Predictive Representations for Policy Gradient in POMDPs

Abdeslam Boularias and Brahim Chaib-draa

Laval University, Canada

{boularias;chaib}@damas.ift.ulaval.ca

DAMAS
Décision, Adaptation, Multi-AgentS
www.damas.ift.ulaval.ca

UNIVERSITÉ LAVAL

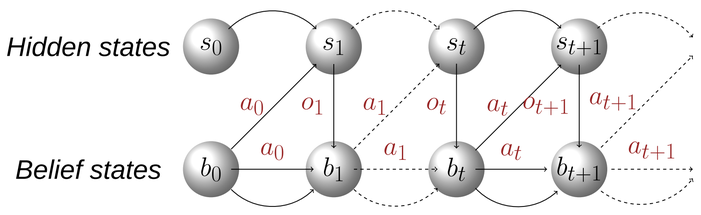
1. Overview

We propose to use Predictive State Representations (PSRs) to represent parametric stochastic policies in Partially Observable Markov Decision Processes (POMDPs). We provide a simple Gradient algorithm for learning the parameters of a PSR policy. Interestingly, we show that the value function of a PSR policy can have less local optima compared to the equivalent Finite State Controller (FSC).

2. Partially Observable Markov Decision Processes

A POMDP is a 8-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{O}, \{T^a\}, \{O^{a,o}\}, r, \gamma, H)$, where:

- $\mathcal{S}$ is a set of states, $\mathcal{A}$ is a set of actions, and $\mathcal{O}$ is a set of observations.
- $\{T^a\}$ is a set of transition functions, where $T^a(s, s')$ is the probability that the agent will end up in state s' after taking action a in state s.
- $\{O^{a,o}\}$ is a set of observation functions, where $O^{a,o}(s)$ gives the probability that the agent receives observation o after taking action a and getting to state s'.
- r is a reward function, such that $r(s, a)$ is the immediate reward received when the agent executes action a in state s.
- $\gamma \in [0, 1]$ is a discount factor.
- H is the planning horizon.



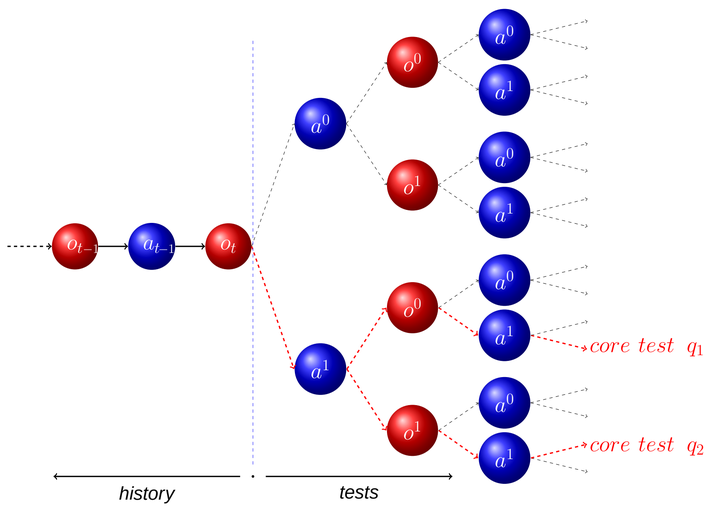
- A history h_t is a sequence of past actions and observations.

$$h_t = a_0 o_1 a_1 o_2 a_2 o_3 \dots a_{t-1} o_t$$

- The belief state is a vector b_t where $b_t(s) = Pr(s_t = s|h_t), s \in \mathcal{S}$.

3. Predictive State Representations [Littman et al., 2002]

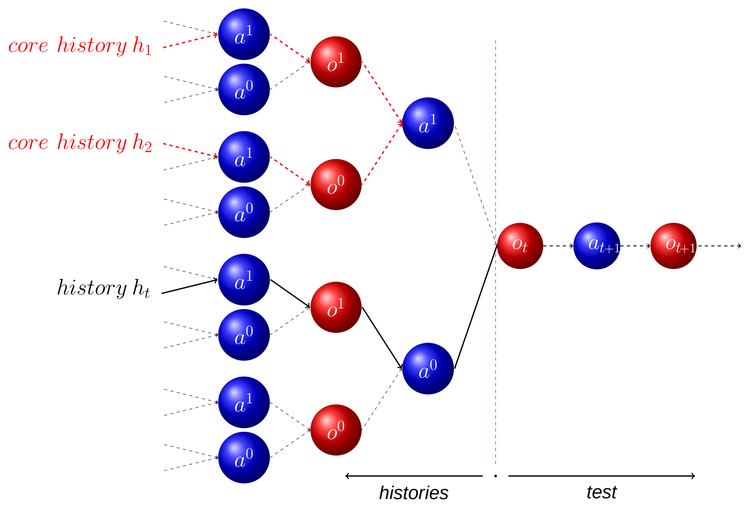
- The probabilities on states are replaced by probabilities on particular future trajectories, called *core tests*.
- A test q is a sequence of actions and observations: $q = a^1 o^1 \dots a^k o^k$
- The probability of a test q starting after a history h_t is defined by: $Pr(q^o|h_t, q^a) = Pr(o_{t+1} = o^1, \dots, o_{t+k} = o^k | h_t, a_{t+1} = a^1, \dots, a_{t+k} = a^k)$
- The probability of any test q is a linear combination of the probabilities of core tests. The belief state b_t is a vector containing the probabilities of core tests.



$$\forall q \in \{\mathcal{A} \times \mathcal{O}\}^* : Pr(q|h_t) = \alpha Pr(q_1|h_t) + \beta Pr(q_2|h_t)$$

4. Predictive State Representations with Core Histories

- The probability of any test q after a history h_t depends on the probabilities of the same test after different core histories.
- We use $\mathcal{H}$ to indicate the set of core histories.
- The PSR belief state 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: $Pr(q^o|h_t, q^a) = \sum_{h \in \mathcal{H}} b_t(h) Pr(q^o|h, q^a) = b_t^T m^q$, where q^a are the actions of test q, and q^o are its observations.



$$\forall q \in \{\mathcal{A} \times \mathcal{O}\}^* : Pr(q|h_t) = \alpha_t Pr(q|h_1) + \beta_t Pr(q|h_2)$$

5. Finite State Controllers (FSCs)

- FSCs map histories (or belief states) into decision equivalence classes called *internal states* (I-states).
- A finite-number of internal states can remember an infinite number of past events.

A Finite State Controller is defined by:

- $\mathcal{G}$: a finite set of internal states (*I-states*).
- $\{\mu^{o,a}\}$: a set of action-selection functions, where $\mu^{o,a}(g, \theta)$ is the probability that the agent will execute the action a if its I-state is g, its last observation was o.
- $\{\omega^o\}$: a set of transition functions, 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.

6. Finite State Controllers with Internal Belief States

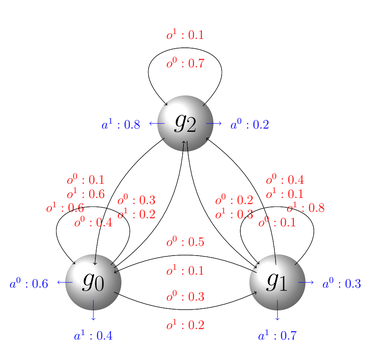
A trajectory (history) h_t is a sequence of actions, observations and sampled internal states. However:

✗ The cumulated reward of a given history depends only on the executed actions and the perceived observations.

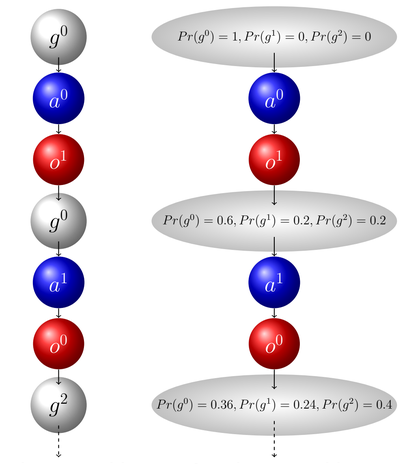
✗ The same trajectory h_t may have been generated by different internal states, with different probabilities.

✗ The estimator of the performance gradient has an unnecessary high variance.

✓ Shelton (2001), Aberdeen and Baxter (2002) proposed to reduce the variance of the gradient estimator by calculating an **internal belief state** at each step instead of sampling a single I-state.



A Finite State Controller



A trajectory with internal states *A trajectory with internal belief states*

7. PSR based Policies

- A stochastic policy is defined as a PSR where the role of actions and observations are switched [Wiewiora, 2005].
- A test q is redefined as a sequence of observations and actions couples, i.e. $q = o^1 a^1 \dots o^k a^k$.
- The probability of q starting after h_t is redefined as: $Pr(q^a|h_t, q^o, \theta) = Pr(a_{t+1} = a^1, \dots, a_{t+k} = a^k | h_t, o_{t+1} = o^1, \dots, o_{t+k} = o^k, \theta)$
- 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 $h \in \mathcal{H}$.
- In particular, the probability of executing action a at time t after observing o is given by: $Pr(a|h_t, o) = \sum_{h \in \mathcal{H}} b_t(h, \theta) Pr(a|h, o, \theta) = b_t^T m^{oa}$

8. General Approach

The gradient of both PSR and FSC value functions is given by:

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

where h_0 is the initial empty history, h^a denotes the actions of a history h, and h^o denotes its observations.

9. Gradient Estimation for FSCs [Aberdeen, 2003]

If we use a Finite State Controller to represent the policy, then:

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

Where $W^{o_j a_j}(g, g', \theta) = \omega^{o_j}(g, g', \theta) \mu^{o_j, a_j}(g', \theta)$ and $e^T = (1, 1, \dots, 1)$. Unless the structure of the FSC is provided *a priori*, the graph of the FSC is generally completely connected, i.e. $\forall a, o, g, g' : \omega^o(g, g', \theta) \mu^{o,a}(g', \theta) > 0$. Therefore, the gradient $\frac{\partial Pr(h_t^a | h_t^o, \theta)}{\partial \theta_i}$ is a multivariate polynomial of degree $2t$.

10. Gradient Estimation for PSR Policies

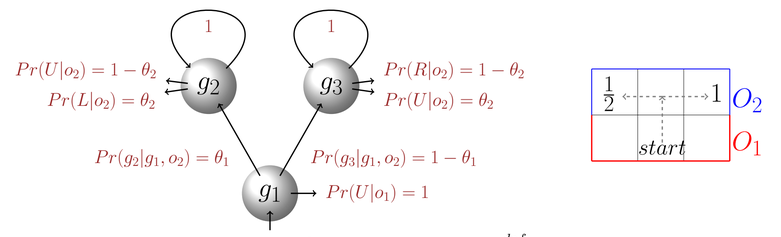
If we use a Predictive Representation of the policy, then:

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

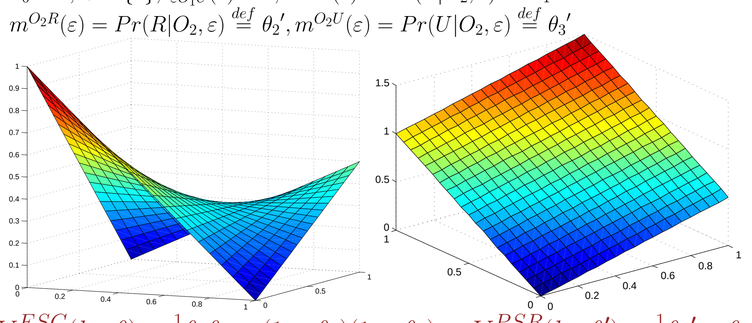
Where $M^{o_j a_j}(h, h', \theta) = m^{o_j a_j}(h, \theta) b_{h o_j a_j}(h', \theta)$.

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 $o_1 a_1 \dots o_i a_i$ of the history h_t corresponds to a core history, then the gradient $\frac{\partial Pr(h_t^a | h_t^o, \theta)}{\partial \theta_i}$ is a multivariate polynomial of degree less than $2t - i$.

11. Small example



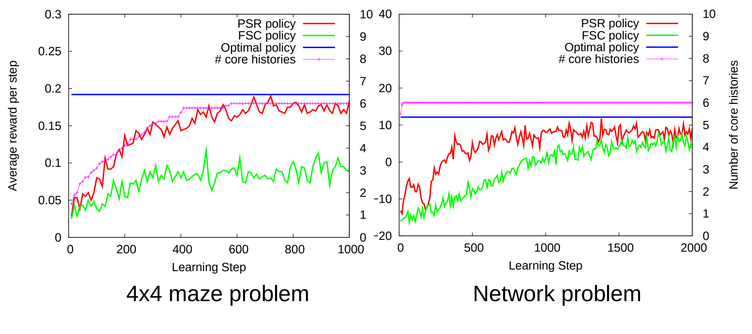
$h_0 = \varepsilon, \mathcal{H} = \{\varepsilon\}, b_{\varepsilon O_1 U}(\varepsilon) = 1, m^{O_2 L}(\varepsilon) = Pr(L|O_2, \varepsilon) \stackrel{def}{=} \theta_1'$
$m^{O_2 R}(\varepsilon) = Pr(R|O_2, \varepsilon) \stackrel{def}{=} \theta_2', m^{O_2 U}(\varepsilon) = Pr(U|O_2, \varepsilon) \stackrel{def}{=}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



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

In this example, the value function of an FSC has two local maxima, while its equivalent PSR policy has only one maxima.

12. Empirical Results



4x4 maze problem Network problem

The results are averaged over 10 independent runs.

13. Conclusion

✓ PSRs are alternative to Finite State Controllers in policy gradient methods for POMDPs.

✓ Internal beliefs of PSRs are based on observable sequences.

✓ The degree of the value function of a PSR policy is reduced by at least the length of the shortest core history.

✗ The discovery of new core histories is based on heuristics.

✗ The belief states of a PSR policy are unstable.

As a future work, we mainly target to study the performance of PSR policies using the natural gradient.

References

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

[Aberdeen, 2003] Aberdeen, D. (2003). Policy-Gradient Algorithms for Partially Observable Markov Decision Processes. Doctoral dissertation, The Australian National University.

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