European Journal of Operational Research 238 (2014) 486–496

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Stochastics and Statistics

Convergence of controlled models and finite-state approximation for discounted continuous-time Markov decision processes with constraints

Xianping Guo [a,*], Wenzhao Zhang [a,b]

[a] School of Mathematics and Computational Science, Sun Yat-Sen University, Guangzhou 510275, PR China
[b] College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350108, PR China

ARTICLE INFO

Article history:
Received 31 December 2012
Accepted 24 March 2014
Available online 5 April 2014

Keywords:
Constrained continuous-time Markov decision processes
Unbounded transition rate
Convergence
Finite approximation

ABSTRACT

In this paper we consider the convergence of a sequence $\{\mathcal{M}_n\}$ of the models of discounted continuous-time *constrained* Markov decision processes (MDP) to the "limit" one, denoted by $\mathcal{M}_\infty$. For the models with denumerable states and unbounded transition rates, under reasonably mild conditions we prove that the (constrained) optimal policies and the optimal values of $\{\mathcal{M}_n\}$ converge to those of $\mathcal{M}_\infty$, respectively, using a technique of occupation measures. As an application of the convergence result developed here, we show that an optimal policy and the optimal value for countable-state continuous-time MDP can be approximated by those of finite-state continuous-time MDP. Finally, we further illustrate such finite-state approximation by solving numerically a controlled birth-and-death system and also give the corresponding error bound of the approximation.

1. Introduction

Constrained Markov decision processes (MDP) form an important class of stochastic control problems with applications in many areas such as telecommunication networks and queueing systems; see, for instance, Guo and Hernández-Lerma (2009), Hordijk and Spieksma (1989), and Sennott (1991). As is well known, the main purpose of studies on constrained MDP is on the existence and computation of optimal policies, see, for instance, the literature on the discrete-time MDP by Feinberg and Shwartz (1999), Feinberg (2000), Hordijk and Spieksma (1989), Hernández-Lerma and González-Hernández (2000), Hernández-Lerma, González-Hernández, and López-Martínez (2003), and Sennott (1991), and the works on continuous-time MDP by Guo (2007), Guo and Hernández-Lerma (2003), Guo and Hernández-Lerma (2009), Guo and Piunovskiy (2011). On the other hand, from a theoretical and practical point of view, it is of interest to analyze the convergence of optimal values and optimal policies for constrained MDP, and such convergence problems have been considered, see, for instance, Altman (1999), Zadorojniy and Shwartz (2006), Alvarez-Mena and Hernández-Lerma (2002) and so on. Alvarez-Mena and Hernández-Lerma (2006) also consider the convergence problem as in Alvarez-Mena and Hernández-Lerma (2002) for the case of more than one controller. To the best of our knowledge, however, these existing works for the convergence problems are on the constrained discrete-time MDP. Most recently, the convergence problem of controlled models for *unconstrained* continuous-time MDP has also been considered by Prieto-Rumeau and Lorenzo (2010) and Prieto-Rumeau and Hernández-Lerma (2012) using an approximation of the optimality equations. However, the similar convergence problem for *constrained* continuous-time MDP has not been considered.

This paper studies the convergence problem for constrained continuous-time MDP. More precisely, in this paper we consider a sequence $\{\mathcal{M}_n\}$ of the models of the constrained continuous-time MDP with the following features: (1) the state space is denumerable, but action space is general; (2) the transition rates and all reward/cost functions are allowed to be *unbounded*; and (3) the optimality criterion is the expected discounted reward/cost, while some constraints are imposed on similar discounted rewards/costs. We aim to give suitable conditions imposed on the models $\{\mathcal{M}_n\}$, under which the optimal policies and the optimal values of $\{\mathcal{M}_n\}$ converge to those of the limit model $\mathcal{M}_\infty$ of the sequence $\{\mathcal{M}_n\}$, respectively.

In general, the approaches to study continuous-time MDP can be roughly classified into two groups: the indirect method and the direct method. For the indirect method, the idea is to convert the continuous-time MDP into equivalent discrete-time MDP. This approach has been justified by Feinberg (2004), Feinberg (2012), and Piunovskiy and Zhang (2012). On the other hand, the most

* Corresponding author. Tel.: +86 020 84113190; fax: +86 020 84037978.
E-mail addresses: mcsgxp@mail.sysu.edu.cn (X. Guo), zhangwenzhao1987@163.com (W. Zhang).

http://dx.doi.org/10.1016/j.ejor.2014.03.037
0377-2217/

common direct method to investigate constrained continuous-time MDP is to establish an equivalent linear program formulation of the original constrained problem, see Guo and Piunovskiy (2011). In this paper, we follow this direct approach without involving discrete-time MDP. First, as in Guo and Piunovskiy (2011), we transform the optimality problem in constrained continuous-time MDP into an *equivalent* optimality problem over a class of some probability measures by introducing an occupation measure of a policy. Then, we analyze the asymptotic characterization of the occupation measure and the expected discounted rewards/costs, which are used to prove that the optimal values and optimal policies of the sequence $\{\mathcal{M}_n\}$ converge to those of $\mathcal{M}_\infty$. Finally, we apply our results to the approximations of the optimal policies and the optimal value of finite-state continuous-time MDP to those of countable-state continuous-time MDP. More precisely, for a model $\mathcal{M}'_\infty$ of constrained countable-state continuous-time MDP satisfying the usual conditions as in Guo and Hernández-Lerma (2009) and Guo and Piunovskiy (2011), we can construct a sequence of models $\{\mathcal{M}'_n\}$ of constrained continuous-time MDP with finite states such that every accumulation point of a sequence of optimal policies of $\{\mathcal{M}'_n\}$ is optimal for $\mathcal{M}'_\infty$ and that the sequence of the optimal values of $\{\mathcal{M}'_n\}$ converge to the optimal value of $\mathcal{M}'_\infty$. Furthermore, we further illustrate such finite-state approximation by solving numerically a controlled birth-and-death system, and also give the corresponding error bound of the approximation. The motivation of providing such approximation is from the following facts: (i) there exist many methods to solve the optimal value and optimal policies for *unconstrained* continuous-time MDP with finite sates, for example, the value iteration algorithm and the policy iteration algorithm by Guo and Hernández-Lerma (2009) and Puterman (1994), the approximation dynamic programming technique by Cervellera and Macciò (2011), and so on. However, these methods, which are all based on the optimality equation, are not applied to constrained continuous-time MDP since the optimality equation no longer exists for the constrained MDP; (ii) the optimal value and optimal policies for finite-state constrained continuous-time MDP with finite actions can be computed by the well known linear programming in Guo and Piunovskiy (2011) and Puterman (1994), whereas in general the optimal value and optimal policies cannot be computed for countable-state continuous-time MDP because the number of states in such MDP is infinite.

The rest of this paper is organized as follows. In Section 2, we introduce the models of constrained continuous-time MDP and the convergence problems. In Section 3, we state our main results, which are proved in Section 6, after technical preliminaries given in Section 5. An application of the main results to finite state approximation and a numerable example are given in Section 4. Finally, we finish this article with a conclusion in Section 7.

2. The models

In this section we introduce the models and convergence problems we are concerned with.

Notation. If X is a Polish space, we denote by $\mathcal{B}(X)$ its Borel σ-algebra, by D^c the complement of a set $D \subseteq X$ (with respect to X), by $\mathcal{P}(X)$ the set of all probability measures on $\mathcal{B}(X)$, endowed with the topology of weak convergence. For a finite set D, we denote by $|D|$ the number of its elements. Let $\mathbb{N} := \{1,2,\ldots\}$ and $\overline{\mathbb{N}} := \mathbb{N} \bigcup \{\infty\}$.

Consider the sequence of models $\{\mathcal{M}_n\}$ for constrained continuous-time MDP:

$$\mathcal{M}_n := \Big\{S_n, (A_n(i), i \in S_n), q_n(\cdot|i,a), c_n^0(i,a), \Big(c_n^l(i,a), d_n^l, 1 \leqslant l \leqslant p\Big), \gamma_n\Big\}, \quad n \in \overline{\mathbb{N}}, \tag{2.1}$$

where S_n are the *state spaces*, which are assumed to be denumerable. The set $A_n(i)$ represents the set of available actions or decisions at state $i \in S_n$ for model $\mathcal{M}_n$. Let

$$K_n := \{(i,a) | i \in S_n, a \in A_n(i)\},$$

represent the set of all feasible state-action pairs for $\mathcal{M}_n$.

In what follows, we assume that $S_n \uparrow S_\infty$, and $S_\infty = \{0,1,\ldots,n,\ldots\}$ without loss of generalization. As a consequence, for each $i \in S_\infty$, we can define $n(i) := \min\{n \geqslant 1, i \in S_n\}$. Furthermore, we assume that $A_n(i) \subseteq A_\infty(i) (n \geqslant n(i), i \in S_\infty)$, and moreover, for each $n \in \overline{\mathbb{N}}$, $A_n(i)$ is in $\mathcal{B}(A_n)$, where A_n is a Polish space, the action space for $\mathcal{M}_n$. Thus, $\mathcal{B}(A_n(i)) = \mathcal{B}(A_\infty(i)) \cap A_n(i)$ and $\mathcal{P}(A_n(i)) \subseteq \mathcal{P}(A_\infty(i))$, for each $i \in S_\infty$ and $n \geqslant n(i)$.

For fixed $n \in \overline{\mathbb{N}}$, the function $q_n(\cdot|i,a)$ in (2.1) refers to the conservative transition rates, that is, $q_n(j|i,a) \geqslant 0$ and $\sum_{j \in S_n} q_n(j|i,a) = 0$ for all $(i,a) \in K_n$ and $i \neq j$. Moreover, $q_n(j|i,a)$ is a measurable function on $A_n(i)$ for each fixed $i,j \in S_n$. Furthermore, $q_n(j|i,a)$ is assumed to be *stable*, that is, $q_n^*(i) := \sup_{a \in A_n(i)} |q_n(i|i,a)| < \infty$ for each $i \in S_n$.

Finally, c_n^0 corresponds to the objective cost function, and $c_n^l (1 \leqslant l \leqslant p)$ correspond to the cost functions on which some constraints are imposed. The real numbers $d_n^l (1 \leqslant l \leqslant p)$ denote the constraints, and γ_n denotes initial distribution on S_n for $\mathcal{M}_n$.

To complete the specification of $\{\mathcal{M}_n\}$ $(n \in \overline{\mathbb{N}})$, we introduce the classes of policies.

A *randomized Markov policy* π for $\mathcal{M}_n$ is a family $(\pi_t, t \geqslant 0)$ of stochastic kernels satisfying: (i) for each $t \geqslant 0$ and $i \in S_n$, $\pi_t(\cdot|i)$ is a probability measure (p.m.) on $A_n(i)$; and (ii) for each $D \in \mathcal{B}(A_n(i))$, and $i \in S_n$, $\pi_t(D|i)$ is a Borel measurable function in $t \geqslant 0$.

Moreover, a policy $\pi = (\pi_t, t \geqslant 0)$ is called (randomized) *stationary* for $\mathcal{M}_n$ if, for each $i \in S_n$, there is a p.m. $\pi(\cdot|i) \in \mathcal{P}(A_n(i))$ such that $\pi_t(\cdot|i) \equiv \pi(\cdot|i)$ for all $t \geqslant 0$. We denote this policy by $(\pi(\cdot|i), i \in S_n)$. We denote by Π_n the family of all randomized Markov policies and by Π_n^s the set of all stationary policies for each $n \in \overline{\mathbb{N}}$.

For each $n \in \overline{\mathbb{N}}$ and policy $\pi = (\pi_t, t \geqslant 0) \in \Pi_n$, let

$$q_n(j|i,\pi_t) := \int_{A_n(i)} q_n(j|i,a)\pi_t(da|i), \; c_n^l(i,\pi_t) := \int_{A_n(i)} c_n^l(i,a)\pi_t(da|i) \tag{2.2}$$

for each $i,j \in S_n, t \geqslant 0$, and $0 \leqslant l \leqslant p$. When π is stationary, we will write $q_n(j|i,\pi_t)$ and $c_n^l(i,\pi_t)$ as $q_n(j|i,\pi)$ and $c_n^l(i,\pi)$, respectively.

Let $Q_n(t,\pi) := [q_n(j|i,\pi_t)]$ be the associated matrix of transition rates with the (i,j)th element $q_n(j|i,\pi_t)$. As the matrix $[q_n(j|i,a)]$ is conservative and stable, so is $Q_n(t,\pi)$. Thus, Proposition C.4 in Guo and Hernández-Lerma (2009) ensures the existence of a so-called minimal transition function (see, Definition C.3 in Guo and Hernández-Lerma (2009)) $p_n(s,i,t,j,\pi)$ for $\mathcal{M}_n$ with $i,j \in S_n$ and $t \geqslant s \geqslant 0$.

To guarantee the regularity condition (i.e. $\sum_{j \in S_n} p_n(s,i,t,j,\pi) = 1$ for all $i \in S_n$ and $t \geqslant s \geqslant 0$), we impose the following so-called *drift conditions*.

Assumption 2.1. There exist a function $1 \leqslant \omega$ on S_∞ and $\omega(i) \uparrow +\infty$ as $i \to \infty$, and constants $\rho, b, L > 0$, such that

(a) $\sum_{j \in S_n} q_n(j|i,a)\omega(j) \leqslant \rho\omega(i) + b$ for all $(i,a) \in K_n, n \in \overline{\mathbb{N}}$;
(b) $q_n^*(i) \leqslant L\omega(i)$ for all $i \in S_n, n \in \overline{\mathbb{N}}$.

For each $\pi \in \Pi_n$, $n \in \overline{\mathbb{N}}$ and $\gamma_n \in \mathcal{P}(S_n)$, under Assumption 2.1, by Proposition C.9 and Theorem 2.3 in Guo and Hernández-Lerma (2009), the corresponding $p_n(s,i,t,j,\pi)$ is unique and regular, and moreover, there exists a unique probability space $(\Omega, \mathcal{F}, P_{\gamma_n}^\pi)$ and a state-action process $\{(x_t, a_t), t \geqslant 0\}$ defined on this space. The