Morfismos, Vol. 3, No. 1, 1999, pp. 41–67

ADAPTIVE POLICIES FOR DISCRETE-TIME MARKOV CONTROL PROCESSES WITH UNBOUNDED COSTS: AVERAGE AND DISCOUNTED CRITERIA [*]

J. ADOLFO MINJÁREZ-SOSA [1]

Abstract

We consider a class of discrete-time Markov control processes with Borel state and action spaces, and possibly unbounded costs. The processes evolve according to the system equation $x_{t+1} = F(x_t, a_t, \xi_t),\ t = 1, 2, ...$ with i.i.d. $\Re^k-$ valued random vectors ξ_t, whose density ρ is unknown. Assuming observability of $\{\xi_t\}$, we introduce two adaptive policies which are, respectively, asymptotically discounted cost optimal and average cost optimal.

1991 Mathematics Subject Classification: 93E20, 90C40.
Keywords and phrases: Markov control process, discounted and average cost criteria, adaptive policy.

1 Introduction

We consider a class of discrete-time Markov control processes (MCPs) evolving according to the system equation

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

[*] Work supported by Consejo Nacional de Ciencia y Tecnologia (CONACyT) under Grant 3871P-E9607.

[1] Graduate student, Universidad Autónoma Metropolitana-Iztapalapa.

where F is a known function, x_t, a_t, and ξ_t are the state, action, and the random disturbance at time t, respectively. The disturbances are independent and identically distributed (i.i.d.) random vectors in $\Re^k$ having density ρ which is unknown to a controller.

Assuming that the realizations of the processes $\{\xi_t\}$ and $\{x_t\}$ are completely observable, our main objective is to introduce adaptive policies which are (1) asymptotically optimal with respect to the discounted criterion, and (2) optimal in the average case. Since ρ is unknown, this adaptive policies combine suitable methods of statistical estimation of ρ and choice of actions a_t as a function of a "history" $(x_0, a_0, \xi_0, ..., x_{t-1}, a_{t-1}, \xi_{t-1}, x_t)$ and of an estimator ρ_t of ρ.

The first adaptive policy is obtained by applying the "Principle of Estimation and Control" described by Mandl in [15] as the method of substituting the estimates into optimal stationary controls. This policy has been studied, for instance, in [11], [13], [3], all them considering bounded one-stage costs.

The average optimality of the second policy is studied via the average cost optimality inequality, and using a variant of the so-called vanishing discount factor approach [2], for which, taking advantage of the results obtained for the discounted case, we fix an appropriate sequence $\{\alpha_t\}$, $\alpha_t \nearrow 1$, of discount factors, and exploit the corresponding α_t- discounted optimality equations, taking limit as $t \to \infty$. This policy was originally introduced in [4] and revised in [11], both considering bounded one- stage costs.

Allowing unbounded costs imposes serious difficulties. First, the nice contractive-operator techniques do not work for both, discounted and average criteria, and so, we are forced to impose Lippman-like conditions ([14], [20]) on the transition probability of the process; we are thus able to use the results in [5]. Second, we need methods of statistical estimation of ρ that provide information about the L_q-norm accuracy $\|\rho_t - \rho\|_q$ of estimator ρ_t, $t = 1, 2, ...$

The paper is organized as follows. In Section 2 we introduce the Markov control model we deal with. Next, in Section 3, we list some preliminary results, proved in previous works, that summarize important facts to be used in Sections 4 and 5, where we construct, respectively, adaptive policies asymptotically discounted cost optimal and average cost optimal. Finally, an example of a queueing system with controllable service rate that satisfies all hypotheses of the paper is described in Section 6.

2 The control Model

We consider a class of discrete-time Markov control models $(X, A, \Re^k, F, \rho, c)$ satisfying the following conditions.

The state space X, and the action space A are both Borel spaces. The dynamics is defined by the system equation (1). Here $F : X \times A \times \Re^k \to X$ is a given (known) 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 $\Re^k$ and a common distribution with an unknown density ρ, that belongs to a given class described below.

For each $x \in X$, $A(x)$ denotes the set of *admissible controls* (*or actions*) when the state is x. The sets $A(x)$ are supposed to be nonempty measurable subsets of A, and the set

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

of admissible state-action pairs is assumed to be a Borel subset of the Cartesian product of X and A. Finally, the cost-per-stage $c(x, a)$ is a nonnegative measurable real-valued function on $\mathbf{K}$, possibly unbounded.

For each density μ on $\Re^k$, $Q_\mu(\cdot \mid \cdot)$ is a stochastic kernel on X given $\mathbf{K}$, defined as

$$Q_\mu(B \mid x, a) := \int\limits_{\Re^k} 1_B[F(x, a, s)]\mu(s)ds, \;\; B \in \mathbf{B}(X), \;\; (x, a) \in \mathbf{K}, \qquad (2)$$

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

We define the spaces of admissible histories up to time t by $\mathbf{H}_0 := X$ and $\mathbf{H}_t := (\mathbf{K} \times \Re^k)^t \times X$, $t \in \mathbf{N} := \{1, 2, ...\}$. A generic element of $\mathbf{H}_t$ is written as $h_t = (x_0, a_0, \xi_0, ..., x_{t-1}, a_{t-1}, \xi_{t-1}, x_t)$. A control policy $\pi = \{\pi_t\}$ is a sequence of measurable functions $\pi_t : \mathbf{H}_t \to A$ such that

$\pi_t(h_t) \in A(x_t),\ h_t \in \mathbf{H}_t,\ t \geq 0$. By Π we denote the set of all control policies and by $\mathbf{F} \subset \Pi$ the subset of stationary policies. As usual, every stationary policy $\pi \in \mathbf{F}$ is identified with a measurable function $f : X \to A$ such that $f(x) \in A(x)$ for every $x \in X$, so that π is of the form $\pi = \{f,\ f,\ f, ...\}$. In this case we use the notation f for π and we write

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

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

$$V_\alpha(\pi, x) := E_x^\pi \left[\sum_{t=0}^{\infty} \alpha^t c(x_t, a_t) \right], \tag{3}$$

where $\alpha \in (0, 1)$ is the so-called discount factor, and 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., [1], p. 140). We also define the long run expected average cost as

$$J(\pi, x) := \limsup_{n\to\infty} n^{-1} E_x^\pi \left[\sum_{t=0}^{n-1} c(x_t, a_t) \right]. \tag{4}$$

The functions

$$V_\alpha(x) := \inf_{\pi\in\Pi} V_\alpha(\pi, x) \quad \text{and} \quad J(x) := \inf_{\pi\in\Pi} J(\pi, x),\ x \in X, \tag{5}$$

are the optimal $\alpha-$ discounted cost and the optimal average cost, respectively, when the initial state is x. A policy $\pi^* \in \Pi$ is said to be $\alpha-$discounted optimal (or simply $\alpha-$ optimal) if $V_\alpha(x) = V_\alpha(\pi^*, x)$ for all $x \in X$. Similarly, a policy $\pi^* \in \Pi$ is said to be average cost optimal (AC- optimal) if $J(x) = J(\pi^*, x)$ for all $x \in X$.

For a given measurable function $W : X \to [1, \infty)$, L_W^∞ denotes the normed linear space of all measurable functions $u : X \to \Re$ with

$$\|u\|_W := \sup_{x\in X} \frac{|u(x)|}{W(x)} < \infty. \tag{6}$$

To guarantee the existence of "measurable minimizers" we need appropriate (semi-) continuity and (σ-) compactness conditions on some components of the Markov control model, as follows.

Assumption 2.1 *a) For every $x \in X$ the function $a \to c(x,a)$ is lower semicontinuous (l.s.c.) and $\sup\limits_{A(x)} |c(x,a)| \leq W(x)$;*

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

Remark 2.2 *Throughout the paper, we fix an arbitrary $\varepsilon \in (0,1/2)$ and denote $q := 1+2\varepsilon$. Also we choose and fix a nonnegative function $\bar{\rho}$: $\Re^k \to \Re$ which is used as a known majorant of the unknown density ρ of the r.v.'s ξ_t in (1).*

Assumption 2.3 *a) For every $s \in \Re^k$,*

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

b) $\int\limits_{\Re^k} \varphi^2(s) \left|\bar{\rho}\ (s)\right|^{1-2\varepsilon} ds < \infty$.

The function φ in (7) might be nonmeasurable. In this case we suppose the existence of a measurable majorant $\bar{\varphi}$ of φ for which Assumption 2.3(b) holds.

3 Preliminaries on the discounted criterion

For construction of the adaptive policy under the discounted criterion, we suppose that ρ belongs to the set of densities D_0 defined as follows.

Definition 3.1 *The set $D_0 = D_0(\bar{\rho}, L, \beta_0, b_0, p, q)$ consists of the densities μ on $\Re^k$ for which the following conditions hold.*
a) $\mu \in L_q(\Re^k)$;

b) there exists a constant L such that for each $z \in \Re^k$

$$\left\|\Delta_z \mu\right\|_{L_q} \leq L\left|z\right|^{1/q}, \tag{8}$$

where $\Delta_z\mu(x) := \mu(x+z) - \mu(x)$, $x \in \Re^k$ and $|\cdot|$ is the Euclidean norm in $\Re^k$;

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

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

$$\int\limits_{\Re^k} W^p[F(x,a,s)]\mu(s)ds \leq \beta_0 W^p(x) + b_0, \tag{9}$$

where $p > 1,\ \beta_0 < 1,\ b_0 < \infty$ are arbitrary but fixed.

In Section 6 we give an example of a queueing system with a controllable service rate for which all assumptions presented in this paper hold.

Remark 3.2 *When $k = 1$ it is not difficult (see [17], p. 13) to show that a sufficient condition for (8) is the following. There are a finite set $G \subset \Re$ (possibly empty) and a constant $M \geq 0$ such that:*

i) μ has a bounded derivative μ' on $\Re \backslash G$ which belongs to L_q;

ii) the function $|\mu'(x)|$ is nonincreasing for $x \geq M$ and nondecreasing for $x \leq -M$.

Note that G includes points of discontinuity of μ if such points exist.

Now we state some results that will be useful in the next section. Each of these results is provided with references for its proof.

Lemma 3.3 *[7] Suppose that Assumption 2.1(a) holds and ρ satisfies the condition (9). Then*

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

$$\int\limits_{\Re^k} W[F(x,a,s)]\rho(s)ds \leq \beta W(x) + b, \tag{10}$$

where $\beta = \beta_0^{1/p},\ \ b = b_0^{1/p}$;

b) $\sup_{t\geq 1} E_x^\pi[W^p(x_t)] < \infty$ and $\sup_{t\geq 1} E_x^\pi[W(x_t)] < \infty$ for each $\pi \in \Pi,\ \ x \in X$.

Lemma 3.4 *Let $\alpha \in (0,1)$ be an arbitrary but fixed discount factor. Then,*

a) [12] if ρ satisfies either (9) or (10), then, under Assumption 2.1(a), we have that $V_\alpha(x) \leq CW(x)/(1-\alpha)$ for some constant $C>0$, and $V_\alpha(\cdot)$ satisfies the dynamic programming equation, i.e.,

$$V_\alpha(x) = \inf_{a\in A(x)} \left[c(x,a) + \alpha \int\limits_{\Re^k} V_\alpha[F(x,a,s)]\rho(s)ds \right], \quad x \in X; \qquad (11)$$

*b) under Assumption 2.1, for each $\delta > 0$ there exists a policy $f \in$**F** such that*

$$c(x,f) + \alpha \int\limits_{\Re^k} V_\alpha[F(x,f,s)]\rho(s)ds \leq V_\alpha(x) + \delta, \quad x \in X. \qquad (12)$$

From the fact that $Q_\rho(\cdot \mid \cdot)$ is a stochastic kernel [see (2)], it is easy to prove that for every non-negative function $u \in L_W^\infty$, and every $r \in \Re$, the set

$$\left\{ (x,a) : \int\limits_{\Re^k} u[F(x,a,s)]\rho(s)ds \leq r \right\}$$

is Borel in **K**. Using this fact, part (b) of Lemma 3.4 is a consequence of Corollary 4.3 in [18].

Density Estimation. To conclude this section, we present a procedure for the statistical estimation of ρ, for which we suppose $\rho \in D_0$.

Denote by $\xi_0, \xi_1, ..., \xi_{t-1}$ the independent realizations (observed up to the moment $t-1$) of r.v.'s with the unknown density $\rho \in D_0$. Let $\hat{\rho}_t := \hat{\rho}_t(s; \xi_0, \xi_1, ..., \xi_{t-1})$, $s \in \Re^k$, be an arbitrary estimator of ρ belonging to L_q, such that for some $\gamma > 0$

$$E \left\| \rho - \hat{\rho}_t \right\|_q^{\frac{qp'}{2}} = \mathbf{O}(t^{-\gamma}) \text{ as } t \to \infty, \qquad (13)$$

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

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

$$D_1 := \{\mu : \mu \text{ is a density on } \Re^k, \mu \in L_q \text{ and } \mu(s) \leq \bar{\rho}\ (s)\ a.e.\};$$

$$D_2 := \left\{\mu : \mu \text{ is a density on } \Re^k, \mu \in L_q, \int W[F(x,a,s)]\mu(s)ds \right. \\ \left. \leq \beta W(x) + b,\ (x,a) \in \mathbf{K}\right\} \tag{14}$$

[see Lemma 3.3 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 [7]. In fact, we have

$$\left\|\rho_t - \hat{\rho}_t\right\|_q = \inf_{\mu \in D} \left\|\mu - \hat{\rho}_t\right\|_q, \qquad t \in \mathbf{N}, \tag{15}$$

that is, the density $\rho_t \in D$ is the "best approximation" of the estimator $\hat{\rho}_t$ on the set D. The fact that $\rho \in D_0$ and Lemma 3.3(a) yield $\rho \in D_0 \subset D$. Examples of estimators satisfying (13) are given in [9].

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

$$\|\mu\| := \sup_X [W(x)]^{-1} \sup_{A(x)} \int_{\Re^k} W[F(x,a,s)]\mu(s)ds. \tag{16}$$

Lemma 3.5 *[7], [8] Suppose that Assumption 2.3 holds and $\rho \in D_0$. Then*

$$E\left\|\rho_t - \rho\right\|^{p'} = \mathbf{O}(t^{-\gamma}) \quad as\ t \to \infty. \tag{17}$$

Throughout the paper we will repeatedly use the following inequalities:

$$|u(x)| \leq \|u\|_W W(x) \tag{18}$$

and

$$\int_{\Re^k} u[F(x,a,s)]\mu(s)ds \leq \|u\|_W \left[\beta W(x) + b\right] \tag{19}$$

for all $u \in L_W^{\infty},\ \mu \in D,\ x \in X,\ a \in A(x)$. The relation (18) is a consequence of the definition of $\|\cdot\|_W$, and (19) holds because of (10) and the definition of D.

4 Adaptive policies in the discounted case

The optimality of adaptive policies constructed under the discounted criterion is studied in the sense of the following definition.

Definition 4.1 *a) [19] A policy $\pi \in \Pi$ is said to be asymptotically discount optimal if, for each $x \in X$,*

$$E_x^{\pi}\left[\Phi(x_t, a_t)\right] \to 0 \ \ as \ \ t \to \infty,$$

where $a_t = \pi_t(h_t)$ and

$$\Phi(x,a) := c(x,a) + \alpha \int\limits_{\Re^k} V_\alpha[F(x,a,s)]\rho(s)ds - V_\alpha(x), \qquad (20)$$

*for $(x,a) \in$ **K**, is the so-called discounted discrepancy function, which is nonnegative in view of Lemma 3.4.*

b) Let $\delta \geq 0$. A policy π is $\delta-$asymptotically discount optimal if, for each $x \in X$,

$$\limsup_{t\to\infty} E_x^{\pi}\left[\Phi(x_t, a_t)\right] \leq \delta.$$

For the construction of adaptive policies we replace the unknown density ρ by its estimates ρ_t and exploit the corresponding optimality equations [11]. To do this we need to extend some assertions in the previous section on the densities ρ_t.

The proof of Lemmas 3.3 and 3.4 (partly given in [12]) shows that the following assertions hold true (because only (10) is used here).

Proposition 4.2 *a) Suppose that Assumption 2.1(a) holds. Then, for each $t \in$**N** there is a unique function $V_t \in L_W^\infty$ such that*

$$V_t(x) = \inf_{A(x)} \left\{ c(x,a) + \alpha \int\limits_{\Re^k} V_t[F(x,a,s)]\rho_t(s)ds \right\}, \quad x \in X, \qquad (21)$$

*Moreover, $V_t(x) \leq C/(1-\alpha)W(x)$, $t \in$**N**, $x \in X$.*

b) Under Assumption 2.1, for each $t \in \boldsymbol{N}$ and $\delta_t^* > 0$ there exists a stationary policy $f_t \in \boldsymbol{F}$ such that

$$c(x, f_t) + \alpha \int\limits_{\Re^k} V_t[F(x, f_t, s)]\rho_t(s)ds \leq V_t(x) + \delta_t^*, \quad x \in X. \tag{22}$$

The minimization in (21) is done for every $\omega \in \Omega$. Similarly, in the following we suppose that the minimization of a term including the estimator ρ_t is done for every $\omega \in \Omega$.

Now we introduce an adaptive policy π^*, which is a slight extension of "The Principle of Estimation and Control" policy [15].

Definition 4.3 *Let $\{\delta_t^*\}$ be an arbitrary sequence of positive numbers, and $\{f_t\}$ a sequence of functions satisfying (22) for each $t \in \boldsymbol{N}$. We define the adaptive policy $\pi^* = \{\pi_t^*\}$ as follows:*

$$\pi_t^*(h_t) = \pi_t^*(h_t; \rho_t) := f_t(x_t), \quad h_t \in \boldsymbol{H}_t, \quad t \in \boldsymbol{N},$$

while $\pi_0^*(x)$ is any fixed action in $A(x)$.

We are now ready to state our first main result. Supposing that $\{\delta_t^*\}$ converges, we denote $\delta^* := \lim_{t\to\infty} \delta_t^*$.

Theorem 4.4 *Suppose that Assumptions 2.1 and 2.3 hold, and $\rho \in D_0$. Then the adaptive policy π^* is δ^*-asymptotically discount optimal. In particular, if $\delta^* = 0$ then the policy π^* is asymptotically discount optimal.*

Proof: For every $\mu \in D$ let us define the operator

$$T_\mu u(x) = \inf_{A(x)} \left\{ c(x, a) + \alpha \int\limits_{\Re^k} u[F(x, a, s)]\mu(s)ds \right\}, \tag{23}$$

$x \in X,\ u \in L_W^\infty$. By Assumption 2.1(a), the definition of D and (19), T maps L_W^∞ into itself.

Let us fix an arbitrary number $\theta \in (\alpha, 1)$ and set $\bar{W}\ (x) := W(x) + d,\ x \in X$, where $d := b\,(\theta/\alpha - 1)^{-1}$. Also we define the space $L^{\infty}_{\bar{W}}$ of measurable functions $u : X \to \Re$ with the norm

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

It is easy to see that

$$\|u\|_{\bar{W}} \leq \|u\|_{W} \leq \|u\|_{\bar{W}}\,(1 + d)\,. \tag{24}$$

Hence $L^{\infty}_{W} = L^{\infty}_{\bar{W}}$ and the norms $\|\cdot\|_{W}$ and $\|\cdot\|_{\bar{W}}$ are equivalent.

In Lemma 2 of [20] it was proved that the inequality

$$\int\limits_{\Re^k} W[F(x, a, s)]\mu(s)ds \leq W(x) + b$$

implies that the operator T_μ in (23) is a contraction with respect to the norm $\|\cdot\|_{\bar{W}}$ with modulus θ, that is,

$$\|T_\mu v - T_\mu u\|_{\bar{W}} \leq \theta\,\|v - u\|_{\bar{W}}\ , \quad v,\, u \in L_W\,. \tag{25}$$

By virtue of (11) and (25) the function V_α is a unique (in L^{∞}_{W}) fixed point of the operator T_ρ, while V_t is a fixed point (unique in L^{∞}_{W}) of T_{ρ_t} for each $t \in \mathbf{N}$, that is

$$T_\rho V_\alpha = V_\alpha, \quad T_{\rho_t} V_t = V_t. \tag{26}$$

We also have

$$\|V_\alpha - V_t\|_{\bar{W}} = \left\|T_\rho V_\alpha - T_{\rho_t} V_t\right\|_{\bar{W}} \leq \left\|T_\rho V_\alpha - T_{\rho_t} V_\alpha\right\|_{\bar{W}}$$

$$+ \left\|T_{\rho_t} V_\alpha - T_{\rho_t} V_t\right\|_{\bar{W}} \leq \left\|T_\rho V_\alpha - T_{\rho_t} V_\alpha\right\|_{\bar{W}} + \theta\,\|V_\alpha - V_t\|_{\bar{W}}\,,$$

or

$$\|V_\alpha - V_t\|_{\bar{W}} \leq \frac{1}{1 - \theta}\left\|T_\rho V_\alpha - T_{\rho_t} V_\alpha\right\|_{\bar{W}}\ , \quad t \in \mathbf{N}. \tag{27}$$

Now, from definition (16), Lemma 3.4 and the fact $[\bar{W}\ (\cdot)]^{-1} < [W(\cdot)]^{-1}$, we obtain

$$\left\|T_\rho V_\alpha - T_{\rho_t} V_\alpha\right\|_{\bar{W}} \leq \alpha \sup_X [\bar{W}\ (x)]^{-1} \sup_{A(x)} \int\limits_{\Re^k} V_\alpha[F(x,a,s)]\,|\rho(s)-\rho_t(s)|\,ds$$

$$\leq \frac{\alpha C}{1-\alpha} \sup_X [W(x)]^{-1} \sup_{A(x)} \int\limits_{\Re^k} W[F(x,a,s)]\,|\rho(s)-\rho_t(s)|\,ds$$

$$\leq \frac{C}{1-\alpha}\left\|\rho-\rho_t\right\|, \ \ t\in\mathbf{N}. \tag{28}$$

From (24) and combining (27) and (28), for each $t \in$**N**, we get

$$\left\|V_\alpha - V_t\right\|_W \leq (1+d)\left\|V_\alpha - V_t\right\|_{\bar{W}} \leq \frac{C(1+d)}{(1-\theta)(1-\alpha)}\left\|\rho-\rho_t\right\|. \tag{29}$$

On the other hand, we define for each $t \in$**N** the function $\Phi_t^* :$**K**$\to \Re$ as:

$$\Phi_t^*(x,a) := c(x,a) + \alpha \int\limits_{\Re^k} V_t[F(x,a,s)]\rho_t(s)ds - V_t(x), \ (x,a)\in\mathbf{K}.$$

By the definition of Φ (see (20)) we get (by adding and subtracting the term $\alpha\int_{\Re^k} V_t[F(x,a,s)]\rho(s)ds$)

$$|\Phi_t^*(x,a)-\Phi(x,a)| \leq$$

$$|V_\alpha(x)-V_t(x)| + \alpha\int\limits_{\Re^k} V_t[F(x,a,s)]\,|\rho_t(s)-\rho(s)|\,ds$$

$$+\alpha\int\limits_{\Re^k} |V_t[F(x,a,s)] - V_\alpha[F(x,a,s)]|\,\rho(s)ds$$

$$\leq \left\|V_\alpha - V_t\right\|_W W(x) + \frac{\alpha C}{1-\alpha}\int\limits_{\Re^k} W[F(x,a,s)]\,|\rho_t(s)-\rho(s)|\,ds$$

$$+\alpha[\beta W(x)+b]\left\|V_t - V_\alpha\right\|_W$$

for each (x,a) $\in$**K**, t $\in$**N** (see Proposition 4.2(a)). Hence, from the definition of $\|\cdot\|$ in (16) and the inequalities (29),

$$\sup_X [W(x)]^{-1} \sup_{A(x)} |\Phi_t^*(x,a)-\Phi(x,a)| \leq C'\left\|\rho_t-\rho\right\|, \tag{30}$$

where $C' = \frac{C}{1-\alpha}\left[1+\frac{1+\beta+b}{1-\theta}\right]$. Moreover, by the definition of the policy π^* (see Definition 4.3) and of the functions f_t in (22), we have $\Phi_t^*(\cdot,\pi_t^*(\cdot)) \leq \delta_t,\ t \in \mathbf{N}$. Thus

$$\begin{aligned}
\Phi(x_t,\pi_t^*(h_t)) &\leq |\Phi(x_t,\pi_t^*(h_t)) - \Phi_t^*(x_t,\pi_t^*(h_t)) + \delta_t| \\
&\leq \sup_{A(x_t)} |\Phi(x_t,a) - \Phi_t^*(x_t,a)| + \delta_t \\
&\leq W(x_t)\sup_X [W(x)]^{-1} \sup_{A(x)} |\Phi(x,a) - \Phi_t^*(x,a)| + \delta_t \\
&\leq C'W(x_t)\,\|\rho_t - \rho\| + \delta_t\ ,\quad t \in \mathbf{N}. \qquad (31)
\end{aligned}$$

The latter inequality implies

$$E_x^{\pi^*}[\Phi(x_t,a_t)] \leq C'E_x^{\pi^*}[W(x_t)\,\|\rho_t-\rho\|] + \delta_t,$$

and, therefore, to prove $\delta-$optimality of the policy π^* it is enough to show that $E_x^{\pi^*}[W(x_t)\,\|\rho_t-\rho\|] \to 0$ as $t \to \infty$. For this, denoting $\bar{C} := \left(E_x^{\pi^*}[W^p(x_t)]\right)^{1/p} < \infty$ [see Lemma 3.3(b)], and applying Holder's inequality, we have

$$E_x^{\pi^*}[W(x_t)\,\|\rho_t-\rho\|] \leq \bar{C}\left(E_x^{\pi^*}\left[\|\rho_t-\rho\|^{p'}\right]\right)^{1/p'}.$$

Observing that $E_x^{\pi}\left[\|\rho_t-\rho\|^{p'}\right] = E\left[\|\rho_t-\rho\|^{p'}\right]$ (since ρ_t do not depend on x and π), Lemma 3.5 yield the desired results. ■

5 Adaptive policies in the average case

Throughout this section, we suppose that ρ belongs to the set of densities D_0' defined as follows.

Definition 5.1 *Let q and $\bar{\rho}$ be as in Remark 2.2. We define the set $D_0' := D_0'(\bar{\rho}, L, \beta_0, b_0, p, q, m, \psi, \bar{\psi})$ as the set consisting of all densities μ on $\Re^k$ with the following properties:*

a) The conditions (a)-(c) in Definition 3.1 hold.

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

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

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

ii) $\int\limits_{\Re^k} W^p[F(x,f,s)]\mu(s)ds \leq \beta_0 W^p(x) + \psi_f(x) b_0$ *for some* $p > 1$, *with* $b_0 := \int\limits_X W^p(y) m(dy) < \infty$;

iii) $\inf\limits_{f\in\boldsymbol{F}} \int\limits_X \psi_f(x) m(dx) :=\bar{\psi} > 0.$

Remark 5.2 *a) The set D_0' is more restrictive than the set of densities D_0 used for the discounted criterion because of additional difficulties in the asymptotic analysis of the average cost.*

b) Observe that to define the set D_0 for the discounted criterion it was only necessary to impose the conditions 5.1(a) together with

$$\int\limits_{\Re^k} W^p[F(x,f,s)]\mu(s)ds \leq \beta_0 W^p(x) + b_0, \;\; x \in X,\, a \in A(x), \qquad (32)$$

where $p > 1$, $\beta_0 < 1$, $b_0 < \infty$. But, as was observed in Remark 2.2(b) in [8], the relation (32) follows from conditions c(i) and c(ii) using the same p, β_0 and b_0.

c) Considering Remark 5.2(b) , under Assumption 2.1 and supposing that ρ satisfies the relation (32) [see (9)], the results of Lemmas 3.2 and 3.3 hold.

The optimality of the policy constructed in this section is studied via the so- called average cost optimality inequality, which is stated in the following results.

Lemma 5.3 *[5] Suppose that Assumption 2.1 holds and $\rho \in D_0'$. Then, there exist a constant j^* and a function ϕ in L_W^∞ such that*

$$j^* + \phi(x) \geq \inf_{A(x)} \left[c(x,a) + \int\limits_{\Re^k} \phi[F(x,a,s)]\rho(s)ds \right], \qquad (33)$$

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

Remark 5.4 *a) In [5] 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 latter assertion, we can get also that* $j^* = \liminf_{\alpha \nearrow 1} j_\alpha$. *Hence,*

$$\lim_{t\to\infty} j_{\alpha_t} = j^*, \tag{34}$$

for any sequence $\{\alpha_t\}$ *of factor discount such that* $\alpha_t \nearrow 1$. *In fact* (j^*, ϕ), *with* $\phi(x) := \lim_{t\to\infty} \phi_{\alpha_t}(x)$, $x \in X$, *satisfies the optimality inequality (33), where* $\phi_\alpha(x) := V_\alpha(x) - V_\alpha(z)$. *Furthermore, also in [5] was proved that,*

$$\sup_{\alpha\in(0,1)} \|\phi_\alpha\|_W < \infty. \tag{35}$$

b) From the definition of j_α *and* ϕ_α, *it is easy to see that the equation (11) and the inequality (12) are equivalent, respectively, to*

$$j_\alpha + \phi_\alpha(x) = \inf_{A(x)} \left[c(x,a) + \alpha \int\limits_{\Re^k} \phi_\alpha[F(x,a,s)]\rho(s)ds \right], \tag{36}$$

and

$$c(x,f) + \alpha \int\limits_{\Re^k} \phi_\alpha[F(x,f,s)]\rho(s)ds \le j_\alpha + \phi_\alpha(x) + \delta. \tag{37}$$

for $x \in X,\ \alpha \in (0,1)$.

Remark on density estimation. Observe that the density estimation scheme for $\rho \in D_0$ proposed for the discounted criterion, only depends on the conditions (a)-(d) in Definition 3.1. In view of condition 5.1(a) in Definition 5.1 and Remark 5.2(b), we can also apply this procedure to construct an estimator ρ_t of a density $\rho \in D_0'$ such that $\rho_t \in D_1 \cap D_2$ [see (14)] and (17)] holds.

Let ν be an arbitrary real number such that $0 < \nu < \gamma/(3p')$ where γ and p' are from (13). We fix an arbitrary nondecreasing sequence of discount factors $\{\alpha_t\}$, such that $1 - \alpha_t = \mathbf{O}(t^{-\nu})$ as $t \to \infty$, and

$$\lim_{n\to\infty} \frac{\kappa(n)}{n} = 0, \tag{38}$$

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

For a fixed t, let $V_{\alpha_t}^{(\rho_t)}(\pi,x) := E_x^{\pi,\rho_t}\left[\sum_{n=0}^{\infty}\alpha_t^n c(x_n,a_n)\right]$ be the total expected α_t-discount cost for the process (1) in which all the r.v.'s $\xi_1,\xi_2,...,$ have the same density ρ_t, and let $V_{\alpha_t}^{(\rho_t)}(x) := \inf_{\pi\in\Pi} V_{\alpha_t}^{(\rho_t)}(\pi,x)$, $x\in X$, be the corresponding value function. The sequences $\phi_{\alpha_t}^{(\rho_t)}(\cdot)$ and $j_{\alpha_t}^{(\rho_t)}$ are defined accordingly [see Remark 5.4]. Thus [see (36)],

$$j_{\alpha_t}^{(\rho_t)} + \phi_{\alpha_t}^{(\rho_t)}(x) = \inf_{A(x)}\left[c(x,a) + \alpha_t \int\limits_{\Re^k} \phi_{\alpha_t}^{(\rho_t)}[F(x,a,s)]\rho_t(s)ds\right], \quad (39)$$

for all $x\in X,\ t\in$**N**. For each $t\in$**N** and $\mu\in D$, let us define the operator $T_{\mu,\alpha_t}\equiv T_\mu : L_W^\infty \to L_W^\infty$ as

$$T_\mu u(x) := \inf_{A(x)}\left\{c(x,a) + \alpha_t \int\limits_{\Re^k} u[F(x,a,s)]\mu(s)ds\right\}, \quad (40)$$

for $x\in X,\ u\in L_W^\infty$.

The proof of Lemma 3.4 and Proposition 4.2 shows that the following assertions hold.

Proposition 5.5 *a) Suppose that Assumption 2.1(a) holds and that ρ satisfies the condition (32). Then, for each $t\in$**N**, we have $T_\rho V_{\alpha_t} = V_{\alpha_t}$, $T_{\rho_t}V_{\alpha_t}^{(\rho_t)} = V_{\alpha_t}^{(\rho_t)}$, and*

$$V_{\alpha_t}(x) \le \frac{C}{1-\alpha_t}W(x), \qquad V_{\alpha_t}^{(\rho_t)}(x) \le \frac{C}{1-\alpha_t}W(x),\ x\in X. \quad (41)$$

*b) Under Assumptions 2.1, for each $t\in$**N**, $x\in X$ and $\delta_t>0$, there exists a policy $\hat f_t\in$**F** such that*

$$c(x,\hat f_t) + \alpha_t \int\limits_{\Re^k} V_{\alpha_t}^{(\rho_t)}[F(x,\hat f_t,s)]\rho_t(s)ds \le V_{\alpha_t}^{(\rho_t)}(x) + \delta_t, \quad (42)$$

or [see Remark 5.4(b)]

$$c(x,\hat f_t) + \alpha_t \int\limits_{\Re^k} \phi_{\alpha_t}^{(\rho_t)}[F(x,\hat f_t,s)]\rho_t(s)ds \le j_{\alpha_t}^{(\rho_t)} + \phi_{\alpha_t}^{(\rho_t)}(x) + \delta_t. \quad (43)$$

Definition 5.6 *Let* $\{\delta_t\}$ *be an arbitrary sequence of positive numbers and* $\left\{\hat{f}_t\right\}$ *be a sequence of functions in* $\boldsymbol{F}$ *satisfying (42) or (43) for each* $t \in \boldsymbol{N}$. *The adaptive policy* $\hat{\pi} = \{\hat{\pi}_t\}$ *is defined as* $\hat{\pi}_t(h_t) = \hat{\pi}_t(h_t; \rho_t) := \hat{f}_t(x_t),\ t \in \boldsymbol{N}$, *where* $\hat{\pi}_0(x)$ *is any fixed action.*

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

Theorem 5.7 *Suppose that Assumptions 2.1 and 2.3 hold, and* $\rho \in D_0'$. *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 5.3.*

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 5.8 *Under assumptions 2.1 and 2.3, for each* $x \in X$ *and* $\pi \in \Pi$, *we have*

$$a) \lim_{t\to\infty} E_x^{\pi} \left\| \phi_{\alpha_t} - \phi_{\alpha_t}^{(\rho_t)} \right\|_W^{p'} = 0, \quad b) \lim_{t\to\infty} E_x^{\pi} \left[\left\| \phi_{\alpha_t} - \phi_{\alpha_t}^{(\rho_t)} \right\|_W W(x_t) \right] = 0.$$

Part a) is proved observing first that

$$\left\| \phi_{\alpha_t} - \phi_{\alpha_t}^{(\rho_t)} \right\|_W \leq 2 \left\| V_{\alpha_t} - V_{\alpha_t}^{(\rho_t)} \right\|_W,$$

and so applying similar ideas to those used to prove (29), (see [16] for details) we get

$$\lim_{t\to\infty} E_x^{\pi} \left\| V_{\alpha_t} - V_{\alpha_t}^{(\rho_t)} \right\|_W^{p'} = 0, \quad x \in X,\ \pi \in \Pi.$$

For the part b), denoting $\bar{C} := \left(E_x^{\pi}\left[W^p(x_t)\right]\right)^{1/p} < \infty$ [see Lemma 3.3(b)], applying the Holder's inequality and using part a), we have as $t \to \infty$,

$$E_x^{\pi}\left\|\phi_{\alpha_t}-\phi_{\alpha_t}^{(\rho_t)}\right\|_W W(x_t)\leq \bar{C}\left(E_x^{\pi}\left[\left\|\phi_{\alpha_t}-\phi_{\alpha_t}^{(\rho_t)}\right\|_W^{p'}\right]\right)^{1/p'}\to 0.$$

Proof of Theorem 5.7.

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

$$L_t:=c(k_t)+\alpha_t\int\limits_{\Re^k}\phi_{\alpha_t}[F(k_t,s)]\rho(s)ds-j_{\alpha_t}-\phi_{\alpha_t}(x_t) \tag{44}$$

$$=c(k_t)+\alpha_t E_x^{\hat{\pi}}\left[\phi_{\alpha_t}(x_{t+1})\mid k_t\right]-j_{\alpha_t}-\phi_{\alpha_t}(x_t).$$

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

$$n^{-1}E_x^{\hat{\pi}}\left[\sum_{t=k}^{n}c(k_t)-j_{\alpha_t}\right]=n^{-1}E_x^{\hat{\pi}}\left[\sum_{t=k}^{n}\left(\phi_{\alpha_t}(x_t)-\alpha_t\phi_{\alpha_t}(x_{t+1})\right)\right]$$

$$+n^{-1}E_x^{\hat{\pi}}\left[\sum_{t=k}^{n}L_t\right]. \tag{45}$$

On the other hand, from (35), Lemma 3.3(b) [see Remark 5.2(c)] and (18) we have $E_x^{\hat{\pi}}\left[\phi_\alpha(x_t)\right]<C'$, $\alpha\in(0,1)$, for a constant $C'<\infty$. Thus, denoting $\alpha_1^*,\alpha_2^*,...,\alpha_{\kappa(n)}^*$, $n\geq 1$, the different values of α_t for $t\leq n$, and using that $\{\alpha_t\}$ is a nondecreasing sequence we have [see condition (38) and the definition of ϕ_α]

$$n^{-1}E_x^{\hat{\pi}}\left[\sum_{t=k}^{n}\left(\phi_{\alpha_t}(x_t)-\alpha_t\phi_{\alpha_t}(x_{t+1})\right)\right]$$

$$=n^{-1}E_x^{\hat{\pi}}\left[\sum_{t=k}^{n}\left(\phi_{\alpha_t}(x_t)-\alpha_t\phi_{\alpha_t}(x_t)\right)\right]$$

$$+n^{-1}E_x^{\hat{\pi}}\left[\sum_{t=k}^{n}\alpha_t\left(\phi_{\alpha_t}(x_t)-\phi_{\alpha_t}(x_{t+1})\right)\right]\leq(1-\alpha_k)C'+n^{-1}2C'\sum_{i=1}^{\kappa(n)}\alpha_i^*$$

$$\leq(1-\alpha_k)C'+2C'\kappa(n)n^{-1},\ x\in X. \tag{46}$$

Now, from (44) and (36) we have

$$L_t = c(k_t) + \alpha_t \int\limits_{\Re^k} \phi_{\alpha_t}[F(k_t,s)]\rho(s)ds$$

$$- \inf_{A(x_t)} \left[c(x_t,a) + \alpha_t \int\limits_{\Re^k} \phi_{\alpha_t}[F(x_t,a,s)]\rho(s)ds \right]$$

$$\leq |I_1(t)| + |I_2(t)| + |I_3(t)|,$$

where

$$I_1(t) := \alpha_t \int\limits_{\Re^k} \phi_{\alpha_t}[F(k_t,s)]\rho(s)ds - \alpha_t \int\limits_{\Re^k} \phi_{\alpha_t}^{(\rho_t)}[F(k_t,s)]\rho(s)ds,$$

$$I_2(t) := \alpha_t \int\limits_{\Re^k} \phi_{\alpha_t}^{(\rho_t)}[F(k_t,s)]\rho(s)ds - \alpha_t \int\limits_{\Re^k} \phi_{\alpha_t}^{(\rho_t)}[F(k_t,s)]\rho_t(s)ds,$$

$$I_3(t) := c(k_t) + \alpha_t \int\limits_{\Re^k} \phi_{\alpha_t}^{(\rho_t)}[F(k_t,s)]\rho_t(s)ds$$

$$- \inf_{A(x_t)} \left[c(x_t,a) + \alpha_t \int\limits_{\Re^k} \phi_{\alpha_t}[F(x_t,a,s)]\rho(s)ds \right]$$

Using (18) and (10) [see Remark 5.2(c)],

$$|I_1(t)| \leq \alpha_t \int\limits_{\Re^k} \left| \phi_{\alpha_t}[F(k_t,s)] - \phi_{\alpha_t}^{(\rho_t)}[F(k_t,s)] \right| \rho(s)ds$$

$$\leq \alpha_t \left\| \phi_{\alpha_t} - \phi_{\alpha_t}^{(\rho_t)} \right\|_W [\beta W(x_t) + b]. \tag{47}$$

Taking expectation $E_x^{\hat{\pi}}$ on both sides of (47) and using the Lemma 5.8, we get

$$E_x^{\hat{\pi}} |I_1(t)| \to 0, \text{ as } t \to \infty. \tag{48}$$

To show that $E_x^{\hat{\pi}} |I_2(t)| \to 0$, first we have, from definition of α_t and (41),

$$\left\|\phi_{\alpha_t}^{(\rho_t)}\right\|_W \leq 2\left\|V_{\alpha_t}^{(\rho_t)}\right\|_W \leq \frac{2C}{1-\alpha_t} = \mathbf{O}(t^{\nu}).$$

Thus, from definition (16),

$$\begin{aligned} |I_2(t)| &\leq \alpha_t \int\limits_{\Re^k} \phi_{\alpha_t}^{(\rho_t)}[F(k_t,s)]\,|\rho(s)-\rho_t(s)|\,ds \\ &\leq \alpha_t W(x_t)\left\|\phi_{\alpha_t}^{(\rho_t)}\right\|_W \|\rho-\rho_t\|. \qquad (49) \end{aligned}$$

Hence, taking expectation and applying Holder's inequality in (49) we get

$$\begin{aligned} E_x^{\hat{\pi}}\,|I_2(t)| &\leq \Big([\mathbf{O}(t^{\nu})]^{p'}\, E_x^{\hat{\pi}}\,\|\rho-\rho_t\|^{p'}\Big)^{1/p'} \\ &= \Big[\mathbf{O}(t^{\nu p'-\gamma})\Big]^{1/p'} \to 0 \text{ as } t\to\infty, \qquad (50) \end{aligned}$$

due to the fact $\nu < \gamma/p'$ [see definition of α_t].

For the term $|I_3(t)|$, from the definition of the policy $\hat{\pi}$ and combining (39) and (43), adding and substracting the term

$$\inf_{A(x_t)}\left\{c(x_t,a)+\alpha_t \int\limits_{\Re^k} \phi_{\alpha_t}^{(\rho_t)}[F(x_t,a,s)]\rho_t(s)ds\right\}$$

in $I_3(t)$, we get

$$\begin{aligned} |I_3(t)| \leq \delta_t + \alpha_t \sup_{A(x_t)} \Bigg| &\int\limits_{\Re^k} \phi_{\alpha_t}^{(\rho_t)}[F(x_t,a,s)]\rho_t(s)ds \\ &- \int\limits_{\Re^k} \phi_{\alpha_t}[F(x_t,a,s)]\rho(s)ds\Bigg| \end{aligned}$$

The latter inequality yields

$$|I_3(t)| \leq \delta_t + \alpha_t \sup_{A(x_t)} \int\limits_{\Re^k} \phi_{\alpha_t}^{(\rho_t)}[F(x_t,a,s)]\,|\rho(s)-\rho_t(s)|\,ds$$

$$+\alpha_t \sup_{A(x_t)} \int\limits_{\Re^k} \left| \phi_{\alpha_t}^{(\rho_t)}[F(x_t,a,s)] - \phi_{\alpha_t}[F(x_t,a,s)] \right| \rho(s)ds.$$

Thus, from (16),

$$|I_3(t)| \leq \delta_t + \alpha_t W(x_t) \left\| \phi_{\alpha_t}^{(\rho_t)} \right\|_W \|\rho - \rho_t\|$$

$$+\alpha_t \left\| \phi_{\alpha_t} - \phi_{\alpha_t}^{(\rho_t)} \right\|_W [\beta W(x) + b].$$

Hence, from (47), (48), (49) and (50), we get $E_x^{\hat{\pi}} |I_3(t)| \to \delta$, as $t \to \infty$. Therefore

$$E_x^{\hat{\pi}} [L_t] \to \delta, \text{ as } t \to \infty. \tag{51}$$

Finally, from (45), (46) and (51), we have for any $k \geq 1$ and $n \to \infty$,

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

It follows that [from (34), the fact that $\lim_{t\to\infty} \alpha_t = 1$, and (4)]

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

This completes the proof of the theorem. ■

6 Example

We consider a control process of the form

$$x_{t+1} = (x_t + a_t - \xi_t)^+, \ \ t = 0, 1, ..., \tag{52}$$

$x_0 = x$ given, with state space $X = [0, \infty)$ and actions set $A(x) = A$ for every $x \in X$, where A is a compact subset of some interval $(0, \theta]$, with $\theta \in A$.

Equations (52) describe, in particular, the model of a single server queueing system of type $GI/D/1/\infty$ with controlled service rates $a_t \in A$. In this case x_t denotes the waiting time of the $t-$th customer, while ξ_t denotes the interarrival time between the $t-$th and the $(t+1)-$th customers.

The random variables $\xi_0, \xi_1, ...,$ are supposed to be i.i.d. with unknown density $\rho \in L_q(\Re)$ satisfying the inequality

$$\left\| \triangle_z \rho \right\|_{L_q} \leq L \left| z \right|^{1/q},$$

for some given constants $L < \infty$, $q > 1$; or the hypotheses mentioned in Remark 3.2.

The following assumption ensures ergodicity of the system when using the slowest services: $a_t = \theta$, $t \geq 0$.

Assumption 6.1. $E(\xi_0)$ exist, and moreover $E(\xi_0) > \theta$.

Considering the function $\Psi(s) := e^{\theta s} E(e^{-s\xi_0})$ we find that Assumption 6.1 implies $\Psi'(0) < 0$, and so there is $\lambda > 0$ for which $\Psi(\lambda) < 1$. Also, by continuity of Ψ we can choose $p > 1$ such that

$$\Psi(p\lambda) := \beta_0 < 1. \tag{53}$$

Let us set $W(x) = \bar{b}\, e^{\lambda x}$, for all $x \in [0, \infty)$, where $\bar{b}$ is an arbitrary positive constant. Easy calculations shows that $\varphi(s) = \max\{1, e^{\lambda(\theta - s)}\} < \infty$ for every $s \in [0, \infty)$ [see definition of φ in (7)]. Therefore, to satisfy Assumption 2.3 we can take, for example,

$$\bar{\rho}\ (s) := M \min\{1, 1/s^{1+r}\}, \tag{54}$$

for all $s \in [0, \infty)$, where $r > 0$.

For $r < 1$ and choosing enough large M in (54), Assumption 6.1 implies $\rho \leq \bar{\rho}$ in a wide class of densities (see conditions (c) and (a) in Definitions 3.1 and 5.1, respectively).

On the other hand, in [6] were taken advantages of (53) and definition $\psi_f(x) := P[x + f(x) - \xi_0 \leq 0]$, $f \in \mathbb{F}$, to verify, for this example, the conditions (b) and (c) in Definition 5.1. Thus, we have $\rho \in D_0'$, and according to Remark 5.2(b), also $\rho \in D_0$.

Finally, to meet Assumption 2.1, the one-stage cost $c(x, a)$ can be chosen as any nonnegative measurable function which is l.s.c. in a and satisfying

$$\sup_A c(x, a) \leq \bar{b}\, e^{\lambda x}, \text{ for all } x \in [0, \infty).$$

7 Concluding remarks

It is well-known that $\alpha-$ optimal and AC$-$ optimal stationary policies exist if the minimum on the right-hand side of (11) and (33) is attained for each $x \in X$, respectively. Thus, to guarantee the existence of such policies it is necessary to impose rather restrictive continuity conditions on the one-stage cost c and the transition probability of the process, as well as compactness of the sets $A(x)$ (see, e.g., [10], pp. 18, 53). Hence, it can happen that under the assumptions made in this paper, stationary policies for the discounted and average criteria do not exist for the process (1) with a known density ρ, while Theorems 4.4 and 5.7 guarantee the existence of, respectively, asymptotically discounted optimal and average cost optimal adaptive policies. The latter theorems thus extend previous results in that they give conditions for the existence of ε-optimal policies with $\varepsilon = \delta^*$ and $\varepsilon = \delta$ in Theorems 4.4 and 5.7, respectively.

Acknowledgement

This article is part of the author's doctoral research under the direction of Dr. Evgueni Gordienko at the Universidad Autónoma Metropolitana-Iztapalapa. The author wishes to thank him for his guidance. The author also wishes to thank Dr. Onésimo Hernández-Lerma and anonymous referees for their very valuable comments and suggestions.

J. Adolfo Minjárez Sosa
Departamento de Matemáticas
Universidad de Sonora
Rosales s/n, Col. Centro
83000, Hermosillo, Son., México.
aminjare@gauss.mat.uson.mx

References

[1] D.P. Bertsekas and S.E. Shreve, Stochastic Optimal Control: The Discrete Time Case, Academic Press, New York, 1978.

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

[3] R. Cavazos-Cadena, *Nonparametric adaptive control of discounted stochastic system with compact state space*, J. Optim. Theory Appl., 65 (1990), 191-207.

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

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

[6] E.I. Gordienko and O. Hernández-Lerma, *Average cost Markov control processes with weighted norms: value iteration*, Applicationes Math., 23 (1995), 219-237.

[7] E.I. Gordienko and J.A. Minjárez-Sosa, *Adaptive control for discrete-time Markov processes with unbounded costs: discounted criterion,* Kybernetika, 34 (1998a), 2, 217-234.

[8] E.I. Gordienko and J.A. Minjárez-Sosa, *Adaptive control for discrete-time Markov processes with unbounded costs: average criterion*, ZOR- Math. Methods of Oper. Res., 48, Iss. 2, 1998b.

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

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

[11] 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.

[12] O. Hernández-Lerma, *Infinite-horizon Markov control processes with undiscounted cost criteria: from average to overtaking optimality*, Reporte Interno 165, Departamento de Matemáticas, CINVESTAV-IPN, A.P. 14-740, 07000, México, D.F., México, 1994.

[13] O. Hernández-Lerma and S.I. Marcus, *Adaptive policies for discrete-time stochastic control system with unknown disturbance distribution*, Systems and Control letters, 9 (1987), 307-315.

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

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

[16] J.A. Minjárez-Sosa, *Nonparametric adaptive control for discrete-time Markov processes with unbounded costs under average criterion.* To appear in Appl. Math. (Warsaw).

[17] J.A. Minjárez-Sosa, *Adaptive Control for Markov Processes with Unbounded Costs,* Doctoral Thesis, UAM-Iztapalapa, México, 1998. (In Spanish).

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

[19] M. Schal, *Estimation and control in discounted stochastic dynamic programming*, Stochastics 20 (1987), 51-71.

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