APPLICATIONES MATHEMATICAE
26,3 (1999), pp. 267–280

J. A. MINJÁREZ-SOSA (Hermosillo)

NONPARAMETRIC ADAPTIVE CONTROL FOR DISCRETE-TIME MARKOV PROCESSES WITH UNBOUNDED COSTS UNDER AVERAGE CRITERION

Abstract. We introduce average cost optimal adaptive policies in a class of discrete-time Markov control processes with Borel state and action spaces, allowing unbounded costs. The processes evolve according to the system equations $x_{t+1} = F(x_t, a_t, \xi_t)$, $t = 1, 2, \ldots$, with i.i.d. $\mathbb{R}^k$-valued random vectors ξ_t, which are observable but whose density ϱ is unknown.

1. Introduction. We consider a class of discrete-time Markov control processes (MCPs) of the form

$$x_{t+1} = F(x_t, a_t, \xi_t), \quad t = 0, 1, \ldots, \tag{1}$$

where F is a known function, x_t and a_t represent, respectively, the state and control (action) at time t, taking values in Borel spaces, and $\{\xi_t\}$ (the "driving process") are independent and identically distributed random vectors in $\mathbb{R}^k$ having an unknown density ϱ. Assuming that realizations $\xi_0, \xi_1, \xi_2, \ldots$ of the driving process and the states $x_0, x_1, x_2, \ldots$ are completely observable, we introduce an optimal adaptive policy with respect to the long run expected average cost with a possibly unbounded one-stage cost. These assumptions are satisfied in some applied problems, for instance in production-inventory systems, control of water reservoirs, certain controlled queueing systems, etc. (see, for example, [2], [8] and references therein).

Since ϱ is unknown, to construct an adaptive policy in this paper, we introduce first a suitable method of statistical estimation of ϱ, and then apply

1991 *Mathematics Subject Classification*: 93E20, 90C40.

Key words and phrases: Markov control process, discounted and average cost criterion, adaptive policy.

Work supported by Consejo Nacional de Ciencia y Tecnología (CONACyT) under Grants 3871P-E9607 and 980147.

the "principle of estimation and control" proposed by Mandl in [12]. This is not easy because of unbounded cost. Indeed, the nice contractive operator techniques do not work for the average criterion, and so we are forced to impose Lippman-like conditions ([11], [14]) and ergodicity assumptions on the class of MCPs considered, to be able to use the results in [4]. Moreover, we need methods of statistical estimation of ϱ such that provide information about the L_q-norm accuracy $\|\varrho_t - \varrho\|_q$ of the estimators ϱ_t, $t = 1, 2, \ldots$

Our work is motivated mostly by recent papers of Gordienko and Minjárez-Sosa [5], [6], in which there were constructed, respectively, asymptotically discounted optimal and average cost optimal adaptive policies, for the same class of processes (1), allowing unbounded one-stage cost.

The main difference between the results presented in this paper and those in [6] concerns the restrictions on the control model and the approach used.

For instance, the assumptions on the set of densities that define the admissible class of control processes for which the adaptive policy constructed in [6] is applicable are more restrictive than our conditions (see Assumptions 2.1(c), (d), and condition (f) for densities used in [6]). In fact, to prove the optimality of the adaptive policy constructed in this paper, we only need to impose conditions that ensure the existence of a solution to an optimality inequality, while in [6] average cost optimality equations play an important role.

As regards the approach, the adaptive policy in [6] was defined by means of an iterative procedure, which is an obvious advantage from the point of view of its implementation. But this gain is rather limited since the proof of the average optimality for that policy relies strongly on the convergence of the so-called value iteration algorithm, for which a very restrictive additional condition was imposed (see Proposition 3.4 in [6]). Instead, the average optimality of the adaptive policy proposed here is studied by means of a variant of the so-called vanishing discount factor approach [1] without additional conditions.

This procedure consists in choosing an appropriate sequence $\{\alpha_t\}$, $\alpha_t \nearrow 1$, of discount factors, then replace the unknown density ϱ by its estimators ϱ_t, which are obtained using the procedure of statistical estimation proposed in [5], [6], and finally exploit the corresponding α_t-discounted optimality equations, taking the limit as $t \to \infty$.

The policy studied here was originally introduced in [3] and revised in [10], both considering bounded one-stage cost.

The paper is organized as follows. In Sections 2 and 3 we introduce the Markov control model and the assumptions considered. Next, in Section 4 we list some preliminary results, which are used to prove the optimality of the adaptive policy in Section 5.

2. The control model. We consider a class of discrete-time Markov control models $(X, A, \mathbb{R}^k, F, \varrho, c)$ in which the state space X and the control A are both Borel. The dynamics is defined by the system equations (1). Here, $F : X \times A \times \mathbb{R}^k \to X$ is a given (measurable) function, and $\{\xi_t\}$ is a sequence of independent and identically distributed (i.i.d.) random vectors (r.v.'s) on a probability space $(\Omega, \mathcal{F}, P)$, with values in $\mathbb{R}^k$ and a common unknown distribution with a density ϱ (unknown), that belongs to a given class described in the next section.

With each $x \in X$, we associate a nonempty set $A(x)$ whose elements are the feasible controls (or actions) when the state of the system is x. The set

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

is assumed to be a Borel subset of $X \times A$, and the one-stage cost c is a nonnegative real-valued measurable function on $\mathbb{K}$, possibly unbounded.

Let Π be the set of all control policies and $\mathbb{F} \subset \Pi$ be the set of all deterministic stationary policies [2]. As usual, every stationary policy $\pi \in \mathbb{F}$ is identified with some measurable function $f : X \to A$ such that $f(x) \in A(x)$ for every $x \in X$, taking the form $\pi = \{f, f, f, \ldots\}$. In this case we use the notation $\mathbf{f}$ for π and we write

$$c(x, f) := c(x, f(x)) \quad \text{and} \quad F(x, f, s) := F(x, f(x), s), \quad x \in X,\ s \in \mathbb{R}^k.$$

Given the initial state $x_0 = x$, when using a policy $\pi \in \Pi$, we define the *total expected α-discount cost* as

$$V_\alpha(\pi, x) := E_x^\pi\Big[\sum_{t=0}^{\infty} \alpha^t c(x_t, a_t)\Big],$$

$\alpha \in (0, 1)$ being the so-called *discount factor*; and the *long run expected average cost* as

$$J(\pi, x) := \limsup_{n\to\infty} n^{-1} E_x^\pi\Big[\sum_{t=0}^{n-1} c(x_t, a_t)\Big], \tag{2}$$

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 $x_0 = x$ (see, e.g., [2]).

A policy $\pi^* \in \Pi$ is said to be *α-discounted optimal* (*α-optimal*) if

$$V_\alpha(x) := \inf_{\pi\in\Pi} V_\alpha(\pi, x) = V_\alpha(\pi^*, x), \qquad x \in X.$$

Similarly, $\pi^* \in \Pi$ is called *average cost optimal* (*AC-optimal*) if

$$J(x) := \inf_{\pi\in\Pi} J(\pi, x) = J(\pi^*, x), \qquad x \in X.$$

3. Assumptions. For a given measurable function $W : X \to [1,\infty)$, we denote by L_W^∞ the normed linear space of all measurable functions $u : X \to \mathbb{R}$ with

$$\|u\|_W := \sup_{x\in X} |u(x)|/W(x) < \infty; \tag{3}$$

and for a density μ on $\mathbb{R}^k$, $Q_\mu(\cdot \mid \cdot)$ is a stochastic kernel on X given $\mathbb{K}$, defined as

$$Q_\mu(B\,|\,x,a) := \int_{\mathbb{R}^k} 1_B[F(x,a,s)]\mu(s)\,ds, \qquad B\in\mathbb{B}(X),\ (x,a)\in\mathbb{K}, \tag{4}$$

where $1_B(\cdot)$ stands for the indicator function of the set B, and $\mathbb{B}(X)$ is the Borel σ-algebra of X.

Assumption 3.1. (a) For every $x \in X$, the function $a \mapsto c(x,a)$ is lower semicontinuous (l.s.c.) and $\sup_{a\in A(x)} |c(x,a)| \le W(x)$;

(b) for each $x\in X$, $A(x)$ is a σ-compact set.

Now, we define a set of densities ϱ of the r.v.'s ξ_t in (1) that describes an admissible class of control processes for which the adaptive policy constructed in this paper is applicable. For this, fix $\varepsilon\in(0,1/2)$ and a nonnegative measurable function $\overline{\varrho} : \mathbb{R}^k \to \mathbb{R}$ which is used as a known majorant of the unknown densities ϱ.

Setting $q := 1+2\varepsilon$, we define the set $D_0 = D_0(\overline{\varrho}, L, \beta_0, b_0, p, q, m, \psi, \overline{\psi})$ to consist of all densities μ on $\mathbb{R}^k$ for which the following holds.

(a) $\mu\in L_q(\mathbb{R}^k)$.

(b) There exists a constant L such that for each $z\in\mathbb{R}^k$,

$$\|\varDelta_z\mu\|_{L_q} \le L|z|^{1/q}, \tag{5}$$

where $\varDelta_z\mu(s) := \mu(s+z)-\mu(s)$ for $s\in\mathbb{R}^k$ and $|\cdot|$ is the Euclidean norm in $\mathbb{R}^k$.

(c) $\mu(s)\le\overline{\varrho}(s)$ almost everywhere with respect to the Lebesgue measure.

(d) For every $f\in\mathbb{F}$ the Markov x_t^f process with transition probability $Q_\mu(B\,|\,x,f)$, $B\in\mathbb{B}(X)$, is positive Harris-recurrent.

(e) There exists a probability measure m on $(X,\mathbb{B}(X))$ and a nonnegative number $\beta_0<1$ and for every $f\in\mathbb{F}$ a nonnegative function $\psi_f : X\to\mathbb{R}$ such that for any $x\in X$ and $B\in\mathbb{B}(X)$,

(i) $Q_\mu(B\,|\,x,f)\ge\psi_f(x)m(B)$;

(ii) $\int_{\mathbb{R}^k} W^p[F(x,f,s)]\mu(s)\,ds \le \beta_0 W^p(x) + \psi_f(x)\int_X W^p(y)\,m(dy)$ for some $p>1$, and $b_0 := \int_X W^p(y)\,m(dy)<\infty$;

(iii) $\inf_{f\in\mathbb{F}}\int_X \psi_f(x)\ m(dx) =: \overline{\psi} > 0$.

Remark 3.2. The set D_0 is more restrictive than the set of densities used in [5] for the discounted criterion because in that work it was only necessary to impose the conditions (a)–(c) together with

$$\int\limits_{\mathbb{R}^k} W^p[F(x,f,s)]\mu(s)\,ds \le \beta_0 W^p(x) + b_0, \qquad x \in X,\ a \in A(x), \tag{6}$$

where $p > 1$, $\beta_0 < 1, b_0 < \infty$. But, as was observed in ([6], Remark 2.2(b)), the relation (6) follows from conditions (e)(i) and (e)(ii) using the same p, β_0 and b_0.

Assumption 3.3. (a) The density ϱ belongs to D_0.

(b) For every $s \in \mathbb{R}^k$,

$$\varphi(s) := \sup_{x\in X}[W(x)]^{-1} \sup_{a\in A(x)} W[F(x,a,s)] < \infty. \tag{7}$$

(c) $\int_{\mathbb{R}^k} \varphi^2(s)|\overline{\varrho}(s)|^{1-2\varepsilon}\,ds < \infty$.

Remark 3.4. The function φ in (7) can be nonmeasurable. In this case we suppose the existence of a measurable majorant $\overline{\varphi}$ of φ for which Assumption 3.3(c) holds.

Assumptions 3.1 and 3.3 were used in [6], where an example of a queueing system with a controllable service rate satisfying those assumptions was given.

4. Preliminary results. In this section we state some preliminary results, proved in previous works, that will be useful in the next sections.

Lemma 4.1 (see [5]). *Suppose that Assumption* 3.1(a) *holds and ϱ satisfies the condition* (6). *Then*:

(a) *for every $x \in X$ and $a \in A(x)$*,

$$\int\limits_{\mathbb{R}^k} W[F(x,a,s)]\varrho(s)\,ds \le \beta W(x) + b, \tag{8}$$

where $\beta = \beta_0^{1/p}$ and $b = b_0^{1/p}$ [see Remark 3.2];

(b) $\sup_{t\ge1} E_x^\pi[W^p(x_t)] < \infty$ *and* $\sup_{t\ge1} E_x^\pi[W(x_t)] < \infty$ *for each $\pi \in \Pi$ and $x \in X$.*

Lemma 4.2. *Let $\alpha \in (0,1)$ be an arbitrary but fixed discount factor. Then*:

(a) (see [9]) *if ϱ satisfies the condition* (6) *or* (8), *then under Assumption* 3.1(a), *we have $V_\alpha(x) \le CW(x)/(1-\alpha)$ for some constant $C > 0$, and $V_\alpha(\cdot)$ satisfies the dynamic programming equation*, *i.e.*,

$$V_\alpha(x) = \inf_{a\in A(x)}\Big[c(x,a) + \alpha\int\limits_{\mathbb{R}^k} V_\alpha[F(x,a,s)]\varrho(s)\,ds\Big], \qquad x \in X; \tag{9}$$

(b) *under Assumption* 3.1, *for each* $\delta > 0$, *there exists a policy* $\mathbf{f} \in \mathbb{F}$ *such that*

$$c(x,f) + \alpha \int_{\mathbb{R}^k} V_\alpha[F(x,f,s)]\varrho(s)\,ds \le V_\alpha(x) + \delta, \qquad x \in X. \tag{10}$$

From the fact that $Q_\varrho(\cdot \mid \cdot)$ is a stochastic kernel [see (4)], it is easy to prove that for a nonnegative function $u \in L_W^\infty$, and every $r \in \mathbb{R}$, the set

$$\Big\{(x,a) : \int_{\mathbb{R}^k} u[F(x,a,s)]\varrho(s)\,ds \le r\Big\}$$

is Borel in $\mathbb{K}$. Hence part (b) of Lemma 4.2 is a consequence of Corollary 4.3 in [13].

Lemma 4.3 (see [4]). *Suppose that Assumption* 3.1 *holds and* $\varrho \in D_0$. *Then there exist a constant* j^* *and a function* ϕ *in* L_W^∞ *such that*

$$j^* + \phi(x) \ge \inf_{a\in A(x)} \Big[c(x,a) + \int_{\mathbb{R}^k} \phi[F(x,a,s)]\varrho(s)\,ds\Big], \tag{11}$$

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

Remark 4.4. (a) In [4] it has been shown that $j^* = \limsup_{\alpha\nearrow 1} j_\alpha$ where j^* is the optimal average cost and, for $z \in X$ fixed, $j_\alpha := (1-\alpha)V_\alpha(z)$, $\alpha \in (0,1)$. Using the same arguments as in the proof of the last assertion, we can also show that $j^* = \liminf_{\alpha\nearrow 1} j_\alpha$. Hence,

$$\lim_{t\to\infty} j_{\alpha_t} = j^* \tag{12}$$

for any sequence $\{\alpha_t\}$ of discount factors such that $\alpha_t \nearrow 1$ (see also [3]). In fact (j^*,ϕ), with $\phi(x) := \lim_{t\to\infty} \phi_{\alpha_t}(x)$, $x \in X$, satisfies the optimality inequality (11), where $\phi_\alpha(x) := V_\alpha(x) - V_\alpha(z)$. Furthermore, also in [4] it was proved that

$$\sup_{\alpha\in(0,1)} \|\phi_\alpha\|_W < \infty. \tag{13}$$

(b) From the definition of j_α and ϕ_α, it is easy to see that the equation (9) and the inequality (10) are equivalent, respectively, to

$$\begin{aligned} &j_\alpha + \phi_\alpha(x) \\ &\quad = \inf_{a\in A(x)} \Big[c(x,a) + \alpha \int_{\mathbb{R}^k} \phi_\alpha[F(x,a,s)]\varrho(s)\,ds\Big], \qquad x \in X,\ \alpha \in (0,1), \end{aligned} \tag{14}$$

and

$$\begin{aligned} &c(x,f) + \alpha \int_{\mathbb{R}^k} \phi_\alpha[F(x,f,s)]\varrho(s)\,ds \\ &\qquad\qquad \le j_\alpha + \phi_\alpha(x) + \delta, \qquad x \in X,\ \alpha \in (0,1). \end{aligned} \tag{15}$$

A key point in the construction of the average cost optimal adaptive policy in the next section is the use of the density estimation scheme proposed originally in [5] for the discounted criterion and used again in [6] (see Remark 3.2) to construct an average optimal iterative adaptive policy. We present a shortened version of this estimation procedure.

Denote by $\xi_0, \xi_1, \ldots, \xi_{t-1}$ the independent realizations (observed up to time $t-1$) of a r.v. with unknown density $\varrho \in D_0$. Let $\widehat{\varrho}_t := \widehat{\varrho}_t(s; \xi_0, \xi_1, \ldots \ldots, \xi_{t-1})$, $s \in \mathbb{R}^k$, be an arbitrary estimator of ϱ belonging to L_q, such that for some $\gamma > 0$,

$$E\|\varrho - \widehat{\varrho}_t\|_q^{qp'/2} = \mathbf{O}(t^{-\gamma}) \quad \text{as } t \to \infty, \tag{16}$$

where $1/p + 1/p' = 1$.

Then we estimate ϱ by the projection ϱ_t of $\widehat{\varrho}_t$ on the set of densities $D := D_1 \cap D_2$ in L_q where

$$\begin{aligned} D_1 &:= \{\mu : \mu \text{ is a density on } \mathbb{R}^k,\ \mu \in L_q \text{ and } \mu(s) \le \overline{\varrho}(s) \text{ a.e.}\}, \\ D_2 &:= \Big\{\mu : \mu \text{ is a density on } \mathbb{R}^k,\ \mu \in L_q, \\ &\qquad\qquad \int W[F(x,a,s)]\mu(s)\,ds \le \beta W(x) + b,\ (x,a) \in \mathbb{K}\Big\} \end{aligned} \tag{17}$$

[see Lemma 4.1 for the constants β and b].

The existence (and uniqueness) of the estimator ϱ_t is guaranteed because the set D is convex and closed in L_q ([5], [6]). In fact, we have

$$\|\varrho_t - \widehat{\varrho}_t\|_q = \inf_{\mu \in D} \|\mu - \widehat{\varrho}_t\|_q, \qquad t \in \mathbb{N}, \tag{18}$$

that is, the density $\varrho_t \in D$ is a "best approximation" of the estimator $\widehat{\varrho}_t$ on the set D. Assumption 3.3(a) and Lemma 4.1(a) yield $\varrho \in D_0 \subset D$.

In the rest of the paper we use densities $\varrho_t(\cdot) := \varrho_t(\cdot; \xi_0, \xi_1, \ldots, \xi_{t-1})$, $t \in \mathbb{N}$, satisfying (16) and (18) as estimators of a density ϱ. Examples of estimators satisfying (16) are given in [7].

Now we define the pseudo-norm $\|\cdot\|$ (possibly taking infinite values) on the space of all densities μ on $\mathbb{R}^k$ by setting

$$\|\mu\| := \sup_{x \in X} [W(x)]^{-1} \sup_{a \in A(x)} \int_{\mathbb{R}^k} W[F(x,a,s)]\mu(s)\,ds. \tag{19}$$

Lemma 4.5 (see [5], [6]). *Suppose that Assumption* 3.3 *holds. Then*

$$E\|\varrho_t - \varrho\|^{p'} = \mathbf{O}(t^{-\gamma}) \quad \textit{as } t \to \infty.$$

5. Adaptive policy as a limit of discounted programs.

Let ν be an arbitrary real number such that $0 < \nu < \gamma/(3p')$ where γ and p' are from (16). We fix an arbitrary nondecreasing sequence $\{\alpha_t\}$ of discount factors such that $1-\alpha_t=\mathbf{O}(t^{-\nu})$ as $t\to\infty$, and

$$\lim_{n\to\infty}\kappa(n)/n=0, \tag{20}$$

where $\kappa(n)$ is the number of changes of value of $\{\alpha_t\}$ on $[0,n]$.

To construct the adaptive policy, we will use similar ideas to [5], [6] and [10]. For this purpose we need to extend some assertions of the previous sections to the densities $\varrho_t\in D$.

For a fixed t, let $V_{\alpha_t}^{(\varrho_t)}(\pi,x):=E_x^{\pi,\varrho_t}[\sum_{n=0}^{\infty}\alpha_t^n c(x_n,a_n)]$ be the total expected α_t-discount cost for the process (1) in which all the r.v.'s $\xi_1,\xi_2,\ldots$ have the same density ϱ_t, and $V_{\alpha_t}^{(\varrho_t)}(x):=\inf_{\pi\in\Pi}V_{\alpha_t}^{(\varrho_t)}(\pi,x)$, $x\in X$, be the corresponding value function. For these, we define [see Remark 4.4] the sequences $\phi_{\alpha_t}^{(\varrho_t)}(\cdot)$ and $j_{\alpha_t}^{(\varrho_t)}$. Thus [see (14)],

$$\begin{aligned}(21)\quad & j_{\alpha_t}^{(\varrho_t)}+\phi_{\alpha_t}^{(\varrho_t)}(x)\\ &=\inf_{a\in A(x)}\Big[c(x,a)+\alpha_t\int_{\mathbb{R}^k}\phi_{\alpha_t}^{(\varrho_t)}[F(x,a,s)]\varrho_t(s)\,ds\Big],\qquad x\in X,\ t\in\mathbb{N},\end{aligned}$$

where the minimization is done for every $\omega\in\Omega$. In the following, we suppose that the minimization of a term including the estimator ϱ_t is done for every $\omega\in\Omega$.

For each $t\in\mathbb{N}$ and $\mu\in D$, define the operator $T_{\mu,\alpha_t}\equiv T_\mu : L_W^\infty\to L_W^\infty$ as

$$\begin{aligned}(22)\quad & T_\mu u(x)\\ &:=\inf_{a\in A(x)}\Big\{c(x,a)+\alpha_t\int_{\mathbb{R}^k}u[F(x,a,s)]\mu(s)\,ds\Big\},\qquad x\in X,\ u\in L_W^\infty.\end{aligned}$$

The proof of Lemmas 4.1 and 4.2 (partly given in [9]) shows that the following assertions hold true (because only (8) is used here).

Proposition 5.1. (a) *Suppose that Assumption* 3.1(a) *holds and ϱ satisfies* (6) *or* (8). *Then, for each $t\in\mathbb{N}$, $T_\varrho V_{\alpha_t}=V_{\alpha_t}$, $T_{\varrho_t}V_{\alpha_t}^{(\varrho_t)}=V_{\alpha_t}^{(\varrho_t)}$ and*

$$V_{\alpha_t}(x)\le\frac{C}{1-\alpha_t}W(x),\qquad V_{\alpha_t}^{(\varrho_t)}(x)\le\frac{C}{1-\alpha_t}W(x),\qquad x\in X. \tag{23}$$

(b) *Under Assumption* 3.1, *for each $t\in\mathbb{N}$ and $\delta_t>0$, there exists a policy $\widehat{\mathbf{f}}_t\in\mathbb{F}$ such that*

$$c(x,\widehat{f}_t)+\alpha_t\int_{\mathbb{R}^k}V_{\alpha_t}^{(\varrho_t)}[F(x,\widehat{f}_t,s)]\varrho_t(s)\,ds\le V_{\alpha_t}^{(\varrho_t)}(x)+\delta_t,\qquad x\in X, \tag{24}$$

or [see Remark 4.4(b)]

$$\begin{aligned}(25)\quad & c(x,\widehat{f}_t)+\alpha_t\int_{\mathbb{R}^k}\phi_{\alpha_t}^{(\varrho_t)}[F(x,\widehat{f}_t,s)]\varrho(s)\,ds\\ &\qquad\le j_{\alpha_t}^{(\varrho_t)}+\phi_{\alpha_t}^{(\varrho_t)}(x)+\delta_t,\qquad x\in X.\end{aligned}$$

For $t \in \mathbb{N}$, we set $h_t := (x_0, a_0, s_0, \ldots, x_{t-1}, a_{t-1}, s_{t-1}, x_t)$, the history up to time t, where $(x_n, a_n) \in \mathbb{K}, s_n \in \mathbb{R}^k$, $n = 0, 1, \ldots, t-1$ and $x_t \in X$.

DEFINITION 5.2. Let $\{\delta_t\}$ be an arbitrary sequence of positive numbers and $\{\widehat{f}_t\}$ be a sequence of functions (selectors) satisfying (24) or (25) for each $t \in \mathbb{N}$. The adaptive policy $\widehat{\pi} = \{\widehat{\pi}_t\}$ is defined as $\widehat{\pi}_t(h_t) = \widehat{\pi}_t(h_t; \varrho_t) := \widehat{f}_t(x_t)$, $t \in \mathbb{N}$, where $\widehat{\pi}_0(x)$ is any fixed action.

Supposing that $\delta := \lim_{t\to\infty} \delta_t < \infty$, we state our main result:

THEOREM 5.3. *Suppose that Assumptions* 3.1 *and* 3.3 *hold. Then the adaptive policy $\widehat{\pi}$ is δ-average cost optimal*, *i.e.*, *for each $x \in X$, $J(\widehat{\pi}, x) \le j^* + \delta$, where j^* is the optimal average cost as in Lemma* 4.3. *In particular*, *if $\delta = 0$ then the policy $\widehat{\pi}$ is average cost optimal.*

The proof of this theorem is based on the following lemma:

LEMMA 5.4. *Under Assumptions* 3.1 *and* 3.3, *for each $x \in X$ and $\pi \in \Pi$*, *as $t \to \infty$*,

$$\text{(a) } E_x^{\pi} \|\phi_{\alpha_t} - \phi_{\alpha_t}^{(\varrho_t)}\|_W^{p'} \to 0, \qquad \text{(b) } E_x^{\pi}[\|\phi_{\alpha_t} - \phi_{\alpha_t}^{(\varrho_t)}\|_W W(x_t)] \to 0.$$

Proof. (a) Observing that $\|\phi_{\alpha_t} - \phi_{\alpha_t}^{(\varrho_t)}\|_W \le 2\|V_{\alpha_t} - V_{\alpha_t}^{(\varrho_t)}\|_W$ it is sufficient to prove

$$\lim_{t\to\infty} E_x^{\pi} \|V_{\alpha_t} - V_{\alpha_t}^{(\varrho_t)}\|_W^{p'} = 0, \qquad x \in X,\ \pi \in \Pi. \tag{26}$$

For each $t \in \mathbb{N}$, we define $\theta_t := (1 + \alpha_t)/2 \in (\alpha_t, 1)$, and $W_t(x) := W(x) + d_t$, $x \in X$, where $d_t := b(\theta_t/\alpha_t - 1)^{-1}$. Let $L_{W_t}^{\infty}$ be the space of measurable functions $u : X \to \mathbb{R}$ with the norm

$$\|u\|_{W_t} := \sup_{x\in X} |u(x)|/W_t(x) < \infty, \qquad t \in \mathbb{N}.$$

Using the fact that $d_t \le 2b/(1 - \alpha_t)$, $t \in \mathbb{N}$, it is easy to see that

$$\|u\|_{W_t} \le \|u\|_W \le l_t \|u\|_{W_t}, \qquad t \in \mathbb{N},$$

where $l_t := 1 + 2b/[(1 - \alpha_t)\inf_{x\in X} W(x)]$. Thus, (26) will be proved if we show

$$l_t^{p'} E_x^{\widehat{\pi}} \|V_{\alpha_t} - V_{\alpha_t}^{(\varrho_t)}\|_{W_t}^{p'} \to 0 \quad \text{as } t \to \infty. \tag{27}$$

A consequence of Lemma 2 in [14] is that, for each $t \in \mathbb{N}$ and $\mu \in D$, the inequality $\int_{\mathbb{R}^k} W[F(x, a, s)]\mu(s)\,ds \le W(x) + b$ implies that the operator T_μ defined in (22) is a contraction with respect to the norm $\|\cdot\|_{W_t}$ with constant θ_t, i.e.,

$$\|T_\mu v - T_\mu u\|_{W_t} \le \theta_t \|v - u\|_{W_t}, \qquad v, u \in L_W^{\infty},\ t \in \mathbb{N}. \tag{28}$$

Hence, from (28) and Proposition 5.1(a) we can see that

$$\|V_{\alpha_t} - V_{\alpha_t}^{(\varrho_t)}\|_{W_t} \le \|T_\varrho V_{\alpha_t} - T_{\varrho_t} V_{\alpha_t}\|_{W_t} + \theta_t \|V_{\alpha_t} - V_{\alpha_t}^{(\varrho_t)}\|_{W_t},$$

which implies that

$$l_t\|V_{\alpha_t}-V_{\alpha_t}^{(\varrho_t)}\|_{W_t}\le\frac{l_t}{1-\theta_t}\|T_\varrho V_{\alpha_t}-T_{\varrho_t}V_{\alpha_t}\|_{W_t},\qquad t\in\mathbb{N}.\tag{29}$$

On the other hand, from definition (19), (23) and the fact that $[W_t(\cdot)]^{-1} < [W(\cdot)]^{-1}$, $t\in\mathbb{N}$, we obtain

$$\begin{aligned}(30)\quad &\|T_\varrho V_{\alpha_t}-T_{\varrho_t}V_{\alpha_t}\|_{W_t}\\ &\le\alpha_t\sup_{x\in X}[W_t(x)]^{-1}\sup_{a\in A(x)}\int\limits_{\mathbb{R}^k}V_{\alpha_t}[F(x,a,s)]|\varrho(s)-\varrho_t(s)|\,ds\\ &\le\frac{C\alpha_t}{1-\alpha_t}\sup_{x\in X}[W(x)]^{-1}\sup_{a\in A(x)}\int\limits_{\mathbb{R}^k}W[F(x,a,s)]|\varrho(s)-\varrho_t(s)|\,ds\\ &\le\frac{C}{1-\alpha_t}\|\varrho-\varrho_t\|.\end{aligned}$$

Now, observe that [see definition of α_t and θ_t]

$$\frac{1}{(1-\theta_t)(1-\alpha_t)^2}=\mathbf{O}(t^{3\nu})\qquad\text{as } t\to\infty.\tag{31}$$

Combining (29)–(31) and using the definition of l_t we get

$$\begin{aligned}(32)\quad &l_t^{p'}\|V_{\alpha_t}-V_{\alpha_t}^{(\varrho_t)}\|_{W_t}^{p'}\\ &\le C^{p'}\left[\frac{1}{(1-\theta_t)(1-\alpha_t)}+\frac{2b}{(1-\theta_t)(1-\alpha_t)^2\inf_{x\in X}W(x)}\right]^{p'}\|\varrho-\varrho_t\|^{p'}\\ &=C^{p'}\mathbf{O}(t^{3p'\nu})\|\varrho-\varrho_t\|^{p'}\qquad\text{as } t\to\infty.\end{aligned}$$

Finally, taking the expectation E_x^π on both sides of (32) and observing that $E_x^\pi\|\varrho-\varrho_t\|^{p'}=E\|\varrho-\varrho_t\|^{p'}$ (since ϱ_t does not depend on x and π), we obtain (27) by virtue of Lemma 4.5 and the fact $3\nu p'<\gamma$ [see definition of α_t]. This proves (a).

(b) Defining $\overline{C}:=(E_x^\pi[W^p(x_t)])^{1/p}<\infty$ [see Lemma 4.1(b)], applying Hölder's inequality and (a), we have

$$\begin{aligned}(33)\quad &E_x^\pi\|\phi_{\alpha_t}-\phi_{\alpha_t}^{(\varrho_t)}\|_W W(x_t)\\ &\le\overline{C}(E_x^\pi[\|\phi_{\alpha_t}-\phi_{\alpha_t}^{(\varrho_t)}\|_W^{p'}])^{1/p'}\to 0\qquad\text{as } t\to\infty.\end{aligned}$$

This completes the proof of Lemma 5.4. ■

Proof of Theorem 5.3. Let $\{k_t\}:=\{(x_t,a_t)\}$ be a sequence of state-action pairs corresponding to applications of the adaptive policy $\widehat{\pi}$. We define

$$\begin{aligned}(34)\quad \mathfrak{L}_t&:=c(k_t)+\alpha_t\int\limits_{\mathbb{R}^k}\phi_{\alpha_t}[F(k_t,s)]\varrho(s)\,ds-j_{\alpha_t}-\phi_{\alpha_t}(x_t)\\ &=c(k_t)+\alpha_t E_x^{\widehat{\pi}}[\phi_{\alpha_t}(x_{t+1})\,|\,k_t]-j_{\alpha_t}-\phi_{\alpha_t}(x_t).\end{aligned}$$

Hence, for $n \ge k \ge 1$,

$$(35)\quad n^{-1}E_x^{\widehat{\pi}}\Big[\sum_{t=k}^{n} c(k_t) - j_{\alpha_t}\Big]$$
$$= n^{-1}E_x^{\widehat{\pi}}\Big[\sum_{t=k}^{n}(\phi_{\alpha_t}(x_t) - \alpha_t\phi_{\alpha_t}(x_{t+1}))\Big] + n^{-1}E_x^{\widehat{\pi}}\Big[\sum_{t=k}^{n}\mathfrak{L}_t\Big].$$

On the other hand, from (13), Lemma 4.1(b) and the fact $|u(x)| \le \|u\|_W W(x)$, $u \in L_W^\infty$, $x \in X$, we have $E_x^{\widehat{\pi}}[\phi_\alpha(x_t)] < C'$, $\alpha \in (0,1)$, for a constant $C' < \infty$. Thus, denoting by $\alpha_1^*, \dots, \alpha_{\kappa(n)}^*$, $n \ge 1$, the different values of α_t for $t \le n$, and using the fact that $\{\alpha_t\}$ is a nondecreasing sequence we have [see condition (20) and the definition of ϕ_α]

$$(36)\quad n^{-1}E_x^{\widehat{\pi}}\Big[\sum_{t=k}^{n}(\phi_{\alpha_t}(x_t) - \alpha_t\phi_{\alpha_t}(x_{t+1}))\Big]$$
$$= n^{-1}E_x^{\widehat{\pi}}\Big[\sum_{t=k}^{n}(\phi_{\alpha_t}(x_t) - \alpha_t\phi_{\alpha_t}(x_t))\Big]$$
$$+ n^{-1}E_x^{\widehat{\pi}}\Big[\sum_{t=k}^{n}\alpha_t(\phi_{\alpha_t}(x_t) - \phi_{\alpha_t}(x_{t+1}))\Big]$$
$$\le (1-\alpha_k)C' + n^{-1}2C'\sum_{i=1}^{\kappa(n)}\alpha_i^*$$
$$\le (1-\alpha_k)C' + 2C'\kappa(n)n^{-1}, \qquad x \in X.$$

Now, from (34) and (14) we have

$$\mathfrak{L}_t = c(k_t) + \alpha_t \int_{\mathbb{R}^k} \phi_{\alpha_t}[F(k_t,s)]\varrho(s)\,ds$$
$$- \inf_{a\in A(x_t)}\Big[c(x_t,a) + \alpha_t\int_{\mathbb{R}^k}\phi_{\alpha_t}[F(x_t,a,s)]\varrho(s)\,ds\Big]$$
$$\le \Big|\alpha_t\int_{\mathbb{R}^k}\phi_{\alpha_t}[F(k_t,s)]\varrho(s)\,ds - \alpha_t\int_{\mathbb{R}^k}\phi_{\alpha_t}^{(\varrho_t)}[F(k_t,s)]\varrho(s)\,ds\Big|$$
$$+ \Big|\alpha_t\int_{\mathbb{R}^k}\phi_{\alpha_t}^{(\varrho_t)}[F(k_t,s)]\varrho(s)\,ds - \alpha_t\int_{\mathbb{R}^k}\phi_{\alpha_t}^{(\varrho_t)}[F(k_t,s)]\varrho_t(s)\,ds\Big|$$
$$+ \Big|c(k_t) + \alpha_t\int_{\mathbb{R}^k}\phi_{\alpha_t}^{(\varrho_t)}[F(k_t,s)]\varrho_t(s)\,ds$$
$$- \inf_{a\in A(x_t)}\Big[c(x_t,a) + \alpha_t\int_{\mathbb{R}^k}\phi_{\alpha_t}[F(x_t,a,s)]\varrho(s)\,ds\Big]\Big|$$
$$=: |I_1(t)| + |I_2(t)| + |I_3(t)|.$$

Using the fact that $|u(x)| \le \|u\|_W W(x)$, $u \in L_W^\infty$, $x \in X$, and (8) gives

$$\begin{aligned} |I_1(t)| &\le \alpha_t \int_{\mathbb{R}^k} |\phi_{\alpha_t}[F(k_t,s)] - \phi_{\alpha_t}^{(\varrho_t)}[F(k_t,s)]|\varrho(s)\,ds \\ &\le \alpha_t \|\phi_{\alpha_t} - \phi_{\alpha_t}^{(\varrho_t)}\|_W [\beta W(x_t) + b]. \end{aligned} \tag{37}$$

Taking $E_x^{\widehat{\pi}}$ on both sides of (37) and using Lemma 5.4, we get

$$E_x^{\widehat{\pi}} |I_1(t)| \to 0 \quad \text{as } t \to \infty. \tag{38}$$

To show that $E_x^{\widehat{\pi}} |I_2(t)| \to 0$, first we have, from the definition of α_t and (23),

$$\|\phi_{\alpha_t}^{(\varrho_t)}\|_W \le 2\|V_{\alpha_t}^{(\varrho_t)}\|_W \le \frac{2C}{1-\alpha_t} = \mathbf{O}(t^\nu).$$

Thus, from definition (19),

$$\begin{aligned} |I_2(t)| &\le \alpha_t \int_{\mathbb{R}^k} \phi_{\alpha_t}^{(\varrho_t)}[F(k_t,s)]|\varrho(s) - \varrho_t(s)|\,ds \\ &\le \alpha_t W(x_t) \|\phi_{\alpha_t}^{(\varrho_t)}\|_W \|\varrho - \varrho_t\|. \end{aligned} \tag{39}$$

Hence, taking expectation and applying Hölder's inequality we get

$$\begin{aligned} E_x^{\widehat{\pi}} |I_2(t)| &\le ([\mathbf{O}(t^\nu)]^{p'} E_x^{\widehat{\pi}} \|\varrho - \varrho_t\|^{p'})^{1/p'} \\ &= [\mathbf{O}(t^{\nu p' - \gamma})]^{1/p'} \to 0 \quad \text{as } t \to \infty, \end{aligned} \tag{40}$$

since $\nu < \gamma/p'$ [see definition of α_t].

For the term $|I_3(t)|$, from the definition of the policy $\widehat{\pi}$ combined with (25) and (21),

$$\begin{aligned} |I_3(t)| \le \Big| & c(k_t) + \alpha_t \int_{\mathbb{R}^k} \phi_{\alpha_t}^{(\varrho_t)}[F(k_t,s)]\varrho_t(s)\,ds \\ & - \inf_{a \in A(x_t)} \Big\{ c(x_t,a) + \alpha_t \int_{\mathbb{R}^k} \phi_{\alpha_t}^{(\varrho_t)}[F(x_t,a,s)]\varrho_t(s)\,ds \Big\} \Big| \\ & + \Big| \inf_{a \in A(x_t)} \Big\{ c(x_t,a) + \alpha_t \int_{\mathbb{R}^k} \phi_{\alpha_t}^{(\varrho_t)}[F(x_t,a,s)]\varrho_t(s)\,ds \Big\} \\ & - \inf_{a \in A(x_t)} \Big\{ c(x_t,a) + \alpha_t \int_{\mathbb{R}^k} \phi_{\alpha_t}[F(x_t,a,s)]\varrho(s)\,ds \Big\} \Big| \\ \le{} & \delta_t + \alpha_t \sup_{a \in A(x_t)} \Big| \int_{\mathbb{R}^k} \phi_{\alpha_t}^{(\varrho_t)}[F(x_t,a,s)]\varrho_t(s)\,ds \\ & - \int_{\mathbb{R}^k} \phi_{\alpha_t}[F(x_t,a,s)]\varrho(s)\,ds \Big|. \end{aligned}$$

Hence, from definition (19),

$$
\begin{aligned}
|I_3(t)| &\le \delta_t + \alpha_t \sup_{a\in A(x_t)} \int_{\mathbb{R}^k} \phi_{\alpha_t}^{(\varrho_t)}[F(x_t,a,s)]|\varrho(s)-\varrho_t(s)|\,ds \\
&\quad + \alpha_t \sup_{a\in A(x_t)} \int_{\mathbb{R}^k} |\phi_{\alpha_t}^{(\varrho_t)}[F(x_t,a,s)] - \phi_{\alpha_t}[F(x_t,a,s)]|\varrho(s)\,ds \\
&\le \delta_t + \alpha_t W(x_t)\|\phi_{\alpha_t}^{(\varrho_t)}\|_W\|\varrho-\varrho_t\| + \alpha_t\|\phi_{\alpha_t} - \phi_{\alpha_t}^{(\varrho_t)}\|_W[\beta W(x)+b].
\end{aligned}
$$

Hence, from (37)–(40), we get $E_x^{\widehat{\pi}}|I_3(t)| \to \delta$ as $t\to\infty$. Therefore

$$E_x^{\widehat{\pi}}[\mathfrak{L}_t] \to \delta \quad \text{as } t\to\infty. \tag{41}$$

Finally, from (35), (36) and (41), for any $k\ge 1$ and $n\to\infty$ we have

$$n^{-1}E_x^{\widehat{\pi}}\Big[\sum_{t=k}^{n} c(k_t) - j_{\alpha_t}\Big] = (1-\alpha_k)C' + \mathbf{o}(1) + \delta, \qquad x\in X.$$

Hence, from (12), the fact that $\lim_{t\to\infty}\alpha_t = 1$ and (2),

$$J(\widehat{\pi},x) \le j^* + \delta, \qquad x\in X.$$

This completes the proof of the theorem. ■

Comments. We have presented a construction of an average optimal adaptive policy, the basic idea being to use the so-called vanishing discount factor approach, and ensure the existence of δ-minimizers. On the other hand, it is well known (see, for instance, [4], [9]) that an optimal stationary policy exists if the minimum on the right-hand side of (11) is attained for each $x\in X$. Therefore, it can happen that under the assumptions made in this paper, such a policy does not exist for the process (1) with a known density ϱ.

References

[1] D. Blackwell, *Discrete dynamic programming*, Ann. Math. Statist. 33 (1962), 719–726.

[2] E. B. Dynkin and A. A. Yushkevich, *Controlled Markov Processes*, Springer, New York, 1979.

[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 J. A. Minjárez-Sosa, *Adaptive control for discrete-time Markov processes with unbounded costs*: *discounted criterion*, Kybernetika 34 (1998), no. 2, 217–234.

[6] —, —, *Adaptive control for discrete-time Markov processes with unbounded costs*: *average criterion*, Math. Methods Oper. Res. 48 (1998), 37–55.

[7] R. Hasminskii and I. Ibragimov, *On density estimation in the view of Kolmogorov's ideas in approximation theory*, Ann. Statist. 18 (1990), 999–1010.

[8] O. Hernández-Lerma, *Adaptive Markov Control Processes*, Springer, New York, 1989.

[9] —, *Infinite-horizon Markov control processes with undiscounted cost criteria: from average to overtaking optimality*, Reporte Interno 165, Departamento de Matemáticas, CINVESTAV-IPN, México, 1994.

[10] 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.

[11] S. A. Lippman, *On dynamic programming with unbounded rewards*, Manag. Sci. 21 (1975), 1225–1233.

[12] P. Mandl, *Estimation and control in Markov chains*, Adv. Appl. Probab. 6 (1974), 40–60.

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

[14] 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
Departamento de Matemáticas
Universidad de Sonora
Rosales s/n Col. Centro
C.P. 83000, Hermosillo, Son., México
E-mail: aminjare@gauss.mat.uson.mx

Received on 4.8.1998