European Journal of Operational Research 224 (2013) 333–339

Contents lists available at SciVerse ScienceDirect

European Journal of Operational Research

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

Stochastics and Statistics

A basic formula for performance gradient estimation of semi-Markov decision processes

Yanjie Li [a,*], Fang Cao [b]

[a] Harbin Institute of Technology, Shenzhen Graduate School, 518055 Shenzhen, China
[b] State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, 100044 Beijing, China

ARTICLE INFO

Article history:
Received 6 August 2011
Accepted 17 August 2012
Available online 10 September 2012

Keywords:
Markov processes
Semi-Markov decision processes
Sample-path-based gradient estimation
Perturbation analysis

ABSTRACT

This paper presents a basic formula for performance gradient estimation of semi-Markov decision processes (SMDPs) under average-reward criterion. This formula directly follows from a sensitivity equation in perturbation analysis. With this formula, we develop three sample-path-based gradient estimation algorithms by using a single sample path. These algorithms naturally extend many gradient estimation algorithms for discrete-time Markov systems to continuous time semi-Markov models. In particular, they require less storage than the algorithm in the literature.

1. Introduction

The sample-path-based gradient estimation algorithms have been widely studied for Markov decision processes (MDPs). The early studies about perturbation analysis (PA) (Ho and Cao, 1991, 1983) and score function or likelihood-ratio method (Aleksandrov et al., 1968; Rubinstein, 1969) showed that the performance gradient can be obtained by analyzing a single sample path of special processes. Many efficient algorithms have been developed therein. The PA idea about performance gradient estimation was extended to Markov systems in Cao et al. (1996) and Cao and Chen (1997). Moreover, extensions of the likelihood-ratio method to regenerative processes was given in Glynn (1990) and Glynn and L'Ecuyer (1995). REINFORCE algorithm (Williams, 1992) provided a gradient-based algorithm to optimize average reward for partially observable Markov systems. A simulation-based gradient estimation algorithm was presented in Marbach and Tsitsiklis (2001) for optimizing the average reward in a finite-state Markov reward process that depends on a set of parameters. The gradient estimation algorithms with value function approximation, e.g. VAPS (Value And Policy Search) algorithm (Baird and Moore, 1998), actor-critic algorithm (Konda and Tsitsiklis, 2003), attempt to combine the advantage of gradient estimation and value function approximation. A GPOMDP (Gradient of Partially Observable Markov Decision Process) algorithm (Baxter and Bartlett, 2001) was proposed for (partially observable) Markov systems with infinite-horizon average reward.

Recently, the extensions of gradient estimation algorithms to Semi-Markov decision processes (SMDPs) have become the research focus. PA theory was extended to SMDPs and performance sensitivity formulas were given in Cao (2003). A policy gradient method for SMDPs with application to call admission control (CAC) was introduced in Singh et al. (2007) and an actor-critic algorithm applied to resource allocation was developed in Usaha and Barria (2007). Our work is inspired by the work in Cao (2005). The research results in Cao (2005) showed that a sensitivity formula from PA plays an important role in the gradient estimation algorithms for Markov decision processes (MDPs) and many gradient estimation algorithms (Marbach and Tsitsiklis, 2001; Baxter and Bartlett, 2001; Cao and Wan, 1998) for MDPs can be easily developed with the gradient formula. In this paper, we extend the results in Cao (2005) to SMDPs with continuous time.

Our main contribution of this paper is to present a basic formula for gradient estimation of SMDPs. In the earlier conference version (Li and Cao, 2011), we provide a self-contained proof for this formula. In this paper, we derive the formula by using the infinitesimal-generator based sensitivity formula in Cao (2003) and show the equivalence between the basic formula and the infinitesimal-generator based gradient formula. Based on this basic formula, we develop three gradient estimation algorithms. These gradient estimation algorithms naturally extend many gradient estimation algorithms for discrete-time MDPs to continuous time semi-Markov models and particularly these algorithms only need about half memory requirement of the algorithm that has appeared in Singh et al. (2007).

2. Semi-Markov decision process

Consider a SMDP (Puterman, 1994) on state space $\mathcal{S} = \{s_1, s_2, \ldots, s_K\}$ with a finite action space denoted by $\mathcal{A}$. Let $\tau_0, \tau_1, \ldots, \tau_n, \ldots$, with $\tau_0 = 0$, be the decision epochs and X_n, $n = 0, 1, 2, \ldots$, denote

* Corresponding author. Tel.: +86 755 26033788; fax: +86 755 26033588.
E-mail addresses: whylyj@ustc.edu (Y. Li), caofang@bjtu.edu.cn (F. Cao).

0377-2217/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ejor.2012.08.010

the state at decision epoch τ_n. At each decision epoch τ_n, if the system is in state $X_n = i \in \mathcal{S}$, an action $A_n = a$ is taken from an available action set $A(i) \subset \mathcal{A}$ according to the current policy. As a consequence of choosing a, the next decision epoch occurs within t time units, and the system state at that decision epoch equals j with probability $p(j,t|i,a)$, which means $p(j,t|i,a) = \mathcal{P}(X_{n+1} = j, \tau_{n+1} - \tau_n \leqslant t | X_n = i, A_n = a)$. The probabilities $p(j,t|i,a), i,j \in \mathcal{S}, a \in A(i)$, are called semi-Markov kernel. We refer to $\{X_0, X_1, \ldots\}$ as an embedded Markov chain of SMDP. Let $p(j|i,a)$ denote the probability that the embedded Markov chain occupies state j at the subsequent decision epoch when a is chosen in state i at the current decision epoch. Then, we have $p(j|i,a) = p(j,\infty|i,a)$. Let $F(t|i,a)$ denote the probability that the next decision epoch occurs within t time units after the current decision epoch given that action a is chosen from $A(i)$ in state i at the current decision epochs. Then, we have

$$F(t|i,a) = \sum_{j\in\mathcal{S}} p(j,t|i,a). \tag{1}$$

Between any two sequent decision epochs τ_n and τ_{n+1}, the system state may vary. The evolution process is called natural process, denoted by W_s, $\tau_n \leqslant s < \tau_{n+1}$. At each decision epoch τ_n, the system generates a fixed reward $f(X_n,A_n)$ and it accumulates additional rewards at rate $c(W_s,X_n,A_n)$ until τ_{n+1}. Let $r(i,a)$ denote the expected reward between two decision epochs, given that the system occupies state i and action a is taken at the first decision epoch. Then, we have

$$r(i,a) = f(i,a) + E\left\{\int_{\tau_n}^{\tau_{n+1}} c(W_s,X_n,A_n)ds | X_n = i, A_n = a\right\}. \tag{2}$$

For each $i \in \mathcal{S}$ and $a \in A(i)$, define $\tau(i,a)$ by

$$\tau(i,a) = E\{\tau_{n+1} - \tau_n | X_n = i, A_n = a\} = \int_0^\infty tF(dt|i,a), \tag{3}$$

which denotes the expected length of time until the next decision epoch given that action a is taken in state i at the current decision epoch.

In general, there are some parameters in the above expected rewards, expected sojourn times and transition probabilities. These parameters may come from a policy. For example, we consider a class of stationary Markov policies Π_m parameterized by θ, where θ is a tuning parameter vector. If $\mu(\theta) \in \Pi_m$, then it chooses an action a from $A(i)$ with probability $\mu(a|i,\theta)$ when state is in $i \in \mathcal{S}$ at any decision epoch. Thus, $\sum_{a\in A(i)} \mu(a|i,\theta) = 1$. We assume that $\mu(a|i,\theta), i \in \mathcal{S}, a \in A(i)$ are differentiable with respect to θ. Naturally, for a given θ, the SMDP evolves according to semi-Markov kernel $p(j,t|i,\theta) = \sum_{a\in A(i)} \mu(a|i,\theta)p(j,t|i,a), i,j \in \mathcal{S}$, and the embedded Markov chain evolves according to the transition probabilities

$$p(j|i,\theta) = \sum_{a\in A(i)} \mu(a|i,\theta)p(j|i,a), \quad i,j \in \mathcal{S}. \tag{4}$$

Denote as $P(\theta)$ the transition probability matrix of embedded Markov chain, whose (i,j)th component is $p(j|i,\theta), i,j \in \mathcal{S}$. Moreover, the expected total reward and the expected length of time between two decision epochs under policy $\mu(\theta) \in \Pi_m$ are

$$r(i,\theta) = \sum_{a\in A(i)} \mu(a|i,\theta)r(i,a), \tag{5}$$

$$\text{and} \quad \tau(i,\theta) = \sum_{a\in A(i)} \mu(a|i,\theta)\tau(i,a), \tag{6}$$

respectively, given that the system occupies state i at the current decision epoch. Let $r(\theta)$ and $\tau(\theta)$ denote their corresponding column vectors, respectively.

We assume that the embedded Markov chain is ergodic under any θ. Let $\pi(\theta) = (\pi(s_1,\theta), \pi(s_2,\theta), \ldots, \pi(s_K,\theta))$ denote the (row) vector representing the steady-state probability of embedded Markov chain, then we have $\pi(\theta)P(\theta) = \pi(\theta)$ and $\pi(\theta)e = 1$, where e denotes a column vector whose all components are 1. Let σ_s denote the number of decision epochs up to time s. The infinite-horizon average-reward is defined as

$$\eta(i,\theta) = \lim_{t\to\infty} \frac{1}{t} E\left\{\int_0^t c(W_s,X_{\sigma_s},A_{\sigma_s})ds + \sum_{n=0}^{\sigma_t - 1} f(X_n,A_n) | X_0 = i\right\}, \quad \forall i \in \mathcal{S}. \tag{7}$$

It has been shown (Puterman, 1994) that the average reward (7) is independent of initial state i under the ergodic assumption and equals

$$\eta(i,\theta) = \eta(\theta) := \frac{\pi(\theta)r(\theta)}{\pi(\theta)\tau(\theta)}, \quad \forall i \in \mathcal{S}. \tag{8}$$

Our goal is to estimate the gradient of average reward (8) with respect to parameter vector θ based on a single sample path of SMDP.

3. Performance gradient formula

In this section, on the basis of reviewing the results about performance sensitivity of SMDPs (Cao, 2003, 2007), we present a performance gradient formula for SMDPs under average-reward performance.

For the SMDP in the above section, an infinitesimal generator $B(\theta)$ was defined in Cao (2003, 2007). Its (i,j)th component is $b(j|i,\theta) = \frac{1}{\tau(i,\theta)}[p(j|i,\theta) - \delta_{ij}]$, where δ_{ij} is a function with $\delta_{ij} = 1$ when $i = j$, otherwise 0. Thus, the infinitesimal generator can be described in vector form as follows:

$$B(\theta) = \Gamma(\theta)[P(\theta) - I], \tag{9}$$

where $\Gamma(\theta)$ is a diagonal matrix with diagonal components $\frac{1}{\tau(s_1,\theta)}, \ldots, \frac{1}{\tau(s_K,\theta)}$ and I is an identity matrix. Assume the parameter vector θ is perturbed to θ', which corresponds to another policy $\mu(\theta')$ and thus corresponds to another semi-Markov process. Based on the infinitesimal generator in (9), the average-reward performance difference under two different policies $\mu(\theta) \in \Pi_m$ and $\mu(\theta') \in \Pi_m$ can be given by the following formula (Cao, 2003),

$$\eta(\theta') - \eta(\theta) = p(\theta')[\Gamma(\theta')r(\theta') - \Gamma(\theta)r(\theta) + (B(\theta') - B(\theta))g(\theta)], \tag{10}$$

where $p(\theta') = (p(1,\theta'), \ldots, p(s_K,\theta'))$ is the steady-state distribution of $B(\theta')$, i.e., $p(\theta')B(\theta') = 0$, $p(\theta')e = 1$, and $g(\theta) = (g(1,\theta), \ldots, g(s_K,\theta))^T$ is the performance potential satisfying the Poisson equation

$$B(\theta)g(\theta) = -\Gamma(\theta)r(\theta) + \eta(\theta)e, \tag{11}$$

where T denotes the transpose. Since the steady state distribution $p(\theta')$ of infinitesimal generator can be described by the steady state distribution of embedded Markov chain as follows (Ross, 1996):

$$p(i,\theta') = \frac{\pi(i,\theta')\tau(i,\theta')}{\pi(\theta')\tau(\theta')},$$

then from (10), we have

$$\begin{aligned}
&\eta(\theta') - \eta(\theta) \\
&= \sum_{i\in\mathcal{S}} \frac{\pi(i,\theta')\tau(i,\theta')}{\pi(\theta')\tau(\theta')} \left\{\frac{r(i,\theta')}{\tau(i,\theta')} - \frac{r(i,\theta)}{\tau(i,\theta)} + \sum_{j\in\mathcal{S}} [b(j|i,\theta') - b(j|i,\theta)]g(j,\theta)\right\} \\
&= \sum_{i\in\mathcal{S}} \frac{\pi(i,\theta')}{\pi(\theta')\tau(\theta')} \left\{r(i,\theta') + \sum_{j\in\mathcal{S}} p(j|i,\theta')g(j,\theta) - g(i,\theta)\right. \\
&\quad \left. - \left[r(i,\theta) + \sum_{j\in\mathcal{S}} p(j|i,\theta)g(j,\theta) - g(i,\theta)\right] \frac{\tau(i,\theta')}{\tau(i,\theta)}\right\} \\
&= \sum_{i\in\mathcal{S}} \frac{\pi(i,\theta')}{\pi(\theta')\tau(\theta')} \left\{r(i,\theta') - r(i,\theta) + \sum_{j\in\mathcal{S}} [p(j|i,\theta') - p(j|i,\theta)]g(j,\theta)\right. \\
&\quad \left. - \left[r(i,\theta) + \sum_{j\in\mathcal{S}} p(j|i,\theta)g(j,\theta) - g(i,\theta)\right] \frac{\tau(i,\theta') - \tau(i,\theta)}{\tau(i,\theta)}\right\} \\
&= \sum_{i\in\mathcal{S}} \frac{\pi(i,\theta')}{\pi(\theta')\tau(\theta')} \left\{r(i,\theta') - r(i,\theta) + \sum_{j\in\mathcal{S}} [p(j|i,\theta') - p(j|i,\theta)]g(j,\theta)\right. \\
&\quad \left. - \eta(\theta)[\tau(i,\theta') - \tau(i,\theta)]\right\}.
\end{aligned}$$