Automatica 50 (2014) 994–998

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Technical communique

Maximizing the set of recurrent states of an MDP subject to convex constraints☆

Eduardo Arvelo, Nuno C. Martins [1]

Department of Electrical and Computer Engineering, University of Maryland, College Park, MD, 20742, USA

ARTICLE INFO

Article history:
Received 30 August 2012
Received in revised form
31 August 2013
Accepted 3 January 2014
Available online 16 February 2014

Keywords:
Maximum entropy
Markov decision problems
Markov models
Convex optimization
Optimal control

ABSTRACT

This paper focuses on the design of time-homogeneous fully observed Markov decision processes (MDPs), with finite state and action spaces. The main objective is to obtain policies that generate the maximal set of recurrent states, subject to convex constraints on the set of invariant probability mass functions. We propose a design method that relies on a finitely parametrized convex program inspired on principles of entropy maximization. A numerical example is provided to illustrate these ideas.

1. Introduction

The formalism of Markov decision processes (MDPs) is widely used to describe the behavior of systems whose state transitions probabilistically among different configurations over time. The impact of a control policy is felt through the actions that dictate the state transition probabilities. Often, but not always, approaches hinge on dynamic programming principles and presume costs that are linear on the time-varying vector of probabilities of the states (Altman, 1999; Bertsekas, 2005; Borkar, 1990, 2002; Fox, 1966; Garg, Kumar, & Marcus, 1999; Hernandez-Lerma & Lasserre, 1996; Hordijk & Kallenberg, 1979; Kumar & Varaiya, 1986; Puterman, 1994; Wolfe & Dantzig, 1962). For an extensive survey, see Arapostathis, Borkar, Fernandez-Guacherand, Ghosh, and Marcus (1993) and the references therein.

We focus on the design of time-homogeneous control policies for fully observable MDPs with finite state and action spaces, represented as $\mathbb{X}$ and $\mathbb{U}$, respectively. In particular, we focus on finding a policy that leads to the maximal set of recurrent states, subject to convex constraints on the set of invariant joint probability mass function (pmf) on the state and action. This framework can be applied to formalize the problem of designing policies for robots tasked with surveillance, where it is desirable that the largest number of states are persistently visited, subject to constraints (Arvelo, Kim, & Martins, 2013).[2] In this setup, the states that are persistently surveilled are the recurrent states of the underlying Markov chain.

1.1. Comparison with existing work

A similar framework has been studied in a series of papers by Arapostathis et al., where the state probability distribution is restricted to be bounded above and below by safety vectors at all times. In Arapostathis, Kumar, and Hsu (2005); Arapostathis, Kumar, and Tangirala (2003); Hsu, Arapostathis, and Kumar (2010), the authors propose algorithms to find the set of distributions whose evolution under a given control policy respect the safety constraint. In Wu, Arapostathis, and Kumar (2004), an augmented Markov chain is used to find the maximal set of probability distributions whose evolution respects the safety constraint over all admissible non-stationary control policies.

☆ This work was partially supported by NSF Grant CNS 0931878, AFOSR Grant FA95501110182, ISR/Maryland Robotics Center Seed Grant, ONR AppEl/UMD Center and Multiscale Systems Center (FCRP). The material in this paper was partially presented at the 2013 IEEE International Conference on Robotics and Automation (ICRA 2013), May 6–10, 2013, Karlsruhe, Germany. This paper was recommended for publication in revised form by Associate Editor Henrik Sandberg under the direction of Editor André L. Tits.

E-mail addresses: earvelo@umd.edu (E. Arvelo), nmartins@umd.edu (N.C. Martins).

[1] Tel.: +1 301 405 9198; fax: +1 301 314 9281.

http://dx.doi.org/10.1016/j.automatica.2014.01.002
0005-1098/

[2] The paper Arvelo et al. (2013) describes an application of the results presented here.

Here we obtain the policy that leads to the maximal set of recurrent states, subject to convex constraints on the set of invariant pmfs. The main contribution of this paper is to solve this problem via a finitely parametrized convex program. Our approach is rooted in entropy maximization, and the proposed solution can be easily implemented using standard convex optimization tools, such as the ones described in Grant and Boyd (2011).

1.2. Paper organization

The remainder of this paper is organized as follows. Section 2 provides notation, basic definitions and the problem statement. The convex program that solves the problem is presented in Section 3. Numerical examples are given in Section 4, and conclusions are discussed in Section 5.

2. Preliminaries and problem statement

The following notation is used throughout the paper:

$\mathbb{X}$	state space of the MDP
$\mathbb{U}$	set of control actions
X_k	state of the MDP at time k
U_k	control action at time k
$\mathbb{P}_{\mathbb{X}}$	set of all pmfs with support in $\mathbb{X}$
$\mathbb{P}_{\mathbb{U}}$	set of all pmfs with support in $\mathbb{U}$
$\mathbb{P}_{\mathbb{XU}}$	set of all joint pmfs with support in $\mathbb{X}\times\mathbb{U}$
$\mathbb{S}_f$	support of a pmf f

The recursion of the MDP is given by the (conditional) pmf of X_{k+1} given the previous state X_k and the control action U_k, and is denoted as:

$$\mathcal{Q}(x^+,x,u)\stackrel{def}{=}P(X_{k+1}=x^+|X_k=x,U_k=u).$$

We denote any time-homogeneous control policy by

$$\mathcal{K}(u,x)\stackrel{def}{=}P(U_k=u|X_k=x),\quad u\in\mathbb{U},x\in\mathbb{X}$$

where $\sum_{u\in\mathbb{U}}\mathcal{K}(u,x)=1$ for all x in $\mathbb{X}$. The set of all such policies is denoted as $\mathbb{K}$.

Assumption. Throughout the paper, we assume that the MDP $\mathcal{Q}$ is given. Hence, all quantities and sets that depend on the closed loop behavior are indexed only by the underlying control policy $\mathcal{K}$.

A pmf f_{XU} in $\mathbb{P}_{\mathbb{XU}}$ is said to be *invariant* under control policy $\mathcal{K}$ if it satisfies the following invariance relation:

$$f_{XU}(x^+,u^+)=\mathcal{K}(u^+,x^+)\sum_{x\in\mathbb{X},u\in\mathbb{U}}\mathcal{Q}(x^+,x,u)f_{XU}(x,u),\tag{1}$$

for all x^+ in $\mathbb{X}$ and u^+ in $\mathbb{U}$. The set of invariant pmfs associated with control policy $\mathcal{K}$ is given by:

$$\mathbb{I}_{\mathcal{K}}\stackrel{def}{=}\Big\{f_{XU}\in\mathbb{P}_{\mathbb{XU}}:(1)\text{ holds with control policy }\mathcal{K}\Big\}.$$

Finally, the set of all invariant pmfs are given by:

$$\mathbb{I}\stackrel{def}{=}\bigcup_{\mathcal{K}\in\mathbb{K}}\mathbb{I}_{\mathcal{K}}.$$

Problem 2.1. Given $\mathbb{W}$, which is a convex subset of $\mathbb{P}_{\mathbb{XU}}$, find a joint pmf f^*_{XU} in $\mathbb{I}\cap\mathbb{W}$ and a corresponding control policy $\mathcal{K}^*$ such that the following inclusion holds:

$$\mathbb{S}^{\mathbb{X}}_{f_{XU}}\subseteq\mathbb{S}^{\mathbb{X}}_{f^*_{XU}},\quad f_{XU}\in\mathbb{I}\cap\mathbb{W};\tag{2}$$

where $\mathbb{S}^{\mathbb{X}}_f=\{x\in\mathbb{X}|\sum_{u\in\mathbb{U}}f(x,u)>0\}$.

Remark. Note that a pmf f^*_{XU} that satisfies (2) has maximal support among all members of the set $\mathbb{I}\cap\mathbb{W}$.

2.1. Comparison with graph-based methods and reachability concepts

Once a control policy is applied to an MDP, one can construct a directed graph of transitions for the resulting Markov chain. Here, the vertices of the graph are the states and an edge from i to j indicates that the transition from i to j has positive probability. In this case, the set of recurrent states is the union of the strongly connected components that are closed, each representing a recurrent class. Hence, one could use standard algorithms (Cormen, Leiserson, Rivest, & Stein, 1990), such as Kosaraju's or Tarjan's, to efficiently find the strongly connected components of the graph and perform the union of the ones that are closed. However, finding a control policy that "maximizes" the set of recurrent states cannot be easily obtained using this method because the graph of transitions may change for different candidate solutions. Furthermore, because we consider constraints on the set of invariant pmfs, both the maximal set of recurrent states and the corresponding control policies will, in general, depend on the actual values of the entries of the transition probability matrices. Examples of constraints of interest include lower and upper bounds on invariant pmfs evaluated at pre-selected state and action pairs, or on the expected value of a function of the state and (or) action.

Broadly speaking, reachability is concerned with the determination of whether a set of states can be reached from another via an appropriate control policy (Tabuada, 2010). There are two main reasons why our formulation cannot be cast as a reachability problem. The first is that reachability is in general distinct from recurrence, such as, for instance, when a reachable state is transient. The second follows from the discussion above, where we emphasize that optimal solutions may depend on the probabilities of the transitions, and not only on whether which ones occur with nonzero probability.

Numerical examples are provided in Section 4 that illustrate how different convex constraints on the set of invariant pmfs can induce changes on both the optimal control policies and the resulting maximal sets of recurrent states.

3. Solution via convex optimization

We propose the following optimization program to solve Problem 2.1:

$$f^*_{XU}=\arg\max_{f_{XU}\in\mathbb{W}}\mathcal{H}(f_{XU})\tag{3}$$

subject to:

$$\sum_{u^+\in\mathbb{U}}f_{XU}(x^+,u^+)=\sum_{x\in\mathbb{X},u\in\mathbb{U}}\mathcal{Q}(x^+,x,u)f_{XU}(x,u),\quad x^+\in\mathbb{X}\tag{4}$$

where $\mathcal{H}:\mathbb{P}_{\mathbb{XU}}\to\Re_{\geq 0}$ is the entropy of f_{XU}, and is given by

$$\mathcal{H}(f_{XU})=-\sum_{u\in\mathbb{U}}\sum_{x\in\mathbb{X}}f_{XU}(x,u)\ln(f_{XU}(x,u)),$$

where we adopt the standard convention that $0\ln(0)=0$.

Remark. The dimension of the search space is given by the number of elements in the product set $\mathbb{X}\times\mathbb{U}$. In many applications, such as in persistent surveillance (Arvelo et al., 2013), the number of possible control actions $\mathbb{U}$ is small compared to $\mathbb{X}$, which may represent the possible locations and orientations of a robot. Since the optimization program (3)–(4) is convex, it can be efficiently solved even when the search space is of very large dimension. There are as many linear equality constraints in (3)–(4) as there are elements in $\mathbb{X}$, and additional constraints can be included such as in the examples of Section 4.