Operations Research Letters 42 (2014) 123–129

Contents lists available at ScienceDirect

Operations Research Letters

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

Constrained optimality for finite horizon semi-Markov decision processes in Polish spaces

Yonghui Huang [a], Zhongfei Li [b], Xianping Guo [a,*]

[a] School of Mathematics and Computational Science, Sun Yat-Sen University, Guangzhou 510275, China

[b] Business School, Sun Yat-Sen University, Guangzhou 510275, China

ARTICLE INFO

Article history:
Received 17 June 2013
Received in revised form
24 December 2013
Accepted 25 December 2013
Available online 2 January 2014

Keywords:
Semi-Markov decision processes
Expected finite horizon reward
Occupancy measure
Constrained-optimal policy
Linear program

ABSTRACT

This paper focuses on solving a finite horizon semi-Markov decision process with multiple constraints. We convert the problem to a constrained absorbing discrete-time Markov decision process and then to an equivalent linear program over a class of occupancy measures. The existence, characterization and computation of constrained-optimal policies are established under suitable conditions. An example is given to demonstrate our results.

1. Introduction

Constrained Markov decision processes (MDPs) form an important class of stochastic control problems, which frequently arise in the real-world situations such as computer networks, data communications, resource-sharing systems and so on. A large amount of literature has been developed for solving constrained MDPs; see, for instance, [1,2,5,9,13] on the discrete-time MDPs (DTMDPs), [6,10,11,19] on the continuous-time jump MDPs (CTJMDPs), and [4,7,6] on the semi-Markov decision processes (SMDPs). In these references, most of them deal with constrained problems with *infinite horizons*, while only a few are related to those with *finite horizons* [5,6,17,18]. For the issue of finite horizon constrained optimality, the studies in [5,17,18] are limited to the case of DTMDPs, and the one in [6] treats *finite-step* SMDPs, which can be directly reduced to finite horizon DTMDPs due to that the number of jumps is fixed therein. Therefore, to our knowledge, the current works on finite horizon constrained MDPs are concentrated on the case of *discrete-time* processes, whereas the case for *continuous-time* processes has not been explored yet. In the meantime, however, the finite horizon optimality in continuous-time often provides more realistic models than the discrete-time ones because it takes into account the fact that the time may evolve continuously, which has received great concerns and interests in many applications [15,16]. Moreover, since the number of jumps during a fixed time horizon becomes random in continuous-time processes, the associated arguments and techniques in [5,6,17,18] for discrete-time processes are *not* suitable to the continuous-time ones. Hence, finite horizon constrained optimality for continuous-time processes is an unsolved but desirable problem.

This paper is devoted to investigating continuous-time finite horizon constrained SMDPs with multiple constraints. The state and action sets are Polish spaces, while the reward functions are assumed to be bounded. Our aim is to obtain a so-called constrained-optimal policy that maximizes an objective reward over a finite horizon in a set of policies satisfying given constraints. We focus on establishing the existence, characterization and computation of constrained-optimal policies. Compared with the well-known reduction of *finite-step* SMDPs to *finite horizon* DTMDPs as in [6], we convert the *continuous-time finite horizon* SMDPs to *infinite horizon* DTMDPs with a two dimension state space of pairs of jump times and original states, where the associated one-stage reward functions and transition probabilities include the current jump time and are very different from those in the reduction of *finite-step* SMDPs; see Section 3.1 for details. Since the transition probabilities of the converted DTMDPs are sub-stochastic, we further reformulate the converted infinite horizon DTMDPs as absorbing DTMDPs under Condition 3.1. In the framework of absorbing DTMDPs [8], we introduce occupancy measures and discuss their properties (see Lemma 3.1), via

* Corresponding author. Tel.: +86 20 34022183.
E-mail addresses: hyongh5@mail.sysu.edu.cn (Y. Huang), lnslzf@mail.sysu.edu.cn (Z. Li), mcsgxp@mail.sysu.edu.cn (X. Guo).

0167-6377/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.orl.2013.12.009

which the continuous-time finite horizon constrained SMDPs are eventually transformed to an equivalent linear program (LP). Moreover, following the results in [8] for absorbing DTMDPs, we propose suitable conditions and prove the existence of a randomized stationary constrained-optimal policy and an N-deterministic constrained-optimal policy (see Theorem 3.2). We also introduce assumptions that involve a weight function, under which the equivalent LP is solvable and constrained-optimal policies can be derived (see Theorem 3.3). Finally, we apply our results to a simple maintenance problem, which is described by a finite horizon constrained SMDP with one constraint, finite states and actions. The constrained-optimal value and a 1-randomized constrained-optimal policy have been computed by the numerical experiment.

The structure of this paper is as follows. Section 2 formulates the control model and the constrained problem we are interested in. Our main results on the existence, characterization and computation of constrained-optimal policies are stated in Section 3. An example is given to illustrate our results in Section 4.

2. The control model

Notation. If X is a Polish space (that is, a complete and separable metric space), we denote by $\mathcal{B}(X)$ the Borel σ-algebra, by $\mathcal{P}(X)$ the set of all probability measures on $\mathcal{B}(X)$, by $\mathcal{M}(X)$ the space of all finite nonnegative measures on $\mathcal{B}(X)$, and moreover, by $\mathbb{C}_b(X)$ the set of all bounded continuous functions on X.

We consider a constrained SMDP model with the following objects:

$$\big\{E, A, (A(x), x \in E), Q(\cdot, \cdot \mid x, a), r_0(x, a), \{r_k(x, a), d_k, k = 1, \ldots, N\}\big\}, \tag{2.1}$$

where E is the *state space* and A is the *action set*, which are assumed to be Polish spaces, respectively; $A(x) \in \mathcal{B}(A)$ denotes the *set of admissible actions* at state $x \in E$. The transition mechanism of the SMDP is defined by the *semi-Markov kernel* $Q(\cdot, \cdot|x, a)$ on $R_+ \times E$ given K, where $R_+ = [0, +\infty)$, and $K = \{(x, a) \mid x \in E, a \in A(x)\}$ denotes the set of feasible state-action pairs and is assumed to be in $\mathcal{B}(E \times A)$. If an action $a \in A(x)$ is selected at state x, then for all $t \in R_+$ and $D \in \mathcal{B}(E)$, $Q(t, D \mid x, a)$ is the joint probability that the sojourn time in state x is not greater than $t \in R_+$, and the next state is in D. Furthermore, the function r_0 on K denotes the *objective reward rate*, while the functions $r_1, \ldots, r_N$ on K represent the *constrained reward rates*. We assume that all of the reward functions r_k are bounded. Finally, the real numbers d_k are constraint constants.

Remark 2.1. The semi-Markov kernel Q can be partitioned as shown below:

$$Q(t, D|x, a) = \int_D F(t|x, a, y)p(dy|x, a) \quad \forall t \in R_+, D \in \mathcal{B}(E), (x, a) \in K, \tag{2.2}$$

where $F(\cdot|x, a, y)$ denotes the sojourn time distribution in state x when action a is chosen and the next state is to be y, and $p(\cdot|x, a)$ is the transition law of the system states.

We now describe how an SMDP evolves. In an SMDP, the controller observes the system states continuously in time. Suppose that the system occupies state $x_0 \in E$ at the initial time $t_0 \in R_+$, then the controller chooses an action $a_0 \in A(x_0)$ according to some rule. As a consequence of this action choice, the system jumps to state $x_1 \in E$ after a sojourn time $\theta_1 \in R_+$ in x_0, in which the transition law is subject to the semi-Markov kernel Q. At time $(t_0 + \theta_1)$, the controller chooses an action $a_1 \in A(x_1)$ according to some rule and the same sequence of events occur. The SMDP evolves in this way and we obtain an admissible history h_n of the SMDP up to the nth jump time, i.e., $h_n = (t_0, x_0, a_0, \theta_1, x_1, a_1, \ldots, \theta_n, x_n)$. We set $t_k = t_{k-1} + \theta_k$, $k = 1, 2, \ldots, n$, representing the jump times of the SMDP. Then, a history h_n can be equivalently rewritten as $h_n = (t_0, x_0, a_0, t_1, x_1, a_1, \ldots, t_n, x_n)$.

To specify rules for the controller to choose actions, policies are needed. To this end, for each $n \geq 0$, let $H_n = (R_+ \times E \times A)^n \times (R_+ \times E)$ denote the set of all admissible histories of the SMDP up to the nth jump, which is endowed with the Borel σ-algebra

$$\sigma(\{(a_1, b_1) \times D_1 \times \Gamma_1 \times \cdots \times (a_n, b_n) \times D_n \times \Gamma_n : (a_i, b_i) \subset [0, +\infty), D_i \in \mathcal{B}(E), \Gamma_i \in \mathcal{B}(A), 1 \leq i \leq n\}).$$

Then, a randomized history-dependent policy (or simply a policy) $\pi = \{\pi_n, n \geq 0\}$ is a sequence of stochastic kernels π_n on A given H_n satisfying $\pi_n(A(x_n) \mid h_n) = 1$ for every $h_n \in H_n$, $n = 0, 1, 2, \ldots$. The set of all policies is denoted by Π. In most cases, however, some special policies as below are useful.

Definition 2.1. (a) A randomized Markov policy $\pi = \{\phi_n\}$ is a sequence of stochastic kernels ϕ_n on A given $R_+ \times E$ such that $\phi_n(A(x_n) \mid t_n, x_n) = 1$ for every $(t_n, x_n) \in R_+ \times E$ and $n \geq 0$, where (t_n, x_n) represents the nth jump time and the post-jump state. If, furthermore, ϕ_n are independent of n, it is said to be randomized stationary. In this case, we write $\pi = \{\phi, \phi, \ldots\}$ as ϕ for simplicity.

(b) A deterministic Markov policy $\pi = \{f_n\}$ is a sequence of measurable functions $f_n : R_+ \times E \to A$ such that $f_n(t_n, x_n)$ is in $A(x_n)$ for every $(t_n, x_n) \in R_+ \times E$ and $n \geq 0$. If, furthermore, f_n are independent of n, it is said to be deterministic stationary, and $\pi = \{f, f, \ldots\}$ is written as f.

We denote by Π_{RM}, Π_{RS}, Π_{DM} and Π_{DS} the families of all randomized Markov, randomized stationary, deterministic Markov and deterministic stationary policies, respectively. Obviously, $\Pi_{RS} \subset \Pi_{RM} \subset \Pi$, and $\Pi_{DS} \subset \Pi_{DM} \subset \Pi$.

For each $n \geq 0$, we denote by T_n the random variable (r.v.) of the nth jump time of the SMDP, by Θ_{n+1} the r.v. of the sojourn time between the nth and $(n + 1)$th jumps, by X_n the r.v. of the post-jump state at T_n, and by A_n the r.v. of the action chosen at T_n. Note that T_n, Θ_{n+1}, X_n and A_n are r.v.s defined on a same measurable space $(\Omega, \mathcal{F})$, where the sample space $\Omega = (R_+ \times E \times A)^\infty \bigcup \{(t_0, x_0, a_0, \theta_1, x_1, a_1, \ldots, \theta_k, x_k, a_k, \infty, x_\infty, a_\infty, \infty, x_\infty, a_\infty, \ldots) | t_0 \in R_+, \theta_1 \in R_+, \ldots, \theta_k \in R_+, x_l \in E, a_l \in A$, for each $0 \leq l \leq k, k \geq 0\}$ with an isolated state x_∞ and an isolated action a_∞, and the Borel σ-algebra $\mathcal{F} = \mathcal{B}(\Omega)$, the smallest σ-algebra which contains all finite products of sets of the form $((a, b) \times D \times \Gamma)$ with $(a, b) \subset [0, +\infty]$, $D \in \mathcal{B}(E \cup \{x_\infty\})$, $\Gamma \in \mathcal{B}(A \cup \{a_\infty\})$. In fact, for each $n \geq 0$ and any trajectory $\omega = (t_0, x_0, a_0, \theta_1, x_1, a_1, \ldots, \theta_n, x_n, a_n, \ldots) \in \Omega$, we can define

$$T_0(\omega) = t_0, \qquad T_{n+1}(\omega) = t_0 + \theta_1 + \cdots + \theta_{n+1},$$

$$\Theta_{n+1}(\omega) = \theta_{n+1}, \qquad X_n(\omega) = x_n, \qquad A_n(\omega) = a_n.$$

It is worth noting that since the sojourn time distribution $F(\cdot|x, a, y)$ is concentrated on R_+ for each $(x, a) \in K$ and $y \in E$, almost all the trajectories have finite sojourn times. Thus, we ignore the trajectories that have infinite sojourn times in the following. Given the semi-Markov kernel Q, an initial time-state pair $(t, x) \in R_+ \times E$ and a policy $\pi \in \Pi$, by the Ionescu Tulcea theorem, there exists a unique probability measure $P^\pi_{(t,x)}$ on $(\Omega, \mathcal{F})$ such that

$$P^\pi_{(t,x)}(T_0 = t, X_0 = x) = 1, \tag{2.3}$$

$$P^\pi_{(t,x)}(A_n \in \Gamma \mid T_0, X_0, A_0, \ldots, T_n, X_n) = \pi_n(\Gamma \mid T_0, X_0, A_0, \ldots, T_n, X_n), \tag{2.4}$$