Automatica 67 (2016) 77–84

Contents lists available at ScienceDirect

Automatica

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

Brief paper

A unified approach to time-aggregated Markov decision processes☆

Yanjie Li [a,1], Xinyu Wu [b]

[a] Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China
[b] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

ARTICLE INFO

Article history:
Received 30 December 2013
Received in revised form
26 October 2015
Accepted 7 December 2015
Available online 4 February 2016

Keywords:
Time aggregation
Performance sensitivity
Markov decision process
Semi-Markov decision process

ABSTRACT

This paper presents a unified approach to time-aggregated Markov decision processes (MDPs) with an average cost criterion. The approach is based on a framework in which a time-aggregated MDP constitutes a semi-Markov decision process (SMDP). By analyzing the performance sensitivity formulas of this SMDP, a number of optimization algorithms for time aggregated MDPs, including those previously reported in the literature, can be developed in a simple and intuitive way.

1. Introduction

Markov decision processes (MDPs) often serve as common models, and are widely applied in a variety of fields, including control, artificial intelligence and operations research (Puterman, 1994; Sutton & Barto, 1998). The major difficulty in solving MDPs is a problem called the "curse of dimensionality" (Puterman, 1994). Reducing the dimensionality of a state space can substantially improve the computational efficiency of MDP solutions. The time aggregation approach (Cao, Ren, Bhatnagar, Fu, & Marcus, 2002) affords MDPs state reduction by dividing the process into time segments according to certain state subsets. Performance gradient estimation for Markov processes with time aggregation was presented using the stochastic recursive method and likelihood ratio in Zhang and Ho (1991). In addition, a number of optimization algorithms, including policy iteration (Cao et al., 2002; Ren & Krogh, 2005) and value iteration algorithms (Arruda & Fragoso, 2011; Ren & Krogh, 2005; Sun, Zhao, & Luh, 2007), have been developed for time-aggregated MDPs. However, the aforementioned algorithms were proposed independently of one another, and the relationship between them remains unclear. The objectives of this paper are to provide a unified formulation from the performance sensitivity point of view proposed in Cao (2007) to relate the previously reported algorithms systematically and to propose new optimization algorithms for time-aggregated MDPs.

We first show that a time-aggregated MDP essentially constitutes a semi-Markov decision process (SMDP). Then, we present a unified approach to time-aggregated MDP using performance sensitivity of this SMDP. Our approach is motivated by the sensitivity-based approach (Cao, 2007; Cao & Chen, 1997), where performance sensitivity formulas provide a unified framework for MDPs. An infinitesimal generator-based performance sensitivity formula was proposed for SMDPs in Cao (2003), which we call a continuous time-type formula in this paper. We then present a discrete time-type performance sensitivity formula. By analyzing these performance sensitivity formulas, we propose a unified approach to time-aggregated MDPs from the continuous-time and discrete-time perspectives. The proposed approach unifies and develops a number of optimization algorithms for time-aggregated MDPs, including those previously reported in the literature, in an intuitive and simple way. This approach is an extension of the sensitivity-based approach (Cao, 2007), and provides new insights to time-aggregated MDPs. Its significance can be described as follows: (1) A unified formulation for policy iteration algorithms is obtained by directly comparing two types of performance difference formulas, an approach that is more intuitive and simple than those in the previous literature. (2) Different value iteration algorithms are investigated in a unified way. This unification demonstrates the differences in the development of value iteration algorithms using two types of Bellman optimality equations. On this basis, we present a stochastic shortest path (SSP) value iteration and a generalized standard value iteration. The SSP value iteration preserves

☆ The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Bart De Schutter under the direction of Editor Ian R. Petersen.

E-mail addresses: lyj@hitsz.edu.cn (Y. Li), xy.wu@siat.ac.cn (X. Wu).

[1] Tel.: +86 755 26033788; fax: +86 755 26033774.

http://dx.doi.org/10.1016/j.automatica.2015.12.022
0005-1098/

the weighted sup-norm contraction property (Bertsekas, 1998), which is helpful for developing asynchronous iterations. The generalized standard value iteration can be more intuitively understood as a traditional value iteration than the data-transformation method in Arruda and Fragoso (2011) or Puterman (1994), and it obviates the need to solve several average-cost MDPs or SSPs during the process of value iteration. (3) Finally, the proposed approach provides a performance gradient-based optimization algorithm that can be applied to cases in which the transition probabilities are unknown.

2. Time-aggregated Markov decision processes

We briefly describe the standard MDP and time-aggregated MDP by following the notations in Cao et al. (2002). Consider a time-homogeneous discrete-time MDP $\mathbf{X} = \{X_t, t = 0, 1, \ldots\}$ on a finite state space $\mathcal{S} = \{1, 2, \ldots, M\}$. At any transition time t with $X_t = i \in \mathcal{S}$, action a is taken from a feasible action space $\mathcal{A}$. That action determines the transition probabilities $p^a(i, j)$ from state i to state j, and a cost $f(i, a)$ is incurred. In this paper, we consider a set of stationary policies Π_s, which means that a policy $\mathcal{L} \in \Pi_s$ is a mapping from state space $\mathcal{S}$ to action space $\mathcal{A}$. Thus, policy $\mathcal{L}$ determines the evolution of the MDP by following the transition probability matrix $P^{\mathcal{L}}$ with $P^{\mathcal{L}} = [p^{\mathcal{L}(i)}(i, j)]_{i,j=1}^{M}$ and cost vector $f^{\mathcal{L}} = [f(1, \mathcal{L}(1)), \ldots, f(M, \mathcal{L}(M))]^T$, where the superscript "T" denotes the transpose.

Assume that the MDP is ergodic under any policy. Let $\pi^{\mathcal{L}} = [\pi^{\mathcal{L}}(1), \ldots, \pi^{\mathcal{L}}(M)]$ be the row vector representing the steady-state probability of Markov chain $\mathbf{X}$ under policy $\mathcal{L}$. Then, we have

$$\pi^{\mathcal{L}} P^{\mathcal{L}} = \pi^{\mathcal{L}} \quad \text{and} \quad \pi^{\mathcal{L}} e = 1,$$

where $e = [1, 1, \ldots, 1]^T$. We consider the following average cost performance, which is well defined and does not depend on the initial state,

$$\eta^{\mathcal{L}} = \lim_{T\to\infty} \frac{1}{T} E^{\mathcal{L}} \left[\sum_{t=0}^{T-1} f(X_t, A_t) | X_0 = i \right] = \pi^{\mathcal{L}} f^{\mathcal{L}}, \tag{1}$$

for $\forall\, i \in \mathcal{S}$, where $E^{\mathcal{L}}$ denotes the expectation under policy $\mathcal{L}$ and A_t denotes the action at time t. The objective is to find an optimal policy $\mathcal{L}^*$ that minimizes the average cost $\eta^{\mathcal{L}}$, i.e., $\mathcal{L}^* \in \arg\min_{\mathcal{L}\in\Pi_s} \eta^{\mathcal{L}}$.

The time aggregation approach (Cao et al., 2002) assumes that state space $\mathcal{S}$ can be divided into subset $\mathcal{S}_1$ and its complementary set $\mathcal{S}_2 = \mathcal{S} - \mathcal{S}_1$. Actions can be taken only for the states in $\mathcal{S}_1$. Without loss of generality, let $\mathcal{S}_1 = \{1, 2, \ldots, M_1\}$ and $\mathcal{S}_2 = \{M_1 + 1, \ldots, M\}$. Under these assumptions, $P^{\mathcal{L}}$ and $f^{\mathcal{L}}$ can be partitioned according to $\mathcal{S}_1$ and $\mathcal{S}_2$, as follows

$$P^{\mathcal{L}} = \begin{bmatrix} P^{\mathcal{L}}_{\mathcal{S}_1\mathcal{S}_1} & P^{\mathcal{L}}_{\mathcal{S}_1\mathcal{S}_2} \\ P_{\mathcal{S}_2\mathcal{S}_1} & P_{\mathcal{S}_2\mathcal{S}_2} \end{bmatrix} \quad \text{and} \quad f^{\mathcal{L}} = \begin{bmatrix} f^{\mathcal{L}}_{\mathcal{S}_1} \\ f_{\mathcal{S}_2} \end{bmatrix}.$$

Define $\tau_0 = 0$ and $\tau_l = \min\{t > \tau_{l-1} | X_t \in \mathcal{S}_1\}, l = 1, 2, \ldots$, as the time points at which the process arrives in $\mathcal{S}_1$, and denote by $\mathbf{Y} = \{Y_l, l = 0, 1, \ldots\}$ the aggregated chain that records the states visited at those points. Thus, the relation between $\mathbf{X}$ and $\mathbf{Y}$ is given by $Y_l = X_{\tau_l}$. Let $\tilde{P}^{\mathcal{L}}$ and $\tilde{\pi}^{\mathcal{L}}$ be the transition matrix and steady-state probability row vector of aggregated chain $\mathbf{Y}$ under policy $\mathcal{L}$. From Cao et al. (2002),

$$\tilde{P}^{\mathcal{L}} = P^{\mathcal{L}}_{\mathcal{S}_1\mathcal{S}_1} + P^{\mathcal{L}}_{\mathcal{S}_1\mathcal{S}_2} (I - P_{\mathcal{S}_2\mathcal{S}_2})^{-1} P_{\mathcal{S}_2\mathcal{S}_1}. \tag{2}$$

Time points $\tau_l, l = 0, 1, 2, \ldots$, divide the process into numerous segments $(X_{\tau_l}, X_{\tau_l+1}, \ldots, X_{\tau_{l+1}-1})$. In each segment, the action needs to be chosen only in the first state and no decisions are made in other states. Thus, the expected total cost of a segment that starts from $i \in \mathcal{S}_1$ and $a \in \mathcal{A}$ is

$$H_f(i, a) = E\left[R_l \middle| X_{\tau_l} = i, A_{\tau_l} = a \right], \quad i \in \mathcal{S}_1,\ l = 0, 1, 2, \ldots,$$

where $R_l = f(X_{\tau_l}, A_{\tau_l}) + \sum_{k=1}^{\tau_{l+1}-\tau_l-1} f(X_{\tau_l+k})$. Let $H_f^{\mathcal{L}} = [H_f(1, \mathcal{L}(1)), \ldots, H_f(M_1, \mathcal{L}(M_1))]^T$ denote the expected total cost vector under policy $\mathcal{L}$, which can be computed by (see Cao et al., 2002)

$$H_f^{\mathcal{L}} = f^{\mathcal{L}}_{\mathcal{S}_1} + P^{\mathcal{L}}_{\mathcal{S}_1\mathcal{S}_2} (I - P_{\mathcal{S}_2\mathcal{S}_2})^{-1} f_{\mathcal{S}_2}. \tag{3}$$

Let $H_1^{\mathcal{L}} = [H_1(1, \mathcal{L}(1)), \ldots, H_1(M_1, \mathcal{L}(M_1))]^T$ denote the case that $f(i, a) = 1, i \in \mathcal{S}_1, a \in \mathcal{A}$ and $f(i) = 1, i \in \mathcal{S}_2$, and thus $H_1(i, \mathcal{L}(i))$ is the expected length of a segment that starts from $i \in \mathcal{S}_1$ under policy $\mathcal{L}$. The average cost of the original MDP defined in (1) can be computed by (see Cao et al., 2002)

$$\eta^{\mathcal{L}} = \frac{\tilde{\pi}^{\mathcal{L}} H_f^{\mathcal{L}}}{\tilde{\pi}^{\mathcal{L}} H_1^{\mathcal{L}}}, \quad \text{for any } \mathcal{L} \in \Pi_s. \tag{4}$$

After applying the time aggregation technique, the original MDP essentially constitutes an SMDP as depicted in Fig. 1. The corresponding SMDP has a state space $\mathcal{S}_1$ and an action space $\mathcal{A}$. The time points $\tau_0, \tau_1, \ldots$ in the MDP are the successive decision epochs of the SMDP. The state evolution between τ_l and τ_{l+1}, $l = 0, 1, 2, \ldots$, is the natural process in the SMDP, whereas aggregated chain $\mathbf{Y}$ constitutes an embedded Markov chain of the SMDP. Time intervals $\tau_{l+1} - \tau_l, l = 0, 1, 2 \ldots$ are the sojourn times of the SMDP, and its expected value is $H_1(i, a)$ if $a \in \mathcal{A}$ is taken in the starting state i of the segment. The total cost $H_f(i, a)$ is the accumulated expected cost between two successive decision epochs, given that the SMDP occupies state i, and action a is taken in the first decision epoch.

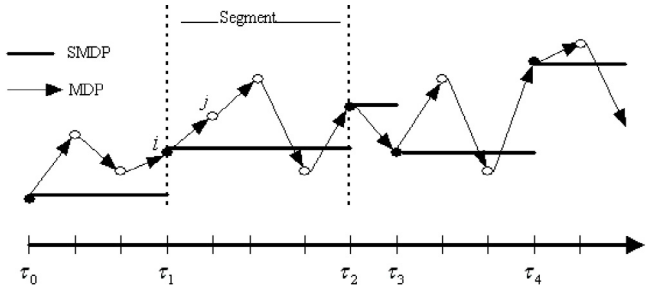

Fig. 1. Time-aggregated MDP and its corresponding SMDP.

3. Performance sensitivity

In this section, we analyze the structure of performance sensitivity of the time-aggregated MDP using the SMDP.

An infinitesimal generator $A^{\mathcal{L}} = \Lambda^{\mathcal{L}}(\tilde{P}^{\mathcal{L}} - I)$ is defined in Cao (2003), where $\Lambda^{\mathcal{L}} = \text{diag}\{\frac{1}{H_1(1,\mathcal{L}(1))}, \ldots, \frac{1}{H_1(M_1,\mathcal{L}(M_1))}\}$. Let $p^{\mathcal{L}}$ be the steady-state probability row vector of the SMDP, and then $p^{\mathcal{L}}$ satisfies $p^{\mathcal{L}} A^{\mathcal{L}} = 0$, $p^{\mathcal{L}} e = 1$. From Ross (1996), we have

$$p^{\mathcal{L}} = \frac{\tilde{\pi}^{\mathcal{L}} (\Lambda^{\mathcal{L}})^{-1}}{\tilde{\pi}^{\mathcal{L}} H_1^{\mathcal{L}}}.$$

Define a cost-rate function vector under policy $\mathcal{L}$ as $\Lambda^{\mathcal{L}} H_f^{\mathcal{L}}$. Then, we have

$$\eta^{\mathcal{L}} = \frac{\tilde{\pi}^{\mathcal{L}} (\Lambda^{\mathcal{L}})^{-1}}{\tilde{\pi}^{\mathcal{L}} H_1^{\mathcal{L}}} \Lambda^{\mathcal{L}} H_f^{\mathcal{L}} = p^{\mathcal{L}} \Lambda^{\mathcal{L}} H_f^{\mathcal{L}}.$$

Thus, performance (4) is equivalent to the average cost performance with cost-rate function $\Lambda^{\mathcal{L}} H_f^{\mathcal{L}}$ in Cao (2003).