APPLICATIONES MATHEMATICAE
42,2-3 (2015), pp. 205–218

J. Adolfo Minjárez-Sosa and José A. Montoya (Hermosillo)

BAYESIAN ESTIMATION OF THE MEAN HOLDING TIME IN AVERAGE SEMI-MARKOV CONTROL PROCESSES

Abstract. We consider semi-Markov control models with Borel state and action spaces, possibly unbounded costs, and holding times with a generalized exponential distribution with unknown mean θ. Assuming that such a distribution does not depend on the state-action pairs, we introduce a Bayesian estimation procedure for θ, which combined with a variant of the vanishing discount factor approach yields average cost optimal policies.

1. Introduction. This paper deals with a class of semi-Markov control models (SMCM) in which the holding time process $\{\delta_n\}$ is a sequence of independent and identically distributed (i.i.d.) random variables with a generalized exponential density g with unknown mean θ, independent of the state-action pairs. Assuming that the costs are possibly unbounded, our objective is to construct optimal policies under the long-run average criterion.

As is observed in Section 3, because g does not depend on the state-action process, the corresponding semi-Markov control problem becomes a Markov control problem. Furthermore, as usual in the study of the average criterion for SMCM (see, e.g., [12, 13, 16]), the cost-per-stage depends on the holding times δ_n, and by using properties of conditional expectation, it can be written in terms of the mean θ. Hence, since θ is unknown, we are facing a Markov optimal control problem with unknown cost-per-stage. In this sense, before choosing the action at the nth decision epoch, the controller gets an estimate θ_n of the parameter θ, and therefore of the cost, and combines this with the history of the system to select an action $a = a_n(\theta_n)$.

The estimation of the parameter θ is obtained by the application of a Bayesian procedure, which is based on the minimization of the so-called

2010 *Mathematics Subject Classification*: Primary 93E10; Secondary 90C40.
Key words and phrases: semi-Markov control processes, average cost criterion, optimal policy, Bayes estimator.

DOI: 10.4064/am42-2-7

Bayes' risk function. This scheme includes a posterior model of the unknown parameter given the observation and a cost-of-error function.

It is worth observing that the hypotheses of independence of the density g and the state-action process could be strong in some application problems. However, it is satisfied, for instance, in some class of queueing systems (see, e.g., [16]).

The average optimality is studied by means of a combination of a variant of the so-called vanishing discount factor approach (see, e.g., [4, 8]) with the Bayesian estimation procedure. Specifically, to construct the average cost optimal policy, we analyze the $\hat{\alpha}_n$-discount optimality equation for a suitable sequence of discount factors $\hat{\alpha}_n \nearrow 1$, and replace the unknown parameter θ by the estimator θ_n obtained at each decision time. The idea of such an approach was originally introduced by Gordienko in [3] and revised in [7], both for Markov control processes with bounded cost and unknown transition law (see also [14]). In this work we use some of their main ideas, but extended to the unbounded and unknown one-stage cost case.

There are similar papers dealing Bayesian estimation in Markov control models under average criterion (see, e.g., [1, 2]). However, unlike our work, in both papers it is assumed that the transition probability among the states depends on an unknown parameter which must be estimated in order to obtain nearly optimal policies.

The paper is organized as follows. In Section 2 we describe the SMCM we will be dealing with. Next, Section 3 contains preliminary results on the average optimality criterion and the assumptions required. The Bayesian estimation procedure is introduced in Section 4, whereas the construction of the optimal policies is presented in Section 5. Finally, the proofs are given in Section 6.

Notation. Given a Borel space $\mathbb{X}$ (that is, a Borel subset of a complete and separable metric space) its Borel σ-algebra is denoted by $\mathcal{B}(\mathbb{X})$, and "measurable", for either sets or functions, means "Borel measurable". Let $\mathbb{X}$ and $\mathbb{Y}$ be Borel spaces. Then a stochastic kernel $Q(dx \mid y)$ on $\mathbb{X}$ given $\mathbb{Y}$ is a function such that $Q(\cdot \mid y)$ is a probability measure on $\mathbb{X}$ for each fixed $y \in \mathbb{Y}$, and $Q(B \mid \cdot)$ is a measurable function on $\mathbb{Y}$ for each fixed $B \in \mathcal{B}(\mathbb{X})$. We denote by $\mathbb{N}$ (respectively $\mathbb{N}_0$) the set of positive (resp. nonnegative) integers; $\mathbb{R}$ (respectively $\mathbb{R}_+$) denotes the set of real (resp. nonnegative real) numbers.

2. The control model. The semi-Markov control model we are concerned with is described by the following elements.

(a) The *state space* $\mathbb{X}$ and the *control set* $\mathbb{A}$ are assumed to be Borel spaces.

To each $x \in \mathbb{X}$, we associate a nonempty measurable subset $A(x)$ of $\mathbb{A}$ denoting the set of *admissible controls* (or *actions*) when the system is in state x. The set

$$\mathbb{K} = \{(x,a) : x \in \mathbb{X},\, a \in A(x)\}$$

of *admissible state-action pairs* is a Borel subset of the Cartesian product of $\mathbb{X}$ and $\mathbb{A}$. We assume that for each $x \in \mathbb{X}$, the set $A(x)$ is compact.

(b) The *transition law* among the states $Q(\cdot \mid \cdot)$ is a stochastic kernel on $\mathbb{X}$ given $\mathbb{K}$. That is, if x is the state at the nth decision time and the controller selects the action a, $Q(\cdot \mid x,a)$ is the distribution of the next state of the system:

$$Q(B \mid x,a) := \Pr[x_{n+1} \in B \mid x_n = x,\, a_n = a], \qquad B \in \mathcal{B}(\mathbb{X}).$$

We assume that Q is strongly continuous on $A(x)$ for every $x \in \mathbb{X}$, that is, for each bounded measurable function $v : \mathbb{X} \to \mathbb{R}$, the function

$$a \mapsto \int_{\mathbb{X}} v(y)\, Q(dy \mid x,a)$$

is continuous on $A(x)$.

(c) The time of the nth decision is denoted by T_n where $T_0 = 0$. Thus $\delta_{n+1} := T_{n+1} - T_n$, $n \in \mathbb{N}_0$, are random variables defined on a probability space $(\Omega, \mathcal{F}, P)$ representing the sojourn or holding time at state x_n. Moreover, observe that $T_n = \sum_{k=1}^{n} \delta_k$. We assume that $\{\delta_n\}$ is a sequence of independent and identically distributed (i.i.d.) random variables such that $\delta_n \geq m$ a.s. for some $m > 0$, with a generalized exponential density of the form

$$g(s|\lambda) = \lambda^{-1} \exp[-(s-m)\lambda^{-1}]\mathbb{1}_{[m,\infty)}(s),$$

which is independent of the state-action pairs, where $\lambda > 0$ is an unknown parameter. Therefore, the mean holding time $\theta := E(\delta_n) = \lambda + m$ is unknown by the controller.

(d) Finally, the cost-per-stage C is a measurable and possibly unbounded real-valued function on $\mathbb{K}$. In particular we assume that C is the sum of an immediate cost $D(x,a)$ incurred at the moment when the controller chooses a decision, plus a cost rate $d(x,a)$ imposed until the transition to a new state of the system occurs. As is specified below (see (6) and (9)–(11)), the cost C takes the form

$$C(x,a) = C_\theta(x,a) = D(x,a) + \theta d(x,a), \qquad (x,a) \in \mathbb{K}. \tag{1}$$

We assume that both functions $D(x,\cdot)$ and $d(x,\cdot)$ are lower semicontinuous (l.s.c.) on $A(x)$ for every $x \in \mathbb{X}$, and there exists a measurable function $W : \mathbb{X} \to [1,\infty)$ such that for all $(x,a) \in \mathbb{K}$,

$$|D(x,a)| \leq W(x) \quad \text{and} \quad |d(x,a)| \leq W(x). \tag{2}$$

Therefore

$$|C_\theta(x,a)| \le (1+\theta)W(x). \tag{3}$$

In addition, we suppose that the mapping

$$(x,a) \mapsto \int_{\mathbb{X}} W(y)\, Q(dy \mid x,a)$$

is continuous on $\mathbb{K}$, and

$$\int_{\mathbb{X}} W^2(y)\, Q(dy \mid x,a) \le \bar{\beta} W^2(x) + \bar{d}, \qquad (x,a)\in\mathbb{K}, \tag{4}$$

for some constants $\bar{\beta}\in(0,1)$ and $\bar{d}\ge 0$. Observe that from Jensen's inequality, (4) yields

$$\int_{\mathbb{X}} W(y)\, Q(dy \mid x,a) \le \beta W(x) + d, \qquad (x,a)\in\mathbb{K}, \tag{5}$$

where $\beta = (\bar{\beta})^{1/2}$ and $d = (\bar{d})^{1/2}$.

The evolution in time of the system is as follows. At time T_n of the nth decision, the system is in state $x_n = x$. Since θ (and therefore the cost function C_θ) is unknown, by using the historical observations of the holding times $\delta_1,\ldots,\delta_n$, the controller implements a Bayesian inference procedure to construct an estimator θ_n of the mean holding time θ, and combines this with the control objectives to select a control $a = a_n(\theta_n) \in A(x)$. Then the system remains in state x during a nonnegative random time δ_{n+1} with density function g, the system jumps to a new state $x_{n+1} = y$ according to the transition law $Q(\cdot \mid x,a)$, and the cost $C_\theta(x,a)$ is incurred. Next the process is repeated. Furthermore, the costs are accumulated throughout the evolution of the system in an infinite time horizon using an average cost criterion defined below.

3. Average optimality criterion. The actions applied by the controller are selected according to rules known as control policies. We denote by Π the set of all control policies and by $\mathbb{F}\subset\Pi$ the subset of stationary policies. If necessary, see for instance [8, 9] for further information about those policies. Following a standard convention, every stationary policy $\pi\in\mathbb{F}$ is identified with some measurable function $f:\mathbb{X}\to A$ such that $f(x)\in A(x)$, $x\in\mathbb{X}$, so that π is of the form $\pi=\{f,f,\ldots\}$. In this case we denote π by f, and moreover, to ease the notation, we write

$$C_\theta(x,f) := C_\theta(x,f(x)) \quad \text{and} \quad Q(\cdot \mid x,f) := Q(\cdot \mid x,f(x)), \qquad x\in\mathbb{X}.$$

For $x\in\mathbb{X}$ and $\pi\in\Pi$, we define the *long-run expected average cost* as

$$J(\pi,x) := \limsup_{n\to\infty} \frac{E_x^\pi[\sum_{k=0}^{n-1}\{D(x_k,a_k)+\delta_{k+1}d(x_k,a_k)\}]}{E_x^\pi[T_n]}, \tag{6}$$

where E_x^π denotes the expectation operator with respect to the probability measure P_x^π induced by the policy π, given the initial state $x_0 = x$. The semi-Markov optimal control problem (OCP) is then to find an optimal policy $\pi^* \in \Pi$ such that

$$J(x) = J(\pi^*, x) \quad \text{for all } x \in \mathbb{X}, \tag{7}$$

where

$$J(x) := \inf_{\pi \in \Pi} J(\pi, x) \tag{8}$$

is the *optimal average cost function.* A policy $\pi^* \in \Pi$ satisfying (7) and (8) is said to be *average cost optimal* (AC-optimal).

It is worth noting that the semi-Markov OCP (6)–(8) can be seen as a Markov OCP. Indeed, first observe that considering the cost C_θ defined in (1) and using the properties of conditional expectation, we can write the performance index (6) as

$$J(\pi, x) = \limsup_{n\to\infty} \frac{E_x^\pi[\sum_{k=0}^{n-1} C_\theta(x_k, a_k)]}{n\theta}. \tag{9}$$

Thus,

$$J(\pi, x) = \limsup_{n\to\infty} \frac{E_x^\pi[\sum_{k=0}^{n-1} \bar{C}_\theta(x_k, a_k)]}{n}, \tag{10}$$

where

$$\bar{C}_\theta(x, a) := \frac{C_\theta(x, a)}{\theta}, \tag{11}$$

which represents the average performance index of a Markov control process. However, since θ is unknown, we have a Markov OCP with unknown one-stage cost (11). In this sense, our approach consists in combining a suitable statistical estimation process of the parameter θ with a control procedure to construct an AC-optimal control policy.

Observe that, because $\theta > m$, from (3) we get, for all $(x, a) \in \mathbb{K}$,

$$|\bar{C}_\theta(x, a)| \le MW(x),$$

where $M := 1/m + 1$.

The average optimality is studied by means of the vanishing discount factor approach. For each $\pi \in \Pi$ and $x \in \mathbb{X}$, we define the total expected α-discounted cost as

$$V_\alpha(\pi, x) := E_x^\pi\Big[\sum_{n=0}^{\infty} \alpha^n \bar{C}_\theta(x_n, a_n)\Big], \tag{12}$$

where $\alpha \in (0, 1)$ is the so-called *discount factor*. The corresponding α-value function is defined as

$$V_\alpha(x) := \inf_{\pi\in\Pi} V_\alpha(\pi, x), \quad x \in \mathbb{X}. \tag{13}$$

To analyze the asymptotic behavior for the average optimality criterion we need the following ergodicity condition.

For every $f \in \mathbb{F}$, $B \in \mathcal{B}(\mathbb{X})$, $x \in \mathbb{X}$, and $n \geq 0$, we define

$$Q_f^n(B \mid x) := Q^n(B \mid x, f) = P_x^f(x_n \in B).$$

Observe that $Q_f^0(B \mid x) = \mathbb{1}_B(x)$. In addition, we denote by $\mathbb{B}_W(\mathbb{X})$ the normed linear space of all measurable functions $u : \mathbb{X} \to \mathbb{R}$ with norm

$$\|u\|_W := \sup_{x\in\mathbb{X}} \frac{|u(x)|}{W(x)} < \infty. \tag{14}$$

Assumption 1 (W-geometric ergodicity). For every $f \in \mathbb{F}$, $x \in \mathbb{X}$, and $u \in \mathbb{B}_W(\mathbb{X})$, there exists a probability measure μ_f on $\mathbb{X}$ such that

$$\Big|\int_{\mathbb{X}} u(y)\, Q_f^n(dy \mid x) - \mu_f(u)\Big| \leq \|u\|_W R\rho^n W(x), \qquad n \geq 0,$$

where $\mu_f(u) := \int_{\mathbb{X}} u\, d\mu_f$, and $R > 0$ and $0 < \rho < 1$ are constants independent of f.

In [4, 5, 9, 10, 11] sufficient conditions for the geometric ergodicity property are given, and as a consequence we have the following results.

Proposition 2. *There exists a constant $M' > 0$ such that*

$$V_\alpha(x) \leq \frac{M'W(x)}{1-\alpha}, \qquad x \in X.$$

In addition, $V_\alpha \in \mathbb{B}_W$ satisfies the discounted optimality equation $V_\alpha(x) = T_\alpha^\theta V_\alpha(x)$, where

$$T_\alpha^\theta u(x) = \inf_{a\in A(x)} \Big[\bar{C}_\theta(x,a) + \alpha \int_{\mathbb{X}} u(y)\, Q(dy \mid x, a)\Big], \qquad x \in \mathbb{X},\ u \in \mathbb{B}_W. \tag{15}$$

Moreover, if Assumption 1 holds, there exist a constant j^* and a function $h \in \mathbb{B}_W$ such that

$$j^* + h(x) \geq \inf_{a\in A(x)} \Big[\bar{C}_\theta(x,a) + \int_{\mathbb{X}} h(y)\, Q(dy \mid x, a)\Big], \qquad x \in \mathbb{X}, \tag{16}$$

and j^* is the optimal average cost, i.e.,

$$j^* = \inf_{\pi\in\Pi} J(\pi, x) \quad \textit{for all } x \in \mathbb{X}.$$

Fix an arbitrary state $z \in X$, and define, for $\alpha \in (0,1)$,

$$j_\alpha := (1-\alpha)V_\alpha(z) \ \text{ and } \ h_\alpha(x) := V_\alpha(x) - V_\alpha(z), \qquad x \in \mathbb{X}. \tag{17}$$

It is easy to see that the discounted optimality equation (15) is equivalent to

$$j_\alpha + h_\alpha(x) = T_\alpha^\theta h_\alpha(x), \qquad x \in \mathbb{X},\ \alpha \in (0,1).$$

Following standard arguments in the literature on average cost Markov control processes (e.g., [4, 5, 9, 11]) we deduce that for any sequence $\{\alpha_n\}$ of discount factors such that $\alpha_n \nearrow 1$,

$$\lim_{n\to\infty} j_{\alpha_n} = j^*, \tag{18}$$

and

$$\sup_{\alpha\in(0,1)} \|h_\alpha\|_W < \infty. \tag{19}$$

4. Bayesian estimation. Let $\delta_1,\ldots,\delta_n$ be independent random variables observed up to the nth decision time with density g and unknown mean $\theta = \lambda + m$. We assume that the prior distribution on λ is Inv-Gamma(μ_0,β_0), with density

$$g^*(\lambda) = \frac{\beta_0^{\mu_0}}{\Gamma(\mu_0)}\lambda^{-(\mu_0+1)}\exp\left(-\frac{\beta_0}{\lambda}\right)\mathbb{1}_{(0,\infty)}(\lambda), \tag{20}$$

and hyperparameters $\mu_0,\beta_0 > 0$, where $\Gamma(\mu_0) = \int_0^\infty z^{\mu_0-1}\exp(-z)\,dz$. The likelihood function of the observed sample $\delta_n^o = (\delta_1,\ldots,\delta_n)$ is

$$\begin{aligned} L(\delta_n^o \mid \lambda) &= \prod_{i=1}^n g(\delta_i \mid \lambda) \\ &= \prod_{i=1}^n \lambda^{-1}\exp[-(\delta_i - m)\lambda^{-1}] \\ &= \lambda^{-n}\exp[-(n\bar{\delta} - nm)\lambda^{-1}], \end{aligned} \tag{21}$$

where $\bar{\delta} = (1/n)\sum_{i=1}^n \delta_i$. According to (20)–(21), the posterior distribution of λ given the data is

$$\begin{aligned} g^*(\lambda|\delta_n^o) &= \frac{L(\delta_n^o \mid \lambda)\times g^*(\lambda)}{\int_0^\infty L(\delta_n^o \mid \lambda)\times g^*(\lambda)\,d\lambda} \\ &= \frac{(\beta_0 + n\bar{\delta} - nm)^{(\mu_0+n)}}{\Gamma(\mu_0+n)}\lambda^{-[(\mu_0+n)+1]}\exp[-(\beta_0 + n\bar{\delta} - nm)\lambda^{-1}], \end{aligned}$$

which is Inv-Gamma$(\mu_0 + n, \beta_0 + n\bar{\delta} - nm)$.

Using the mean squared error (MSE) as risk, the Bayes estimator of the unknown parameter λ is simply the mean of the posterior distribution of λ,

$$\lambda_n = \frac{\beta_0 + n\bar{\delta} - nm}{\mu_0 + n - 1} = k_n\beta_0 + k_n Y,$$

where $k_n = (n + \mu_0 - 1)^{-1} > 0$ and $Y = n\bar{\delta} - nm = \sum_{i=1}^{n}(\delta_i - m)$ is Gamma(n, λ).

The moment generating function of λ_n is

$$M_{\lambda_n}(t) = \exp(k_n\beta_0 t)\left(\frac{1}{1 - k_n\lambda t}\right)^n, \qquad t < \frac{1}{\lambda}.$$

Thus,

$$\begin{aligned} E(\lambda_n) &= \frac{d}{dt}M_{\lambda_n}(t)\bigg|_{t=0} = k_n\beta_0 + nk_n\lambda, \\ E(\lambda_n^2) &= \frac{d^2}{dt^2}M_{\lambda_n}(t)\bigg|_{t=0} = k_n^2\beta_0^2 + nk_n^2(2\lambda\beta_0 + \lambda^2) + n^2k_n^2\lambda^2. \end{aligned}$$

We then have

$$\begin{aligned} E[(\lambda_n - \lambda)^2] &= E(\lambda_n^2) - 2\lambda E(\lambda_n) + \lambda^2 \\ &= k_n^2\beta_0^2 + nk_n^2(2\lambda\beta_0 + \lambda^2) + n^2k_n^2\lambda^2 - 2\theta(k_n\beta_0 + nk_n\lambda) + \lambda^2 \\ &= k_n^2\beta_0^2 + nk_n^2(2\lambda\beta_0 + \lambda^2) - 2k_n\lambda\beta_0 + n^2k_n^2\lambda^2 - 2nk_n\lambda^2 + \lambda^2 \\ &= k_n^2\beta_0^2 + nk_n^2(2\lambda\beta_0 + \lambda^2) - 2k_n\lambda\beta_0 + (nk_n - 1)^2\lambda^2. \end{aligned} \tag{22}$$

Now we obtain the rate of convergence of $E[(\lambda_n - \lambda)^2]$. For $\mu_0 \in (0, 1)$ and $n > (1 - \mu_0^2)\mu_0^{-1}$, we have

$$\begin{aligned} k_n &< (1 + \mu_0)\left(\frac{1}{n}\right), \\ k_n^2 &< (1 + \mu_0)^2\left(\frac{1}{n}\right)^2, \\ nk_n^2 &< (1 + \mu_0)^2\left(\frac{1}{n}\right), \\ (nk_n - 1)^2 &< (1 + \mu_0)^2\left(\frac{1}{n}\right)^2. \end{aligned} \tag{23}$$

On the other hand, for $\mu_0, n \geq 1$,

$$\begin{aligned} k_n &\leq \frac{1}{n}, \\ k_n^2 &\leq \left(\frac{1}{n}\right)^2, \\ nk_n^2 &\leq \frac{1}{n}, \\ (nk_n - 1)^2 &\leq (1 - \mu_0)^2\left(\frac{1}{n}\right)^2. \end{aligned} \tag{24}$$

Therefore, combining (22)–(24) we obtain

$$E[(\lambda_n - \lambda)^2] = O(n^{-2}) \quad \text{as } n \to \infty.$$

Hence, we define the Bayes estimator of the parameter θ as $\theta_n = \lambda_n + m$. Furthermore, observe that θ_n satisfies

$$\theta_n > m,$$

and

$$\begin{aligned} E[(\theta_n - \theta)^2] &= E[((\lambda_n + m) - (\lambda + m))^2] \\ &= E[(\lambda_n - \lambda)^2] \\ &= O(n^{-2}) \quad \text{as } n \to \infty. \end{aligned} \tag{25}$$

5. Estimation and control. In this section we present the construction of an AC-optimal policy. The procedure consists in the combination of the Bayes estimation scheme of the mean holding time θ with a variant of the vanishing discount factor approach.

Let $\nu \in (0, 1/2)$ be arbitrary. Now we fix a nondecreasing sequence $\{\hat{\alpha}_n\}$ of discount factors with the following properties:

S.1. $(1 - \hat{\alpha}_n)^{-1} = O(n^\nu)$ as $n \to \infty$.

S.2. $\lim_{n\to\infty} \kappa(n)/n = 0$, where $\kappa(n)$ is the number of changes of value of $\{\hat{\alpha}_n\}$ among the first n terms.

Let $\bar{C}_{\theta_n}(\cdot,\cdot)$ be the corresponding estimator of the one-stage cost $C_\theta(\cdot,\cdot)$ (see (11)), that is,

$$\bar{C}_{\theta_n}(x,a) := \frac{C_{\theta_n}(x,a)}{\theta_n} = \frac{D(x,a) + \theta_n d(x,a)}{\theta_n}, \qquad (x,a) \in \mathbb{K}. \tag{26}$$

Now, for a fixed n, let $V_{\hat{\alpha}_n}^{\theta_n}(\pi, x) := E_x^\pi[\sum_{t=0}^\infty \hat{\alpha}_n^t \bar{C}_{\theta_n}(x_t, a_t)]$ and $V_{\hat{\alpha}_n}^{\theta_n}(x) := \inf_{\pi\in\Pi} V_{\hat{\alpha}_n}^{\theta_n}(\pi, x)$, $x \in X$, be the total expected $\hat{\alpha}_n$-discounted cost and the corresponding optimal value function under the one-stage cost $\bar{C}_{\theta_n}$. We define accordingly the sequences $j_{\hat{\alpha}_n}^{\theta_n}$, $h_{\hat{\alpha}_n}^{\theta_n}(\cdot)$ and $T_{\hat{\alpha}_n}^{\theta_n}$ (see (17) and (15)). Then, from Proposition 2 we have, for each $x \in \mathbb{X}$ and $n \geq 0$,

$$V_{\hat{\alpha}_n}^{\theta_n}(x) = T_{\hat{\alpha}_n}^{\theta_n} V_{\alpha_n}^{\theta_n}(x) \quad \text{and} \quad j_{\hat{\alpha}_n}^{\theta_n} + h_{\hat{\alpha}_n}^{\theta_n}(x) = T_{\hat{\alpha}_n}^{\theta_n} h_{\hat{\alpha}_n}^{\theta_n}(x) \text{ a.s.} \tag{27}$$

Hence, by applying standard arguments on the existence of minimizers (see, e.g., [17]), for each $n \geq 0$ there exists $f_n \in \mathbb{F}$ such that

$$j_{\hat{\alpha}_n}^{\theta_n} + h_{\hat{\alpha}_n}^{\theta_n}(x) = \bar{C}_{\theta_n}(x, f_n) + \hat{\alpha}_n \int_{\mathbb{X}} h_{\hat{\alpha}_n}^{\theta_n}(y)\, Q(dy \mid x, f_n) \qquad \text{a.s., } x \in \mathbb{X}. \tag{28}$$

Finally, we state our main result.

THEOREM 3. *Under Assumption* 1, *the control policy* $\hat{\pi} = \{f_n\}$ *determined by the minimizers* f_n *in* (28) *is average cost optimal, that is,* $J(\hat{\pi}, x) = j^*$.

6. Proofs. We first introduce some useful consequences of our assumptions which are summarized in the following remark.

Remark 4. (a) By means of an iterative process (see, e.g., [4, 6, 9, 10, 11]) it is easy to see that the inequalities (4) and (5) yield

$$\sup_{n>0} E_x^\pi[W(x_n)] < \infty \text{ and } \sup_{n>0} E_x^\pi[W^2(x_n)] < \infty, \qquad \forall x \in \mathbb{X},\ \pi \in \varPi. \tag{29}$$

Then, from (19), $E_x^\pi[h_\alpha(x_n)] < M_1$ for all $\alpha \in (0,1)$ and some constant $M_1 < \infty$. In addition, observe that Condition S.2 implies that the sequence $\{\hat\alpha_n\}$ remains constant for long time periods. Hence, denoting by $\alpha_1^*, \ldots, \alpha_{\kappa(n)}^*$, $n \geq 1$, the different values of $\hat\alpha_t$ for $t \leq n$, and using the fact that $\{\hat\alpha_n\}$ is nondecreasing, we have, for any $k \in \mathbb{N}$,

$$\begin{aligned} (30)\quad & n^{-1}E_x^\pi\Big[\sum_{t=k}^{n}(h_{\hat\alpha_t}(x_t) - \hat\alpha_t h_{\hat\alpha_t}(x_{t+1}))\Big] \\ &= n^{-1}E_x^\pi\Big[\sum_{t=k}^{n}(h_{\hat\alpha_t}(x_t) - \hat\alpha_t h_{\hat\alpha_t}(x_t))\Big] \\ &\quad + n^{-1}E_x^\pi\Big[\sum_{t=k}^{n}\hat\alpha_t(h_{\hat\alpha_t}(x_t) - h_{\hat\alpha_t}(x_{t+1}))\Big] \\ &\leq (1-\alpha_k)M_1 + n^{-1}2M_1\sum_{i=1}^{\kappa(n)}\alpha_i^* \\ &\leq (1-\alpha_k)M_1 + 2M_1\kappa(n)n^{-1}, \qquad x \in \mathbb{X},\ \pi \in \varPi. \end{aligned}$$

(b) (cf. Van Nunen and Wessels [18], [15]) Let d be the constant in (5). For each $n \in \mathbb{N}$, we define $\rho_n := (1+\hat\alpha_n)/2 \in (\hat\alpha_n, 1)$, $e_n := d(\rho_n/\hat\alpha_n - 1)^{-1}$, and the function $W_n(x) := W(x) + e_n$ for $x \in \mathbb{X}$. Now, consider the space $\mathbb{B}_{W_n}(\mathbf{X})$ of functions $u : \mathbf{X} \to \mathbb{R}$ with finite W_n-norm, that is,

$$\|u\|_{W_n} := \sup_{x \in \mathbf{X}} \frac{|u(x)|}{W_n(x)} < \infty.$$

As is shown in [18, Lemma 2], the inequality (5) implies that the operators $T_{\hat\alpha_n}^\theta$ and $T_{\hat\alpha_n}^{\theta_n}$ are contractions with respect to the W_n-norm with ratio ρ_n, that is, for all $v, u \in \mathbb{B}_W(\mathbb{X})$ and $n \in \mathbb{N}$,

$$\begin{aligned} \|T_{\hat\alpha_n}^\theta v - T_{\hat\alpha_n}^\theta u\|_{W_n} &\leq \rho_n\|v-u\|_{W_n} \\ \|T_{\hat\alpha_n}^{\theta_n} v - T_{\hat\alpha_n}^{\theta_n} u\|_{W_n} &\leq \rho_n\|v-u\|_{W_n}. \end{aligned} \tag{31}$$

Moreover, observe that for each $n \in \mathbb{N}$,

$$\|u\|_{W_n} \leq \|u\|_W \leq l_n\|u\|_{W_n}, \tag{32}$$

where

$$l_n := 1 + \frac{2d}{1-\hat\alpha_n}. \tag{33}$$

It is worth noting that from condition S.1, $\hat{\alpha}_n$ and ρ_n satisfy the relation

$$\frac{1}{(1-\rho_n)(1-\hat{\alpha}_n)} = O(n^{2\nu}) \quad \text{as } n \to \infty. \tag{34}$$

(c) From (1), (11), and (26),

$$\begin{aligned} \sup_{a\in A(x)} |\bar{C}_\theta(x,a) - \bar{C}_{\theta_n}(x,a)| &= \sup_{a\in A(x)} \left| \frac{D(x,a)}{\theta} - \frac{D(x,a)}{\theta_n} \right| \\ &\le W(x) \left| \frac{1}{\theta} - \frac{1}{\theta_n} \right| \\ &\le W(x) \frac{|\theta - \theta_n|}{m^2} \quad \text{a.s., } \forall n \in \mathbb{N},\ x \in \mathbb{X}, \end{aligned} \tag{35}$$

where the last inequality follows from the fact that $\theta, \theta_n > m$.

Lemma 5. *Suppose that Assumption* 1 *holds. Then, for each $x \in \mathbb{X}$ and $\pi \in \Pi$,*

$$\lim_{n\to\infty} E_x^\pi \|h_{\hat{\alpha}_n} - h_{\hat{\alpha}_n}^{\theta_n}\|_W^2 = 0.$$

Proof. (a) From Proposition 2, (27), and (31) we have

$$\|V_{\hat{\alpha}_n} - V_{\hat{\alpha}_n}^{\theta_n}\|_{W_n} \le \|T_{\hat{\alpha}_n}^{\theta} V_{\hat{\alpha}_n} - T_{\hat{\alpha}_n}^{\theta_n} V_{\hat{\alpha}_n}\|_{W_n} + \rho_n \|V_{\hat{\alpha}_n} - V_{\hat{\alpha}_n}^{\theta_n}\|_{W_n}.$$

Thus,

$$l_n \|V_{\hat{\alpha}_n} - V_{\hat{\alpha}_n}^{\theta_n}\|_{W_n} \le \frac{l_n}{1-\rho_n} \|T_{\hat{\alpha}_n}^{\theta} V_{\hat{\alpha}_n} - T_{\hat{\alpha}_n}^{\theta_n} V_{\hat{\alpha}_n}\|_{W_n}, \qquad n \in \mathbb{N}. \tag{36}$$

On the other hand, from (35),

$$\begin{aligned} |T_{\hat{\alpha}_n}^{\theta} V_{\hat{\alpha}_n}(x) - T_{\hat{\alpha}_n}^{\theta_n} V_{\hat{\alpha}_n}(x)| &\le \sup_{a\in A(x)} |\bar{C}_\theta(x,a) - \bar{C}_{\theta_n}(x,a)| \\ &\le W(x) \frac{|\theta-\theta_n|}{m^2} \quad \text{a.s., } \forall n \in \mathbb{N},\ x \in \mathbb{X}. \end{aligned} \tag{37}$$

Hence,

$$\|T_{\hat{\alpha}_n}^{\theta} V_{\hat{\alpha}_n} - T_{\hat{\alpha}_n}^{\theta_n} V_{\hat{\alpha}_n}\|_{W_n} \le \frac{|\theta-\theta_n|}{m^2}, \qquad \forall n \in \mathbb{N}. \tag{38}$$

Then, combining (36)–(38) and using (25), (33), and (34), we get

$$\begin{aligned} l_n^2 E_x^\pi \|V_{\hat{\alpha}_n} - V_{\hat{\alpha}_n}^{\theta_n}\|_{W_n}^2 &\le \frac{1}{m^4} \left[\frac{1+2d}{(1-\rho_n)(1-\hat{\alpha}_n)} \right]^2 |\theta - \theta_n|^2 \\ &= O(n^{4\nu}) O(n^{-2}) \quad \text{as } n \to \infty. \end{aligned} \tag{39}$$

Taking expectation E_x^π on both sides of (39) and using the fact that $4\nu < 2$ (see condition S.1) yields

$$l_n^2 E_x^\pi \|V_{\hat{\alpha}_n} - V_{\hat{\alpha}_n}^{\theta_n}\|_{W_n}^2 \to 0 \quad \text{as } n \to \infty.$$

Finally observe that from (17), $\|h_{\hat{\alpha}_n} - h_{\hat{\alpha}_n}^{\theta_n}\|_W \le 2\|V_{\hat{\alpha}_n} - V_{\hat{\alpha}_n}^{\theta_n}\|_W$ for each $n \in \mathbb{N}$, which in turn, by (32), proves the result. ■

Remark 6. Let $\hat{M} := \sup_{n\geq 0}(E_x^\pi[W^2(x_n)])^{1/2} < \infty$ (see (29)). Then applying Hölder's inequality and Lemma 5 we get

$$(40)\quad E_x^\pi\|h_{\hat{\alpha}_n} - h_{\hat{\alpha}_n}^{\theta_n}\|_W W(x_n) \leq \hat{M}(E_x^\pi[\|h_{\hat{\alpha}_n} - h_{\hat{\alpha}_n}^{\theta_n}\|_W^2])^{1/2} \to 0 \quad \text{as } n \to \infty.$$

Lemma 7. *Under Assumption* 1,

$$\lim_{t\to\infty} E_x^{\hat{\pi}}\eta_{\hat{\alpha}_t}(x_t, a_t) = 0,$$

where

$$(41)\quad \eta_\alpha(x,a) := \bar{C}_\theta(x,a) + \alpha\int_{\mathbb{X}} h_\alpha(y)\,Q(dy\mid x,a) - j_\alpha - h_\alpha(x), \qquad \alpha\in(0,1),\ (x,a)\in\mathbb{K}.$$

Proof. Let $k_t = \{(x_t, a_t)\}$ be a sequence of state-action pairs corresponding to application of the policy $\hat{\pi}$, and denote

$$(42)\qquad \eta_t := \eta_{\hat{\alpha}_t}(x_t, a_t).$$

We will show that

$$\lim_{t\to\infty} E_x^{\hat{\pi}}\eta_t = 0.$$

Adding and subtracting the term $\bar{C}_{\theta_t}(k_t) + \hat{\alpha}_t\int_{\mathbb{X}} h_{\hat{\alpha}_t}^{\theta_t}(y)\,Q(dy\mid k_t)$ we get

$$\begin{aligned}(43)\qquad \eta_t &= \bar{C}_\theta(k_t) - \bar{C}_{\theta_t}(k_t)\\ &\quad + \hat{\alpha}_t\int_{\mathbb{X}} h_{\hat{\alpha}_t}(y)\,Q(dy\mid k_t) - \hat{\alpha}_t\int_{\mathbb{X}} h_{\hat{\alpha}_t}^{\theta_t}(y)\,Q(dy\mid k_t)\\ &\quad + \bar{C}_{\theta_t}(k_t) + \hat{\alpha}_t\int_{\mathbb{X}} h_{\hat{\alpha}_t}^{\theta_t}(y)\,Q(dy\mid k_t) - j_{\hat{\alpha}_t} - h_{\hat{\alpha}_t}(x_t).\end{aligned}$$

Furthermore, using the fact that $|u(x)| \leq \|u\|_W W(x)$ for $u \in \mathbb{B}_W(\mathbb{X})$ and $x\in\mathbb{X}$, (28), and (35), we obtain

$$\begin{aligned}(44)\qquad \eta_t &\leq \sup_{a\in A(x)}|\bar{C}_\theta(x_t,a) - \bar{C}_{\theta_t}(x_t,a)|\\ &\quad + \hat{\alpha}_t\|h_{\hat{\alpha}_t} - h_{\hat{\alpha}_t}^{\theta_t}\|_W[\beta W(x_t)+d] + j_{\hat{\alpha}_t}^{\theta_t} + h_{\hat{\alpha}_t}^{\theta_t}(x_t) - j_{\hat{\alpha}_t} - h_{\hat{\alpha}_t}(x_t)\\ &\leq W(x_t)\frac{|\theta-\theta_t|}{m^2} + \hat{\alpha}_t\|h_{\hat{\alpha}_t} - h_{\hat{\alpha}_t}^{\theta_t}\|_W[\beta W(x_t)+d]\\ &\quad + j_{\hat{\alpha}_t}^{\theta_t} + h_{\hat{\alpha}_t}^{\theta_t}(x_t) - j_{\hat{\alpha}_t} - h_{\hat{\alpha}_t}(x_t).\end{aligned}$$

Now, note that from (25) and (29), for each $x\in\mathbb{X}$,

$$(45)\qquad E_x^{\hat{\pi}}\left[W(x_t)\frac{|\theta-\theta_t|}{m^2}\right] \leq \frac{\hat{M}}{m^2}(E_x^{\hat{\pi}}|\theta-\theta_t|^2)^{1/2} \to 0 \quad \text{as } t\to\infty$$

(see Remark 6 for the constant $\hat{M}$), and from Lemma 5 and (40),

$$(46)\qquad \lim_{t\to\infty} E_x^{\hat{\pi}}\left[\hat{\alpha}_t\|h_{\hat{\alpha}_t} - h_{\hat{\alpha}_t}^{\theta_t}\|_W[\beta W(x_t)+d]\right] = 0.$$

In addition, taking into account that for each $t \in \mathbb{N}$,

$$|j^{\theta_t}_{\hat{\alpha}_t} - j_{\hat{\alpha}_t}| \le (1-\hat{\alpha}_t)\|V^{\theta_t}_{\hat{\alpha}_t} - V_{\hat{\alpha}_t}\|_W W(z)$$

and

$$|h^{\theta_t}_{\hat{\alpha}_t}(x_t) - h_{\hat{\alpha}_t}(x_t)| \le \|h^{\theta_t}_{\hat{\alpha}_t} - h_{\hat{\alpha}_t}\|_W W(x_t),$$

again Lemma 5 and (40) yield

$$\lim_{t\to\infty} E^{\hat{\pi}}_x (j^{\theta_t}_{\hat{\alpha}_t} - j_{\hat{\alpha}_t}) = 0 \tag{47}$$

and

$$\lim_{t\to\infty} E^{\hat{\pi}}_x (h^{\theta_t}_{\hat{\alpha}_t}(x_t) - h_{\hat{\alpha}_t}(x_t)) = 0. \tag{48}$$

Therefore, since η_t is nonnegative, the combination of (43)–(48) proves the desired result. ∎

Proof of Theorem 3. Observe that from (41) and (42),

$$\eta_t := \bar{C}_\theta(x_t, f_t) + \hat{\alpha}_t E^{\hat{\pi}}_x[h_{\hat{\alpha}_t}(x_{t+1}) \mid h_t] - j_{\hat{\alpha}_t} - h_{\hat{\alpha}_t}(x_t).$$

Hence

$$E^{\hat{\pi}}_x \bar{C}_\theta(x_t, a_t) = j_{\hat{\alpha}_t} + E^{\hat{\pi}}_x[h_{\hat{\alpha}_t}(x_t) - \hat{\alpha}_t h_{\hat{\alpha}_t}(x_{t+1})] + E^{\hat{\pi}}_x(\eta_t),$$

which implies, for $n \ge k \ge 1$,

$$\begin{aligned} n^{-1}E^{\hat{\pi}}_x\Big[\sum_{t=0}^{n-1}\bar{C}_\theta(x_t,a_t)\Big] &= n^{-1}\sum_{t=0}^{n-1} j_{\hat{\alpha}_t} + n^{-1}E^{\hat{\pi}}_x\Big[\sum_{t=0}^{k-1}(h_{\hat{\alpha}_t}(x_t) - \hat{\alpha}_t h_{\hat{\alpha}_t}(x_{t+1}))\Big] \\ &\quad + n^{-1}E^{\hat{\pi}}_x\Big[\sum_{t=k}^{n-1}(h_{\hat{\alpha}_t}(x_t) - \hat{\alpha}_t h_{\hat{\alpha}_t}(x_{t+1}))\Big] + n^{-1}E^{\hat{\pi}}_x\Big[\sum_{t=0}^{n-1}\eta_t\Big]. \end{aligned}$$

Now, from (18),

$$\lim_{n\to\infty} n^{-1}\sum_{t=0}^{n-1} j_{\hat{\alpha}_t} = j^* = \inf_{\pi\in\Pi} J(\pi, x).$$

Therefore, from (10), condition S.1, (30), and Lemma 7 we get

$$\lim_{n\to\infty} n^{-1}E^{\hat{\pi}}_x\Big[\sum_{t=0}^{n-1}\bar{C}_\theta(x_t,a_t)\Big] = j^*, \qquad \forall x \in \mathbb{X},$$

that is, $\hat{\pi}$ is an average cost optimal policy. ∎

References

[1] V. S. Borkar and S. M. Mundra, *Bayesian parameter estimation and adaptive control of Markov processes with time-averaged cost*, Appl. Math. (Warsaw) 25 (1998), 339–358.

[2] G. B. Di Masi and Ł. Stettner, *Bayesian ergodic adaptive control of discrete time Markov processes*, Stochastics Stochastics Rep. 54 (1995), 301–316.

[3] E. I. Gordienko, *Adaptive strategies for certain classes of controlled Markov processes*, Theory Probab. Appl. 29 (1985), 504–518.

[4] E. I. Gordienko and O. Hernández-Lerma, *Average cost Markov control processes with weighted norms: existence of canonical policies*, Appl. Math. (Warsaw) 23 (1995), 199–218.

[5] E. I. Gordienko and O. Hernández-Lerma, *Average cost Markov control processes with weighted norms: value iterations*, Appl. Math. (Warsaw) 23 (1995), 219–237.

[6] E. I. Gordienko and J. A. Minjárez-Sosa, *Adaptive control for discrete-time Markov processes with unbounded costs: discounted criterion*, Kybernetika 34 (1998), 217–234.

[7] O. Hernández-Lerma and R. Cavazos-Cadena, *Density estimation and adaptive control of Markov processes: average and discounted criteria*, Acta Appl. Math. 20 (1990), 285–307.

[8] O. Hernández-Lerma and J. B. Lasserre, *Discrete-Time Markov Control Processes: Basic Optimality Criteria*, Springer, New York, 1996.

[9] O. Hernández-Lerma and J. B. Lasserre, *Further Topics on Discrete-Time Markov Control Processes*, Springer, New York, 1999.

[10] A. Jaśkiewicz, *Zero-sum semi-Markov games*, SIAM J. Control Optim. 41 (2002), 723–739.

[11] A. Jaśkiewicz and A. S. Nowak, *On the optimality equation for average cost Markov control processes with Feller transition probabilities*, J. Math. Anal. Appl. 316 (2006), 495–509.

[12] F. Luque-Vásquez and O. Hernández-Lerma, *Semi-Markov models with average costs*, Appl. Math. (Warsaw) 26 (1999), 315–331.

[13] F. Luque-Vásquez and J. A. Minjárez-Sosa, *Semi-Markov control processes with unknown holding times distribution under a discounted criterion*, Math. Methods Oper. Res. 61 (2005), 455–468.

[14] J. A. Minjárez-Sosa, *Nonparametric adaptive control for discrete-time Markov processes with unbounded costs under average criterion*, Appl. Math. (Warsaw) 26 (1999), 267–280.

[15] J. A. Minjárez-Sosa and O. Vega-Amaya, *Optimal strategies for adaptive zero-sum average Markov games*, J. Math. Anal. Appl. 402 (2013), 44–56.

[16] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, Wiley, New York, 1994.

[17] U. Rieder, *Measurable selection theorems for optimization problems*, Manuscripta Math. 24 (1978), 115–131.

[18] J. A. E. E. Van Nunen and J. Wessels, *A note on dynamic programming with unbounded rewards*, Manag. Sci. 24 (1978), 576–580.

J. Adolfo Minjárez-Sosa, José A. Montoya
Departamento de Matemáticas
Universidad de Sonora
Rosales s/n, Col. Centro, 83000 Hermosillo, Sonora, Mexico
E-mail: aminjare@gauss.mat.uson.mx
montoya@gauss.mat.uson.mx

Received on 8.4.2015;
revised version on 15.10.2015 (2265)