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 \( x_{t+1} = F(x_t, a_t, \xi_t), \quad t = 1, 2, \ldots, \) with i.i.d. \( \mathbb{R}^k \)-valued random vectors \( \xi_t \), which are observable but whose density \( \rho \) is unknown.

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

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

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 \( \rho \). 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 \( \rho \) is unknown, to construct an adaptive policy in this paper, we introduce first a suitable method of statistical estimation of \( \rho \), and then apply

[267]
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 $\hat{\rho}$ such that provide information about the $L_q$-norm accuracy $\|\hat{\rho}_t - \rho\|_q$ of the estimators $\hat{\rho}_t, t = 1, 2, \ldots$.

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 $\rho$ by its estimators $\hat{\rho}_t$, which are obtained using the procedure of statistical estimation proposed in [5], [6], and finally exploit the corresponding $\alpha_t$-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 equation (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 \(\varrho\) (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 \(K = \{(x, a) : x \in X, a \in A(x)\}\) 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 \(K\), possibly unbounded.

Let \(\Pi\) be the set of all control policies and \(\mathcal{F} \subset \Pi\) be the set of all deterministic stationary policies \([2]\). As usual, every stationary policy \(\pi \in \mathcal{F}\) 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 \(f\) for \(\pi\) 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 \(\alpha\)-discount cost as

\[
V_\alpha(\pi, x) := E^\pi_x \left[ \sum_{t=0}^{\infty} \alpha^t c(x_t, a_t) \right],
\]

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

\[
J(\pi, x) := \limsup_{n \to \infty} n^{-1} E^\pi_x \left[ \sum_{t=0}^{n-1} c(x_t, a_t) \right],
\]

where \(E^\pi_x\) denotes the expectation operator with respect to the probability measure \(P^\pi_x\) induced by the policy \(\pi\), given the initial state \(x_0 = x\) (see, e.g., [2]).

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

\[
V_\alpha(x) := \inf_{\pi \in \Pi} V_\alpha(\pi, x) = V_\alpha(\pi^*, x), \quad x \in X.
\]

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

\[
J(x) := \inf_{\pi \in \Pi} J(\pi, x) = J(\pi^*, x), \quad x \in X.
\]
3. Assumptions. For a given measurable function $W : X \to [1, \infty)$, we denote by $L_{\infty}^W$ the normed linear space of all measurable functions $u : X \to \mathbb{R}$ with

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

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

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

where $1_B(\cdot)$ stands for the indicator function of the set $B$, and $\mathcal{B}(X)$ is the Borel $\sigma$-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 $\sup_{a \in A(x)} |c(x, a)| \leq W(x)$;

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

Now, we define a set of densities $\rho$ of the r.v.'s $\xi_t$ 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 $\overline{\rho} : \mathbb{R}^k \to \mathbb{R}$ which is used as a known majorant of the unknown densities $\rho$.

Setting $q := 1 + 2\varepsilon$, we define the set $D_0 = D_0(\overline{\rho}, L, \beta_0, b_0, p, q, m, \psi, \overline{\psi})$ to consist of all densities $\mu$ 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 $z \in \mathbb{R}^k$,

$$
\|\Delta_z \mu\|_{L_q} \leq L|z|^{1/q},
$$

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

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

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

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

(i) $Q_\mu(B \mid x, f) \geq \psi_f(x)m(B)$;

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

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

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

$$
\int_{\mathbb{R}^k} W^p[F(x,f,s)]\mu(s) \, ds \leq \beta_0 W^p(x) + b_0, \quad x \in X, \ a \in A(x),
$$

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 $p$, $\beta_0$ and $b_0$.

Assumption 3.3. (a) The density $\varrho$ belongs to $D_0$.

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

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

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

Remark 3.4. The function $\phi$ in (7) can be nonmeasurable. In this case we suppose the existence of a measurable majorant $\overline{\phi}$ of $\phi$ 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 $\varrho$ satisfies the condition (6). Then:

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

$$
\int_{\mathbb{R}^k} W[F(x,a,s)]\varrho(s) \, ds \leq \beta W(x) + b,
$$

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

(b) $\sup_{t \geq 1} E_\pi [W_\pi(x_t)] < \infty$ and $\sup_{t \geq 1} E_x^\pi [W(x_t)] < \infty$ 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 $\varrho$ satisfies the condition (6) or (8), then under Assumption 3.1(a), we have $V_\alpha(x) \leq CW(x)/(1-\alpha)$ for some constant $C > 0$, and $V_\alpha(\cdot)$ satisfies the dynamic programming equation, i.e.,

$$
V_\alpha(x) = \inf_{a \in A(x)} \left[ c(x,a) + \alpha \int_{\mathbb{R}^k} V_\alpha[F(x,a,s)]\varrho(s) \, ds \right], \quad x \in X;
$$
(b) under Assumption 3.1, for each $\delta > 0$, there exists a policy $f \in F$ such that
\begin{equation}
    c(x, f) + \alpha \int_{\mathbb{R}^k} V_{\alpha}[F(x, f, s)]q(s) \, ds \leq V_{\alpha}(x) + \delta, \quad x \in X.
\end{equation}

From the fact that $Q_q(\cdot | \cdot)$ is a stochastic kernel [see (4)], it is easy to prove that for a nonnegative function $u \in L^\infty_W$, and every $r \in \mathbb{R}$, the set
\begin{equation*}
    \{(x, a) : \int_{\mathbb{R}^k} u[F(x, a, s)]q(s) \, ds \leq r\}
\end{equation*}
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 $q \in D_0$. Then there exist a constant $j^*$ and a function $\phi$ in $L^\infty_W$ such that
\begin{equation}
    j^* + \phi(x) \geq \inf_{a \in A(x)} \left[ c(x, a) + \int_{\mathbb{R}^k} \phi[F(x, a, s)]q(s) \, ds \right],
\end{equation}
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 $j^* = \limsup_{\alpha \to 1} j_\alpha$ 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 $j^* = \liminf_{\alpha \to 1} j_\alpha$. Hence,
\begin{equation}
    \lim_{t \to \infty} j_{\alpha_t} = j^*
\end{equation}
for any sequence $\{\alpha_t\}$ of discount factors such that $\alpha_t \not\to 1$ (see also [3]). In fact $(j^*, \phi)$, with $\phi(x) := \lim_{t \to \infty} \phi_{\alpha_t}(x), \quad x \in X,$ satisfies the optimality inequality (11), where $\phi_{\alpha_t}(x) := V_{\alpha_t}(x) - V_{\alpha_t}(z)$. Furthermore, also in [4] it was proved that
\begin{equation}
    \sup_{\alpha \in (0, 1)} \|\phi_{\alpha}\|_W < \infty.
\end{equation}

(b) From the definition of $j_\alpha$ and $\phi_{\alpha}$, it is easy to see that the equation (9) and the inequality (10) are equivalent, respectively, to
\begin{equation}
    j_\alpha + \phi_{\alpha}(x) = \inf_{a \in A(x)} \left[ c(x, a) + \alpha \int_{\mathbb{R}^k} \phi_{\alpha}[F(x, a, s)]q(s) \, ds \right], \quad x \in X, \ \alpha \in (0, 1),
\end{equation}
and
\begin{equation}
    c(x, f) + \alpha \int_{\mathbb{R}^k} \phi_{\alpha}[F(x, f, s)]q(s) \, ds \leq j_\alpha + \phi_{\alpha}(x) + \delta, \quad x \in X, \ \alpha \in (0, 1).
\end{equation}
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 $\widehat{\varrho}_t := \widehat{\varrho}_t(s; \xi_0, \xi_1, \ldots, \xi_{t-1}), s \in \mathbb{R}^k$, be an arbitrary estimator of $\varrho$ belonging to $L_q$, such that for some $\gamma > 0$,

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

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

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

$$D_1 := \{\mu : \mu \text{ is a density on } \mathbb{R}^k, \mu \in L_q \text{ and } \mu(s) \leq \overline{\varrho}(s) \text{ a.e.}\},$$

$$D_2 := \{\mu : \mu \text{ is a density on } \mathbb{R}^k, \mu \in L_q, \int W[F(x,a,s)]\mu(s) \, ds \leq \beta W(x) + b, (x,a) \in K\}$$

[see Lemma 4.1 for the constants $\beta$ and $b$].

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

$$\|\varrho_t - \varrho_t\|_q = \inf_{\mu \in D} \|\mu - \varrho_t\|_q, \quad t \in \mathbb{N},$$

that is, the density $\varrho_t$ in $D$ is a “best approximation” of the estimator $\varrho_t$ 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 $\varrho$. 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 $\mu$ on $\mathbb{R}^k$ by setting

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

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

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

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

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

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 $\hat{\varrho}_t \in D.$

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

$$j^{(\varrho_t)} + \phi^{(\varrho_t)}(x)$$

$$= \inf_{a \in A(x)} \left[ c(x, a) + \alpha_t \int_{\mathbb{R}^k} \phi^{(\varrho_t)}[F(x, a, s)] \varrho_t(s) \, ds \right], \quad x \in X, \ t \in \mathbb{N},$$

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

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

$$T_{\mu} u(x)$$

$$:= \inf_{a \in A(x)} \left\{ c(x, a) + \alpha_t \int_{\mathbb{R}^k} u[F(x, a, s)] \mu(s) \, ds \right\}, \quad x \in X, \ u \in L^\infty_{\mathbb{W}}.$$

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 $\varrho$ satisfies (6) or (8). Then, for each $t \in \mathbb{N},$ $T_{\varrho} V_{\alpha_t} = V_{\alpha_t},$ $T_{\varrho_t} V^{(\varrho_t)}_{\alpha_t} = V^{(\varrho_t)}_{\alpha_t}$ and

$$V_{\alpha_t}(x) \leq \frac{C}{1 - \alpha_t} W(x), \quad V^{(\varrho_t)}_{\alpha_t}(x) \leq \frac{C}{1 - \alpha_t} W(x), \quad x \in X.$$

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

$$c(x, \tilde{f}_t) + \alpha_t \int_{\mathbb{R}^k} V^{(\varrho_t)}_{\alpha_t}[F(x, s)] \varrho_t(s) \, ds \leq V^{(\varrho_t)}_{\alpha_t}(x) + \delta_t, \quad x \in X,$$

or [see Remark 4.4(b)]

$$c(x, \tilde{f}_t) + \alpha_t \int_{\mathbb{R}^k} \phi^{(\varrho_t)}[F(x, s)] \varrho(s) \, ds$$

$$\leq j^{(\varrho_t)} + \phi^{(\varrho_t)}(x) + \delta_t, \quad x \in X.$$
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 $\{\hat{f}_t\}$ be a sequence of functions (selectors) satisfying (24) or (25) for each $t \in \mathbb{N}$. The adaptive policy $\hat{\pi} = \{\hat{\pi}_t\}$ is defined as $\hat{\pi}_t(h_t) = \hat{\pi}_t(h_t; \theta_t) := \hat{f}_t(x_t), t \in \mathbb{N}$, where $\hat{\pi}_0(x)$ 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 $\hat{\pi}$ is $\delta$-average cost optimal, i.e., for each $x \in X, J(\hat{\pi}, x) \leq j^* + \delta$, where $j^*$ is the optimal average cost as in Lemma 4.3. In particular, if $\delta = 0$ then the policy $\hat{\pi}$ 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 $t \to \infty$,

\[(a) \ E_x^\pi \| \phi_{\alpha_t} - \phi_{\alpha_t}^{(g)} \|^p_W \to 0, \quad (b) \ E_x^\pi \| \phi_{\alpha_t} - \phi_{\alpha_t}^{(g)} \|^W_{W(x)} \to 0.\]

**Proof.** (a) Observing that $\| \phi_{\alpha_t} - \phi_{\alpha_t}^{(g)} \|_W \leq 2 \| V_{\alpha_t} - V_{\alpha_t}^{(g)} \|^W_\|_W$, it is sufficient to prove

\[(26) \ \lim_{t \to \infty} E_x^\pi \| V_{\alpha_t} - V_{\alpha_t}^{(g)} \|^p_W = 0, \quad x \in X, \ \pi \in \Pi.\]

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

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

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

\[\|u\|_{W_t} \leq \|u\|_{W} \leq l_t \|u\|_{W_t}, \quad t \in \mathbb{N},\]

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

\[(27) \ E_x^\pi \| V_{\alpha_t} - V_{\alpha_t}^{(g)} \|^p_{W_t} \to 0 \quad \text{as } t \to \infty.\]

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

\[(28) \ |T_\mu v - T_\mu u|_{W_t} \leq \theta_t |v - u|_{W_t}, \quad v, u \in L_{W_t}^\infty, \quad t \in \mathbb{N}.\]

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

\[\| V_{\alpha_t} - V_{\alpha_t}^{(g)} \|_{W_t} \leq \| T_\mu V_{\alpha_t} - T_\mu V_{\alpha_t}^{(g)} \|_{W_t} + \theta_t \| V_{\alpha_t} - V_{\alpha_t}^{(g)} \|_{W_t},\]
which implies that

\[(29) \quad l_t \|V_{\alpha_t} - V_{\alpha_t}^{(g_t)}\|_{W_t} \leq \frac{l_t}{1 - \theta_t} \|T_gV_{\alpha_t} - T_{\theta_t} V_{\alpha_t}\|_{W_t}, \quad t \in \mathbb{N}.\]

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

\[(30) \quad \|T_gV_{\alpha_t} - T_{\theta_t} V_{\alpha_t}\|_{W_t}
\leq \alpha_t \sup_{x \in X} \|W_t(x)\|^{-1} \sup_{a \in A(x)} \int_{\mathbb{R}_+} V_{\alpha_t}[F(x, a, s)]|\varrho(s) - \varrho_t(s)| \, ds
\leq \frac{C\alpha_t}{1 - \alpha_t} \sup_{x \in X} \|W(x)\|^{-1} \sup_{a \in A(x)} \int_{\mathbb{R}_+} W[F(x, a, s)]|\varrho(s) - \varrho_t(s)| \, ds
\leq \frac{C}{1 - \alpha_t} \|\varrho - \varrho_t\|.
\]

Now, observe that [see definition of \(\alpha_t\) and \(\theta_t\)]

\[(31) \quad \frac{1}{(1 - \theta_t)(1 - \alpha_t)^2} = O(t^{3\nu}) \quad \text{as} \quad t \rightarrow \infty.
\]

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

\[(32) \quad l_t^\nu \|V_{\alpha_t} - V_{\alpha_t}^{(g_t)}\|_{W_t}^{p'} \leq C\|\varrho - \varrho_t\|^{p'}
\leq C\nu \nu' O(t^{3\nu'}) \|\varrho - \varrho_t\|^{p'} \quad \text{as} \quad t \rightarrow \infty.
\]

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

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

\[(33) \quad E_x^\pi \|\phi_{\alpha_t} - \phi_{\alpha_t}^{(g_t)}\|_{W}W(x_t) \leq \bar{C}(E_x^\pi \|\phi_{\alpha_t} - \phi_{\alpha_t}^{(g_t)}\|_{W}W(x_t))^{1/p'} \rightarrow 0 \quad \text{as} \quad t \rightarrow \infty.
\]

This completes the proof of Lemma 5.4. ■

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

\[(34) \quad \Sigma_t := c(k_t) + \alpha_t \int_{\mathbb{R}_+} \phi_{\alpha_t}[F(k_t, s)]\varrho(s) \, ds - j_{\alpha_t} - \phi_{\alpha_t}(x_t)
\leq c(k_t) + \alpha_t E_x^\pi [\phi_{\alpha_t}(x_{t+1}) | k_t] - j_{\alpha_t} - \phi_{\alpha_t}(x_t).
\]
Hence, for \( n \geq k \geq 1 \),

\[
(35) \quad n^{-1}E_x^n\left[ \sum_{t=k}^{n} c(k_t) - j_{\alpha_t} \right]
= n^{-1}E_x^n\left[ \sum_{t=k}^{n} (\phi_{\alpha_t}(x_t) - \alpha_t \phi_{\alpha_t}(x_{t+1})) \right] + n^{-1}E_x^n\left[ \sum_{t=k}^{n} \Sigma_t \right].
\]

On the other hand, from (34) and (14) we have

\[
\left| \left| E_x^n \left[ \phi_{\alpha_t}(x_t) - \alpha_t \phi_{\alpha_t}(x_{t+1}) \right] \right| \right|_W \leq \|u\|_W \|W(x)\|, \quad x \in X,
\]

we have \( E_x^n [\phi_{\alpha_t}(x)] < C', \quad \alpha \in (0, 1) \), for a constant \( C' < \infty \). Thus, denoting by \( \alpha_1^*, \ldots, \alpha_{k(n)}^* \), \( n \geq 1 \), the different values of \( \alpha_t \) for \( t \leq n \), and using the fact that \( \{\alpha_t\} \) is a nondecreasing sequence we have [see condition (20) and the definition of \( \phi_{\alpha_t} \)]

\[
(36) \quad n^{-1}E_x^n\left[ \sum_{t=k}^{n} (\phi_{\alpha_t}(x_t) - \alpha_t \phi_{\alpha_t}(x_{t+1})) \right] = n^{-1}E_x^n\left[ \sum_{t=k}^{n} (\phi_{\alpha_t}(x_t) - \alpha_t \phi_{\alpha_t}(x_{t+1})) \right] + n^{-1}E_x^n\left[ \sum_{t=k}^{n} \alpha_t (\phi_{\alpha_t}(x_t) - \phi_{\alpha_t}(x_{t+1})) \right] \leq (1 - \alpha_k)C' + n^{-1}2C' \sum_{i=1}^{k(n)} \alpha_t^* \leq (1 - \alpha_k)C' + 2C'k(n)n^{-1}, \quad x \in X.
\]

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

\[
\Sigma_t = c(k_t) + \alpha_t \int_{R^k} \phi_{\alpha_t} [F(k_t, s)] \varrho(s) \, ds
- \inf_{a \in A(x_t)} \left[ c(x_t, a) + \alpha_t \int_{R^k} \phi_{\alpha_t} [F(x_t, a, s)] \varrho(s) \, ds \right]
\leq \left| \alpha_t \int_{R^k} \phi_{\alpha_t} [F(k_t, s)] \varrho(s) \, ds - \alpha_t \int_{R^k} \phi_{\alpha_t}^{(\varphi)} [F(k_t, s)] \varrho(s) \, ds \right|
+ \left| \alpha_t \int_{R^k} \phi_{\alpha_t}^{(\varphi)} [F(k_t, s)] \varrho(s) \, ds - \alpha_t \int_{R^k} \phi_{\alpha_t}^{(\varphi)} [F(k_t, s)] \varrho(s) \, ds \right|
+ \left| c(k_t) + \alpha_t \int_{R^k} \phi_{\alpha_t}^{(\varphi)} [F(k_t, s)] \varrho(s) \, ds \right|
- \inf_{a \in A(x_t)} \left[ c(x_t, a) + \alpha_t \int_{R^k} \phi_{\alpha_t} [F(x_t, a, s)] \varrho(s) \, ds \right]
=: |I_1(t)| + |I_2(t)| + |I_3(t)|.
\]
Using the fact that $|u(x)| \leq \|u\|_W W(x)$, $u \in L^\infty_W$, $x \in X$, and (8) gives

$$|I_1(t)| \leq \alpha_t \int_{\mathbb{R}^k} |\phi_{\alpha_t}[F(k_t, s)] - \phi_{\alpha_t}^{(\nu_1)}[F(k_t, s)]| \varrho(s) \, ds$$

$$\leq \alpha_t \|\phi_{\alpha_t} - \phi_{\alpha_t}^{(\nu_1)}\|_W |\beta W(x_t) + b|.$$

Taking $E^\pi_x$ on both sides of (37) and using Lemma 5.4, we get

$$E^\pi_x|I_1(t)| \to 0 \quad \text{as } t \to \infty. \quad (38)$$

To show that $E^\pi_x|I_2(t)| \to 0$, first we have, from the definition of $\alpha_t$ and (23),

$$\|\phi_{\alpha_t}^{(\nu_1)}\|_W \leq 2\|V_{\alpha_t}^{(\nu_1)}\|_W \leq \frac{2C}{1 - \nu_1} = O(t^{\nu_1}).$$

Thus, from definition (19),

$$|I_2(t)| \leq \alpha_t \int_{\mathbb{R}^k} \phi_{\alpha_t}^{(\nu_1)}[F(k_t, s)]| \varrho(s) - \varrho_t(s)\| \, ds$$

$$\leq \alpha_t W(x_t)\|\phi_{\alpha_t}^{(\nu_1)}\|_W \|\varrho - \varrho_t\|.$$

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

$$E^\pi_x|I_2(t)| \leq \left(\|O(t^{\nu_1})\|^{p'} E^\pi_x \|\varrho - \varrho_t\|^{p'}\right)^{1/p'}$$

$$= \left[O(t^{\nu_1 - \gamma})\right]^{1/p'} \to 0 \quad \text{as } t \to \infty,$$

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

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

$$|I_3(t)| \leq \left|c(k_t) + \alpha_t \int_{\mathbb{R}^k} \phi_{\alpha_t}^{(\nu_1)}[F(k_t, s)]\varrho_t(s) \, ds\right|$$

$$- \inf_{a \in A(x_t)} \left\{ c(x_t, a) + \alpha_t \int_{\mathbb{R}^k} \phi_{\alpha_t}^{(\nu_1)}[F(x_t, a, s)]\varrho_t(s) \, ds \right\}$$

$$+ \left| \inf_{a \in A(x_t)} \left\{ c(x_t, a) + \alpha_t \int_{\mathbb{R}^k} \phi_{\alpha_t}^{(\nu_1)}[F(x_t, a, s)]\varrho_t(s) \, ds \right\} \right|$$

$$- \inf_{a \in A(x_t)} \left\{ c(x_t, a) + \alpha_t \int_{\mathbb{R}^k} \phi_{\alpha_t}[F(x_t, a, s)]\varrho_t(s) \, ds \right\}$$

$$\leq \delta_t + \alpha_t \sup_{a \in A(x_t)} \left| \int_{\mathbb{R}^k} \phi_{\alpha_t}^{(\nu_1)}[F(x_t, a, s)]\varrho_t(s) \, ds \right|$$

$$- \int_{\mathbb{R}^k} \phi_{\alpha_t}[F(x_t, a, s)]\varrho_t(s) \, ds \right|.$$

Hence, from definition (19),
\[ |I_3(t)| \leq \delta_t + \alpha_t \sup_{a \in A(x_t)} \int \phi_{\alpha_t}^{(g_t)}[F(x_t, a, s)]g(s) - g_t(s) \, ds \]

\[ + \alpha_t \sup_{a \in A(x_t) \not\in \mathcal{B}} \int \phi_{\alpha_t}^{(g_t)}[F(x_t, a, s)] - \phi_{\alpha_t}[F(x_t, a, s)]g(s) \, ds \]

\[ \leq \delta_t + \alpha_t W(x_t)\|\phi_{\alpha_t}^{(g_t)}\|_W - g_t + \alpha_t\|\phi_{\alpha_t} - \phi_{\alpha_t}^{(g_t)}\|_W[\beta W(x) + b]. \]

Hence, from (37)–(40), we get \( E^{\hat{x}}_x|I_3(t)| \to \delta \) as \( t \to \infty \). Therefore

(41)

\[ E^{\hat{x}}_x[I_3] \to \delta \quad \text{as} \quad t \to \infty. \]

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

\[ n^{-1}E^{\hat{x}}_x\left[ \sum_{i=k}^n c(k_i) - j_{\alpha_t} \right] = (1 - \alpha_k)C' + o(1) + \delta, \quad x \in X. \]

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

\[ J(\hat{\pi}, x) \leq j^* + \delta, \quad x \in X. \]

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 \( \delta \)-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 \( g \).

**References**


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