KYBERNETIKA — VOLUME 34 (1998), NUMBER 2, PAGES 217–234

ADAPTIVE CONTROL FOR DISCRETE-TIME MARKOV PROCESSES WITH UNBOUNDED COSTS: DISCOUNTED CRITERION[1]

EVGUENI I. GORDIENKO AND J. ADOLFO MINJÁREZ-SOSA

We study the adaptive control problem for discrete-time Markov control processes with Borel state and action spaces and possibly unbounded one-stage costs. The processes are given by recurrent equations $x_{t+1} = F(x_t, a_t, \xi_t)$, $t = 0, 1, \ldots$ with i.i.d. $\Re^k$-valued random vectors ξ_t whose density ρ is unknown. Assuming observability of ξ_t we propose the procedure of statistical estimation of ρ that allows us to prove discounted asymptotic optimality of two types of adaptive policies used early for the processes with bounded costs.

1. INTRODUCTION

The paper deals with finding of adaptive policies for Markov control processes of the following type:

$$x_{t+1} = F(x_t, a_t, \xi_t), \quad t = 0, 1, \ldots \tag{1.1}$$

defined on general state and action spaces. As a performance criterion we use asymptotic optimality with respect to discounted expected total cost with possibly unbounded nonnegative one-stage costs $c(x, a)$. We suppose "driving process" ξ_t in (1.1) to be independent and identically distributed random vectors in $\Re^k$ having a density ρ which is unknown to a controller. For the latter, the *adaptive policies* given in this paper combine suitable estimation of ρ and choice of actions a_t as a function of "a history" $(x_0, a_0, \ldots, x_{t-1}, a_{t-1}, x_t)$ and of an estimator ρ_t of ρ. To provide a sequence of estimators ρ_t, $t = 1, 2, \ldots$ we suppose the random vectors ξ_t, as well as the states x_t, are observable. These assumptions are satisfied in some applied problems, for instance in production-inventory systems, control of water reservoirs, some controlled queueing systems, etc. (see for example, [4, 12, 15]).

A similar problem has been solved for Markov control processes with bounded one-stage costs in [3, 7, 8, 14, 16, 17, 22]. The adaptive policies called Principle of Estimation and Control (PEC) and Nonstationary Value Iteration (NVI) proposed in these works and some techniques of proofs of their optimality were the starting points

[1] Research supported by Fondo del Sistema de Investigación del Mar de Cortés under Grant SIMAC/94/CT-005 and by Consejo Nacional de Ciencia y Tecnología (CONACyT) under grant 0635P-E9506.

for the present paper. On the other hand, unbounded costs impose serious additional obstacles. First, the operator in the optimality equation is no longer contractive in general, and so, we are forced to impose Lippman's like hypothesis [21,27] on transition probability of processes. Second, to be able to keep values of $c(x_t, a_t)$ from "escape to infinity" and to insure their uniform integrability when "quasioptimal" actions corresponded to the estimator ρ_t are chosen, we use projection of some estimator $\hat{\rho}_t$ on a certain "good" subset of densities in $L_q(\Re^k)$, $q > 1$. Also we need the convergence of $\hat{\rho}_t$ to an unknown ρ in L_q-norm. To meet the last condition we apply the recent results on density estimation in L_q [11]. Making of projection can be unpleasant operation when one tries to apply the control policies found here; but we do not known how overcome this difficulty. In any case, examples such as given in Section 6 make it reasonable that something as projection cannot be skipped completely.

The assumptions beyond briefly discussed, allows us to prove asymptotic optimality in the sense of [24] of both PEC and NVI adaptive policies for the unbounded cost case. To the best of our knowledge except of well-developed theory of linear stochastic systems, there are no papers which treat adaptive optimal policies for Markov control processes with unbounded costs. In adaptive control of systems given by linear equations the random vectors ξ_t in (1.1) are interpreted as "disturbance" of the system, and optimal policies, as a rule, depend on two first moments of ξ_t, but not on its distribution (if a cost is quadratic; see, for example, [20]). The situation is different for non-linear processes, as in the simple example of section 6, where a more refined information about a distribution of ξ_t is crucial for construction of optimal policies. Moreover, methods of theory of linear system depend heavily on specific structure of linear processes, and so they are not applicable in our case.

There are some new papers on adaptive control of Markov processes. We mention, for instance, [1,5,6,25,26]. All of them work with either finite state-action processes or Borel spaces and bounded cost functions.

The remainder of the paper is organized as follows. In Section 2 we introduce the Markov control model we deal with, whereas in Section 3 we list the assumptions on control model together with some preliminaries results. Assumptions and results related to density estimation are given in Section 4. Next, in section 5, we present the adaptive policies and the optimality result. An example of a queueing system with controllable service rate that satisfies all hypothesis of the paper is described in Section 6.

2. MARKOV CONTROL PROCESSES

We consider a class of discrete-time Markov control models $(X, A, \Re^k, F, \rho, c)$ with Borel spaces X, A, of states and actions, whose dynamic is defined by system equation (1.1).

In (1.1) $F : XA\Re^k \to X$ is a given (known) measurable function and $\{\xi_t\}$, so-called driving process, is a sequence of independent and identically distributed (i.i.d.) random vectors (r.v.'s) with values in $\Re^k$ and a common unknown distribution.

We suppose that the distribution of ξ_t has a density ρ which is unknown, but belongs to a given class described in next section. Moreover, we assume 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 (see the discussion of this hypothesis in the Introduction and Section 6).

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

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

of admissible state-action pairs is assumed to be a Borel subset of the Cartesian product of X and A. The last element of the model is a given one-stage cost, c, which is a nonnegative real-valued measurable function on $\mathbb{K}$ (possibly unbounded).

We define the spaces of admissible histories up the time t by $\mathbb{H}_0 := X$ and $\mathbb{H}_t := (\mathbb{K}\, \Re^k)^t X$, $t \in \mathbb{N} := \{1, 2, \ldots\}$. A control policy $\pi = \{\pi_t\}$ is a sequence of measurable functions $\pi_t : \mathbb{H}_t \to A$ such that $\pi_t(h_t) \in A(x_t)$, $h_t \in \mathbb{H}_t$, $t \geq 0$. By Π we denote the set of all control policies and by $\mathbb{F} \subset \Pi$ the set of all stationary policies. Every stationary policy $\pi \in \mathbb{F}$ is identified with some measurable functions $f : X \to A$ 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 $\boldsymbol{f}$ for π.

For an arbitrary policy $\pi \in \Pi$ and initial state $x \in X$, there exists a unique probability measure P_x^π on $\Omega := (XA\,\Re^k)^\infty$ (see e.g. [4,18]). Moreover, P_x^π satisfies $P_x^\pi[x_0 = x] = 1$ and for every $h_t \in \mathbb{H}_t$, $t = 0, 1, 2, \ldots$, and Borel set B in X,

$$P_x^\pi[x_{t+1} \in B | h_t] = \int\limits_{\Re^k} 1_B[F(x_t, a_t, s)]\, \rho(s)\, \mathrm{d}s, \tag{2.1}$$

where $1_B(\cdot)$ stands for the indicator function of the set B.

The expectation operator with respect to P_x^π is denoted by E_x^π, and for the stationary policy $\boldsymbol{f} \in \mathbb{F}$ we write

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

For every policy $\pi \in \Pi$ and initial state $x \in X$, let

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

be the α-discounted expected total cost, where $\alpha \in (0, 1)$ is discount factor. The function

$$V^*(x) := \inf_{\pi \in \Pi} V(\pi, x), \quad x \in X,$$

is the optimal α-discounted cost when the initial state is x. A policy $\pi \in \Pi$ is said to be α-optimal (or optimal) if $V^*(x) = V(\pi, x)$ for all $x \in X$.

3. ASSUMPTIONS AND PRELIMINARY RESULTS

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

$$||u||_W := \sup_{x \in X} \frac{|u(x)|}{W(x)} < \infty. \tag{3.1}$$

Assumption 3.1.

a) For each $u \in L_W^\infty$ the set

$$\left\{ (x,a) : \int_{\Re^k} u[F(x,a,s)]\, \rho(s)\, \mathrm{d}s \le r \right\}$$

is Borel in $\mathbb{K}$ for every $r \in \Re$;

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

c) $\sup_{A(x)} |c(x,a)| \le W(x)$ for every $x \in X$.

We suppose the function W to be fixed in what follows. Remark that the abridged notations: $\sup_X$, $\inf_X$, $\sup_{A(x)}$, $\inf_{A(x)}$ we will currently use in place of complete ones: $\sup_{x \in X}$, $\inf_{x \in X}$, $\sup_{a \in A(x)}$, $\inf_{a \in A(x)}$.

A set D_0, described below, of densities ρ of r.v. ξ_t in (1.1) defines an admissible class of control processes for which adaptive policies constructed in Section 5 are applicable.

Let us fix an arbitrary $\varepsilon \in (0, 1/2)$ and denote $q := 1 + 2\varepsilon$. We will use these parameters throughout the paper without additional explanation. Also we choose and fix throughout the following a nonnegative measurable function $\overline{\rho} : \Re^k \to \Re$ which is used as a known majorant of unknown densities ρ.

We define the set $D_0 = D_0(\overline{\rho}, L, \beta_0, b_0, p, q)$ as a set consisting of all densities μ on $\Re^k$ for which the following holds.

a) $\mu \in L_q(\Re^k)$;

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

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

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

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

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

$$\int_{\Re^k} W^p[F(x,a,s)]\, \mu(s)\, \mathrm{d}s \le \beta_0 W^p(x) + b_0; \tag{3.2}$$

where $p > 1$, $\beta_0 < 1$, $b_0 < \infty$ are arbitrary but fixed for the defined set D_0 constants.

Remark 3.2. When $k = 1$ it is not difficult to show that a sufficient condition for part (b) is the following. There are a finite set $G \subset \Re$ (possibly empty) and a constant $M \ge 0$ such that:

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

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

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

Assumption 3.3.

a) The density ρ of r.v. ξ_t in (1.1) belongs to D_0.

b) The function

$$\varphi(s) := \sup_X [W(x)]^{-1} \sup_{A(x)} W[F(x,a,s)] \tag{3.3}$$

is finite for every $s \in \Re^k$.

c) $\int\limits_{\Re^k} \varphi^2(s) \, |\overline{\rho}(s)|^{2-q} \, \mathrm{d}s < \infty$.

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

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

Now we state two results that summarize simple but important facts to be used in the later sections.

Lemma 3.5. Suppose that Assumptions 3.1 (c) holds and $\rho \in D_0$. Then

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

$$\int\limits_{\Re^k} W[F(x,a,s)] \, \rho(s) \, \mathrm{d}s \le \beta W(x) + b, \tag{3.4}$$

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

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

c) there exists constant B such that

$$V^*(x) \le BW(x) \qquad x \in X. \tag{3.5}$$

Proof. a) By (3.2) we have

$$\int\limits_{\Re^k} W[F(x,a,s)] \, \rho(s) \, \mathrm{d}s \le \left[\int\limits_{\Re^k} W^p[F(x,a,s)] \, \rho(s) \, \mathrm{d}s \right]^{1/p}$$
$$\le \;\; [\beta_0 W^p(x) + b_0]^{1/p} \le \beta_0^{1/p} W(x) + b_0^{1/p}.$$

b) From (3.2) and (2.1) we get

$$E_x^\pi [W^p(x_t) | h_t] = \int\limits_{\Re^k} W^p[F(x_{t-1}, a_{t-1}, s)] \, \rho(s) \, \mathrm{d}s \le \beta_0 W^p(x_{t-1}) + b_0.$$

Hence

$$E_x^\pi[W^p(x_t)] \le \beta_0 E_x^\pi[W^p(x_{t-1})] + b_0, \quad t \in \mathbb{N}.$$

Iterating this inequality and using the fact $\beta_0 < 1$ we obtain

$$E_x^\pi[W^p(x_t)] \le \beta_0^t W^p(x) + (1+\beta_0+\ldots+\beta_0^{t-1})b_0 \le W^p(x) + b_0/(1-\beta_0).$$

The proof of second inequality in b) is similar because of (3.4).

c) The bound (3.5) for the value function V^* was proved in [13], Lemma 4.2 (d), provided to fulfillment of Assumption 3.1 (c) and (3.4). □

Proposition 3.6. Suppose that Assumption 3.1 holds and $\rho \in D_0$. Then,

a) the value function $V^*(\cdot)$ satisfies the α-discounted cost optimality equation, i.e.

$$V^*(x) = \inf_{a\in A(x)} \left\{ c(x,a) + \alpha \int_{\Re^k} V^*[F(x,a,s)]\,\rho(s)\,(\mathrm{d}s) \right\}, \quad x \in X; \tag{3.6}$$

b) for each $\delta > 0$, there exist a stationary policy $\boldsymbol{f} \in \mathbb{F}$ such that

$$c(x,f) + \alpha \int_{\Re^k} V^*[F(x,f,s)]\,\rho(s)\,(\mathrm{d}s) \le V^*(x) + \delta$$

for each $x \in X$.

Under condition (3.4) the proof of equality (3.6) was given in [13], Theorem 4.1 (b), while the part b) can be easily derived from Corollary 4.3 in [23].

4. DENSITY ESTIMATION

Denote by $\xi_0, \xi_1, \ldots, \xi_{t-1}$ independent realization (observed up to the moment $t-1$) of r.v. with the unknown density $\rho \in D_0$. Let $\hat\rho_t := \hat\rho_t(s;\xi_0,\xi_1,\ldots,\xi_{t-1})$, $s \in \Re^k$ be an arbitrary estimator of ρ belonging to L_q, such that

$$E\,\|\rho - \hat\rho_t\|_q^{\frac{q}{2}} \to 0 \quad as \quad t \to \infty. \tag{4.1}$$

We do not assume estimators $\hat\rho_t$, $t \in \mathbb{N}$, to be densities and even to be nonnegative.

To be able to provide asymptotically optimal adaptive policies we estimate ρ by a projection ρ_t from $\hat\rho_t$ on the set of densities D in L_q defined as follows.

We set $D := D_1 \cap D_2$, where

$$D_1 := \{\mu : \mu \text{ is a density on } \Re^k, \mu \in L_q \text{ and } \mu(s) \le \overline{\rho}(s) \text{ a.e.}\}; \tag{4.2}$$

$$D_2 := \left\{\mu : \mu \text{ is a density on } \Re^k, \mu \in L_q, \int W[F(x,a,s)]\,\mu(s)\,\mathrm{d}s \right.$$
$$\left. \le \beta W(x) + b,\ (x,a) \in \mathbb{K} \right\} \tag{4.3}$$

The constants β and b were defined in Lemma 3.5. Assumption 3.3 and Lemma 3.5 yields $\rho \in D_0 \subset D$. On the other hand, Lemma A in the Appendix shows that the set D is convex and closed in L_q. Thus, by virtue of well-known facts about a best approximation in the space L_q (see, e. g. Propositions 2 and 3 in [19], p. 343) there is an unique density $\rho_t \in D$ satisfying

$$\|\rho_t - \hat{\rho}_t\|_q = \inf_D \|\mu - \hat{\rho}_t\|_q , \quad t \in \mathbb{N}; \tag{4.4}$$

i.e. the density ρ_t is a best approximation of an estimator $\hat{\rho}_t$ on the set D.

Assumption 4.1. The density estimators $\rho_t(\cdot) := \rho_t(\cdot; \xi_0, \xi_1, \xi_2, \dots, \xi_{t-1})$, $t \in \mathbb{N}$, used in what follows, satisfy (4.1) and (4.4).

Examples of estimators with property (4.1) are given in [11]. Let $\{z_t\}$ be a sequence of positive real numbers such that $\lim_{t\to\infty} z_t^k/t = 0$ and $z_t \to \infty$ as $t \to \infty$. Set

$$\hat{\rho}_t(s) = \hat{\rho}_t(s; z_t, \xi_0, \xi_1, \dots, \xi_{t-1}) := \frac{1}{t}\sum_{i=0}^{t-1} V_{z_t}(s - \xi_i), \quad s \in \Re^k, \tag{4.5}$$

where $V_z(y)$ is the kernel of Vallée Poussin type:

$$V_z(y) = \prod_{n=1}^{k} \frac{\cos zy_n - \cos 2zy_n}{\pi z y_n^2}, \qquad y = (y_1, y_2, \dots, y_k) \in \Re^k,\ z > 0.$$

As it is shows in [11] relation (4.1) holds for estimator (4.5), provided that $\rho \in D_0$.

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

$$\|\mu\| := \sup_X [W(x)]^{-1} \sup_{A(x)} \int_{\Re^k} W[F(x,a,s)]\, \mu(s)\, \mathrm{d}s. \tag{4.6}$$

Proposition 4.2. Suppose that Assumptions 3.1, 3.3 and 4.1 hold. Then

$$E\, \|\rho_t - \rho\| \to 0 \quad \text{as } t \to \infty.$$

Proof. From (4.6) and (3.3) we have

$$\begin{aligned} \|\rho_t - \rho\| &\le \int_{\Re^k} \sup_X [W(x)]^{-1} \sup_{A(x)} W[F(x,a,s)]\ |\rho_t(s) - \rho(s)|\ \mathrm{d}s \\ &= \int_{\Re^k} \varphi(s)\, |\rho_t(s) - \rho(s)|\, \mathrm{d}s, \quad t \in \mathbb{N}. \end{aligned}$$

Thus, by Corollary A in the Appendix we obtain:

$$\|\rho_t - \rho\| \le M\, \|\rho_t - \rho\|_q^{q/2}, \quad t \in \mathbb{N}. \tag{4.7}$$

for some constant M.

On the other hand, by (4.4)

$$\|\rho_t - \rho\|_q^{q/2} \le 2^{q/2} \|\hat{\rho}_t - \rho\|_q^{q/2}, \quad t \in \mathbb{N}. \tag{4.8}$$

Combining (4.7), (4.8) and taking into account (4.1) we obtain the desired result. □

Remark 4.3. It can be that $\|\rho_t - \rho\|$ are not random variables. If it occurs then Proposition 4.2 proves the existence of measurable upper bounds b_t for $\|\rho_t - \rho\|$ such that $\lim_{t\to\infty} Eb_t = 0$ (see Remark 3.4).

5. ADAPTIVE POLICIES

For adaptive policies we use optimality criterion in the sense of the following definition.

Definition 5.1. a) [24] A policy π 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

$$\Phi(x,a) := c(x,a) + \alpha \int\limits_{\Re^k} V^*[F(x,a,s)]\, \rho(s)\, \mathrm{d}s - V^*(x), \;\; (x,a) \in \mathbb{K} \tag{5.1}$$

(Φ is a nonnegative function in view of Proposition 3.6.)

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

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

Remark 5.2. Φ is called the discrepancy function because it can be interpreted as measure of "deviation from optimality". For more details see e. g. [12].

For construction of suitable adaptive policies we replace an unknown density ρ by its estimations ρ_t and exploit corresponding optimality equations [14]. For these purposes we need to extend some assertions of Section 3 on densities ρ_t defined in Section 4 (belonging to D).

Assumption 5.3. For each $u \in L_W^\infty$ and $t \in \mathbb{N}$, $r \in \Re$, the set

$$\left\{ (x,a) : \int\limits_{\Re^k} u[F(x,a,s)]\, \rho_t(s)\, \mathrm{d}s \le r \right\}$$

is Borel in $\mathbb{K}$.

The proof of Lemma 3.5 and Proposition 3.6 (partly given in [13]) show that the following assertions hold true (because only inequality (3.4) is used here).

Proposition 5.4. Under Assumption 3.1 b), c) and Assumption 5.3 we have:

a) For each $t \in \mathbb{N}$ there is an unique function $V_t \in L_W^\infty$ such that

$$V_t(x) = \inf_{A(x)} \left\{ c(x,a) + \alpha \int_{\Re^k} V_t[F(x,a,s)]\, \rho_t(s)\, \mathrm{d}s \right\}, \quad x \in X. \tag{5.2}$$

b) For each $t \in \mathbb{N}$, $\delta_t > 0$, there exists a stationary policy $f_t \in \mathbb{F}$ such that

$$c(x,f_t) + \alpha \int_{\Re^k} V_t[F(x,f_t,s)]\, \rho_t(s)\, \mathrm{d}s \le V_t(x) + \delta_t, \quad x \in X. \tag{5.3}$$

c) There is a constant B^* such that $\sup_{t \ge 1} \|V_t\|_W \le B^*$.

d) If $\overline{V}_0 \equiv 0$ and

$$\overline{V}_t(x) = \inf_{A(x)} \left\{ c(x,a) + \alpha \int_{\Re^k} \overline{V}_{t-1}[F(x,a,s)]\, \rho_t(s)\, \mathrm{d}s \right\}, \quad x \in X, \quad t \in \mathbb{N}, \tag{5.4}$$

then $\|\overline{V}_t\|_W \le \overline{B}$ for some constant $\overline{B}$, and for every $\overline{\delta}_t > 0$ there exists a stationary policy $\overline{f}_t \in \mathbb{F}$ such that

$$c(x,\overline{f}_t) + \alpha \int_{\Re^k} \overline{V}_{t-1}[F(x,\overline{f}_t,s)]\, \rho_t(s)\, \mathrm{d}s \le \overline{V}_t(x) + \overline{\delta}_t, \quad x \in X. \tag{5.5}$$

Now we introduce two adaptive policies π^* and $\overline{\pi}$ that are slight extensions of "The Principle of Estimation and Control" policy [22] and of "The Non stationary Value Iteration" policy [16].

Definition 5.5. Let sequences of positive numbers $\{\delta_t\}$ and $\{\overline{\delta}_t\}$ be arbitrary but fixed, and arbitrary sequences $\{f_t\}$ and $\{\overline{f}_t\}$ of stationary policies be chosen such that (5.3) and (5.5) are satisfied.

a) The policy $\pi^* = \{\pi_t^*\}$ is defined as follows

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

b) The policy $\overline{\pi} = \{\overline{\pi}_t\}$ is defined as follows

$$\overline{\pi}_t(h_t) = \overline{\pi}_t(h_t;\rho_t) := \overline{f}_t(x_t), \quad h_t \in \mathbb{H}_t, \quad t \in \mathbb{N}.$$

($\pi_0^*(x)$ and $\overline{\pi}_0(x)$ are any fixed actions).

We are now ready to state our main results. Denote $\delta := \limsup_{t\to\infty} \delta_t$; $\overline{\delta} := \limsup_{t\to\infty} \overline{\delta}_t$.

Theorem 5.6. Suppose that Assumptions 3.1, 3.3, 4.1 and 5.3 hold. Then the adaptive policy π^* is δ-asymptotically discount optimal, and the adaptive policy $\overline{\pi}$ is $\overline{\delta}$-asymptotically discount optimal.

In particularly, if $\delta = \overline{\delta} = 0$ then the policies π^* and $\overline{\pi}$ are asymptotically discount optimal.

Remark 5.7. It is well-known fact that an optimal stationary policy exists if the minimum on the right-hand side of (3.6) is attained for each $x \in X$. Thus to guarantee the existence of such policy one should impose rather restrictive continuity conditions on one-stage cost c and transition probability of process, and suppose something as compactness of $A(x)$ (see e. g. [12]). It can happen that under the assumptions made in this paper, stationary discount optimal policy do not exist for process (1.1) with a known density ρ, while Theorem 5.6 guarantees the existence of asymptotically optimal adaptive policies.

Remark 5.8. In the remainder of this sections we will use repeatedly the following inequalities:

$$u(x) \le \|u\|_W \, W(x) \tag{5.6}$$

and

$$\int_{\Re^k} u[F(x,a,s)]\, \mu(s)\, ds \le \|u\|_W \, [\beta W(x) + b] \tag{5.7}$$

for all $u \in L_W^\infty$, $\mu \in D$, $x \in X$, $a \in A(x)$. The relation (5.6) is a consequence of the definition of $\|\cdot\|_W$, and (5.7) holds because of (3.4) and the definition of D.

The proof of Theorem 5.6 is based on the following result.

Lemma 5.9. Under Assumption 3.1, 3.3, 4.1 and 5.3, for each $x \in X$ and $\pi \in \Pi$

$$\text{a)} \quad \lim_{t\to\infty} E_x^\pi \|V_t - V^*\|_W = 0 \quad \text{and} \quad \text{b)} \quad \lim_{t\to\infty} E_x^\pi \left\|\overline{V}_t - V^*\right\|_W = 0.$$

Proof. a) For every $\mu \in D$ let us define the operator

$$T_\mu u(x) = \inf_{A(x)} \left\{ c(x,a) + \alpha \int_{\Re^k} u[F(x,a,s)]\, \mu(s)\, ds \right\}, \tag{5.8}$$

$x \in X$, $u \in L_W^\infty$. By Assumption 3.1 (c), the definition of D and (5.7), T maps L_W^∞ into itself.

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

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

It is easy to see

$$\|u\|_{\overline{W}} \le \|u\|_W \le \|u\|_{\overline{W}}(1 + d/\inf_X W(x)), \tag{5.9}$$

hence $L_W^\infty = L_{\overline{W}}^\infty$ and the norms $\|\cdot\|_W$ and $\|\cdot\|_{\overline{W}}$ are equivalent.

In Lemma 2 in [27] was proved that the inequality

$$\int_{\Re^k} W[F(x,a,s)]\,\mu(s)\,\mathrm{d}s \le W(x) + b$$

implies the operator T_μ in (5.8) to be a contraction with respect to the norm $\|\cdot\|_{\overline{W}}$, that is

$$\|T_\mu v - T_\mu u\|_{\overline{W}} \le \gamma\,\|v - u\|_{\overline{W}}\,, \quad v,\, u \in L_W\,. \tag{5.10}$$

By virtue of (3.6) and (5.10) the function V^* is an unique (in L_W^∞) fixed point of the operator T_ρ, while V_t are fixed points (unique in L_W^∞) of T_{ρ_t}, $t \in \mathbb{N}$, that is

$$T_\rho V^* = V^*, \quad T_{\rho_t} V_t = V_t \tag{5.11}$$

Because of (5.9) the part a) will be proved if we show that

$$\lim_{t\to\infty} E_x^\pi \|V_t - V^*\|_{\overline{W}} = 0. \tag{5.12}$$

We have

$$\begin{aligned} \|V^* - V_t\|_{\overline{W}} &= \left\|T_\rho V^* - T_{\rho_t} V_t\right\|_{\overline{W}} \le \left\|T_\rho V^* - T_{\rho_t} V^*\right\|_{\overline{W}} + \left\|T_{\rho_t} V^* - T_{\rho_t} V_t\right\|_{\overline{W}} \\ &\le \left\|T_\rho V^* - T_{\rho_t} V^*\right\|_{\overline{W}} + \gamma\,\|V^* - V_t\|_{\overline{W}}, \end{aligned}$$

or

$$\|V^* - V_t\|_{\overline{W}} \le \frac{1}{1-\gamma}\left\|T_\rho V^* - T_{\rho_t} V^*\right\|_{\overline{W}}\,, \quad t \in \mathbb{N}. \tag{5.13}$$

On the other hand, from definition (4.6), (3.5) and the fact $[\overline{W}(\cdot)]^{-1} < [W(\cdot)]^{-1}$, we obtain

$$\begin{aligned} \left\|T_\rho V^* - T_{\rho_t} V^*\right\|_{\overline{W}} &\le \alpha \sup_X [\overline{W}(x)]^{-1} \sup_{A(x)} \int_{\Re^k} V^*[F(x,a,s)]\,|\rho(s) - \rho_t(s)|\,\mathrm{d}s \\ &\le \alpha B \sup_X [W(x)]^{-1} \sup_{A(x)} \int_{\Re^k} W[F(x,a,s)]\,|\rho(s) - \rho_t(s)|\,\mathrm{d}s \\ &= \alpha B\,\|\rho - \rho_t\|, \quad t \in \mathbb{N}. \end{aligned} \tag{5.14}$$

Observing that $E_x^\pi \|\rho - \rho_t\| = E\,\|\rho - \rho_t\|$ (since ρ_t do not depend on π and x) and combining inequalities (5.13), (5.14) with Proposition 4.2 we find that (5.12) holds.

b) Using argument similar to the proof of the part a), from equations (3.6), (5.4), (5.10) and (5.14) we get

$$\begin{aligned} &\left\|V^* - \overline{V}_{t+1}\right\|_{\overline{W}} \\ \le\ &\left\|T_\rho V^* - T_{\rho_t} V^*\right\|_{\overline{W}} + \gamma\left\|V^* - \overline{V}_t\right\|_{\overline{W}} \le \alpha B\,\|\rho - \rho_t\| + \gamma\left\|V^* - \overline{V}_t\right\|_{\overline{W}}. \end{aligned}$$

Therefore,

$$E_x^\pi \left\|V^* - \overline{V}_{t+1}\right\|_{\overline{W}} \leq \alpha B E_x^\pi \|\rho - \rho_t\| + \gamma E_x^\pi \left\|V^* - \overline{V}_t\right\|_{\overline{W}}, \tag{5.15}$$

for each $x \in X$, $\pi \in \Pi$, $t \in \mathbb{N}$.

In view of Lemma 3.5 (c), Proposition 5.4 (d) and equivalence of the norms $\|\cdot\|_W$ and $\|\cdot\|_{\overline{W}}$ we have $\lambda := \limsup_{t\to\infty} E_x^\pi \left\|V^* - \overline{V}_t\right\|_{\overline{W}} < \infty$. Taking limsup as $t \to \infty$ in both sides of (5.15) and applying Proposition 4.2 we see that $\lambda \leq \gamma\lambda$, so $\lambda = 0$. This completes the proof of Lemma 5.9. □

Proof of Theorem 5.6.

First, we define for each $t \in \mathbb{N}$ the following nonnegative functions $\mathbb{K} \to \Re$ by the formulas:

$$\begin{aligned}\Phi_t^*(x,a) &:= c(x,a) + \alpha \int_{\Re^k} V_t[F(x,a,s)]\,\rho_t(s)\,\mathrm{d}s - V_t(x);\\ \overline{\Phi}_t(x,a) &:= c(x,a) + \alpha \int_{\Re^k} \overline{V}_{t-1}[F(x,a,s)]\,\rho_t(s)\,\mathrm{d}s - \overline{V}_t(x).\end{aligned}$$

(see Proposition 5.4 (a), (d) to verify that these functions are nonnegative).

Using the definitions of Φ_t^* and Φ (see (5.1)) we get (by adding and subtracting the term $\alpha \int_{\Re^k} V_t[F(x,a,s)]\,\rho(s)\,\mathrm{d}s$)

$$\begin{aligned}&|\Phi_t^*(x,a) - \Phi(x,a)|\\ \leq\ & |V^*(x) - V_t(x)| + \alpha \int_{\Re^k} V_t[F(x,a,s)]\, |\rho_t(s) - \rho(s)|\,\mathrm{d}s\\ &+\alpha \int_{\Re^k} |V_t[F(x,a,s)] - V^*[F(x,a,s)]|\,\rho(s)\,\mathrm{d}s\\ \leq\ & \|V^* - V_t\|_W\, W(x) + \alpha B^* \int_{\Re^k} W[F(x,a,s)]\, |\rho_t(s) - \rho(s)|\,\mathrm{d}s + \alpha[\beta W(x) + b]\, \|V_t - V^*\|_W\end{aligned}$$

for each $(x,a) \in \mathbb{K}$, $t \in \mathbb{N}$ (see Proposition 5.4 (c) and Lemma 3.5 (a)). Hence (see the definition of $\|\cdot\|$ in (4.6) and inequalities (5.13) and (5.14))

$$\sup_X [W(x)]^{-1} \sup_{A(x)} |\Phi_t^*(x,a) - \Phi(x,a)| \leq B' \|\rho_t - \rho\|, \tag{5.16}$$

where $B' = \alpha B^* + \{1 + \alpha[\beta + b/\inf_X W(x)]\}(1-\gamma)^{-1}\alpha B$.

On the other hand, by the definition of the policy π^* (see Definition 5.5) and of the functions f_t in (5.3) we have $\Phi_t^*(\cdot, \pi_t^*(\cdot)) \leq \delta_t$, $t \in \mathbb{N}$. Thus

$$\begin{aligned}&\Phi(x_t, \pi_t^*(h_t)) \leq |\Phi(x_t, \pi_t^*(h_t)) - \Phi_t^*(x_t, \pi_t^*(h_t)) + \delta_t|\\ \leq\ & \sup_{A(x_t)} |\Phi(x_t, a) - \Phi_t^*(x_t, a)| + \delta_t\end{aligned}$$

$$\begin{aligned} &\leq W(x_t)\sup_X[W(x)]^{-1}\sup_{A(x)}|\Phi(x,a)-\Phi_t^*(x,a)|+\delta_t \\ &\leq W(x_t)\eta_t+\delta_t\ ,\ t\in\mathbb{N}, \end{aligned} \tag{5.17}$$

where $\eta_t := B'\,\|\rho_t-\rho\|$. Inequality (5.17) implies the following one:

$$E_x^{\pi^*}\,[\Phi(x_t,a_t)]\leq E_x^{\pi^*}[W(x_t)\eta_t]+\delta_t,$$

and therefore, to prove δ-optimality of the policy π^* (see Definition 5.1) it is enough to show that

$$\limsup_{t\to\infty}E_x^{\pi^*}[W(x_t)\eta_t]=0. \tag{5.18}$$

First, if $\mu\in D_2$ then by (4.3), (4.6)

$$\|\mu\|\leq\sup_X[W(x)]^{-1}\,[\beta W(x)+b]\leq\beta+b/\inf_X W(x). \tag{5.19}$$

For this inequality $\sup_{t\geq1}\|\rho_t-\rho\|\leq B_1<\infty$ with some constant B_1. Proposition 4.2 yields the convergence in probability:

$$\eta_t\xrightarrow{P_x^{\pi^*}}0\ as\ t\to\infty. \tag{5.20}$$

Furthermore, from Lemma 3.5 (b) we get

$$\sup_{t\geq1}E_x^{\pi^*}[W(x_t)\eta_t]^p<(B')^p\,B_1^p\sup_{t\geq1}E_x^{\pi^*}[W^p(x_t)]<\infty.$$

This means that the sequence $\{W(x_t)\eta_t\}$ is $P_x^{\pi^*}$-uniformly integrable (see Lemma 7.6.9, p. 301 in [2]). In this way, using the known criterion of convergence of integrals (see for instance, Theorem 7.5.2 in [2]), we prove (5.18) if we show that $W(x_t)\eta_t\xrightarrow{P_x^{\pi^*}}0$ as $t\to\infty$. But the latter follows from (5.20), Lemma 3.5 (b) and the inequalities:

$$P_x^{\pi^*}[W(x_t)\eta_t>\gamma]\leq P_x^{\pi^*}[\eta_t>\frac{\gamma}{l}]+P_x^{\pi^*}[W(x_t)>l]\leq P_x^{\pi^*}\left[\eta_t>\frac{\gamma}{l}\right]+\frac{E_x^{\pi^*}[W(x_t)]}{l}$$

with, γ,l being arbitrary positive numbers.

The proof of the second part of the theorem is similar up to minor changes. First we can show that

$$\begin{aligned} &\sup_X[W(x)]^{-1}\sup_{A(x)}\left|\overline{\Phi}_{t+1}(x,a)-\Phi(x,a)\right| \\ \leq\ &\left\|V^*-\overline{V}_{t+1}\right\|_W+\alpha\overline{B}\,\|\rho_{t+1}-\rho\|+\alpha\,[\beta+b/\inf_X W(x)]\left\|\overline{V}_t-V^*\right\|_W \\ :=\ &\overline{\eta}_t,\ t\in\mathbb{N}. \end{aligned}$$

Again, $\lim_{t\to\infty}E_x^{\overline{\pi}}[\overline{\eta}_t]=0$ (see Lemma 5.9 and Proposition 4.2), and the random variables $\overline{\eta}_t$ are uniformly bounded (due to Lemma 3.5 (c), Proposition 5.4 (d) and (5.19)). Repeating the arguments of first part we complete the proof. □

6. EXAMPLE

We consider a particular control system of the form

$$x_{t+1} = (x_t + a_t - \xi_t)^+, \quad t = 0, 1, 2 \dots, \tag{6.1}$$

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

Relations (6.1) describe, in particular, some control models in *storage system* (see, for instance, [4, 15]). Another interpretation of (6.1) which we have in mind is a model of control of deterministic service rate in a single server queueing system of type $GI|D|1|\infty$. In this example x_t denotes the waiting time of the tth customer and ξ_t denotes the interarrival time between the tth and the $(t+1)$th customers. Control actions a_t, $t = 0, 1, 2, \dots$, are service times chosen for corresponding customers among admissible $a \in A$.

Nonnegative random variables $\xi_0, \xi_1, \xi_2 \dots$ are supposed to be i.i.d with a common density $\rho \in L_q(\Re)$ satisfying the inequality

$$\|\Delta_z \rho\|_{L_q} \le L\, |z|^{1/q},$$

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

In the spirit of setting of the problem chosen in this paper, we assume the density ρ to be unknown, but realizations of $\xi_0, \xi_1, \xi \dots, \xi_{t-1}$ and states x_t to be observable at the moment t of taking decision a_t. The latter assumption is met in some communication and computer control system.

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

Assumption 6.1. $E(\xi_0)$ exists, and moreover

$$E(\xi_0) > \theta. \tag{6.2}$$

Considering the function $\Psi(s) := e^{\theta s} E(e^{-s\xi_0})$ we find that (6.2) implies $\Psi'(0) < 0$, so there is $\lambda > 0$ for which $\Psi(\lambda) < 1$. Also by continuity of Ψ we can choose $p > 1$ such that

$$\Psi(p\lambda) = \beta_0 < 1. \tag{6.3}$$

Let us set $W(x) = \bar{b} e^{\lambda x}$, $x \in [0, \infty)$, where $\bar{b}$ is an arbitrary constant. Then

$$W^p(x) = \left(\bar{b}\right)^p e^{\lambda p x}.$$

In [9] was show that (6.3) implies the inequality

$$\int_0^\infty b_0 e^{\lambda p (x+a-s)^+} \rho(s)\, \mathrm{d}s \le \beta_0 b_0 e^{\lambda p x} + b_0,$$

where $b_0 := \left(\bar{b}\right)^p$.

Thus Assumption 3.3 will be satisfied if we find a suitable majorant $\overline{\rho}$ with the properties as in Assumption 3.3 (b), (c).

Straightforward calculations show that $\varphi(s) = \max\left\{1, e^{\lambda(\theta - s)}\right\}$, $s \in [0,\infty)$, thus it is a bounded function. Therefore, to satisfy Assumption 3.3 (c) we can take, for example,

$$\overline{\rho}(s) := M \min\left\{1, 1/s^{1+r}\right\}, \quad s \in [0,\infty), \tag{6.4}$$

where $r > 0$.

For $r < 1$ in (6.4), Assumption 6.1 implies $\rho \le \overline{\rho}$ (choosing enough large M) in a wide class of densities. The fulfillment of measurability conditions from Assumption 3.1 (a) and Assumption 5.3 (a) can be checked easily. Finally, we meet Assumption 3.1 (c) if we endow the considering control model with arbitrary nonnegative measurable one-stage cost function $c : [0,\infty) \times A \to [0,\infty)$ for which

$$\sup_A c(x,a) \le \overline{b} e^{\lambda x}, \quad x \in [0,\infty).$$

7. CONCLUDING REMARKS

The difficult part of application of the adaptive policies π^* and $\overline{\pi}$ is making a projection of estimator $\hat{\rho}_t$ as in (4.5) on the set D defined in (4.2) and (4.3). Notice in view of it that D can be replaced by any closed convex subset of D containing ρ. Thus, sufficient conditions of (4.3) could be used (as in the above example). We are going to present some algorithms of projection in next publications. Also we plan to propose adaptive policies optimal with respect to the average cost criterion with unbounded one-stage costs. Such criterion seems more natural in adaptive control problems, and the findings on exponentially fast approximation of average optimal policies [10] are suitable techniques here. Finally, observe that from the proof of Theorem 5.6, some estimation of rate of convergence of $E_x^{\pi^*}\Phi(x_t, a_t)$ and $E_x^{\overline{\pi}}\Phi(x_t, a_t)$ can be made if one uses the estimations of $E\,\|\rho - \hat{\rho}_t\|_q^{q/2}$ in [11].

APPENDIX

Lemma A. The set $D = D_1 \cap D_2$ defined in (4.2) and (4.3) is closed and convex subset of L_q.

Proof. We start proving that D is closed. Let $\mu_n \in D$ be a sequence such that $\mu_n \xrightarrow{L_q} \mu \in L_q$. Suppose $\mu \notin D_1$, i.e. there is $A \subset \Re^k$ with $m(A) > 0$ such that $\mu(s) > \overline{\rho}(s)$, $s \in A$ (m is the Lebesgue measure on $\Re^k$). Then, for some $\delta > 0$ and $A' \subset A$ with $m(A') > 0$

$$\mu(s) > \overline{\rho}(s) + \delta, \quad s \in A'. \tag{A.1}$$

Since $\mu_n \in D_1$, $n \in \mathbb{N}$, there exists $H \subset \Re^k$ with $m(H) = 0$ such that

$$\mu_n(s) \le \overline{\rho}(s), \quad s \in \Re^k \setminus H, \quad n \in \mathbb{N}. \tag{A.2}$$

Joining (A.1) and (A.2) we get

$$|\mu(s) - \mu_n(s)| \geq \delta, \quad s \in A' \cap (\Re^k \setminus H), \quad n \in \mathbb{N}.$$

Since $m(A' \cap (\Re^k \setminus H)) > 0$ we see that μ_n does not converge to μ in measure, that contradicts to the convergence in L_q.

Now, to prove that $\mu \in D_2$, using the fact

$$\int_{\Re^k} W[F(x,a,s)]\,\mu_n(s)\,ds \leq \beta W(x) + b, \quad (x,a) \in \mathbb{K},\ n \in \mathbb{N},$$

it suffices to show that

$$\int_{\Re^k} W[F(x,a,s)]\,\mu_n(s)\,ds \longrightarrow \int_{\Re^k} W[F(x,a,s)]\,\mu(s)\,ds \quad as \quad n \longrightarrow \infty,$$

for all (x,a) $\in\mathbb{K}$. By (3.3) $W[F(x,a,s)] \leq W(x)\varphi(s)$, (x,a) $\in\mathbb{K}$, $s \in \Re^k$. Hence, for any fixed (x,a) $\in\mathbb{K}$, and $\varepsilon = (q-1)/2$

$$\begin{aligned} I &:= \left|\int_{\Re^k} W[F(x,a,s)][\mu_n(s) - \mu(s)]\,ds\right| \leq W(x)\left|\int_{\Re^k} \varphi(s)\,[\mu_n(s) - \mu(s)]\,ds\right| \\ &\leq W(x)\int_{\Re^k} \varphi(s)\,|\mu_n(s) - \mu(s)|^{(1-2\varepsilon)/2}\,|\mu_n(s) - \mu(s)|^{(1+2\varepsilon)/2}\,ds \qquad \text{(A.3)} \end{aligned}$$

Applying the Hölder Inequality and taking into account that $\mu, \mu_n \in D_1$ we obtain

$$\begin{aligned} I &\leq W(x)\left[\int_{\Re^k} \varphi^2(s)\,|\mu_n(s) - \mu(s)|^{1-2\varepsilon}\,ds\right]^{\frac{1}{2}}\left[\int_{\Re^k} |\mu_n(s) - \mu(s)|^{1+2\varepsilon}\,ds\right]^{\frac{1}{2}} \\ &\leq W(x)\left[\int_{\Re^k} \varphi^2(s)\,|2\overline{\rho}(s)|^{1-2\varepsilon}\,ds\right]^{\frac{1}{2}}\left[\int_{\Re^k} |\mu_n(s) - \mu(s)|^{1+2\varepsilon}\,ds\right]^{\frac{1}{2}} \\ &\leq MW(x)\left[\int_{\Re^k} |\mu_n(s) - \mu(s)|^{1+2\varepsilon}\,ds\right]^{\frac{1}{2}} \qquad \text{(A.4)} \end{aligned}$$

due to Assumption 3.3 (c).

Since $q = 1 + 2\varepsilon$ and $\mu_n \xrightarrow{L_q} \mu$, the right-hand side of inequality (A.4) vanishes as $n \longrightarrow \infty$.

To complete the proof of closeness of D we have to check that μ is a density on $\Re^k$. It is evident that $\mu \geq 0$, almost everywhere.

On the other hand, similarly to (A.4)

$$\begin{aligned}\left|1-\int_{\Re^k}\mu(s)\,\mathrm{d}s\right| &\le \int_{\Re^k}|\mu_n(s)-\mu(s)|\,\mathrm{d}s\\ &\le \left[\int_{\Re^k}[2\bar{\rho}(s)]^{1-2\varepsilon}\,\mathrm{d}s\right]^{\frac{1}{2}}\left[\int_{\Re^k}|\mu_n(s)-\mu(s)|^{1+2\varepsilon}\,\mathrm{d}s\right]^{\frac{1}{2}}\\ &\le M_1\left[\int_{\Re^k}|\mu_n(s)-\mu(s)|^{1+2\varepsilon}\,\mathrm{d}s\right]^{\frac{1}{2}} \to 0, \quad \text{as } n\to\infty,\end{aligned}$$

because $\varphi(s)\ge 1,\ \ s\in\Re^k$.

The convexity of $D_1$and D_2 is verified directly by using definitions (4.2) and (4.3). □

Taking into account the inequalities (A.3) and (A.4) we get the following.

Corollary A. Under Assumptions 3.3 and 4.1

$$\int_{\Re^k}\varphi(s)\,|\rho_t(s)-\rho(s)|\,\mathrm{d}s \le M\|\rho_t-\rho\|_q^{q/2},\ t\in\mathbb{N}.$$

(Received September 27, 1996.)

REFERENCES

[1] R. Agrawal: Minimizing the learning loss in adaptive control of Markov chains under the weak accessibility condition. J. Appl. Probab. *28* (1991), 779–790.

[2] R. B. Ash: Real Analysis and Probability. Academic Press, New York 1972.

[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. B. Dynkin and A. A: Yushkevich: Controlled Markov Processes. Springer–Verlag, New York 1979.

[5] E. Fernández–Gaucherand, A. Arapostathis and S. I. Marcus: A methodology for the adaptive control of Markov chains under partial state information. In: Proc. of the 1992 Conf. on Information Sci. and Systems, Princeton, New Jersey, pp. 773–775.

[6] E. Fernández–Gaucherand, A. Arapostathis and S. I. Marcus: Analysis of an adaptive control scheme for a partially observed controlled Markov chain. IEEE Trans. Automat. Control *38* (1993), 987–993.

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

[8] E. I. Gordienko: Controlled Markov sequences with slowly varying characteristics II. Adaptive optimal strategies. Soviet J. Comput. Systems Sci. *23* (1985), 87–93.

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

[10] E. I. Gordienko, R. Montes-de-Oca and J. A. Minjárez-Sosa: Approximation of average cost optimal policies for general Markov decision processes with unbounded costs. Math. Methods Oper. Res. *45* (1997), 2, to appear.

[11] 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.

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

[13] 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). (Submitted for publication).

[14] 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.

[15] O. Hernández-Lerma and J. B. Lasserre: Discrete-Time Markov Control Processes. Springer-Verlag, New York 1995.

[16] O. Hernández-Lerma and S. I. Marcus: Adaptive control of discounted Markov decision chains. J. Optim. Theory Appl. *46* (1985), 227–235.

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

[18] K. Hinderer: Foundations of Non-Stationary Dynamic Programming with Discrete Time Parameter. (Lecture Notes in Operations Research and Mathematical Systems 33.) Springer-Verlag, Berlin - Heidelberg - New York 1970.

[19] G. Köthe: Topological Vector Spaces I. Springer-Verlag, New York 1969.

[20] P. R. Kumar and P. Varaiya: Stochastic Systems: Estimation, Identification and Adaptive Control. Prentice-Hall, Englewood Cliffs 1986.

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

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

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

[24] M. Schäl: Estimation and control in discounted stochastic dynamic programming. Stochastics *20* (1987), 51–71.

[25] L. Stettner: On nearly self-optimizing strategies for a discrete-time uniformly ergodic adaptive model. J. Appl. Math. Optim. *27* (1993), 161–177.

[26] L. Stettner: Ergodic control of Markov process with mixed observation structure. Dissertationes Math. *341* (1995), 1–36.

[27] J. A. E. E. van Nunen and J. Wessels: A note on dynamic programming with unbounded rewards. Management Sci. *24* (1978), 576–580.

Prof. Dr. Evgueni I. Gordienko, Departamento de Matemáticas, Universidad Autónoma Metropolitana-I. A. P. 55-534, C. P. 09340. México, D. F. México.
e-mail: gord@xanum.uam.mx

Dr. J. Adolfo Minjárez-Sosa, Departamento de Matemáticas, Universidad de Sonora. Rosales s/n Col. Centro, C. P. 83 000. Hermosillo, Son. México.
e-mail: aminjare@fisica.uson.mx