Predictive Representations for Policy Gradient in POMDPs

Abdeslam Boularias BOULARIAS@DAMAS.IFT.ULAVAL.CA
Brahim Chaib-draa CHAIB@DAMAS.IFT.ULAVAL.CA
Department of Computer Science and Software Engineering, Laval University, Quebec, Canada, G1K 7P4

Abstract

We consider the problem of estimating the policy gradient in Partially Observable Markov Decision Processes (POMDPs) with a special class of policies that are based on Predictive State Representations (PSRs). We compare PSR policies to Finite-State Controllers (FSCs), which are considered as a standard model for policy gradient methods in POMDPs. We present a general Actor-Critic algorithm for learning both FSCs and PSR policies. The critic part computes a value function that has as variables the parameters of the policy. These latter parameters are gradually updated to maximize the value function. We show that the value function is polynomial for both FSCs and PSR policies, with a potentially smaller degree in the case of PSR policies. Therefore, the value function of a PSR policy can have less local optima than the equivalent FSC, and consequently, the gradient algorithm is more likely to converge to a global optimal solution.

1. INTRODUCTION

Learning an optimal policy in a partially observable environment is still considered as one of the most difficult challenges in Reinforcement Learning (RL). To achieve this, one must consider two main issues: (1) mapping histories of interaction with the environment into an internal state representation, (2) finding the optimal decisions associated to each internal state. Partially Observable Markov Decision Processes (POMDPs) provide a rich mathematical framework for studying such problems. In POMDPs, the environment is represented by a finite set of states, and the effects of actions on the environment are represented by probabilistic transition functions. After every executed action, the agent perceives a numerical reward and a partial, possibly noisy, observation.

Instead of keeping track of ever growing histories, Finite-States Controllers (FSCs) provide an efficient graphical model for mapping histories into equivalence classes called *internal-states* (I-states). Most of policy gradient methods for POMDPs are based on this family of policies (Meuleau et al., 1999; Baxter & Bartlett, 2000; Shelton, 2001). FSC is like a probabilistic automaton where the observations are the inputs and the actions are the outputs, with stochastic transitions between the I-states. The main drawback of FSCs is that I-states do not directly affect the state of the environment. In fact, only the sequence of executed actions determines the returned outcome, and the same sequence of actions may be generated by different sequences of I-states. Thus, considering the sequence of I-states within the learning data slows down the convergence of RL algorithms (Aberdeen & Baxter, 2002).

The principal contribution of this paper is a new policy gradient method for finite-horizon POMDPs based on a Predictive State Representation (PSR) policy. Aberdeen et al. (2007) have already proposed to use PSRs for reinforcement learning in partially observable domains. However, they used PSRs for learning the dynamics of the environment, and mapped the *belief states* into a distribution over actions through a linear approximator. In our approach, we use PSRs to represent policies directly without learning a model of the environment. The parameters of PSR policies are updated according to the gradient of a *value function.* Our second contribution is a comparison of PSR policies to FSCs. For this purpose, we use the same basic method for learning both FSCs and PSR policies, and show that the value function of a PSR policy is potentially less complex than the value function of FSC policy. This result is due to the fact that the parameters of a PSR policy are based only on observable data. Finally, we empirically demonstrate the effectiveness of PSR policies through several standard benchmarks.

Appearing in *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada, 2009.

2. Background

We review the POMDP, PSR and FSC models, and show how PSRs can be adapted to represent policies.

2.1. POMDPs

Formally, a POMDP is defined by the following components: a finite set of hidden states $\mathcal{S}$; a finite set of actions $\mathcal{A}$; a finite set of observations $\mathcal{O}$; a set of transition functions $\{T^a\}$, where $T^a(s,s')$ is the probability that the agent will end up in state s' after taking action a in state s; a set of observation functions $\{O^{a,o}\}$, where $O^{a,o}(s)$ gives the probability that the agent receives observation o after taking action a and getting to state s'; and a reward function r, such that $r(s,a)$ is the immediate reward received when the agent executes action a in state s. Additionally, there can be a discount factor, $\gamma \in [0,1]$, which is used to weigh less rewards received further into the future. The horizon H indicates the maximum number of actions that an agent can execute while performing the given task.

Instead of memorizing a complete history of actions and observations $h_t = a_1 o_1 \dots a_t o_t$, a probability distribution over the hidden states, called *the belief state*, is sufficient to make predictions about the future observations and rewards. The belief state is a vector b_t, where $b_t(s) = Pr(s_t = s|h_t), s \in \mathcal{S}$. The expected reward of an action a given a history h_t is given by:

$$R(a|h_t) = \sum_{s \in S} b_t(s) r(s,a) \tag{1}$$

2.2. PSRs

PSRs (Littman et al., 2002) are an alternative model to represent partially observable environments without using hidden states. The fundamental idea of PSRs is to replace the probabilities on states by probabilities on particular future trajectories, called *core tests*, or by weights of particular past trajectories, called *core histories*.

A test q is an ordered sequence of (*action*, *observation*) couples, i.e. $q = a^1 o^1 \dots a^k o^k$. The probability of a test q starting after a history h_t is defined by:

$$Pr(q^o|h_t, q^a) \overset{def}{=} Pr(o_{t+1} = o^1, \dots, o_{t+k} = o^k | h_t, \\ a_{t+1} = a^1, \dots, a_{t+k} = a^k) \tag{2}$$

where q^a denotes the actions of q and q^o denotes its observations.

The probability of any test q after a history h_t depends linearly on the probabilities of the same test after the different core histories. We use $\mathcal{H}$ to indicate the set of core histories. The PSR belief state[1] is a vector b_t, where $b_t(h)$ is the weight of the core history $h \in \mathcal{H}$ in the current history h_t. The probability of any test q after a history h_t is given by:

$$Pr(q^o|h_t, q^a) = \sum_{h \in \mathcal{H}} b_t(h) Pr(q^o|h, q^a) = b_t^T m^q \tag{3}$$

where $m^q(h) \overset{def}{=} Pr(q^o|h, q^a)$ is independent of h_t, and $b_t(h)$ is independent of q. The belief state $b_t \in \mathbb{R}^{|\mathcal{H}|}$ is a vector of real-valued weights and not probabilities.

After executing an action a and receiving an observation o, the PSR belief is updated by Bayes' Rule:

$$\begin{aligned} Pr(q^o|h_t, a, o, q^a) &= \frac{Pr(oq^o|h_t, a, q^a)}{Pr(o|h_t, a)} \\ &= \frac{\sum_{h \in \mathcal{H}} b_t(h) Pr(oq^o|h, a, q^a)}{\sum_{h \in \mathcal{H}} b_t(h) Pr(o|h, a)} \\ &= \frac{\sum_{h \in \mathcal{H}} b_t(h) Pr(o|h, a) Pr(q^o|h, a, o, q^a)}{\sum_{h \in \mathcal{H}} b_t(h) Pr(o|h, a)} \\ &= \frac{\sum_{h \in \mathcal{H}} b_t(h) m^{ao}(h) \sum_{h' \in \mathcal{H}} b_{hao}(h') m^q(h')}{\sum_{h \in \mathcal{H}} b_t(h) m^{ao}(h)} \\ &= \sum_{h \in \mathcal{H}} b_{t+1}(h) m^q(h) = b_{t+1}^T m^q \end{aligned} \tag{4}$$

where

$$b_{t+1}(h) = \frac{\sum_{h' \in \mathcal{H}} b_t(h') b_{h'ao}(h) m^{ao}(h')}{\sum_{h' \in \mathcal{H}} b_t(h') m^{ao}(h')} \tag{5}$$

The parameters of a PSR based on core histories are: $\mathcal{A}$, $\mathcal{O}$, the set of core histories $\mathcal{H}$, and the vectors m^{ao} and b_{hao}, $\forall a \in \mathcal{A}, o \in \mathcal{O}, h \in \mathcal{H}$. The vector b_{hao} denotes the belief state at the history hao. Notice that the vectors m^{ao} define a probability distribution over the observations o for each action a and core history h, thus, they can be represented by a softmax function. This latter property is not satisfied in PSRs based on core tests, and it lays behind our choice of using core histories in this paper. However, most of the results that we will present also apply to PSRs with core tests.

2.3. FSCs

The goal of a rational agent is to find an action that maximizes its expected long-term reward after a given history. The function that selects an action a given a history h_t is called *policy*. A parametric stochastic policy π is a function defined by $\pi(a, h_t, \theta) = Pr(a_{t+1} = a|h_t, \theta)$, where θ is a vector of real-valued parameters $\theta = (\theta_1, \theta_2, \dots, \theta_n)$. Finite-State Controllers are an

[1] The same notation is used for different types of belief state; the interpretation depends on the model.

efficient graphical model used to represent stochastic policies without memorizing all the histories (Meuleau et al., 1999). An FSC is defined by: a finite set of internal states $\mathcal{G}$; a set of action-selection functions $\mu^{o,a}$, where $\mu^{o,a}(g,\theta)$ is the probability that the agent will execute action a if its I-state is g, its last observation was o, and the parameters of the policy are θ; a set of transition functions ω^o, where $\omega^o(g,g',\theta)$ is the probability that the agent will select the I-state g if the current I-state is g and the perceived observation is o.

Contrary to the states of a POMDP, the I-states of an FSC are completely observable, thus, a history h_t should be a sequence of actions, observations and sampled I-states. However, the cumulated reward of a given history depends only on the executed actions and the perceived observations. Moreover, the same sequence of actions and observations can be used to update the parameters (μ and ω) of all the I-states that may have generated that sequence, with different likelihoods. Shelton (2001), Aberdeen and Baxter (2002) proposed a Rao-Blackwellized version of FSCs where *internal-belief states* are calculated at each step instead of sampling a single I-state. Rao-Blackwellization technique is known to reduce the variance of Monte-Carlo estimators (Casella & Robert, 1996). The internal belief state is a vector $b_t(.,\theta)$, where $b_t(g,\theta) = Pr(g_t = g|h_t,\theta), g \in \mathcal{G}$. The initial internal belief state b_0 is given in the parameters θ. At each step, the internal belief state is updated by Bayes' Rule as in POMDPs, and used to sample an action to be executed.

2.4. PSR Policies

PSRs can be used to represent policies instead of the environment dynamics by switching the roles of actions and observations (Wiewiora, 2005). In fact, a test q can be redefined as an ordered sequence of (*observation, action*) couples, i.e. $q = o^1a^1\ldots o^ka^k$. Given the vector θ of policy parameters, the probability of q starting after h_t is redefined as:

$$Pr(q^a|h_t,q^o,\theta) \overset{def}{=} Pr(a_{t+1}=a^1,\ldots,a_{t+k}=a^k|h_t, o_{t+1}=o^1,\ldots,o_{t+k}=o^k,\theta) \quad (6)$$

As for the environment dynamics, the probability $Pr(q^a|h_t,q^o,\theta)$ of any test q starting after a history h_t is given by a linear combination of the probabilities of the same test q starting after different core histories $h \in \mathcal{H}$. The internal belief state b_t corresponds to the weights of these core histories in $Pr(q^a|h_t,q^o,\theta)$. In particular, the probability of executing action a at time t after observing o is given by:

$$Pr(a|h_t,o) = \sum_{h\in\mathcal{H}} b_t(h,\theta)Pr(a|h,o,\theta) = b_t^T m^{oa} \quad (7)$$

The core histories of a policy are independent from the core histories of the controlled environment. We will use $\mathcal{H}$ only to indicate the policy core histories, since no model of the environment is used in our approach.

After receiving an observation o and executing an action a, the internal belief state $b_t(.,\theta)$ is updated by:

$$b_{t+1}(h,\theta) = \frac{\sum_{h'\in\mathcal{H}} b_t(h',\theta)b_{h'oa}(h,\theta)m^{oa}(h',\theta)}{\sum_{h'\in\mathcal{H}} b_t(h',\theta)m^{oa}(h',\theta)} \quad (8)$$

A PSR policy is defined by: the set of core histories $\mathcal{H}$, and the vectors m^{oa} and b_{hoa}, $\forall o \in \mathcal{O}, a \in \mathcal{A}, h \in \mathcal{H}$. The initial history $h_0 = \varepsilon$ is always a core history, thus $b_0(\varepsilon,\theta) = 1$ and $\forall h \in \mathcal{H} - \{\varepsilon\} : b_0(h,\theta) = 0$.

3. Reinforcement Learning in POMDPs

There are two major families of reinforcement learning algorithms that can be used to find the optimal parameters of a policy. In the first family, a *value function* is used to estimate the expected long-term reward of each action in every state of the environment (which is a history h_t in our case). The parameters of the policy are adjusted so that in each state, the actions with the highest estimated long-term reward will be executed more frequently. In the second family, the optimal policy is gradually learned by *searching* in the space of parameters. The parameters are updated according to the immediate reward after executing each action, or according to the cumulated reward at the end of each trial. Most Policy Gradient algorithms for POMDPs, such as GAPS (Peshkin, 2001) or GPOMDP (Baxter & Bartlett, 2000) belong to this family.

Actor-critic algorithms (Sutton et al., 2000; Peters & Schaal, 2006; Aberdeen et al., 2007) combine the advantages of both value-function and policy search methods. The actor part maintains a parametric policy that is gradually improved according to the evaluation provided by the critic part. The critic learns a value function where the variables correspond to the parameters of the policy. In the following section, we describe a general method where the critic part learns unbiased estimates of immediate rewards which are used to calculate the gradient of the value function with respect to the policy parameters. We show how history-action Q-values are eliminated from the expression of the gradient and replaced by immediate rewards. Finally, we present the gradient expression for both cases where the policy is represented by an FSC and where the policy is represented by a PSR.

4. A General Actor-Critic Approach

The actor is a stochastic policy with a vector of parameters θ, the critic is a function that returns the gradient of the value function $V(h_0, \theta)$, which is the total, expected, discounted reward when executing a policy parameterized by θ after a starting history h_0.

$$V(h_0, \theta) = \sum_{a \in \mathcal{A}} Pr(a|h_0, \theta) Q(h_0, a, \theta)$$

where the history-action Q-values are given by:

$$Q(h_t, a, \theta) = R(a|h_t) + \gamma \sum_{o \in \mathcal{O}} \sum_{a' \in \mathcal{A}} Pr(oa'|h_t a, \theta) Q(h_t ao, a', \theta) \quad (9)$$

4.1. Gradient Expression

The following derivation is an adaptation to POMDPs of the proof of the policy gradient theorem for MDPs, proposed by Sutton et al. (2000). We indicate by h_t^a the actions of a history h_t and by h_t^o its observations.

$$\frac{\partial V(h_0, \theta)}{\partial \theta_i} = \frac{\partial}{\partial \theta_i} \sum_{a \in \mathcal{A}} Pr(a|h_0, \theta) Q(h_0, a, \theta)$$

$$= \sum_{a \in \mathcal{A}} \frac{\partial Pr(a|h_0, \theta)}{\partial \theta_i} Q(h_0, a, \theta) + Pr(a|h_0, \theta) \frac{\partial Q(h_0, a, \theta)}{\partial \theta_i}$$

Substituting $Q(h_0, a, \theta)$ with Equation (9), and repeating the substitution for H steps yields to:

$$\frac{\partial V(h_0, \theta)}{\partial \theta_i} = \sum_{t=0}^{H-1} \sum_{h_t \in \{\mathcal{A} \times \mathcal{O}\}^t} \sum_{a \in \mathcal{A}} [\gamma^t Pr(h_t|\theta) Q(h_t, a, \theta) \frac{\partial Pr(a|h_t, \theta)}{\partial \theta_i}] \quad (10)$$

Applying Bayes' Rule $Pr(a|h_t, \theta) = \frac{Pr(h_t^a a|h_t^o, \theta)}{Pr(h_t^a|h_t^o, \theta)}$ gives:

$$\frac{\partial Pr(a|h_t, \theta)}{\partial \theta_i} = \frac{1}{Pr(h_t^a|h_t^o, \theta)} \frac{\partial Pr(h_t^a a|h_t^o, \theta)}{\partial \theta_i} - \frac{\partial Pr(h_t^a|h_t^o, \theta)}{\partial \theta_i} \frac{Pr(a|h_t, \theta)}{Pr(h_t^a|h_t^o, \theta)} \quad (11)$$

After replacing this latter formula within Equation (10), and adequately reordering the terms (details are removed due to the lack of space), we find:

$$\frac{\partial V(h_0, \theta)}{\partial \theta_i} = \sum_{t=0}^{H-1} \sum_{h_t \in \{\mathcal{A} \times \mathcal{O}\}^t} \sum_{a \in \mathcal{A}} \gamma^t \frac{\partial Pr(h_t^a a|h_t^o, \theta)}{\partial \theta_i} \frac{Pr(h_t a|\theta)}{Pr(h_t^a a|h_t^o, \theta)} \underbrace{[Q(h_t, a, \theta) - \gamma \sum_{o \in \mathcal{O}} \sum_{a' \in \mathcal{A}} Pr(oa'|h_t a) Q(h_t ao, a', \theta)]}_{R(a|h_t)}$$

Thus:

$$\frac{\partial V(h_0, \theta)}{\partial \theta_i} = \sum_{t=0}^{H-1} \sum_{h_t \in \{\mathcal{A} \times \mathcal{O}\}^t} \sum_{a \in \mathcal{A}} \gamma^t \frac{\partial Pr(h_t^a a|h_t^o, \theta)}{\partial \theta_i} Pr(h_t^o|h_t^a) R(a|h_t) \quad (12)$$

Notice that the term $Pr(h_t^o|h_t^a) R(a|h_t)$ is independent of θ. It can be learned by a simple Monte Carlo method with Importance Sampling, using a look-up table:

$$\hat{Pr}(h_t^o|h_t^a) = \tfrac{1}{Pr(h_t^a|h_t^o, \theta)} \tfrac{\#h_t}{N}$$

$$\hat{R}(a|h_t) = \tfrac{1}{N} \textstyle\sum_{i=1}^{N} r_{i,t} \delta_{h_t, h_{i,t}} \delta_{a, a_{i,t+1}}$$

where N is the number of trials, $a_{i,t}$ and $r_{i,t}$ are respectively the action and the reward observed at time t in trial i, $h_{i,t}$ is the history at time t in trial i.

Now, we will show how to calculate the other term, $\frac{\partial Pr(h_t^a a|h_t^o, \theta)}{\partial \theta_i}$, depending on the model of the policy.

4.2. Gradient Estimation for FSCs

If we use an FSC to represent the policy, then:

$$Pr(h_t^a|h_t^o, \theta) = b_0^T(., \theta) M_\theta^{o_1 a_1} \dots M_\theta^{o_t a_t} e \quad (13)$$

where

$$\begin{cases} M_\theta^{o_j a_j}(g, g') = \omega^{o_j}(g, g', \theta) \mu^{o_j, a_j}(g', \theta) \\ e^T = (1, 1, \dots, 1) \end{cases}$$

Thus:

$$\frac{\partial Pr(h_t^a|h_t^o, \theta)}{\partial \theta_i} = \sum_{j=1}^{t} \sum_{g \in \mathcal{G}} \sum_{g' \in \mathcal{G}} [\alpha_{j-1}(g, \theta) \beta_{j+1}^t(g', \theta) \frac{\partial(\omega^{o_j}(g, g', \theta) \mu^{o_j, a_j}(g', \theta))}{\partial \theta_i}] + \sum_{g \in \mathcal{G}} \beta_0^t(g, \theta) \frac{\partial b_0(g, \theta)}{\partial \theta_i}$$

where

$$\begin{cases} \alpha_0(g, \theta) = b_0(h, \theta) \\ \alpha_i(g, \theta) = b_0^T(., \theta) M_\theta^{o_1 a_1} \dots M_\theta^{o_i a_i}(., g) \\ \beta_i^t(g, \theta) = M_\theta^{o_i a_i}(g, .) M_\theta^{o_{i+1} a_{i+1}} \dots M_\theta^{o_t a_t} e \\ \beta_{t+1}^t(g, \theta) = 1 \end{cases}$$

This latter expression of the gradient corresponds to the same expression proposed in (Shelton, 2001).

4.3. Gradient Estimation for PSR Policies

If we use a PSR to represent the policy, then:

$$Pr(h_t^a|h_t^o, \theta) = b_0^T(., \theta) M_\theta^{o_1 a_1} \dots M_\theta^{o_t a_t} e \quad (14)$$

where

$$\begin{cases} M_\theta^{o_j a_j}(h, h') = m^{o_j a_j}(h, \theta) b_{h o_j a_j}(h', \theta) \\ e^T = (1, 1, \dots, 1) \end{cases}$$

Therefore:

$$\frac{\partial Pr(h_t^a|h_t^o,\theta)}{\partial \theta_i} = \sum_{j=1}^{t}\sum_{h\in\mathcal{H}}\sum_{h'\in\mathcal{H}} [\alpha_{j-1}(h,\theta)\beta_{j+1}^t(h',\theta) \frac{\partial(m^{o_j a_j}(h,\theta) b_{h o_j a_j}(h',\theta))}{\partial\theta_i}]$$

where

$$\begin{cases} \alpha_0(h,\theta) = b_0(h,\theta) \\ \alpha_i(h,\theta) = b_0^T(.,\theta)M_\theta^{o_1a_1}\dots M_\theta^{o_ia_i}(.,h) \\ \beta_i^t(h,\theta) = M_\theta^{o_ia_i}(h,.)M_\theta^{o_{i+1}a_{i+1}}\dots M_\theta^{o_ta_t}e \\ \beta_{t+1}^t(h,\theta) = 1 \end{cases}$$

4.4. Updating the Parameters of the Policy

The gradient calculated by the critic (Equation 12) is used to update the parameters θ_i:

$$\theta_i \leftarrow \theta_i + \eta\frac{\partial V(h_0,\theta)}{\partial\theta_i} \qquad (15)$$

where $\eta \in [0,1]$ is the gradient step size.

Policy gradient methods are proved to converge to a local optimum. The probability of reaching a global optimum depends on the complexity (or the shape) of the value function. In the following section, we show that the value function is a polynomial, with a smaller degree when the policy is represented by a PSR.

5. Complexity of The Value Function

The functions ω^o and $\mu^{o,a}$ return distributions of probabilities, they are usually represented as softmax functions. The analysis of the value function is more difficult with this representation, since it involves exponentials and fractions. For the purpose of comparing to PSRs, and without loss of generality, we consider each $\omega^o(g,g')$ and each $\mu^{o,a}(g)$ as a different parameter, so $\omega^o(g,g',\theta) = \theta_{g,o,g'}$ and $\mu^{o,a}(g,\theta) = \theta_{o,g,a}$, we also consider $b_0(g,\theta) = \theta_g$. For PSRs, the vectors b_{hoa} are not stochastic, therefore each $b_{hoa}(h')$ is a different parameter, so $b_{hoa}(h',\theta) = \theta_{hoa,h'}$. Similarly, we put $m^{oa}(h,\theta) = \theta_{oa,h}$.

Notice now that the function returning $Pr(h_t^a|h_t^o,\theta)$ (Equations 13 and 14) is a multivariate polynomial of degree less than or equal to $2t+1$. The variables of this polynomial correspond to the parameters $\theta_{ho_ia_i,h'}, \theta_{o_ja_j,h}$ for PSR policy, and to $\theta_{g,o_j,g'}, \theta_{o_j,g,a_j}, \theta_g$ for FSC.

Unless the structure of the FSC is provided *a priori*, the graph of the FSC is generally completely connected, i.e. $\forall a,o,g,g' : \omega^o(g,g',\theta)\mu^{o,a}(g',\theta) > 0$. Thus, we have $\alpha_j(g,\theta) > 0$ and $\beta_j^t(g',\theta) > 0$ for every I-state g and step $i \leq t$, the degree of the polynomial $Pr(h_t^a|h_t^o,\theta)$ for the FSC is then equal to $2t+1$.

To reduce the degree of the value function, Aberdeen and Baxter (2002) proposed reducing the outdegree of the FSC graph and augmenting the number of I-states. However, the convergence can be delayed in that case, given that more parameters should be learned.

The main advantage of PSRs comes from the fact that, contrary to I-states, the core histories are contained within the sequence of actions and observations, and no transition probabilities are used to calculate the probability of a core history sequence. Namely, when a prefix sequence $o_1a_1\dots o_ia_i$ of the history $h_t = o_1a_1\dots o_ia_i\dots o_ta_t$ corresponds to a core history h_i, then the term $b_0^T(.,\theta)M_\theta^{o_1a_1}\dots M_\theta^{o_ia_i}$ in Equation (14), can be replaced by a vector $\alpha_i(.,\theta)$ where $\alpha_i(h_i,\theta) = Pr(h_i^a|h_i^o,\theta)$ and $\alpha_i(h,\theta) = 0$ for $h \neq h_i$ (α_i is the "unnormalized" belief at time i). We have:

$$Pr(h_i^a|h_i^o,\theta) = Pr(a_1|h_0,o_1,\theta)\dots Pr(a_i|h_{i-1},o_i,\theta)$$

From the construction of PSRs (Littman et al., 2002), we know that all the prefixes of a core history are also core histories, therefore:

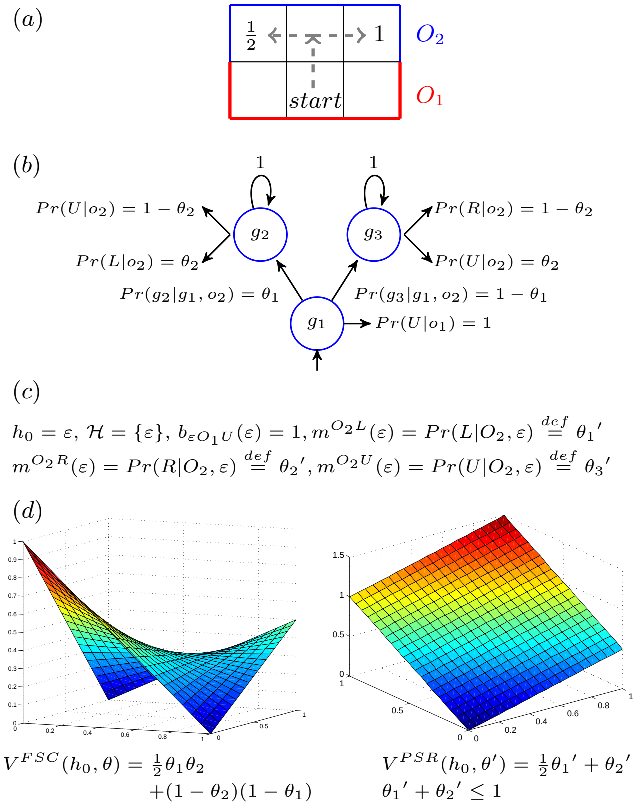
$$Pr(h_i^a|h_i^o,\theta) = \theta_{o_1a_1,h_0}\theta_{o_2a_2,h_1}\dots\theta_{o_ia_i,h_{i-1}}$$

This latter term is a polynomial of degree i. Hence, the degree of the polynomial $Pr(h_t^a|h_t^o,\theta)$ (Equation 14) is at most equal to $2t-i$.

Consequently, the degree of the value function polynomial is generally smaller when the policy is represented by a PSR than when it is represented by an FSC, and the number of local optima is also smaller, as we will see in the following example.

5.1. Example

Figure 1 shows a simple toy problem with 3 deterministic actions: U (*Up*), L (*Left*) and R (*Right*), and two deterministic observations o_1 (*lower states*) and o_2 (*upper states*) (Figure 1.a). There are also two final states marked with positive rewards. Figure 1.b presents a parametric FSC with three states, g_1 is the initial state, g_2 and g_3 are final states. We focus on only two parameters, since representing functions with more than 2 variables graphically is not feasible. The first parameter θ_1 corresponds to the probability of going from state g_1 to state g_2 after observing o_2, $1-\theta_1$ is the probability of going to g_3. The second parameter θ_2 corresponds to the probability of executing



$h_0 = \varepsilon, \mathcal{H} = \{\varepsilon\}, b_{\varepsilon O_1 U}(\varepsilon) = 1, m^{O_2 L}(\varepsilon) = Pr(L|O_2, \varepsilon) \overset{def}{=} {\theta_1}'$
$m^{O_2 R}(\varepsilon) = Pr(R|O_2, \varepsilon) \overset{def}{=} {\theta_2}', m^{O_2 U}(\varepsilon) = Pr(U|O_2, \varepsilon) \overset{def}{=} {\theta_3}'$

$V^{FSC}(h_0, \theta) = \frac{1}{2}\theta_1\theta_2 + (1 - \theta_2)(1 - \theta_1)$

$V^{PSR}(h_0, \theta') = \frac{1}{2}{\theta_1}' + {\theta_2}'$
${\theta_1}' + {\theta_2}' \leq 1$

Figure 1. The value functions of an FSC (polynomial of degree 2) and its equivalent PSR (polynomial of degree 1).

action L in state g_2 after observing o_2, and $1 - \theta_2$ is the probability of executing action R in g_3. The value function of this FSC for horizon 2 is given by $V^{FSC}(h_0, \theta) = \frac{1}{2}\theta_1\theta_2 + (1 - \theta_1)(1 - \theta_2)$, it has one global maximum of value 1 and one local maximum of value $\frac{1}{2}$ (Figure 1.d). The equivalent PSR policy (Figure 1.c) can be represented by one core history: ε, and three parameters: ${\theta_1}' = m^{O_2 L}(\varepsilon)$ and ${\theta_2}' = m^{O_2 R}(\varepsilon)$ and ${\theta_3}' = m^{O_2 U}(\varepsilon)$. The value function of this policy is given by $V^{PSR}(h_0, \theta') = \frac{1}{2}{\theta_1}' + {\theta_2}'$, it has only a global maximum (Figure 1.d) within the simplex of valid parameters, i.e. $\theta'_1 + \theta'_2 \leq 1$. This gain is due to the fact that the history $h_1 = O_1 U$ is equivalent to ϵ (all the histories are equivalent to ϵ when $\mathcal{H} = \{\epsilon\}$), thus $b_{h_1}(\varepsilon) = 1$ (degree 0) and the probability of action L after h_1 and O_2, for example, is $Pr(L|h_1, O_2) = {\theta_1}'$ (degree 1). While for the FSC, we have $b_{h_1}(g_2) = \theta_1$, $b_{h_1}(g_3) = 1 - \theta_1$ and $Pr(L|h_1, O_2) = \theta_1\theta_2$ (degree 2).

6. Constraining PSR Belief States

The belief state of a PSR is not a probability distribution, it is rather subject to a set of constraints that should be satisfied at any time t:

$C_t^1 : \forall o \in \mathcal{O}, \forall a \in \mathcal{A} : \sum_{h \in \mathcal{H}} b_t(h, \theta) m^{oa}(h, \theta) \geq 0;$
$C_t^2 : \forall o \in \mathcal{O} : \sum_{a \in \mathcal{A}} \sum_{h \in \mathcal{H}} b_t(h, \theta) m^{oa}(h, \theta) = 1;$

The number of reachable belief states b_t is exponential w.r.t. the horizon H, while only the vectors m^{oa} and b_{hoa} appear in the parameters of the policy. Thus, we should define another set of constraints on m^{oa} and b_{hoa} to ensure that $\{C_t^1, C_t^2\}$ will be satisfied after an arbitrary number of bayesian updates. The existence of such constraints depends on the convexity of the bayesian update function inside the convex hull of valid belief states, and up to our knowledge, this problem has not been studied before in the context of PSRs.

Nevertheless, we will define a set of weaker constraints on m^{oa} and b_{hoa} that can keep the beliefs b_t within the hull of valid weights for short horizons:

$C_0^1 : \forall o \in \mathcal{O}, \forall a \in \mathcal{A}, \forall h \in \mathcal{H} : m^{oa}(h, \theta) \geq 0;$
$C_0^2 : \forall o \in \mathcal{O}, \forall h \in \mathcal{H} : \sum_{a \in \mathcal{A}} m^{oa}(h, \theta) = 1;$
$C_0^3 : \forall o \in \mathcal{O}, \forall a \in \mathcal{A}, \forall h \in \mathcal{H} : \sum_{h' \in \mathcal{H}} b_{hoa}(h', \theta) = 1;$

Constraints C_0^1 and C_0^2 can be satisfied by using a Boltzmann distribution for each core history h and each observation o. Constraint C_0^3 is justified by the fact that: $\forall t \in \{0, \ldots, H\} : \sum_{h \in \mathcal{H}} b_t(h, \theta) = 1$. Indeed, for any observation o, we have:

$$\begin{aligned}\textstyle\sum_{h \in \mathcal{H}} b_t(h, \theta) &= \textstyle\sum_{h \in \mathcal{H}} b_t(h, \theta) \sum_{a \in \mathcal{A}} m^{oa}(h, \theta)\\ &= \textstyle\sum_{a \in \mathcal{A}} \sum_{h \in \mathcal{H}} b_t(h, \theta) m^{oa}(h, \theta)\\ &= \textstyle\sum_{a \in \mathcal{A}} Pr(a|h_t, o, \theta) = 1\end{aligned}$$

To satisfy C_0^3, we project the gradient $\frac{\partial V(h_0, \theta)}{\partial \theta_{hoa,h'}}$ onto the hyperplane defined by $C_{3,0}$ by solving a system of linear equations, the projected values are given by:

$$\frac{\partial V(h_0, \theta)}{\partial \theta_{hoa,h'}} \leftarrow \frac{\partial V(h_0, \theta)}{2\partial \theta_{hoa,h'}} - \frac{1}{2|\mathcal{H}|} \sum_{h'' \in \mathcal{H}} \frac{\partial V(h_0, \theta)}{\partial \theta_{hoa,h''}} \quad (16)$$

7. Discovery of Core Histories

Initially, the set $\mathcal{H}$ contains only the empty history ε, the vectors m^{oa} are initialized to uniform distributions over actions a for each observation o, and all the histories are considered as equivalent to ε, i.e. $\theta_{\varepsilon oa, \varepsilon} \overset{def}{=} b_{\varepsilon oa}(\varepsilon, \theta) = 1$ (because of the constraint C_0^3). As the parameters θ are updated, new core histories are iteratively discovered and added to $\mathcal{H}$. A history hoa (which is an extension of a core history $h \in \mathcal{H}$) is considered as a core history, if there is at least a test q such that $Pr(q^a|hoa, q^o)$ cannot be written as a linear combination of the probabilities $Pr(q^a|h', q^o), h' \in \mathcal{H}$. Since this criteria cannot be verified for every possible test q, we will rather use a set of heuristic measures as an indicator of the predictability of the history hoa. The first one is based on the distance d_{hoa} between the vector θ_{hoa}, which is updated by using the gra-

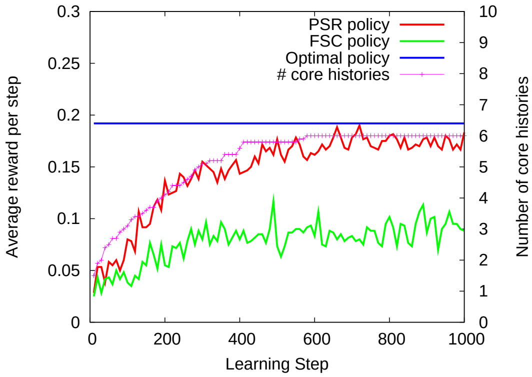


Figure 2. 4x4 maze results.

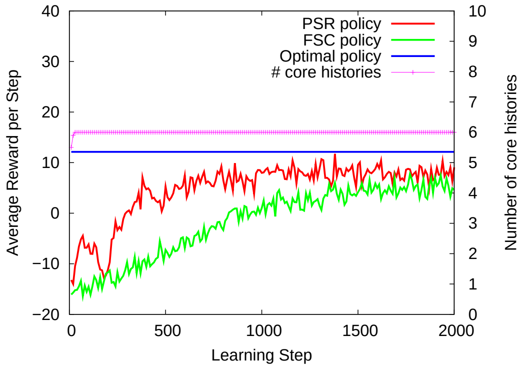


Figure 3. Network results.

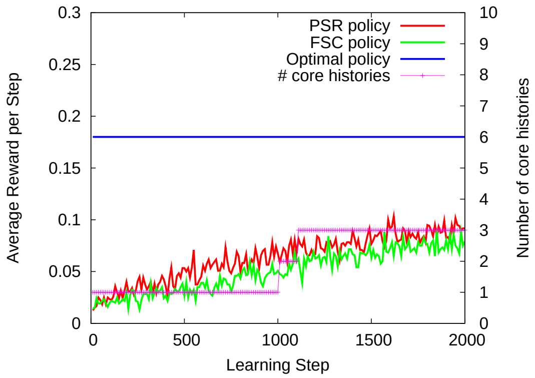


Figure 4. Cheese maze results.

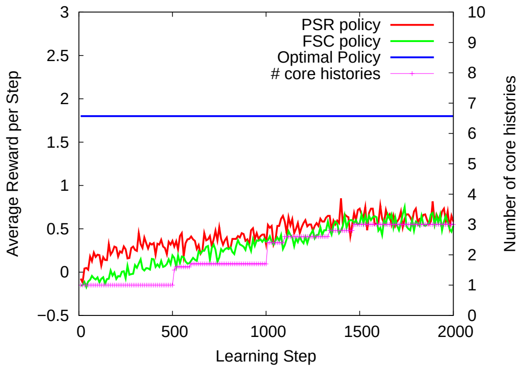


Figure 5. Shuttle results.

dient calculated in Equation (12), and the corrected vector θ'_{hoa}, which is updated by using the projected gradient of Equation (16) (i.e. d_{hoa} is the correction made on θ_{hoa} to make it satisfy the constraint C_0^3),

$$d^2_{hoa} = \sum_{h' \in \mathcal{H}} (\theta_{hoa,h'} - \theta'_{hoa,h'})^2$$

The second measure is the entropy e_{ho} of the distribution over actions a conditioned on h and o,

$$e_{ho} = -\sum_{a \in \mathcal{A}} Pr(a|h,o) \ln Pr(a|h,o)$$

Finally, a history hoa cannot be added to $\mathcal{H}$ if it has not been tested enough. Thus, our third measure is $\hat{Pr}(hoa|\varepsilon) = \frac{\#hoa}{N}$, where N is the number of trials.

A history hoa is considered as a new core history *iff*:

$$(d_{hoa} \geq \epsilon_d) \wedge (e_{ho} \geq \epsilon_e) \wedge (\hat{Pr}(hoa|\varepsilon) \geq \epsilon_p)$$

where ϵ_d,ϵ_e and ϵ_p are predefined thresholds. The probabilities $m^{oa}(h', \theta)$ of a new core history h' are initialized to a uniform distribution, whereas all the extensions of h' are considered as equivalent to h', i.e. $b_{h'o'a'}(h', \theta) = 1$ and $b_{h'o'a'}(h'', \theta) = 0$ for $h'' \neq h'$.

8. Experiments

We test the policy gradient approach described in Section 4 on small standard problems taken from A. R. Cassandra's POMDP data repertoire (on the web), using PSRs and FSCs as models of the policies. We use softmax functions to represent the distributions over actions in both PSRs (with the vectors m^{oa}) and FSCs (with the functions μ^{o_j,a_j}), and to represent the distributions over the next I-states in FSCs (with the functions ω^{o_j}). We use the same temperature τ for these functions, τ is initialized to 0.1 for 4×4 maze problem, to 1 for Cheese maze and Shuttle problems, and to 10 for Network problem. For all these problems, τ is decreasing by a constant factor of 0.999.

The number of I-states $|\mathcal{G}|$ is the only hyper-parameter for FSCs, it is chosen by a cross-validation for each problem. We found that the best value of $|\mathcal{G}|$ is 6 for 4×4 maze, 3 for Cheese maze and Shuttle problems, and 8 for Network problem. The parameters of the transition functions ω^{o_j}, used by the softmax function, are randomly initialized to values between 0 and 1. The parameters of the action-selection functions μ^{o_j,a_j} are initialized to 0 (a uniform distribution). The functions μ^{o_j,a_j} and ω^{o_j} cannot be all initialized to uniform distributions, this results in uniform beliefs and a zero gradient (Aberdeen & Baxter, 2002). The hyper-parameters of PSR policies are the thresholds ϵ_p,ϵ_e, and ϵ_d (see Section 7). ϵ_p is set to 0.1 for all the problems, while ϵ_d is set to 0.1 for 4×4 maze, to 6 for Network, and to 0 for Cheese maze and Shuttle. ϵ_e is set to 0 for 4×4 maze and Network, and to 1.5 for Cheese maze and Shuttle. Notice that these thresholds are used to control the size of $\mathcal{H}$: using smaller thresholds leads to more core histories and vice versa. They are found by a cross-validation, as for $|\mathcal{G}|$ in FSCs. Finally, we use the same constant gradient step η for both PSRs and FSCs, η is set to 1 for 4×4 and Cheese mazes (problems with small rewards), and to 0.1 for Network and Shuttle (problems with high rewards).

Figures 2-5 show the average reward per step of FSC and PSR, and the average number of discovered core histories per step for PSR. The results are averaged over 10 independent runs. For all these problems, the Policy Gradient algorithm converged to locally optimal values for both models of policies. However, we notice that the final value is slightly higher when we use a PSR to represent the policy, which is expected from the theoretical analysis of Section 5. The outperformance of PSR policies is more pronounced in problems with a smaller number of histories ($|\mathcal{A}||\mathcal{O}| = 8$ in 4×4 maze and Network, 15 in Shuttle, and 28 in Cheese maze). Indeed, discovering the accurate set of core histories becomes more difficult when the branching factor $|\mathcal{A}||\mathcal{O}|$ is higher. Finally, we notice that the learning algorithm for PSRs converged rapidly in Network problem, this is explained by the fact that all the core histories were discovered within the first 30 trials.

9. Discussion

We have shown that PSRs are alternative to Finite-State Controllers in policy gradient methods for POMDPs. Internal states of PSR policies are based on observable sequences of interacting with the environment, called core histories. Two main advantages result from this property. The first one is related to finite-horizon problems, where the degree of the value function polynomial is reduced by the length of the longest core history observed in a given trial. The second advantage is the possibility of discovering new core histories, based on the predictability of the extensions of previous core histories. We used different heuristics, based on the entropy of the actions distribution and the correction of the gradient, as an indicator of the predictability, but more sophisticated techniques should be investigated (Makino & Takagi, 2008).

However, it is unclear how PSR policies will perform in infinite-horizon problems where the value function is defined only on the stationary belief state and the core histories cannot be observed. This problem can be treated by using a mechanism for detecting *reset points*. Particularly, one can combine core histories and core tests, and use the first ones to detect the reset points for the second. The other issue related to infinite-horizons is the stability of the belief states. We defined some constraints that can keep the belief within the hull of valid parameters for short horizons, but the general solution remains an open problem.

References

Aberdeen, D., & Baxter, J. (2002). Scaling Internal-State Policy-Gradient Methods for POMDPs. *Proc. 19th Int. Conf. Machine Learning* (pp. 3–10).

Aberdeen, D., Buffet, O., & Thomas, O. (2007). Policy-Gradients for PSRs and POMDPs. *Proc. 11th Int. Conf. Artificial Intelligence and Statistics.*

Baxter, J., & Bartlett, P. (2000). Reinforcement Learning in POMDP's via Direct Gradient Ascent. *Proc. 17th Int. Conf. Machine Learning* (pp. 41–48).

Casella, G., & Robert, C. P. (1996). Rao-blackwellisation of Sampling Schemes. *Biometrika*, *15*, 229–235.

Littman, M., Sutton, R., & Singh, S. (2002). Predictive Representations of State. *Advances in Neural Information Processing Systems 14* (pp. 1555–1561).

Makino, T., & Takagi, T. (2008). On-line Discovery of Temporal-Difference Networks. *Proc. 25th Int. Conf. Machine Learning* (pp. 632–639).

Meuleau, N., Peshkin, L., Kim, K., & Kaelbling, L. (1999). Learning Finite-State Controllers for Partially Observable Environments. *Uncertainty in Artificial Intelligence: Proc. 15th Conf.* (pp. 427–436).

Peshkin, L. (2001). *Reinforcement Learning by Policy Search.* Doctoral dissertation, Massachusetts Institute of Technology.

Peters, J., & Schaal, S. (2006). Policy Gradient Methods for Robotics. *Proc. IEEE Int. Conf. Intelligent Robotics Systems* (pp. 2219–2225).

Shelton, C. R. (2001). *Importance Sampling for Reinforcement Learning with Multiple Objectives.* Doctoral dissertation, Massachusetts Institute of Technology.

Sutton, R. S., Mcallester, D., Singh, S., & Mansour, Y. (2000). Policy Gradient Methods for Reinforcement Learning with Function Approximation. *Advances in Neural Information Processing Systems 12* (pp. 1057–1063).

Wiewiora, E. (2005). Learning Predictive Representations from a History. *Proc. 22nd Int. Conf. Machine Learning* (pp. 964–971).