Mixed Observability Predictive State Representations

Sylvie C.W. Ong and **Yuri Grinberg** and **Joelle Pineau**
School of Computer Science
McGill University, Montreal, Canada

Abstract

Learning accurate models of agent behaviours is crucial for the purpose of controlling systems where the agents' and environment's dynamics are unknown. This is a challenging problem, but structural assumptions can be leveraged to tackle it effectively. In particular, many systems exhibit mixed observability, when observations of some system components are essentially perfect and noiseless, while observations of other components are imperfect, aliased or noisy. In this paper we present a new model learning framework, the mixed observability predictive state representation (MO-PSR), which extends the previously known predictive state representations to the case of mixed observability systems. We present a learning algorithm that is scalable to large amounts of data and to large mixed observability domains, and show theoretical analysis of the learning consistency and computational complexity. Empirical results demonstrate that our algorithm is capable of learning accurate models, at a larger scale than with the generic predictive state representation, by leveraging the mixed observability properties.

Introduction

A central problem in building many autonomous agents— software, robots, or otherwise— is to estimate the dynamics of the agents. Once the dynamics are known, they can be used for tracking, planning and control, simulation, and a multitude of other tasks. This problem is particularly challenging in large, complex systems with many interacting components. In such systems, the observation and action spaces can easily become so large that model learning becomes all but intractable.

In this paper, we focus on *mixed observability* systems where some system components are *fully observable*, while others are *partially observable*. In other words, there are essentially perfect, noiseless observations of some system components while observations of the other components are imperfect, aliased or noisy. Such systems are often encountered in practice, in domains as diverse as resource management (Chades et al. 2012) and robotics (Capitan, Merino, and Ollero 2011). For example, in human-robot interactions, rich, highly accurate sensors placed directly on the robot agents effectively make the robots state fully observable. On the other hand, the sensing of human agents is generally done remotely, for example, via video cameras, hence is much more inaccurate and noisy. We leverage the special structure of mixed observability systems to develop a new framework for learning models of such systems directly from observable quantities.

A rich framework for modelling controlled dynamic systems is the partially observable Markov decision process (POMDP) (Kaelbling, Littman, and Cassandra 1998). To handle mixed observability systems, this has been recently extended to the mixed observability Markov decision processes (MOMDPs) (Ong et al. 2010). And while there are a few methods for learning the parameters of a POMDP, this remains a challenging problem - possibly due to local minima issues and other difficulties inherent to learning approaches based on latent state variables. Furthermore there is no existing work on learning parameters in MOMDPs.

An alternate method for modelling controlled dynamic systems which shows much greater potential for learning and estimation is the predictive state representation (PSR) (Littman, Sutton, and Singh 2002). In this paradigm, the representation is based entirely on observable quantities, and thus predictive models are typically easier to learn from data than latent state-based representations (Singh, James, and Rudary 2004; Boots, Siddiqi, and Gordon 2010). The primary contribution of this paper is to adapt PSRs to modelling mixed observability systems, such that they can be applied to systems with this characteristic. Our novel framework, called *mixed observability predictive state representation (MO-PSR)*, is the first approach for *learning* models of mixed observability dynamic systems.

As part of our contributions, we show that model learning with this framework is theoretically sound, and present a learning algorithm that remains tractable given large amounts of data. We present experimental results and analysis on a simulated domain with mixed observability, comparing our algorithm's performance with model learning using the generic PSR framework.

Technical Background

A controlled, discrete-time, finite dynamical system generates observations from a set $\mathcal{O}$ in response to actions from a set $\mathcal{A}$. At each time step t, an agent interacting with the system takes an action $a_t \in \mathcal{A}$ and receives an ob-

Copyright © 2013, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

servation $o_t \in \mathcal{O}$. A history, $h = a_1 o_1 a_2 o_2 \cdots a_t o_t$, is a sequence of past actions and observations, while a test, $\tau = a_{t+1} o_{t+1} a_{t+2} o_{t+2} \cdots a_{t+k} o_{t+k}$, is a sequence of actions and observations that may occur in the future. The prediction for test τ given history h, $p(\tau|h)$, is the conditional probability that the observation sequence in τ occurs, if the actions specified in τ are executed.

Let $\mathcal{T}$ be the set of all tests and $\mathcal{H}$ the set of all histories. Given an ordering over $\mathcal{H}$ and $\mathcal{T}$, we define the system-dynamics matrix, $\mathcal{D}$, with rows corresponding to histories and columns corresponding to tests. Matrix entries are the predictions of tests, i.e., $D_{i,j} = p(\tau_j|h_i)$. All the information necessary to predict system dynamics is contained in $\mathcal{D}$, so it completely characterizes the system (Singh, James, and Rudary 2004). Suppose the matrix has a finite number of linearly independent columns, and we denote the set of tests corresponding to those columns as the core tests, Q. Then, the prediction of any test is the linear combination of the predictions of Q. Let $p(Q|h)$ be a vector of predictions for Q given history h, then, for all possible tests τ, there exists a set of weights m_τ, such that $p(\tau|h) = m_\tau^\top p(Q|h)$. Thus, the vector $p(Q|h)$ is a sufficient statistic for history, i.e. it represents the system state, and we can regard it as the projection of history h on the PSR state-space. Define M_{ao} as the matrix with rows $m_{aoq}^\top$, for all $q \in Q$, and define a set of weights m_∞, where $m_{ao}^\top = m_\infty^\top M_{ao}$, for all $a \in \mathcal{A}$, $o \in \mathcal{O}$. Given a history h, a new action a, and subsequent observation o, the PSR state vector is updated by,

$$p(Q|hao) = \frac{p(aoQ|h)}{p(ao|h)} = \frac{M_{ao}p(Q|h)}{m_{ao}^\top p(Q|h)} = \frac{M_{ao}p(Q|h)}{m_\infty^\top M_{ao} p(Q|h)} \ . \quad (1)$$

The projection of any history can be calculated by repeatedly applying the above equation, starting from the initial system state, m_0, which is the prediction vector given an initial distribution over histories.

A PSR model is completely specified by the core tests Q, and m_0, m_∞, and M_{ao}, for all $a \in A$, $o \in O$. The usual approach to model learning is to approximate the system-dynamics matrix, $\mathcal{D}$, through interactions with the system, i.e. by sampling action-observation trajectories, then use $\mathcal{D}$ to estimate the model (Singh, James, and Rudary 2004).

The Transformed Predictive State Representation (TPSR) is a recently proposed extension to the PSR, which has shown good model learning performance (Boots, Siddiqi, and Gordon 2010). The TPSR approach estimates $P_{\mathcal{T},\mathcal{H}}$, a matrix containing the joint probabilities of histories and tests $p(\tau, h)$, for all $\tau \in \mathcal{T}$ and $h \in \mathcal{H}$, where the rows correspond to tests and columns to histories. $P_{\mathcal{T},\mathcal{H}}$ characterizes a system in a similar way as the system-dynamics matrix $\mathcal{D}$. TPSR is a spectral approach which obtains a compact representation of history by making use of the matrix of nonzero left singular vectors U, from the singular value decomposition (SVD) of $P_{\mathcal{T},\mathcal{H}}$. For any history h, $p(\mathcal{T}, h)$, the joint probabilities of h and the set of all tests $\mathcal{T}$ is a linear combination of the column vectors in U. The TPSR representation for the history h is $U^\top p(\mathcal{T}, h)$, the weights of that linear combination.

The TPSR model learning algorithm additionally constructs $P_{\mathcal{H}}$, a vector of probabilities of all histories $h \in \mathcal{H}$, and $P_{\mathcal{T},ao,\mathcal{H}}$, matrices containing the joint probabilities of every history $h \in \mathcal{H}$, followed by action a, observation o, and every test $\tau \in \mathcal{T}$, for all $a \in \mathcal{A}$ and $o \in \mathcal{O}$. Then, regression methods are applied to learn model parameters b_0, b_∞, and B_{ao}, for all $a \in \mathcal{A}$, $o \in \mathcal{O}$, which are transformed versions of the PSR parameters m_0, m_∞, and M_{ao}.

The TPSR state is a linear combination of core tests. Due to finite training data, in practice the TPSR algorithm has the potential to learn a more compact state representation than PSR algorithms. At the same time, the TPSR gives predictions which are equivalent to those of PSR, as shown in the consistency results in (Boots, Siddiqi, and Gordon 2010), for the prediction of action-observation sequence $a_1 o_1 \cdots a_t o_t$, given initial system state:

$$\begin{aligned} p(a_1 o_1 \cdots a_t o_t) &= b_\infty^\top B_{a_t o_t} \cdots B_{a_1 o_1} b_0 \\ &= m_\infty^\top M_{a_t o_t} \cdots M_{a_1 o_1} m_0 \ . \end{aligned} \quad (2)$$

Modelling mixed observability dynamical systems with predictive state representations

As in the general PSR, a MO-PSR models a controlled, discrete-time, finite dynamical system which generates observations from a set $\mathcal{O}$ in response to actions from a set $\mathcal{A}$. Here however, we assume that observations are factored as observation variables. Furthermore, in a mixed observability system, a subset of the observation variables provide perfect information on some aspects of the system, while other observation variables are noisy. The fully observable system components are thus directly observed through these perfect information observation variables. In the MO-PSR formulation, we lump these observation variables together, denoted by o^x, while the rest of the observation variables are denoted as o^y. Thus we have, $o^x \in \mathcal{O}^x$, $o^y \in \mathcal{O}^y$ and $\mathcal{O} = \mathcal{O}^x \times \mathcal{O}^y$, with observation at time step t, $\mathbf{o}_t = [o_t^x, o_t^y] \in \mathcal{O}$.

As in the case for a general dynamical system, we could characterize the system with a single system-dynamics matrix $\mathcal{D}$ as in the PSR approach (or equivalently, a single $P_{\mathcal{T},\mathcal{H}}$ in TPSRs), however, we obtain a more parsimonious model by characterizing the system with a *set* of system-dynamics matrices $\mathcal{D}_i$, $i = 1, \cdots |\mathcal{O}^x|$, where the histories represented in $\mathcal{D}_i$ is the set of histories that end with observation variable o^x taking value i. We present our model learning approach in the next section, and then show that the resultant model is more compact, and thus more amenable to efficient learning.

Mixed Observability Predictive State Representations (MO-PSRs)

In the MO-PSR approach, the set of all histories $\mathcal{H}$, is partitioned into sets $\mathcal{H}_i$, $i = 1, \cdots, |\mathcal{O}^x|$, where $\mathcal{H}_i$ is the set of histories that end with observation variable o^x taking value i, i.e., $h \in \mathcal{H}_i$ is of form $a_1[o_1^x, o_1^y]a_2[o_2^x, o_2^y] \cdots a_t[o_t^x, o_t^y]$, with $o_t^x = i$. We then estimate a set of matrices, $\{P_{\mathcal{T},\mathcal{H}_i} | i = 1, \cdots, |\mathcal{O}^x|\}$, where matrix $P_{\mathcal{T},\mathcal{H}_i}$ contains the joint probabilities of every test $\tau \in \mathcal{T}$ and every history $h \in \mathcal{H}_i$.

The MO-PSR system state representation makes use of the matrices of nonzero left singular vectors, U_i, from the SVD decomposition of the $P_{\mathcal{T},\mathcal{H}_i}$ matrices. Whereas the TPSR approach represents system state by projecting all histories $h \in \mathcal{H}$ onto the same matrix U, the MO-PSR approach projects histories from the different sets of histories $\mathcal{H}_i$, onto different matrices U_i. The MO-PSR representation of history $h \in \mathcal{H}_i$ is $U_i^\top p(\mathcal{T}, h)$.

MO-PSR learning algorithm We now present the MO-PSR learning algorithm. For ease of notation, we define a function $f^x(\mathbf{o})$ that returns the value of the o^x variable in $\mathbf{o}$. The symbol † denotes the Moore-Penrose pseudoinverse.

1. Sample action-observation trajectories from the mixed observability system to be modelled, and compute empirical estimates of the following sets of vectors and matrices:
   - $\{P_{\mathcal{H}_i} | i = 1, \cdots, |\mathcal{O}^x|\}$, where $P_{\mathcal{H}_i}$ is a $|\mathcal{H}_i| \times 1$ vector containing the probabilities of all histories $h \in \mathcal{H}_i$.
   - $\{P_{\mathcal{T},\mathcal{H}_i} | i = 1, \cdots, |\mathcal{O}^x|\}$, where $P_{\mathcal{T},\mathcal{H}_i}$ is a $|\mathcal{T}| \times |\mathcal{H}_i|$ matrix containing the joint probabilities of all tests $\tau \in \mathcal{T}$ and all histories $h \in \mathcal{H}_i$.
   - $\{P_{\mathcal{T},a\mathbf{o},\mathcal{H}_i} | i = 1, \cdots, |\mathcal{O}^x|\}$, for all $a \in \mathcal{A}$ and $\mathbf{o} \in \mathcal{O}$. $P_{\mathcal{T},a\mathbf{o},\mathcal{H}_i}$ is a $|\mathcal{T}| \times |\mathcal{H}_i|$ matrix containing the joint probabilities of all histories $h \in \mathcal{H}_i$, followed by action a, observation $\mathbf{o}$, and all tests $\tau \in \mathcal{T}$.
   - $\mathbf{p}$ is a $|\mathcal{O}^x| \times 1$ vector, where the i-th element, $\mathbf{p}_i$, is the probability that a history belongs to the set of history $\mathcal{H}_i$, or equivalently, the probability of observing $o^x = i$ under the sampling policy.
2. Perform SVD on empirically estimated $\hat{P}_{\mathcal{T},\mathcal{H}_i}$, to obtain matrices of left singular vectors, $\hat{U}_i$, for $i = 1, ..., |\mathcal{O}^x|$.
3. Compute model parameters from the empirical estimates.
   - $B^i_{a\mathbf{o}} = \hat{U}_{i'}^\top \hat{P}_{\mathcal{T},a\mathbf{o},\mathcal{H}_i} (\hat{U}_i^\top \hat{P}_{\mathcal{T},\mathcal{H}_i})^\dagger$, for $i = 1, ..., |\mathcal{O}^x|$, and for all $a \in \mathcal{A}$ and $\mathbf{o} \in \mathcal{O}$, where $i' = f^x(\mathbf{o})$. This is analogous to $M_{a\mathbf{o}}$ in the PSR model. Instead of having one linear operator associated with a particular $a\mathbf{o}$ combination, the MO-PSR model learns a set of such operators and applies the appropriate operator according to the history prior to the appearance of $a\mathbf{o}$.
   - Compute $b^i_0 = \frac{1}{\mathbf{p}_i} \hat{U}_i^\top \hat{P}_{\mathcal{T},\mathcal{H}_i} \mathbf{1}$, for $i = 1, ..., |\mathcal{O}^x|$, where $\mathbf{1}$ is a $|\mathcal{H}_i| \times 1$ vector of ones. This set of parameters is analogous to m_0 in the PSR model.
   - Compute $b^i_\infty = (\hat{P}^\top_{\mathcal{T},\mathcal{H}_i} \hat{U}_i)^\dagger \hat{P}_{\mathcal{H}_i}$, for $i = 1, ..., |\mathcal{O}^x|$, analogous to m_∞ in the PSR model.

Predictions

Given the model parameters, we can calculate the prediction of a sequence $a_1\mathbf{o}_1 \cdots a_t\mathbf{o}_t$, given an initial system state:

$$p(a_1\mathbf{o}_1 \cdots a_t\mathbf{o}_t) = \sum_{i_0=1}^{|\mathcal{O}^x|} \mathbf{p}_{i_0} \times \left((b^{i_t}_\infty)^\top B^{i_{t-1}}_{a_t\mathbf{o}_t} \cdots B^{i_1}_{a_2\mathbf{o}_2} B^{i_0}_{a_1\mathbf{o}_1} b^{i_0}_0 \right)$$
$$= (b^{i_t}_\infty)^\top B^{i_{t-1}}_{a_t\mathbf{o}_t} \cdots B^{i_1}_{a_2\mathbf{o}_2} \sum_{i_0=1}^{|\mathcal{O}^x|} \mathbf{p}_{i_0} \times B^{i_0}_{a_1\mathbf{o}_1} b^{i_0}_0 \ . \tag{3}$$

The value of o^x is known for all time steps, $1, \cdots, t$, since it is the observations of the fully observable system components. So for time steps $k = 2, \cdots, t$, we apply the appropriate operator $B^{i_{k-1}}_{a_k\mathbf{o}_k}$, where $i_{k-1} = f^x(\mathbf{o}_{k-1})$, the value of the o^x variable in the previous time step, $k-1$. At system initialization, the value of the o^x variable is undefined. We thus apply operators $B^{i_0}_{a_1\mathbf{o}_1} b^{i_0}_0$, weighted by $\mathbf{p}_{i_0}$, the probability of observing $o^x = i_0$ under the sampling policy, for $i_0 = 1, \cdots, |\mathcal{O}^x|$.

The system state at time step 1 is defined as:

$$b_1 = \frac{\sum_{i_0=1}^{|\mathcal{O}^x|} \mathbf{p}_{i_0} \times B^{i_0}_{a_1\mathbf{o}_1} b^i_0}{(b^{i_1}_\infty)^\top \sum_{i_0=1}^{|\mathcal{O}^x|} \mathbf{p}_{i_0} \times B^{i_0}_{a_1\mathbf{o}_1} b^{i_0}_0} \ , \tag{4}$$

and for time steps $t \geq 2$:

$$b_t = \frac{B^{i_{t-1}}_{a_t\mathbf{o}_t} \cdots B^{i_1}_{a_2\mathbf{o}_2} \sum_{i_0=1}^{|\mathcal{O}^x|} \mathbf{p}_{i_0} \times B^{i_0}_{a_1\mathbf{o}_1} b^{i_0}_0}{(b^{i_t}_\infty)^\top B^{i_{t-1}}_{a_t\mathbf{o}_t} \cdots B^{i_1}_{a_2\mathbf{o}_2} \sum_{i_0=1}^{|\mathcal{O}^x|} \mathbf{p}_{i_0} \times B^{i_0}_{a_1\mathbf{o}_1} b^{i_0}_0}$$
$$= \frac{B^{i_{t-1}}_{a_t\mathbf{o}_t} b_{t-1}}{(b^{i_t}_\infty)^\top B^{i_{t-1}}_{a_t\mathbf{o}_t} b_{t-1}} \ , \tag{5}$$

where $i_t = f^x(\mathbf{o}_t)$, for $t \geq 1$.

Consistency results

In this section, we show that the MO-PSR prediction for action-observation sequence $a_1\mathbf{o}_1 \cdots a_t\mathbf{o}_t$ is equivalent to the PSR prediction. We start by showing the relationship between the MO-PSR and TPSR model parameters (Boots, Siddiqi, and Gordon 2010). In our derivations, we impose an ordering for the matrices used for learning TPSR, so for example, $P_{\mathcal{H}}$ is ordered such that

$$P^\top_{\mathcal{H}} = \begin{bmatrix} P^\top_{\mathcal{H}_1} & P^\top_{\mathcal{H}_2} & \cdots & P^\top_{\mathcal{H}_{|\mathcal{O}^x|}} \end{bmatrix},$$

and similarly for $P_{\mathcal{T},\mathcal{H}}$ and $P_{\mathcal{T},a\mathbf{o},\mathcal{H}}$.

We first derive the relationship of the TPSR representation of the initial system state to the MO-PSR parameters:

$$\begin{aligned} b_0 &= U^\top P_{\mathcal{T},\mathcal{H}} \mathbf{1} \\ &= U^\top \begin{bmatrix} P_{\mathcal{T},\mathcal{H}_1} & P_{\mathcal{T},\mathcal{H}_2} & \cdots & P_{\mathcal{T},\mathcal{H}_{|\mathcal{O}^x|}} \end{bmatrix} \mathbf{1} \\ &= \sum_{i=1}^{|\mathcal{O}^x|} U^\top P_{\mathcal{T},\mathcal{H}_i} \mathbf{1} \\ &= \sum_{i=1}^{|\mathcal{O}^x|} U^\top (U_i U_i^\top P_{\mathcal{T},\mathcal{H}_i}) \mathbf{1} \\ &= \sum_{i=1}^{|\mathcal{O}^x|} \mathbf{p}_i \times (U^\top U_i) b^i_0 \ , \end{aligned} \tag{6}$$

where, since U_i and $P_{\mathcal{T},\mathcal{H}_i}$ have the same column space, $U_i U_i^\top P_{\mathcal{T},\mathcal{H}_i} = P_{\mathcal{T},\mathcal{H}_i}$.

We next derive the relationship between TPSR parameter $B_{a\mathbf{o}}$ and MO-PSR parameters. For the set of histories, $\mathcal{H}_i$,

$$\begin{aligned} B^i_{a\mathbf{o}} &= U_{i'}^\top P_{\mathcal{T},a\mathbf{o},\mathcal{H}_i} (U_i^\top P_{\mathcal{T},\mathcal{H}_i})^\dagger \\ B^i_{a\mathbf{o}} (U_i^\top P_{\mathcal{T},\mathcal{H}_i}) &= U_{i'}^\top P_{\mathcal{T},a\mathbf{o},\mathcal{H}_i} \ , \end{aligned} \tag{7}$$

where $i' = f^x(\mathbf{o})$, and, assuming $P_{\mathcal{T},a\mathbf{o},\mathcal{H}_i}$ and $P_{\mathcal{T},\mathcal{H}_i}$ perfectly characterise the system, we get an exact solution for $B^i_{a\mathbf{o}}$. From the TPSR definition of $B_{a\mathbf{o}}$:

$$\begin{aligned} B_{a\mathbf{o}} &= U^\top P_{\mathcal{T},a\mathbf{o},\mathcal{H}} (U^\top P_{\mathcal{T},\mathcal{H}})^\dagger \\ B_{a\mathbf{o}}(U^\top P_{\mathcal{T},\mathcal{H}}) &= U^\top P_{\mathcal{T},a\mathbf{o},\mathcal{H}} \ , \end{aligned} \tag{8}$$

which assumes that $B_{a\mathbf{o}}$ is an exact solution.

We can pick out the columns corresponding to histories $h \in \mathcal{H}_i$, in the LHS and RHS of (8):

$$\begin{aligned} B_{a\mathbf{o}}(U^\top P_{\mathcal{T},\mathcal{H}_i}) &= U^\top P_{\mathcal{T},a\mathbf{o},\mathcal{H}_i} \\ &= U^\top U_{i'} U_{i'}^\top P_{\mathcal{T},a\mathbf{o},\mathcal{H}_i} \\ &= U^\top U_{i'} (B^i_{a\mathbf{o}} U_i^\top P_{\mathcal{T},\mathcal{H}_i}) \text{ [(7)]}, \end{aligned} \tag{9}$$

where $i' = f^x(\mathbf{o})$. $B_{a\mathbf{o}}$ is a similarity transform of $B^i_{a\mathbf{o}}$ when their application is limited to state representations for histories from the set $\mathcal{H}_i$:

$$\begin{aligned} B_{a\mathbf{o}}(U^\top P_{\mathcal{T},\mathcal{H}_i}) &= U^\top U_{i'} (B^i_{a\mathbf{o}} U_i^\top P_{\mathcal{T},\mathcal{H}_i}) \\ B_{a\mathbf{o}} U^\top &= U^\top U_{i'} B^i_{a\mathbf{o}} U_i^\top \\ B_{a\mathbf{o}} U^\top U &= U^\top U_{i'} B^i_{a\mathbf{o}} U_i^\top U \\ B_{a\mathbf{o}} &= (U^\top U_{i'}) B^i_{a\mathbf{o}} (U_i^\top U) \ , \end{aligned} \tag{10}$$

where, since U has orthogonal columns, $U^\top U$ is identity.

Lastly, we derive the relationship between the TPSR parameter b_∞, and MO-PSR parameters. For the set of histories, $\mathcal{H}_i$, we have the MO-PSR definition:

$$\begin{aligned} (b^i_\infty)^\top &= P^\top_{\mathcal{H}_i} (U_i^\top P_{\mathcal{T},\mathcal{H}_i})^\dagger \\ (b^i_\infty)^\top (U_i^\top P_{\mathcal{T},\mathcal{H}_i}) &= P^\top_{\mathcal{H}_i} \ , \end{aligned} \tag{11}$$

which assumes that b^i_∞ is an exact solution. And from the TPSR definition:

$$\begin{aligned} b_\infty^\top &= P^\top_{\mathcal{H}} (U^\top P_{\mathcal{T},\mathcal{H}})^\dagger \\ b_\infty^\top (U^\top P_{\mathcal{T},\mathcal{H}}) &= P^\top_{\mathcal{H}} \ , \end{aligned} \tag{12}$$

which also assumes that b_∞ is an exact solution.

Next we pick out the columns corresponding to histories $h \in \mathcal{H}_i$, in the LHS and RHS of (12):

$$\begin{aligned} b_\infty^\top (U^\top P_{\mathcal{T},\mathcal{H}_i}) &= P^\top_{\mathcal{H}_i} \\ &= (b^i_\infty)^\top (U_i^\top P_{\mathcal{T},\mathcal{H}_i}) \text{ [(11)]}. \end{aligned} \tag{13}$$

b_∞ is equivalent to a similarity transform of b^i_∞ when their application is limited to state representations for histories from the set $\mathcal{H}_i$:

$$\begin{aligned} b_\infty^\top (U^\top P_{\mathcal{T},\mathcal{H}_i}) &= (b^i_\infty)^\top (U_i^\top P_{\mathcal{T},\mathcal{H}_i}) \\ b_\infty^\top U^\top &= (b^i_\infty)^\top U_i^\top \\ b_\infty^\top U^\top U &= (b^i_\infty)^\top U_i^\top U \\ b_\infty^\top &= (b^i_\infty)^\top (U_i^\top U) \ . \end{aligned} \tag{14}$$

From the prediction with the TPSR model in (2), we substitute in the derivations from (6), (10) and (14):

$$\begin{aligned} p(a_1\mathbf{o}_1 \cdots a_t\mathbf{o}_t) &= b_\infty^\top B_{a_t\mathbf{o}_t} \cdots B_{a_1\mathbf{o}_1} b_0 \\ &= (b^{i_t}_\infty)^\top (U_{i_t}^\top U)(U^\top U_{i_t}) B^{i_{t-1}}_{a_t\mathbf{o}_t} (U^\top_{i_{t-1}} U) \\ &\quad \cdots (U^\top U_{i_2}) B^{i_1}_{a_2\mathbf{o}_2} (U_{i_1}^\top U) \\ &\quad \times \sum_{i_0=1}^{|\mathcal{O}^x|} \mathbf{p}_{i_0} \times (U^\top U_{i_1}) B^{i_0}_{a_1\mathbf{o}_1} (U_{i_0}^\top U)(U^\top U_{i_0}) b_0^{i_0} \\ &= (b^{i_t}_\infty)^\top B^{i_{t-1}}_{a_t\mathbf{o}_t} \cdots B^{i_1}_{a_2\mathbf{o}_2} \sum_{i_0=1}^{|\mathcal{O}^x|} \mathbf{p}_{i_0} \times B^{i_0}_{a_1\mathbf{o}_1} b_0^{i_0}, \end{aligned} \tag{15}$$

and arrive at the MO-PSR prediction equation (3). Since the predictions of the TPSR and PSR models are equivalent (Boots, Siddiqi, and Gordon 2010), the predictions of the MO-PSR and the PSR models are also equivalent.

In the above, $B_{a_k\mathbf{o}_k}$ can be substituted with $(U^\top U_{i_k}) B^{i_{k-1}}_{a_k\mathbf{o}_k} (U^\top_{i_{k-1}} U)$ because the terms are applied to the state representation for a history that ends in $a_{k-1}\mathbf{o}_{k-1}$. (Similarly for the substitution of b_∞ with $b^{i_t}_\infty$). Also, the column space of U encompasses that of U_i, thus $U^\top U_i$ has orthogonal columns, and $(U^\top U_i)^\top (U^\top U_i) = (U_i^\top U)(U^\top U_i)$ is identity.

Complexity Analysis

In the following, we compare the MO-PSR and TPSR approaches in terms of time complexity for learning mixed observability systems. We show that while both give equivalent predictions in terms of accuracy, the MO-PSR approach has lower time complexity and learns models which are more compact.

Computational Complexity

The main computational bottleneck in the spectral approach to model learning is the SVD operation itself. The SVD operation on a matrix of size $m \times n$, where $n < m$, has complexity $O(mn^2)$. Assume that $|\mathcal{H}| = |\mathcal{T}| = n$. The TPSR approach requires an SVD operation on the matrix $P_{\mathcal{T},\mathcal{H}}$, which has complexity $O(n^3)$. The MO-PSR approach, on the other hand, performs $|\mathcal{O}^x|$ of SVD operations on matrices $P_{\mathcal{T},\mathcal{H}_i}$, $i = 1, \cdots, |\mathcal{O}^x|$. Assuming that on the average, the size of each set of history $\mathcal{H}_i$ is $\frac{n}{|\mathcal{O}^x|}$, SVD on each of the matrices $P_{\mathcal{T},\mathcal{H}_i}$ has $O(n(\frac{n}{|\mathcal{O}^x|})^2)$ operations. This gives a total of $O(|\mathcal{O}^x|(\frac{n^3}{|\mathcal{O}^x|^2})) = O(\frac{n^3}{|\mathcal{O}^x|})$ operations, representing a $|\mathcal{O}^x|$ factor reduction in computational complexity.

Model Complexity

Next, we consider the size of the models learned. For the TPSR approach, the dimensions of the model parameters (and hence, the number of entries in the model parameters) are directly related to the number of nonzero left singular vectors from the SVD on matrix $P_{\mathcal{T},\mathcal{H}}$. This is the same as the rank of $P_{\mathcal{T},\mathcal{H}}$, and the rank of the system-dynamics matrix $\mathcal{D}$. For the MO-PSR approach, the dimensions of the model parameters are directly related to the number of nonzero left singular vectors from the SVD on each matrix

$P_{\mathcal{T},\mathcal{H}_i}$, $i = 1, \cdots |\mathcal{O}^x|$. This is the same as the rank of each matrix $\mathcal{D}_i$, $i = 1, \cdots |\mathcal{O}^x|$. Thus, we begin by examining the ranks of matrices $\mathcal{D}$, and $\mathcal{D}_i$, $i = 1, \cdots |\mathcal{O}^x|$.

Rank of system-dynamics matrices As mentioned above, mixed observability systems can be modelled as MOMDPs. The MOMDP formulation utilizes the concept of an underlying or nominal state, s, which for a mixed observability system is factorized as the nominal state of the fully observable system components, x, and the nominal state of the partially observable system components, y. So, for a mixed observability system, the state space is factored as $\mathcal{S} = \mathcal{X} \times \mathcal{Y}$. MOMDPs maintain system state information by keeping track of the probabilities of being in each of the nominal states $s \in \mathcal{S}$ as a function of history. Thus the representation of history h is a $|\mathcal{X}||\mathcal{Y}| \times 1$ belief vector, with vector entries $p(x, y|h)$, for all $x \in \mathcal{X}$ and all $y \in \mathcal{Y}$. The system-dynamics matrix $\mathcal{D}$ of a system modelled as a MOMDP can be generated from estimating $p(\tau|h)$, for all $\tau \in \mathcal{T}$ and $h \in \mathcal{H}$, from $p(\tau|h) = \sum_{x,y} p(\tau|x, y)p(x, y|h)$ (Ong et al. 2010).

Lemma 1. *For any mixed observability system that can be modelled as a MOMDP with fully observable nominal states $x \in \mathcal{X}$ and partially observable nominal states $y \in \mathcal{Y}$, the rank of the matrix $\mathcal{D}$ is no more than $|\mathcal{X}||\mathcal{Y}|$.*

Proof: Matrix $\mathcal{D}$ can be factored as the product of matrices A and C, where A is a $|\mathcal{H}| \times |\mathcal{X}||\mathcal{Y}|$ matrix with entries $p(x, y|h)$, and C is a $|\mathcal{X}||\mathcal{Y}| \times |\mathcal{T}|$ matrix with entries $p(\tau|x, y)$. Both A and C have at most rank $|\mathcal{X}||\mathcal{Y}|$, thus $\mathcal{D}$ also has at most rank $|\mathcal{X}||\mathcal{Y}|$.

Now consider the matrices $\mathcal{D}_i$, $i = 1, \cdots |\mathcal{O}^x|$, where the histories represented in $\mathcal{D}_i$ is the set of histories $\mathcal{H}_i$ which end in the observation variable o^x taking on value i. Note that observation o^x represents the observation that gives perfect information on the fully observable components in the system while in the MOMDP framework, x represents the nominal state of those components. Thus, in the MOMDP model, o^x and x are one and the same, and the observation space $\mathcal{O}^x$ is the same as the nominal state space $\mathcal{X}$.

Theorem 1. *For any mixed observability system that can be modelled as a MOMDP with fully observable nominal states $x \in \mathcal{X}$ and partially observable nominal states $y \in \mathcal{Y}$, the rank of each of the $\mathcal{D}_i$, $i = 1, \cdots |\mathcal{O}^x|$, is no more than $|\mathcal{Y}|$.*

Proof: The matrix $\mathcal{D}_i$ has rows corresponding to the set of histories which end in the observation variable o^x taking on value i. $\mathcal{D}_i$ can be factored as the product of A_i and C, where A_i is $|\mathcal{H}_i| \times |\mathcal{X}||\mathcal{Y}|$ and C is $|\mathcal{X}||\mathcal{Y}| \times |\mathcal{T}|$. Here, as above, C has at most rank $|\mathcal{X}||\mathcal{Y}|$. However, this is not the case for A_i. Each of the histories in the set $\mathcal{H}_i$ ends in the observation variable o^x taking on value i. Thus, for all histories $h \in \mathcal{H}_i$, $p(x, y|h) = 0$ for all $x \neq i$. So only columns in A_i corresponding to $x = i$, for all $y \in \mathcal{Y}$, are nonzero, and thus A_i has at most rank $|\mathcal{Y}|$. $\mathcal{D}_i$ thus also has at most rank $|\mathcal{Y}|$.

Size of model representation The dimensions of the TPSR model parameters, b_0, b_∞ and $B_{a\mathbf{o}}$, for all $a \in \mathcal{A}$ and $\mathbf{o} \in \mathcal{O}$, are functions of the number of columns in U, which is the same as the rank of $\mathcal{D}$. From Lemma 1, for a system that can be modelled by a MOMDP, the model representation size for a TPSR is of order $2|\mathcal{X}||\mathcal{Y}| + |\mathcal{A}||\mathcal{O}|(|\mathcal{X}||\mathcal{Y}|)^2$, or equivalently, $2|\mathcal{O}^x||\mathcal{Y}| + |\mathcal{A}||\mathcal{O}|(|\mathcal{O}^x||\mathcal{Y}|)^2$. From Theorem 1, the upper bound on the rank of $\mathcal{D}_i$ is $|\mathcal{Y}|$. Thus, the upper bound on the dimensions of both b_0^i and b_∞^i is $|\mathcal{Y}| \times 1$, and the upper bound on the dimensions of each of the $|\mathcal{A}||\mathcal{O}|$ matrices $B_{a\mathbf{o}}^i$ is $|\mathcal{Y}| \times |\mathcal{Y}|$. There are $|\mathcal{O}^x|$ sets of such model parameters, thus the MO-PSR model representation size is upper bounded by $2|\mathcal{O}^x||\mathcal{Y}| + (|\mathcal{O}^x||\mathcal{A}||\mathcal{O}||\mathcal{Y}|^2) = 2|\mathcal{O}^x||\mathcal{Y}| + (1/|\mathcal{O}^x|)(|\mathcal{A}||\mathcal{O}|(|\mathcal{O}^x||\mathcal{Y}|)^2)$. In general, the second term is dominant, so the MO-PSR approach will have a factor of $|\mathcal{O}^x|$ reduction compared to the TPSR approach.

Preliminary Results

We present preliminary results to illustrate the performance of MO-PSR compared to TPSR on a problem from the International Probabilistic Planning Competition (IPPC) at ICAPS 2010. The *Elevators* problem consists of one or more elevator agents operating in a building, and human agents waiting on each of the floors. The elevator agents are under MO-PSR/TPSR control and are fully observable, while the human agents follow a stochastic process and are partially observable. We refer the reader to the IPPC's website for detailed problem specification.

Method and Evaluation

We ran experiments on two *Elevators* domains: 1 elevator and 3 floors (*Elev1Floor3*), and, 1 elevator and 4 floors (*Elev1Floor4*). In both domains, the MO-PSR and TPSR learning algorithms were each given 10,000 to 100,000 training trajectories of 8 time steps, sampled from a simulator given uniformly random generated actions. The MO-PSR algorithm learned models from histories of lengths up to 3 and tests of lengths up to 4 (*MO-PSR h3t4*) as well as models from histories and tests of length 1 (*MO-PSR h1t1*). The TPSR algorithm was only able to incorporate histories and tests of length 1 in the model (*TPSR h1t1*) without running out of memory (7 GB) during the SVD operation.

In the *Elev1Floor3* domain, the upper bounds on the ranks of $P_{\mathcal{T},\mathcal{H}}$ and each of $P_{\mathcal{T},\mathcal{H}_i}$ is 768 and 16, respectively. Accordingly, we learned MO-PSR models by keeping the largest 16 left singular vectors from SVD on each $P_{\mathcal{T},\mathcal{H}_i}$ while for TPSR, no more than the largest 250 left singular vectors from the SVD on $P_{\mathcal{T},\mathcal{H}}$ could be used without running out of memory. In the *Elev1Floor4* domain, while the upper bound on the rank of each of $P_{\mathcal{T},\mathcal{H}_i}$ is 64, with 10,000 training trajectories there were only 48 nonzero left singular vectors from SVD on each $P_{\mathcal{T},\mathcal{H}_i}$ so we learned MO-PSR models of dimension 48 throughout. The upper bound on the rank of $P_{\mathcal{T},\mathcal{H}}$ is 4096 but for TPSR, no more than the largest 128 left singular vectors from the SVD on $P_{\mathcal{T},\mathcal{H}}$ could be used without running out of memory.

We evaluated the learned models on 1000 test trajectories of length 4. The prediction error for each test trajectory is the mean squared error = $\frac{1}{4}\sum_{t=1\ldots4}(p_t - \hat{p}_t)^2$, where p_t is the actual probability of the observation at time t according to the true system dynamics, and $\hat{p}_t$ is the probability of the observation as predicted by the learned model.

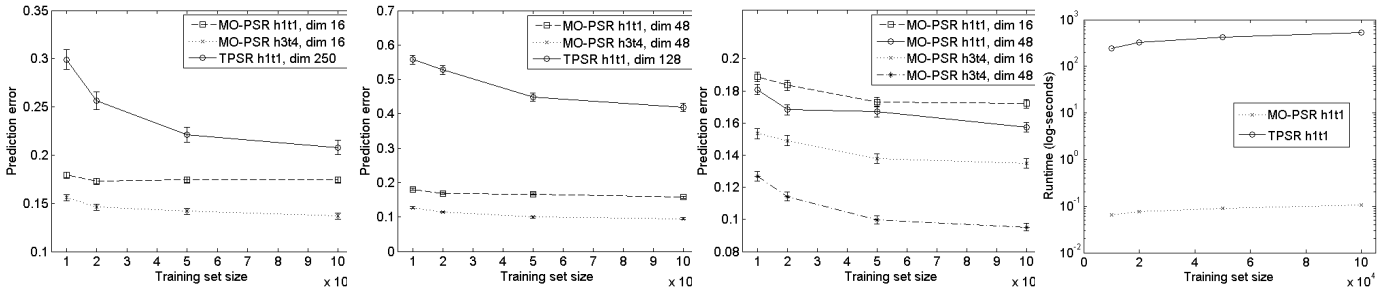

Figure 1: (*a*) Prediction errors of the MO-PSR and TPSR models on *Elev1Floor3*. (*b*) Prediction errors of the MO-PSR and TPSR models on *Elev1Floor4*. (*c*) Comparison of prediction errors on *Elev1Floor4* for MO-PSR models trained with different parameters. (*d*) Run times for SVD operation for the MO-PSR and TPSR models on *Elev1Floor4*.

Results

As shown in Figure 1 (a) and (b), in both the *Elev1Floor3* and *Elev1Floor4* domains, MO-PSR models learned from histories and tests of length 1 (*MO-PSR h1t1*) gave more accurate predictions than the corresponding TPSR models *TPSR h1t1*. MO-PSR models learned from histories of up to length 3 and tests of up length 4 (*MO-PSR h3t4*) further improved in prediction accuracy. A possible explanation for the better performance of *MO-PSR h1t1* versus *TPSR h1t1* is their respective model complexities. Looking at the *Elev1Floor3* domain for example, in the TPSR model, each of the $B_{a\mathbf{o}}$ matrices has dimension 250×250. In the MO-PSR model, for each combination of $a\mathbf{o}$ values, there are $|\mathcal{O}^x| = 48$ of $B^i_{a\mathbf{o}}$ matrices of dimension 16×16 each. Thus, the model representation size for the MO-PSR model is smaller by a factor of approximately $\frac{250^2}{48 \times 16^2} \approx 5$. This results in less training data required to learn a model of comparative or better accuracy.

Furthermore, the MO-PSR algorithm was able to achieve this better performance with much reduced computational complexity, as illustrated in Figure 1 (d). In the MO-PSR algorithm, the run time for SVD operations was at most 0.11 second, for the model learned from 100,000 training trajectories. The corresponding time required in the TPSR algorithm was 529.5 seconds. All experiments were run on a PC with two 3.3 GHz CPUs and 7GB RAM, running Ubuntu 10.04.3 LTS.

Figure 1 (c) compares the performance of MO-PSR models on *Elev1Floor4*, when learning from histories and tests of different lengths, and with different dimensions for model parameters, i.e. by retaining different numbers of left singular vectors from SVD on each $P_{\mathcal{T},\mathcal{H}_i}$. The results show that MO-PSR models with dimension 48 give better predictions as compared to models with dimension 16. This is to be expected as the upper bound on the rank of each $P_{\mathcal{T},\mathcal{H}_i}$ is 64. Given the same model dimensions, MO-PSR models learned from histories and tests of longer lengths perform better than with shorter lengths, because more information is extracted from the training data.

Final Discussion

We have proposed a novel model learning approach for learning agent behaviours in systems with mixed observability. The MO-PSR framework shares the properties of power of expressiveness and ease of model learning as in the general PSR approach. However, the MO-PSR approach takes advantage of structural properties in the system to improve learning and modelling efficiency, as compared to PSRs. While the model learning algorithm itself is relatively straightforward, it has potential impact for a great number of domains, given that many large, complex systems are mixed observability systems. As such, the consistency and complexity results we have shown are important, for ensuring correctness and generalization.

The work presented here shares some similarities with memory-PSRs (James, Wessling, and Vlassis 2006), where the systems-dynamics matrix is partitioned based on memories consisting of arbitrary length $a\mathbf{o}$ sequences. In contrast, MO-PSR partitions the systems-dynamics matrix based on only the fully observable o^x variable, and in the last observation only. While memory-PSRs enable more compact modelling of some dynamical systems, the class of such systems has not been clearly defined, and is not believed to overlap directly with mixed observability systems.

The focus of this paper is on model learning, therefore we don't demonstrate how the learned model can be used for control and planning. However, prior work on PSR planning algorithms is well established, much of which can likely be extended to MO-PSRs (James, Singh, and Littman 2004; Izadi and Precup 2008; James, Wessling, and Vlassis 2006; Boots, Siddiqi, and Gordon 2010; Rafols et al. 2005; Aberdeen, Buffet, and Thomas 2007). We expect computational savings on the order of $|\mathcal{O}^x|$ compared to planning with TPSRs, due to the smaller model representation. Verifying this will be the subject of future work.

Acknowledgments

The authors would like to thank Doina Precup for helpful discussions on this work. Financial support for this research was provided by the NSERC Discovery grant program.

References

Aberdeen, D.; Buffet, O.; and Thomas, O. 2007. Policy-gradients for PSRs and POMDPs. In *AISTATS*.

Boots, B.; Siddiqi, S.; and Gordon, G. 2010. Closing the learning-planning loop with predictive state representations. In *Proceedings of Robotics: Science and Systems*.

Capitan, J.; Merino, L.; and Ollero, A. 2011. Multi-robot coordinated decision making under mixed observability through decentralized data fusion. In *Proceedings of the 11th International Conference on Mobile Robots and Competitions (Robotica 2011)*.

Chades, I.; Carwardine, J.; Martin, T. G.; Nicol, S.; Sabbadin, R.; and Buffer, O. 2012. MOMDPs: a solution for modelling adaptive management problems. In *AAAI*.

Izadi, M., and Precup, D. 2008. Point-based planning for predictive state representations. In *Canadian Conference on AI*.

James, M. R.; Singh, S.; and Littman, M. L. 2004. Planning with predictive state representations. In *International Conference on Machine Learning and Applications*, 304–311.

James, M. R.; Wessling, T.; and Vlassis, N. 2006. Improving approximate value iteration using memories and predictive state representations. In *AAAI*.

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

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

Ong, S. C. W.; Png, S. W.; Hsu, D.; and Lee, W. S. 2010. Planning under uncertainty for robotic tasks with mixed observability. *International Journal of Robotics Research*.

Rafols, E. J.; Ring, M.; Sutton, R.; and Tanner, B. 2005. Using predictive representations to improve generalization in reinforcement learning. In *IJCAI*.

Singh, S.; James, M.; and Rudary, M. 2004. Predictive state representations: A new theory for modeling dynamical systems. In *Proceedings UAI*.