JMLR: Workshop and Conference Proceedings 24 (2012) 1–2 European Workshop on Reinforcement Learning

Stochastic POMDP controllers: How easy to optimize?

Nikos Vlassis NIKOS.VLASSIS@UNI.LU
Luxembourg Centre for Systems Biomedicine, University of Luxembourg

Michael L. Littman MLITTMAN@CS.BROWN.EDU
Dept. of Computer Science, Rutgers University / Brown University

David Barber D.BARBER@CS.UCL.AC.UK
Department of Computer Science, University College London

Editor: M. Deisenroth, C. Szepesvári, J. Peters

Abstract

It was recently shown that computing an optimal stochastic controller in a discounted infinite-horizon partially observable Markov decision process is an NP-hard problem. The reduction (from the INDEPENDENT-SET problem) involves designing an MDP with special state-action rewards. In this note, we show that the case of state-only-dependent rewards is also NP-hard.

Keywords: Markov decision process, POMDP, stochastic controller, NP-hardness.

1. Introduction

Consider the family of γ-discounted ($\gamma < 1$) infinite-horizon POMDPs with n states and k actions, a single observation (aka unobserved MDPs), state-action immediate costs (negative rewards) c_{sa}, transition model $p(\bar{s}|s,a)$, and starting distribution $\boldsymbol{\mu}$. An optimal stochastic 'blind' controller in such a POMDP is a vector $\boldsymbol{\pi}^* = (\pi_a^*)$ in the probability simplex Δ that is a global minimizer of the following bilinear program:

$$\min_{\boldsymbol{\pi}\in\Delta,\, \mathbf{x}\geq 0} \sum_s x_s \sum_a \pi_a c_{sa}, \qquad \text{s.t.} \quad x_{\bar{s}} = (1-\gamma)\mu_{\bar{s}} + \gamma \sum_a \pi_a \sum_s p(\bar{s}|s,a) x_s \quad \forall_{\bar{s}}, \tag{1}$$

where $\mathbf{x} = (x_s)$ is an occupancy distribution over states, and the objective is total discounted cost. Such bilinear programs are in general nonconvex and hence difficult to solve to global optimality. Moreover, an optimal $\boldsymbol{\pi}^*$ may involve irrational numbers (Littman, 1996).

2. The computational decision problem

The corresponding decision problem STOCHASTIC-BLIND-POLICY asks: Given a POMDP in the above family and a target cost r, is there a stochastic blind controller $\boldsymbol{\pi}$ that incurs cost $J(\boldsymbol{\pi}) \leq r$? Here, $J(\boldsymbol{\pi}) = \mathbf{x}^\intercal \mathbf{C}\boldsymbol{\pi}$ is the cost of controller $\boldsymbol{\pi}$ as given in (1), where $\mathbf{C} = (c_{sa})$ is an $n \times k$ matrix containing all state-action costs.

Vlassis et al. (2012) prove that the STOCHASTIC-BLIND-POLICY problem is NP-hard, reducing from the INDEPENDENT-SET problem. They construct a special MDP with n states and n actions, uniform starting distribution $\boldsymbol{\mu}$, cost matrix $\mathbf{C} = \frac{1}{\gamma}(\mathbf{G} + \mathbf{I})$ where $\mathbf{G}$ is the $n \times n$ adjacency matrix of an input cubic graph and $\mathbf{I}$ is the identity matrix, and

© 2012 N. Vlassis, M. Littman & D. Barber.

deterministic transitions $p(\bar{s}|s,a) = 1$ if $\bar{s} = a$ and 0 otherwise. For that MDP, the cost of controller $\boldsymbol{\pi}$ can be shown to be $J(\boldsymbol{\pi}) = \frac{4(1-\gamma)}{n\gamma} + \boldsymbol{\pi}^{\intercal}(\mathbf{G}+\mathbf{I})\boldsymbol{\pi}$, which involves a quadratic form in $\boldsymbol{\pi}$. If the target cost r is chosen carefully, a straightforward application of the Motzkin-Straus theorem (Motzkin and Straus, 1965) establishes the reduction. We refer to the above paper for details and additional results.

3. The case of state-only-dependent costs

In this note we establish the following:

Theorem 1 STOCHASTIC-BLIND-POLICY *is NP-hard even for POMDPs with state-only-dependent costs.*

Proof Consider the class of γ-discounted MDPs with discount factor $\gamma = \delta^2$, for rational $\delta \in (0,1)$. It is easy to verify that STOCHASTIC-BLIND-POLICY remains NP-hard for this class of MDPs by recapitulating the proof of Vlassis et al. (2012). For any MDP $\mathcal{M}$ in this class with state-action costs c_{sa}, we will construct a δ-discounted MDP $\mathcal{M}'$ with state-only-dependent costs such that $\mathcal{M}$ and $\mathcal{M}'$ attain the same total discounted cost under any blind policy. The construction is as follows: $\mathcal{M}'$ inherits all original states of $\mathcal{M}$, plus it has one extra state for each state-action pair of $\mathcal{M}$. From each original state s, action a moves deterministically to the extra state (s,a), and from the latter *any* action moves to original state s' according to the dynamics $p(s'|s,a)$ of $\mathcal{M}$. The costs of $\mathcal{M}'$ are state-only-dependent: they are zero for all original states s and are equal to $\frac{1+\delta}{\delta}c_{sa}$ for each extra state (s,a). The starting distribution in $\mathcal{M}'$ is the natural extension of the one in $\mathcal{M}$, that is, it places zero mass on extra states (s,a). Note that, in both MDPs, the process starts from each state with the same probability, but in $\mathcal{M}'$ each trajectory is 'twice as long': there is always an intermediate extra state between two original states in a trajectory. Since $\gamma = \delta^2$, it is easy to verify that the two MDPs attain the same total discounted cost under any blind policy. Therefore, $\mathcal{M}$ and $\mathcal{M}'$ are equivalent, and a polynomial-time algorithm for deciding STOCHASTIC-BLIND-POLICY in $\mathcal{M}'$ would imply a polynomial-time algorithm for deciding STOCHASTIC-BLIND-POLICY in $\mathcal{M}$, which cannot be the case unless P=NP. ■

Acknowledgments

We thank Peter Dayan and Matthijs Spaan for helpful discussions.

References

M. L. Littman. *Algorithms for Sequential Decision Making*. PhD thesis, Department of Computer Science, Brown University, 1996. CS-96-09.

T. S. Motzkin and E. G. Straus. Maxima for graphs and a new proof of a theorem of Turán. *Canadian Journal of Mathematics*, 17:533–540, 1965.

N. Vlassis, M. L. Littman, and D. Barber. On the computational complexity of stochastic controller optimization in POMDPs. *ACM Transactions on Computation Theory*, 2012. To appear.