J. Math. Anal. Appl. 455 (2017) 864–871

Contents lists available at ScienceDirect

Journal of Mathematical Analysis and Applications

www.elsevier.com/locate/jmaa

Note

Semi-Markov decision processes with limiting ratio average rewards

Sagnik Sinha [a], Prasenjit Mondal [b,*]

[a] *Mathematics Department, Jadavpur University, Kolkata-700032, India*
[b] *Mathematics Department, Government General Degree College, Ranibandh, Bankura-722135, India*

ARTICLE INFO

Article history:
Received 21 July 2016
Available online 12 June 2017
Submitted by V. Pozdnyakov

Keywords:
Semi-Markov decision process
Limiting ratio average payoff
Semi-stationary policy

ABSTRACT

We prove that a finite (state and action spaces) semi-Markov decision process with limiting ratio average (undiscounted) payoff has an optimal pure semi-stationary policy (i.e., a semi-Markov policy independent of decision epoch count). We conclude by showing (with the aid of an example) that the result cannot be strengthened further. A crude but finite step algorithm is given to compute such an optimal policy.

1. Introduction

In the field of dynamic decision problems, we restrict ourselves to the notion of limiting ratio average reward/payoff. Derman [3] has shown that a finite (state and action spaces) Markov decision process (MDP) has a pure stationary optimal policy. If we relax the Markov property in such a dynamic set up and consider the more general (and applicable) semi-Markov decision processes (SMDPs) where the sojourn time is a random variable depending not only on the present state and the action chosen there but also on the next state to which it jumps, there was no existence result in the general multichain case till date (existence of stationary optimals are available under various ergodicity constraints viz. Ross [14], Federgruen, Hordijk and Tijms [6], Schäl [15], Feinberg [7]). SMDPs or Markov renewal programs were introduced in Jewell [10] and Howard [9] and these are powerful and natural tools for the optimization of queues, production scheduling, reliability and maintenance. Jianyong and Xiaobo [11] gave an example of an undiscounted (limiting ratio average) multichain SMDP which does not have any stationary optimal policy. This particular example was solved by Mondal and Sinha [13] where they have shown that it does not have any optimal Markov policy too. Recently, Mondal [12] proved that an undiscounted absorbing SMDP (with the assumption that the single non-absorbing state does not vary with policy) has a pure semi-stationary optimal policy. In its full

* Corresponding author.
E-mail addresses: sagnik62@yahoo.co.in (S. Sinha), prasenjit1044@yahoo.com (P. Mondal).

http://dx.doi.org/10.1016/j.jmaa.2017.06.017
0022-247X/

generality, still it was unknown whether for an undiscounted SMDP, any optimal policy exists for the decision maker. In this paper, we answer this question affirmatively and prove that a general undiscounted SMDP with finite state and action spaces admits a pure semi-stationary optimal policy. The paper is organized as follows. Section 2 presents a brief description of finite SMDP with limiting ratio average payoff. In section 3, we state and prove our main results. In section 4, a finite step algorithm is given for computing a pure semi-stationary optimal policy. An example is provided to describe the algorithm and to show that our result cannot be strengthened further.

2. Finite semi-Markov decision processes

A finite (state and action spaces) SMDP is defined by a collection of objects $\langle S, A = \{A(s) : s \in S\}, p, Q, r\rangle$, where the state space $S = \{1, 2, \ldots, z\}$ is a finite set, $A(s)$ is the finite set of admissible pure actions in state $s \in S$ and for each $s, s' \in S, a \in A(s)$, $p(s'|s, a)$ represents the transition probability (i.e., $p(s'|s, a) \geq 0$ and $\sum_{s' \in S} p(s'|s, a) = 1$), whereas $Q_{ss'}(\cdot \mid a)$ is a distribution function on $[0, \infty)$, called the conditional transition (sojourn) time distribution and $r(s, a)$ is the immediate (expected) reward. The process starts at a state $s_1 \in S$ and the decision maker chooses an action $a_1 \in A(s_1)$. Consequently, he receives an immediate reward $r(s_1, a_1)$ and the system moves to a new state $s_2 \in S$ with probability $p(s_2|s_1, a_1)$ and following the transition time distribution function $Q_{s_1 s_2}(\cdot \mid a_1)$. Once the transition to s_2 occurs on the next decision epoch, the entire process, with s_1 replaced by s_2, is repeated over and over again. Thus, the SMDP proceeds over infinite time. An MDP is a particular case of an SMDP when all the transition times have identical distributions.

A history of the process up to the n-th decision epoch is defined by $h_n = (s_1, a_1, s_2, \ldots, s_{n-1}, a_{n-1}, s_n)$ for $n \geq 2$ and $h_1 = (s_1)$.
A (behavioral) policy π is defined by a sequence $\{\pi^n(\cdot \mid h_n)\}_{n=1}^{\infty}$, where $\pi^n(\cdot \mid h_n)$ specifies a probability distribution on $A(s_n)$ depending on the history h_n.
A policy $g = \{g^n\}_{n=1}^{\infty}$ is called semi-Markov if for each n, g^n depends only on s_1, s_n and the decision epoch number n.
A policy $f = \{f^n\}$ is called Markov if for each n, f^n depends only on s_n and n.
A stationary policy is a time-independent Markov policy. Such a policy f is simply denoted as $f = (f(1), f(2), \ldots, f(z))$, where for each $s \in S$, $f(s)$ specifies a probability distribution on $A(s)$ given by $f(s) = \{f(s, i) : i \in A(s)\}$, such that $f(s, i)$ is the probability of choosing action i in state s.
A semi-stationary policy is a semi-Markov policy which is time-independent i.e., if a semi-Markov policy $g(s_1, s_n, n)$ turns out to be independent of the time count n, we term it a semi-stationary policy. The terminology is justified just as we descend from the class of Markov to stationary policies with time invariance. Thus, it can be represented as $g = (f_1, f_2, \ldots, f_z)$, where f_s is a stationary policy for the initial state $s \in S$.
A policy is called pure if it is non-randomized.
Let $\Pi, \mathcal{G}^{\mathcal{S}}, \mathcal{G}^{\mathcal{SP}}, \mathcal{F}$ and $\mathcal{F}^{\mathcal{P}}$ be respectively the classes of all behavioral, semi-stationary, pure semi-stationary, stationary and pure stationary policies.
Let $(X_1, A_1, X_2, A_2, X_3, \ldots)$ be a coordinate sequence in $S \times (A \times S)^{\infty}$. Given a policy $\pi \in \Pi$ and an initial state $s \in S$, there exists a unique probability measure $P_\pi(\cdot \mid X_1 = s)$ (and hence an expectation $E_\pi(\cdot \mid X_1 = s)$) on the product σ-field of $S \times (A \times S)^{\infty}$ by Kolmogorov's extension theorem.

Definition 1. For a behavioral policy $\pi \in \Pi$, the expected limiting ratio average rewards ϕ_1 and ϕ_2 are defined as

$$\phi_1(s, \pi) = \liminf_{n \to \infty} \frac{E_\pi[\sum_{m=1}^{n} r(X_m, A_m) \mid X_1 = s]}{E_\pi[\sum_{m=1}^{n} \tau(X_m, A_m) \mid X_1 = s]} \quad \text{for all} \quad s \in S, \tag{1}$$

and