Predictive Representations for Policy Gradient in POMDPs

Abdeslam Boularias and Brahim Chaib-draa

Department of Computer Science and Software Engineering,
Laval University, Quebec, Canada

June 17^{th}, ICML 2009

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

UNIVERSITÉ LAVAL

Overview

Motivation

- Many real-world planning problems can only be casted as POMDPs.
- Learning an optimal policy in a partially observable environment is still considered as one of the most difficult challenges in Reinforcement Learning (RL).

Overview

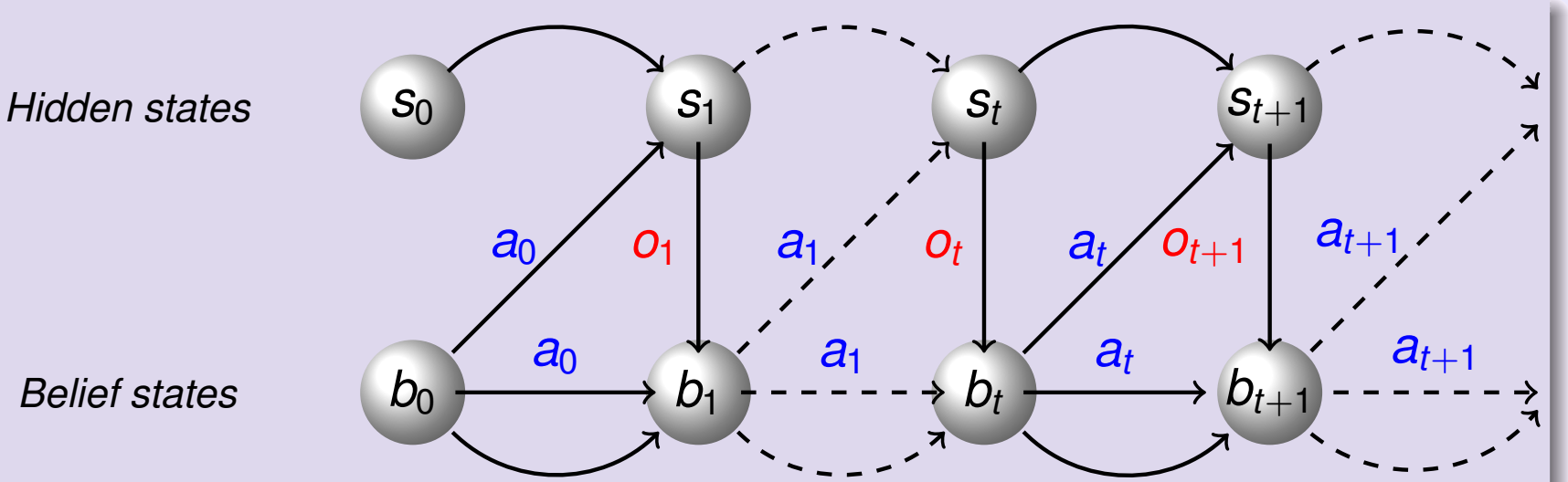
Motivation

- Many real-world planning problems can only be casted as POMDPs.
- Learning an optimal policy in a partially observable environment is still considered as one of the most difficult challenges in Reinforcement Learning (RL).

Contributions

- ✓ A stochastic gradient algorithm for POMDPs, where the control policy is a PSR.
- ✓ A theoretical comparison of the value functions of PSR policies and Finite State Controllers (FSCs).
- ✓ An empirical comparison of PSR policies and FSCs using the same gradient algorithm.

Partially Observable Markov Decision Processes (POMDPs)



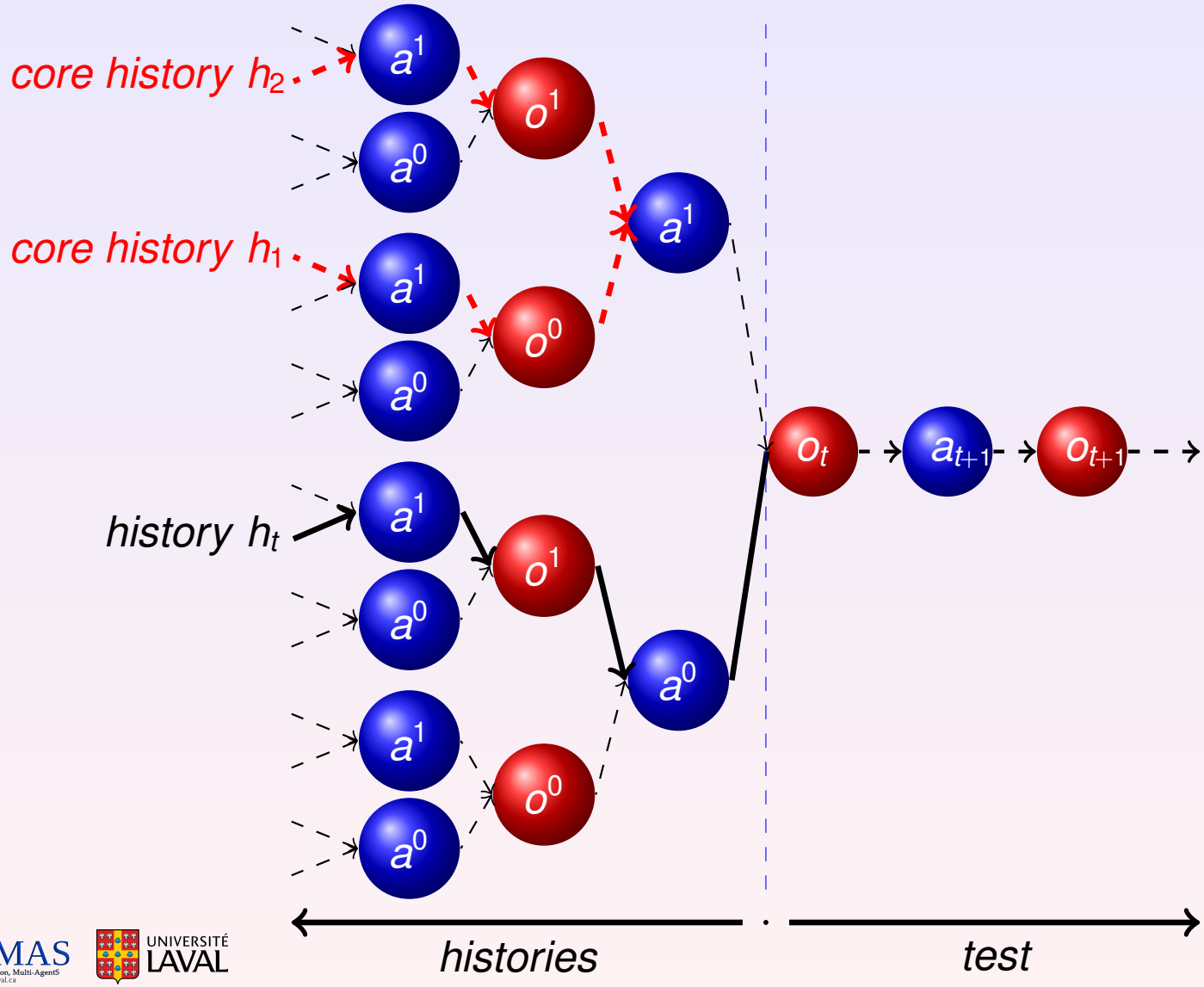
- A history h_t is a sequence of past actions and observations.

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

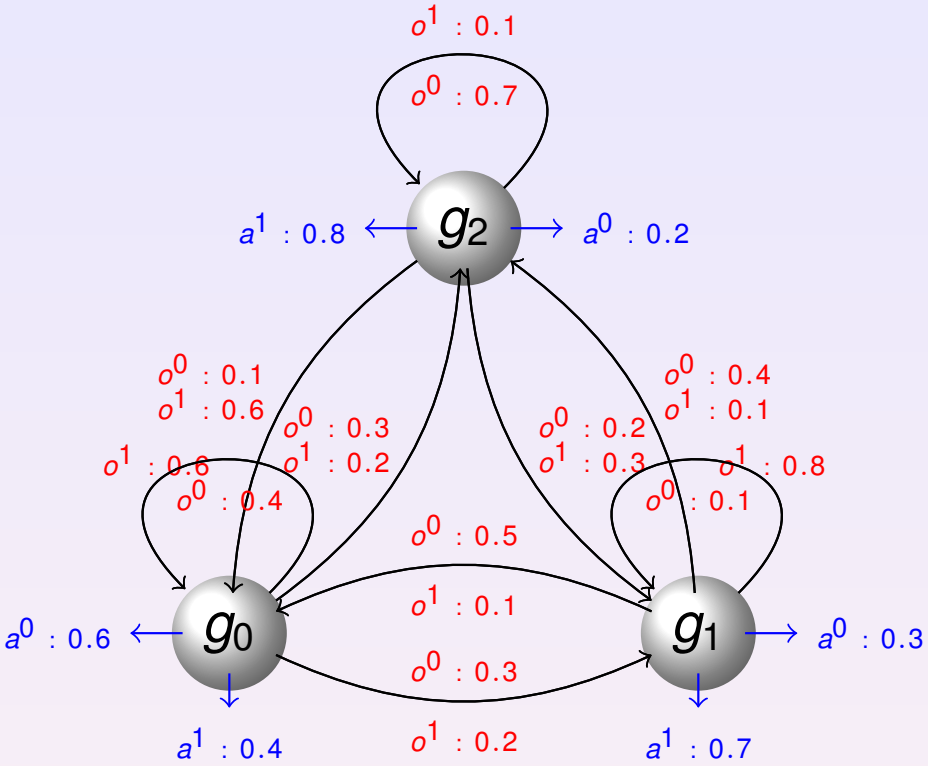
- The belief state is a vector b_t where $b_t(s) = Pr(s_t = s|h_t), s \in \mathcal{S}$.

Predictive State Representations (PSRs) - core histories

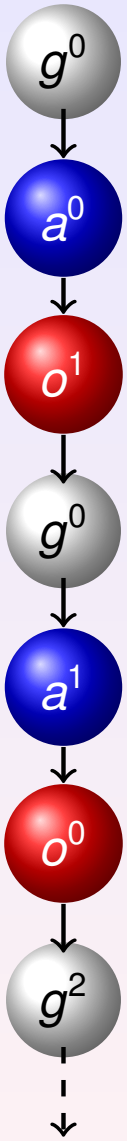
$$\forall q \in \{\mathcal{A} \times \mathcal{O}\}^* : Pr(q|h_t) = \alpha_t Pr(q|h_1) + \beta_t Pr(q|h_2)$$



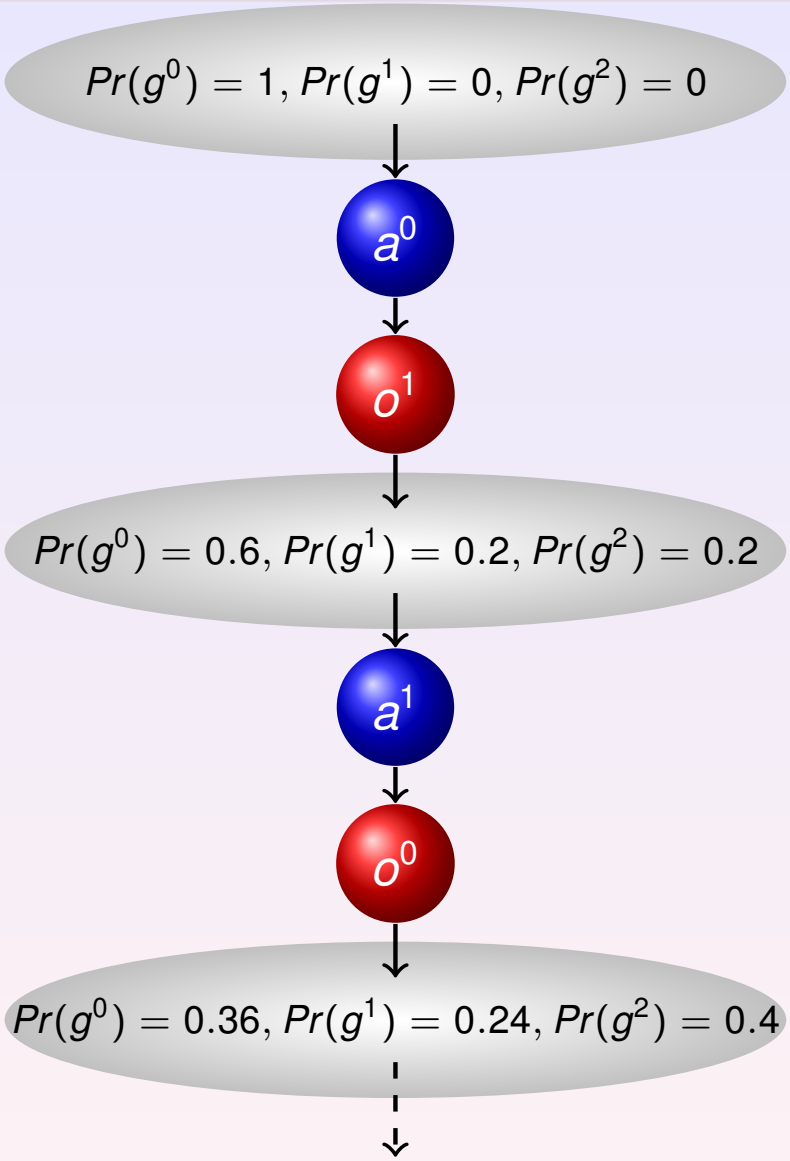
Finite State Controllers with Internal Belief States [Aberdeen & Baxter, 2002]



A Finite State Controller



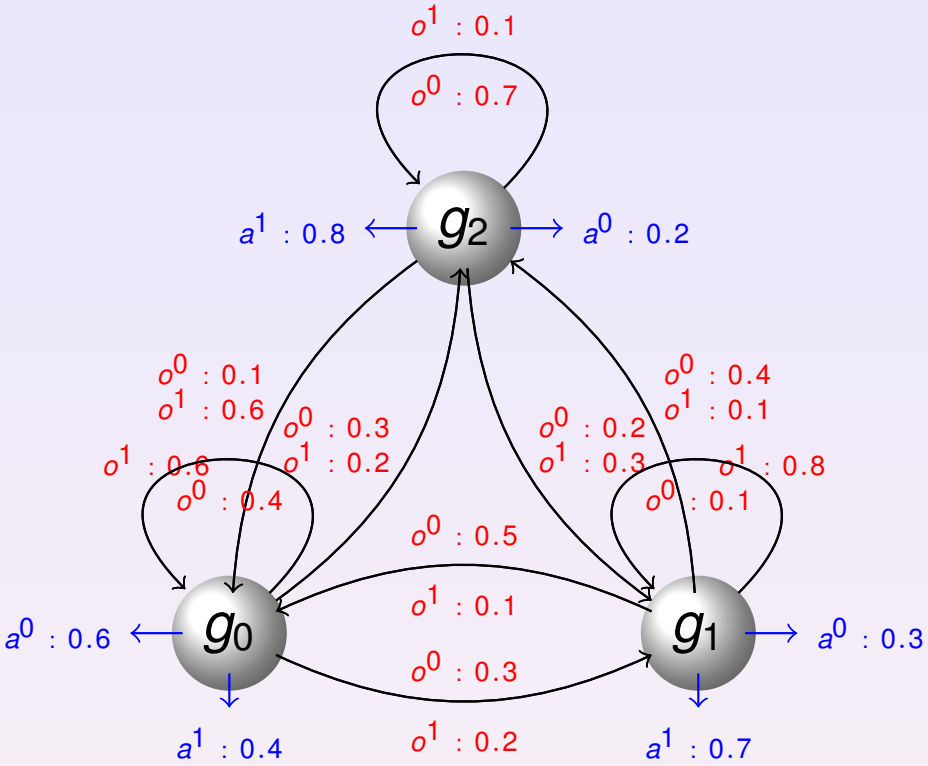
A sampled trajectory



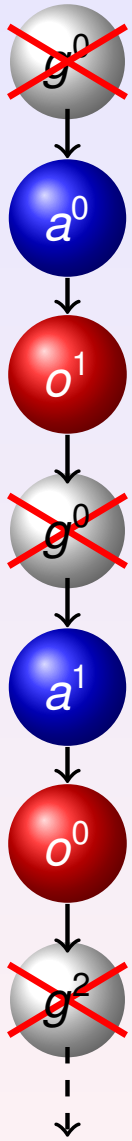
A trajectory with internal belief states

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

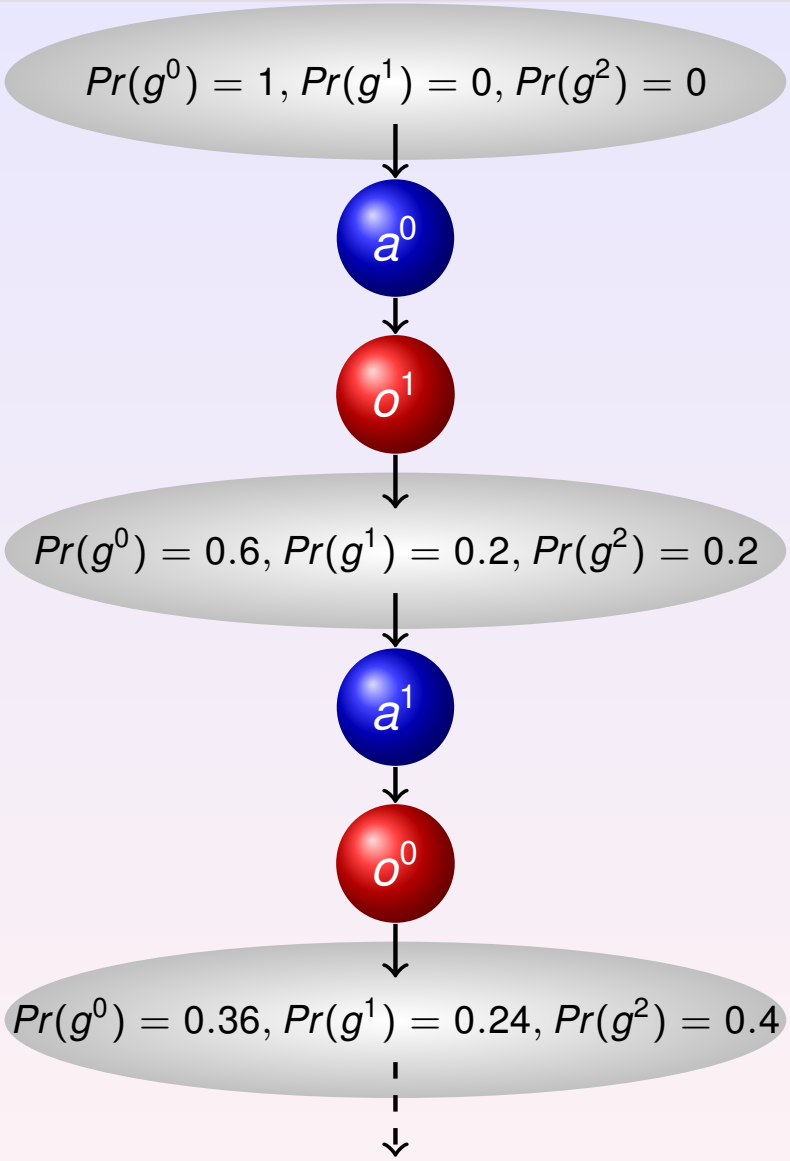
Finite State Controllers with Internal Belief States [Aberdeen & Baxter, 2002]



A Finite State Controller



A sampled trajectory



A trajectory with internal belief states

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

Predictive Representations of Policies

A test q is redefined as a sequence of observations and actions couples, i.e. $q = o^1 a^1 \dots o^k a^k$ [Wiewiora, 2005].

The probability of q starting after a history h_t is redefined as:

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

PSR based policies

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:

$$Pr(a|h_t, o, \theta) = \sum_{h \in \mathcal{H}} b_t(h, \theta) Pr(a|h, o, \theta)$$

PSR based policies

Updating the internal belief state

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

$$b_{t+1}(h) = \frac{\sum_{h' \in \mathcal{H}} b_t(h') b_{h'oa}(h) Pr(a|h', o)}{\sum_{h' \in \mathcal{H}} b_t(h') Pr(a|h', o)}$$

The Gradient of The Value Function

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

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

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

The Gradient of The Value Function

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

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

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

The term $Pr(h_t^o | h_t^a)\mathbb{E}[r|h_t, a]$ can be learned from the trials by an Importance Sampling method, using a look-up table.

$$\hat{Pr}(h_t^o | h_t^a) = \frac{\hat{Pr}(h_t)}{Pr(h_t^a | h_t^o)}$$

Gradient Estimation for FSCs

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

$$Pr(h_t^a | h_t^o, \theta) = b_0^T W_\theta^{o_1 a_1} W_\theta^{o_2 a_2} \dots W_\theta^{o_t a_t} e$$

where

$$\begin{cases} W^{o_j a_j}(g, g', \theta) = Pr(g' | g, o_j, \theta) Pr(a_j | g', o_j, \theta) \\ e^T = (1, 1, \dots, 1) \end{cases}$$

Gradient Estimation for FSCs

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

$$Pr(h_t^a|h_t^o,\theta) = b_0^T W_\theta^{o_1 a_1} W_\theta^{o_2 a_2} \dots W_\theta^{o_t a_t} e$$

where

$$\begin{cases} W^{o_j a_j}(g,g',\theta) = Pr(g'|g,o_j,\theta)Pr(a_j|g',o_j,\theta) \\ e^T = (1,1,\dots,1) \end{cases}$$

Unless the structure of the FSC is provided *a priori*, the graph of the FSC is generally completely connected, i.e.
$\forall a,o,g,g' : Pr(g'|g,o,\theta)Pr(a|g',o,\theta) > 0$.
The gradient $\frac{\partial Pr(h_t^a|h_t^o,\theta)}{\partial \theta_i}$ is a multivariate polynomial of degree $2t$.

Gradient Estimation for PSR Policies

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

$$Pr(h_t^a | h_t^o, \theta) = b_0^T M_\theta^{o_1 a_1} M_\theta^{o_2 a_2} \dots M_\theta^{o_t a_t} e$$

where $M_\theta^{o_j a_j}(h, h') = Pr(a_j | o_j, h, \theta) b_{h o_j a_j}(h', \theta)$.

Gradient Estimation for PSR Policies

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

$$Pr(h_t^a | h_t^o, \theta) = b_0^T M_\theta^{o_1 a_1} M_\theta^{o_2 a_2} \dots M_\theta^{o_t a_t} e$$

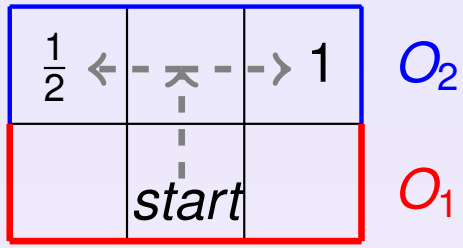
where $M_\theta^{o_j a_j}(h, h') = Pr(a_j | o_j, h, \theta) b_{h o_j a_j}(h', \theta)$.

When a prefix sequence $o_1 a_1 \dots o_i a_i$ of the history h_t corresponds to a core history, then we can write:

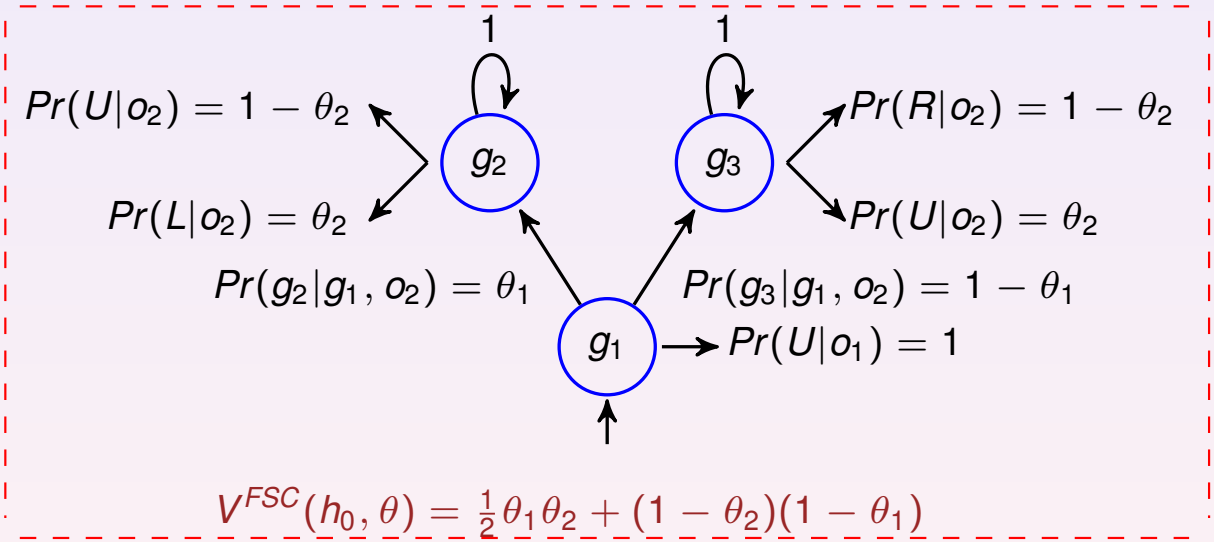
$$Pr(h_t^a | h_t^o, \theta) = \underbrace{Pr(a_1 | o_1, h_0, \theta) \dots Pr(a_i | o_i, h_{i-1}, \theta) b_i^T}_{\text{degree } i} \underbrace{M_\theta^{o_{i+1} a_{i+1}} M_\theta^{o_{i+2} a_{i+2}} \dots M_\theta^{o_t a_t} e}_{\text{degree } 2(t-i)}$$

The gradient $\frac{\partial Pr(h_t^a | h_t^o, \theta)}{\partial \theta_i}$ is a multivariate polynomial of degree $2t - i$.

Small example



A tiny grid-world problem



A finite state controller

$\mathcal{H} = \{h_0\}$

$b_{h_0 o_1 U}(h_0) = 1$

$Pr(L|O_2, h_0) \stackrel{def}{=} \theta_1'$

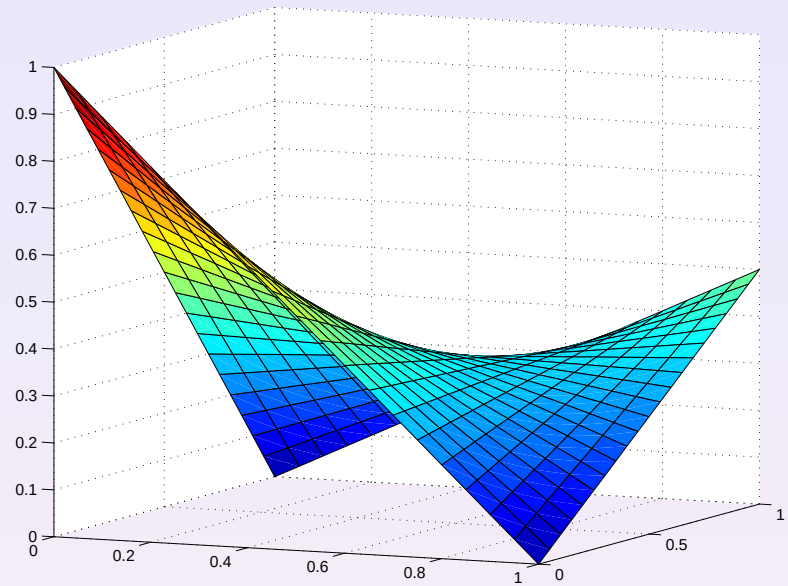
$Pr(R|O_2, h_0) \stackrel{def}{=} \theta_2'$

$Pr(U|O_2, h_0) \stackrel{def}{=} \theta_3'$

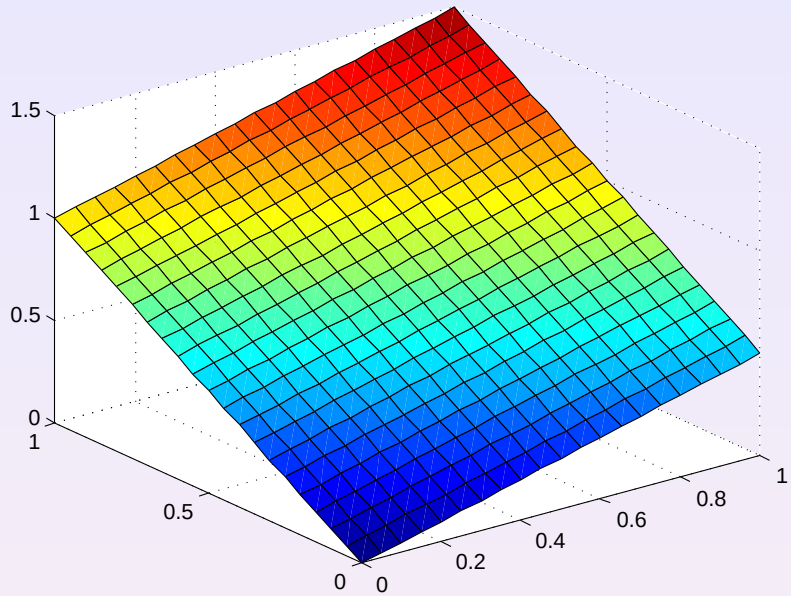
$V^{PSR}(h_0, \theta') = \frac{1}{2}\theta_1' + \theta_2'$

The equivalent PSR policy

Small example



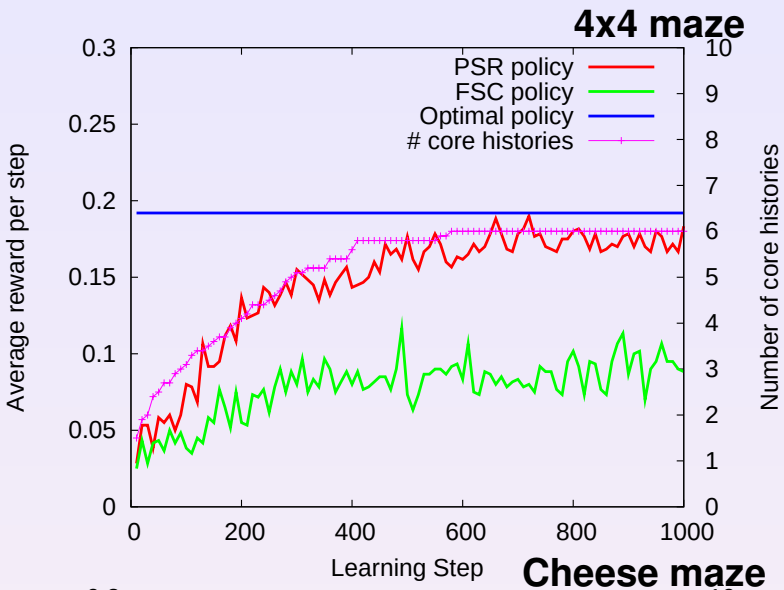
$$V^{FSC}(h_0, \theta) = \frac{1}{2}\theta_1\theta_2 + (1-\theta_2)(1-\theta_1)$$

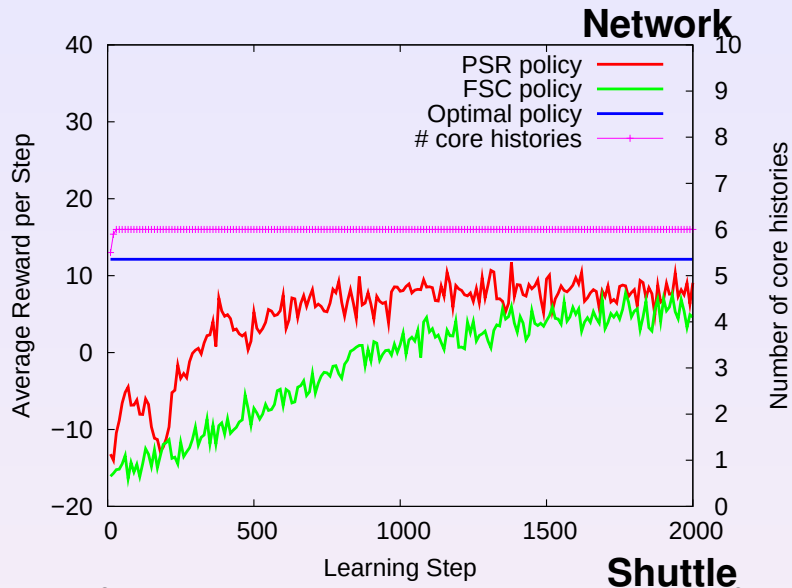


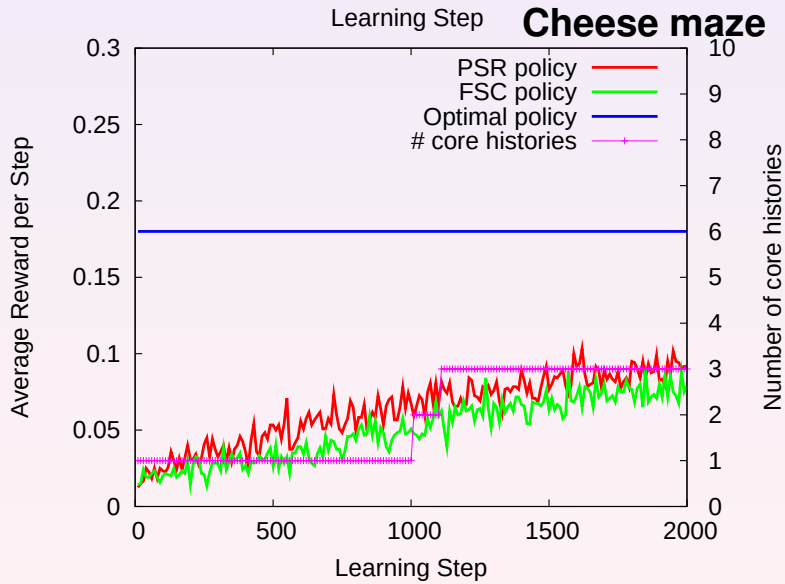
$$V^{PSR}(h_0, \theta') = \frac{1}{2}\theta_1{}' + \theta_2{}'$$
$$\theta_1{}' + \theta_2{}' \leq 1$$

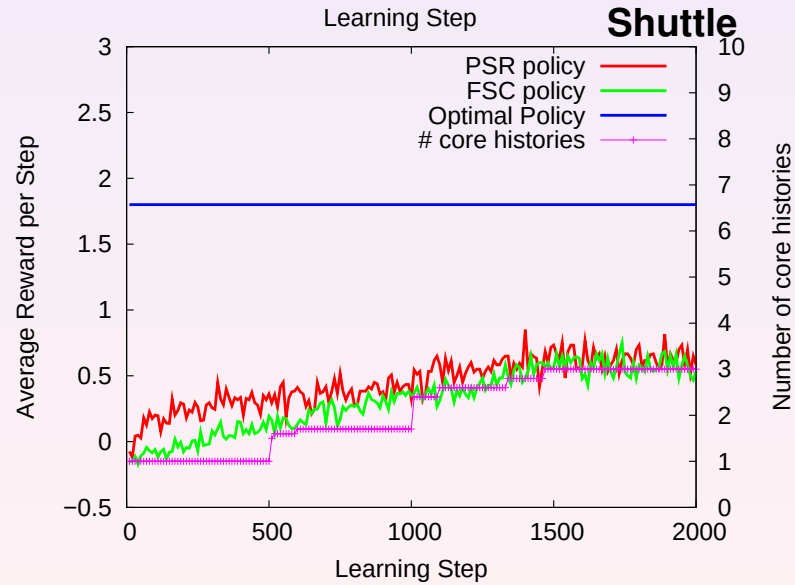
In this example, the value function of an FSC has one local optimum and one global optimum, while its equivalent PSR policy has only one global optimum.

Empirical Results









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

UNIVERSITÉ LAVAL

Conclusion

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.
- ✗ It is unclear how PSR policies will perform in infinite-horizon problems.
- ✗ The discovery of new core histories is based on heuristics.
- ✗ The belief states of a PSR policy are unstable.

Future Work

- Study the performance of PSR policies using the natural gradient [Peters et al., 2008].

For further reading:


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


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


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


Peters, J. and Schaal, S. (2008).
Natural Actor-Critic.
Neurocomputing, 71 (pp. 1180–1190).

Thank you!

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