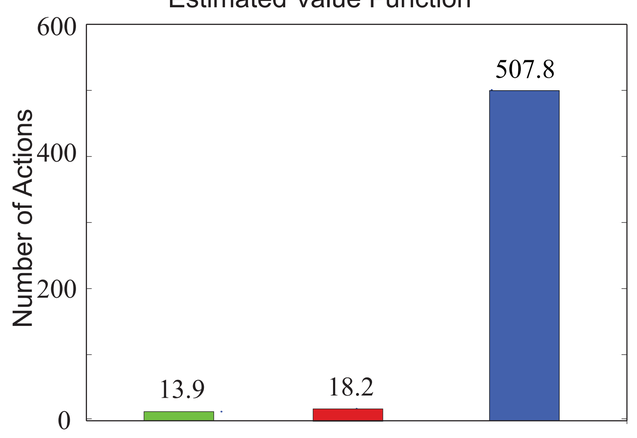
Learned Representation
Mapped to
Geometric Space



Estimated Value Function



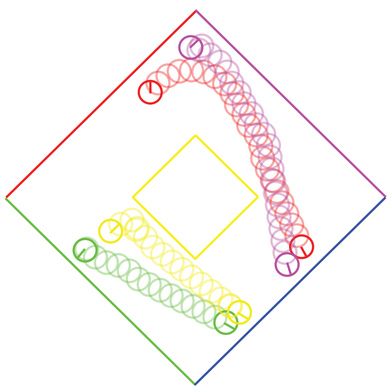
Policies Executed in
Learned Subspace



Optimal
Lower Bound
(A* in C-space)

PBVI

Random
Walk



Paths Taken in
Geometric Space

Summary

We present a novel **consistent** subspace identification algorithm for learning transformed PSR parameters from execution traces.

We extend transformed PSR learning to **features** of tests and histories and to **continuous observations** in a way that preserves consistency.

We assess **accuracy** of learned model on a difficult **planning task** and show that approximate planning in the learned model generates **near-optimal results**.

We believe that subspace identification algorithms for learning PSRs can increase the scope of planning under uncertainty for autonomous agents in previously intractable scenarios.