Math Meth Oper Res (2001) 54:491–505

Mathematical Methods
of Operations Research

Adaptive policies for time-varying stochastic systems under discounted criterion*

Nadine Hilgert[1] **and J. Adolfo Minjárez-Sosa**[2]

[1] Laboratoire de Biométrie, INRA-ENSA.M, 2 place Viala, 34060 Montpellier CEDEX 1, France. (hilgert@ensam.inra.fr). The research of this author was performed while she was visiting the Departamento de Matemáticas, CINVESTAV-IPN, México, DF.
[2] Departamento de Matemáticas, Universidad de Sonora, Rosales s/n, Col. Centro, 83000, Hermosillo, Sonora, México. (aminjare@gauss.mat.uson.mx)

Abstract. We consider a class of time-varying stochastic control systems, with Borel state and action spaces, and possibly unbounded costs. The processes evolve according to a discrete-time equation $x_{n+1} = G_n(x_n, a_n, \xi_n)$, $n = 0, 1, \ldots,$ where the ξ_n are i.i.d. $\Re^k$-valued random vectors whose common density is unknown, and the G_n are given functions converging, in a restricted way, to some function G_∞ as $n \to \infty$. Assuming observability of ξ_n, we construct an adaptive policy which is asymptotically discounted cost optimal for the *limiting control system* $x_{n+1} = G_\infty(x_n, a_n, \xi_n)$.

AMS 1991 subject classifications: 93E20, 90C40.

Key words: Non-homogeneous Markov control processes; discrete-time stochastic systems; discounted cost criterion; optimal adaptive policy

1 Introduction

This paper deals with discrete-time, time-varying stochastic control systems of the form

$$x_{n+1} = G_n(x_n, a_n, \xi_n), \quad n \in \mathbb{N}_0 := \{0, 1, \ldots\}, \tag{1}$$

where x_n and a_n denote the state and control variables respectively, and $\{\xi_n\}$, the so-called "disturbance" or "driving" process, is a sequence of independent and identically distributed (i.i.d.) random vectors in $\Re^k$ having an unknown density ρ. In addition, $\{G_n\}$ is a sequence of given functions such that

* Work supported by Consejo Nacional de Ciencia y Tecnología (CONACyT) under Grant 28309E.

$$E1_B[G_n(x,a,\xi_0)] \to E1_B[G_\infty(x,a,\xi_0)] \quad \text{for all } (x,a) \text{ and Borel set } B, \quad (2)$$

where $1_B(.)$ denotes the indicator function of the set B (See Assumption 2.2 for more details on this condition).

Our main objective in this paper is to introduce asymptotically discounted optimal adaptive policies for the general limiting system

$$x_{t+1} = G_\infty(x_t, a_t, \xi_t), \quad t \in \mathbb{N}_0, \quad (3)$$

considering possibly unbounded one-stage costs.

Systems of the type (1) appear, for instance, in some time-varying controlled biotechnological processes ([1, 12]), taking the particular form

$$x_{n+1} = (H(x_n)g_n(x_n) + G(x_n, a_n) + \xi_n)^+, \quad n \in \mathbb{N}_0.$$

This model represents, for example, the real time evolution of biomasses (microorganisms) and substrates concentrations in bioreactions. Such bioreactions are very common in depollution and in the agro-food industry. This example will be analyzed below (Section 5) to illustrate the main results of this paper.

Our work extends recent results in [4] and [11]. In the former, the adaptive control problem in the discounted case is studied for general time-invariant systems of the type (3). The construction of optimal policies is done by first estimating the density ρ with suitable statistical methods, and then applying the "principle of estimation and control" proposed in [13, 14]. On the other hand, [11] studies time-varying additive-noise systems of the form

$$x_{n+1} = G_n(x_n, a_n) + \xi_n, \quad n \in \mathbb{N}_0,$$

where the density of the random disturbance $\{\xi_n\}$ is supposed to be known, and $\{G_n\}$ is a sequence of given functions converging pointwise to some function G_∞. Conditions are given for the existence of α-discounted optimal stationary policies for the limiting system

$$x_{t+1} = G_\infty(x_t, a_t) + \xi_t, \quad t \in \mathbb{N}_0. \quad (4)$$

The same approach is applied to system (1); that is, we consider the α-discounted problem for the time-invariant system

$$x_{t+1} = G_n(x_t, a_t, \xi_t), \quad t \in \mathbb{N}_0, \quad (5)$$

for each fixed $n \in \mathbb{N}_0$, and then we let $n \to \infty$ to obtain the corresponding result for the limiting system (3). Put in this form, our main result, Theorem 4.5, can also be seen as a further result of [4] on system (3) where the function G_∞ is unknown and estimated by some consistent functional estimator G_n.

The paper is organized as follows. In Section 2 we introduce the Markov control models we are concerned with and the assumptions required. Some preliminary results are given in Section 3. The adaptive policies are introduced in Section 4 together with the main result, Theorem 4.5. Finally, a generic example of a biotechnological process satisfying all the hypotheses of the paper is described in Section 5.

2 Markov control models

For each fixed $n = 0, 1, \ldots, \infty$, we consider the Markov control model

$$M_n := (X, A, \{A(x) \mid x \in X\}, Q_n, c) \tag{6}$$

associated to the system (5), satisfying the following conditions. The state space X and the action space A are Borel spaces. They are endowed with their Borel σ-algebras $\mathbb{B}(X)$ and $\mathbb{B}(A)$. For each state $x \in X$, $A(x)$ is a nonempty Borel subset of A denoting the set of admissible controls when the system is in state x. 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. In addition, Q_n is a stochastic kernel denoting the transition law corresponding to (5), that is, for all $t \in \mathbb{N}_0$, $(x, a) \in \mathbb{K}$ and $B \in \mathbb{B}(X)$,

$$\begin{aligned} Q_n(B|x, a) &:= \text{Prob}[G_n(x_t, a_t, \xi_t) \in B \mid x_t = x, a_t = a] \\ &= E 1_B[G_n(x, a, \xi_t)] \\ &= \int_{\Re^k} 1_B[G_n(x, a, s)]\rho(s)\, ds, \end{aligned} \tag{7}$$

where $\{\xi_t\}$ is a sequence of i.i.d. random vectors (r.v.'s) on a probability space $(\Omega, \mathscr{F}, P)$, with values in $\Re^k$ and a common unknown distribution with a density ρ. Moreover, we assume that the realizations $\xi_0, \xi_1, \ldots$ of the driving process and the states $x_0, x_1, \ldots$ are completely observable. Finally, the cost-per-stage $c(x, a)$ is a nonnegative measurable real-valued function on $\mathbb{K}$, possibly unbounded.

We define the spaces of admissible histories up to time t by $\mathbb{H}_0 := X$ and $\mathbb{H}_t := (\mathbb{K} \times \Re^k)^t \times X$, $t \geq 1$. A generic element of $\mathbb{H}_t$ is written as $h_t = (x_0, a_0, \xi_0, \ldots, x_{t-1}, a_{t-1}, \xi_{t-1}, x_t)$. A control policy $\pi = \{\pi_t\}$ is a sequence of measurable functions $\pi_t : \mathbb{H}_t \to A$ such that $\pi_t(h_t) \in A(x_t)$, $h_t \in \mathbb{H}_t$, $t \in \mathbb{N}_0$. Let Π be the set of all control policies and $\mathbb{F} \subset \Pi$ the subset of stationary policies. If necessary, see for example [3, 4, 5, 7, 8, 9, 10, 11] for further information on those policies. As usual, each stationary policy $\pi \in \mathbb{F}$ 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 $\pi = \{f, f, f, \ldots\}$. In this case we use the notation f for π and we write

$$c(x, f) := c(x, f(x)) \quad \text{and} \quad G_n(x, f, s) := G_n(x, f(x), s)$$

for all $x \in X$, $s \in \Re^k$ and $n = 0, 1, \ldots, \infty$.

For a fixed $n = 0, 1, \ldots, \infty$, let $V_n(\pi, x)$ be the α-discounted cost using the policy $\pi \in \Pi$, given the initial state $x_0 = x$, when the control model is M_n [see (6)]. That is,

$$V_n(\pi, x) := E_x^{(n)\pi}\left[\sum_{t=0}^{\infty} \alpha^t c(x_t, a_t)\right], \tag{8}$$

where $\alpha \in (0,1)$ is the so-called discount factor, and $E_x^{(n)\pi}$ denotes the expectation operator with respect to the probability measure $P_x^{(n)\pi}$ induced by the policy π, given the initial state $x_0 = x$ and the model M_n (see, e.g., [3]). The corresponding value (or optimal cost) function is

$$V_n(x) := \inf_{\pi \in \Pi} V_n(\pi, x), \quad x \in X. \tag{9}$$

A policy $\pi^* \in \Pi$ is said to be α-discounted optimal (or simply α-optimal) for the control model M_n $(n = 0, 1, \ldots, \infty)$ if

$$V_n(x) = V_n(\pi^*, x) \quad \text{for all } x \in X. \tag{10}$$

Throughout the paper, we will use the following assumptions on the Markov control model. Note that Assumption 2.1 allows an unbounded cost-per-stage function $c(x, a)$ provided that it is upper bounded by some function $W(x)$. Next, Assumption 2.2 refers to system (1). Assumptions 2.4 and 2.6 are technical requirements on the unknown density ρ and the function W.

Assumption 2.1 (Bounds and semicontinuity.).
a) For all $x \in X$ the function $a \to c(x, a)$ is lower semicontinuous (l.s.c.) on $A(x)$. Moreover, there exists a measurable function $W : X \to [1, \infty)$ such that $\sup_{A(x)} c(x, a) \le \bar{c} W(x)$, $x \in X$, for some constant $\bar{c} > 0$.
b) For each $x \in X$, $A(x)$ is a σ-compact set.

Assumption 2.2 (On the dynamics of the system.). For each $n \in \mathbb{N}_0$, the function $G_n : \mathbb{K} \times \Re^k \to X$ is continuous, and furthermore, there exists a continuous function $G_\infty : \mathbb{K} \times \Re^k \to X$ such that the transition law $Q_n(B|x, a) = E 1_B \cdot [G_n(x, a, \xi_t)]$ converges (setwise) to $Q_\infty(B|x, a) = E 1_B[G_\infty(x, a, \xi_t)]$ as $n \to \infty$, for each $B \in \mathbb{B}(X)$.

Remark 2.3. Suppose that model (1) is noise additive, i.e. that $x_{n+1} = G_n(x_n, a_n) + \xi_n$ for all n, and that the density ρ of ξ_n is bounded and continuous. Assumption 2.2 then trivially holds if G_n converges pointwise to G_∞. See [11].

In the remainder, we fix an arbitrary $\varepsilon \in (0, 1/2)$ and denote L_q the space $L_q(\Re^k)$ where $q := 1 + 2\varepsilon$. Also we choose and fix a nonnegative and measurable 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_n in (1).

We define the set $\tilde{D} = \tilde{D}(\bar{\rho}, L, \beta_0, b_0, p, q)$ as the set consisting of all densities μ on $\Re^k$ for which the following conditions hold.
a) $\mu \in L_q$;
b) there exists a constant L such that for each $z \in \Re^k$

$$\|\Delta_z \mu\|_q \le L|z|^{1/q}, \tag{11}$$

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

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

d) for all $x \in X$, $n \in \mathbb{N}_0$

$$\sup_{A(x)} \int_{\Re^k} W^p[G_n(x, a, s)]\mu(s)\, ds \le \beta_0 W^p(x) + b_0, \tag{12}$$

for some $p > 1$, $\beta_0 < 1$, $b_0 < \infty$.

Assumption 2.4 (On the density ρ.). The density ρ belongs to $\tilde{D}$.

Remark 2.5. When $k = 1$ it is not difficult (see [4]) to show that a sufficient condition for (11) is the following. There are a finite set $H \subset \Re$ (possibly empty) and a constant $M \ge 0$ such that:

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

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

Note that H might include points of discontinuity of ρ if such points exist. Moreover, from i) and ii) $\rho'(x) \ge 0$ for $x \le -M$ and $\rho'(x) \le 0$ for $x \ge M$.

Assumption 2.6.

a) For all $s \in \Re^k$ the function φ defined by

$$\varphi(s) := \sup_X [W(x)]^{-1} \sup_{a \in A(x), n \in \mathbb{N}_0} W[G_n(x, a, s)] \tag{13}$$

is finite, and verifies

b) $\int_{\Re^k} \varphi^2(s) |\bar{\rho}(s)|^{1-2\varepsilon}\, ds < \infty$.

The function φ in (13) may be nonmeasurable. In this case we suppose the existence of a measurable upper bound $\bar{\varphi}$ of φ for which Assumption 2.6(b) holds. Besides, from (13), note that, for each $n = 0, 1, \ldots, \infty$, Assumption 2.6 holds with φ_n instead of φ, where

$$\varphi_n(s) := \sup_X [W(x)]^{-1} \sup_{a \in A(x)} W[G_n(x, a, s)].$$

In Section 5 we give an example of a controlled process for which all assumptions presented in this section hold.

3 Preliminary results

Let W be the function in Assumption 2.1(a). We denote by L_W^∞ the normed linear space of all measurable functions $u : X \to \Re$ with

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

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 3.1. *Suppose that Assumption 2.1(a) holds and the density ρ satisfies the condition (12). Then, for all $\pi \in \Pi$, $x \in X$ and $n \in \mathbb{N}_0$:*

a) *[4] denoting $\beta = \beta_0^{1/p}$ and $b = b_0^{1/p}$,*

$$\sup_{A(x)} \int_{\Re^k} W[G_n(x, a, s)]\rho(s)\, ds \leq \beta W(x) + b; \tag{15}$$

b) *[4]* $\sup_{t \geq 1} E_x^{(n)\pi}[W^p(x_t)] < \infty$ *and* $\sup_{t \geq 1} E_x^{(n)\pi}[W(x_t)] < \infty$.

If moreover Assumption 2.2 holds, then

c) *[11] for all $x \in X$,*

$$\sup_{A(x)} \int_{\Re^k} W^p[G_\infty(x, a, s)]\rho(s)\, ds \leq \beta_0 W^p(x) + b_0$$

and (from (15))

$$\sup_{A(x)} \int_{\Re^k} W[G_\infty(x, a, s)]\rho(s)\, ds \leq \beta W(x) + b, \tag{16}$$

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

d) *[11] for each $n = 0, 1, \ldots, \infty$, the value function V_n in (9) and the functions*

$$V^*(x) := \limsup_{n \to \infty} V_n(x) \quad \textit{and} \quad V_*(x) := \liminf_{n \to \infty} V_n(x) \quad (x \in X) \tag{17}$$

are in L_W^∞. In fact,

$$0 \leq V_n(x) \leq \bar{c} W(x)/(1 - \alpha), \quad x \in X. \tag{18}$$

From the fact that, for each $n = 0, 1, \ldots, \infty$, $Q_n(\cdot \mid \cdot)$ is a stochastic kernel [see (7)], it is easy to prove that for every nonnegative function $u \in L_W^\infty$, and every $r \in \Re$, the set

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

is Borel in $\mathbb{K}$. Using this fact, the following result is a consequence of Corollary 4.3 in [15].

Lemma 3.2. *Let $\alpha \in (0, 1)$ be an arbitrary but fixed discount factor, and u a nonnegative function in L_W^∞. Under Assumptions 2.1 and 2.2, if ρ satisfies (15), then for any $\delta > 0$ and $n = 0, 1, \ldots, \infty$, there exists a policy $f_{\delta,n} \in \mathbb{F}$ such that*

$$c(x, f_{\delta,n}) + \alpha \int_{\Re^k} u[G_n(x, f_{\delta,n}, s)]\rho(s)\, ds \leq u(x) + \delta \quad \forall x \in X. \tag{19}$$

The selector $f_{\delta,n}$ is also called a δ-minimizer of the function $a \mapsto c(x, a) + \alpha \int u \cdot [G_n(x, a, s)]\rho(s)\, ds$.

Throughout the paper we will repeatedly use the following inequalities. Let μ be a density satisfying (15) and (16), then

$$|u(x)| \leq \|u\|_W W(x) \tag{20}$$

$$\text{and} \quad \int_{\Re^k} u[G_n(x,a,s)]\mu(s)\,ds \leq \|u\|_W[\beta W(x) + b], \tag{21}$$

for all $n = 0, 1, \ldots, \infty$, $u \in L_W^\infty$, $x \in X$, $a \in A(x)$. The relation (20) is a consequence of the definition of $\|\cdot\|_W$ in (14), and (21) holds thanks to (20).

Theorem 3.3. *Suppose that Assumptions 2.1 and 2.2 hold, and that the density ρ satisfies the condition (12). Then, $V_n(x) \to V_\infty(x)$, as $n \to \infty$, for all $x \in X$, and the value function $V_\infty(x) \in L_W^\infty$ satisfies the α-discounted cost optimality equation*

$$V_\infty(x) = \inf_{a \in A(x)} \left[c(x,a) + \alpha \int_{\Re^k} V_\infty[G_\infty(x,a,s)]\rho(s)\,ds \right], \quad x \in X. \tag{22}$$

Theorem 3.3 was proved in [11] supposing, in addition, continuity and boundedness of the density ρ. These stronger assertions are necessary to get a unique solution V_∞ to the optimality equation (22). In the present context, the uniqueness is not required, which allows weaker assumptions. Here we give a sketch of the proof without these conditions, which is a slight modification of [11].

Proof. Let us first fix an arbitrary $n \in \mathbb{N}_0$. Then (see [9, Chapter 8]), Assumptions 2.1 and 2.2, and (12), together with Lemma 3.1(a,d), ensure that the value function V_n in (9) satisfies

$$V_n(x) = \inf_{a \in A(x)} \left[c(x,a) + \alpha \int_{\Re^k} V_n[G_n(x,a,s)]\rho(s)\,ds \right], \quad x \in X. \tag{23}$$

Now, take the limit infimum in (23) as $n \to \infty$. Then, from (18), applying an extension of Fatou's Lemma [8] and a general result on the interchange of limits and minima [10], we get

$$V_*(x) \geq \inf_{a \in A(x)} \left[c(x,a) + \alpha \int_{\Re^k} V_*[G_\infty(x,a,s)]\rho(s)\,ds \right], \quad x \in X. \tag{24}$$

where V_* is as in (17). From Lemma 3.1(b), $V_* \in L_W^\infty$. Let $\varepsilon > 0$ be an arbitrary number. According to Lemma 3.2, there exists an ε-minimizer, $f_\varepsilon \in \mathbb{F}$, of the right hand side of (24), that is

$$c(x, f_\varepsilon) + \alpha \int_{\Re^k} V_*[G_\infty(x, f_\varepsilon, s)]\rho(s)\,ds \leq V_*(x) + \varepsilon, \quad x \in X.$$

Iteration of the latter inequality yields

$$V_*(x) \geq \sum_{t=0}^{N-1} \alpha^t E_x^{(\infty)f_\varepsilon} c(x_t, f_\varepsilon) + \alpha^N E_x^{(\infty)f_\varepsilon} V_*(x_N) - \varepsilon \sum_{t=0}^{N-1} \alpha^t. \tag{25}$$

Letting $N \to \infty$ in (25), observe that from (20) and Lemma 3.1(b,d), we have $\alpha^N E_x^{(\infty)f_\varepsilon} V_*(x_N) \to 0$, which, together with (8), implies $V_*(x) \geq V_\infty(f_\varepsilon, x) - \varepsilon/(1-\alpha) \geq V_\infty(x) - \varepsilon/(1-\alpha)$. As $\varepsilon > 0$ was arbitrary, we conclude that

$$V_*(x) \geq V_\infty(x), \quad x \in X. \tag{26}$$

The remainder of the proof is as in [11], which consists, mainly, in showing that $V^*(x) \leq V_\infty(x)$, for all $x \in X$, which, together with (26) yields that $V_*(x) = V^*(x) = V_\infty(x)$ for all $x \in X$. □

Remark 3.4. Since $V_n \to V_\infty$, it is important to have in mind that $f_{\delta,\infty}$, defined in Lemma 3.2, can be obtained as an "accumulation point" of the δ-minimizers $\{f_{\delta,n}\}$ for the control models M_n with *finite n*. Indeed, by a result of [16], there is a policy $f_{\delta,\infty} \in \mathbb{F}$ such that, for each $x \in X$, $f_{\delta,\infty}(x) \in A(x)$ is an accumulation point of $\{f_{\delta,n}(x)\}$. That is to say, for each $x \in X$, there exists a subsequence $\{n_i(x)\}$ of $\{n\}$ such that

$$f_{\delta,n_i(x)}(x) \to f_{\delta,\infty}(x) \quad \textit{as } i \to \infty.$$

Now fix an arbitrary $x \in X$ and in (19) replace u with V_n and n with $n_i(x)$. Then letting $i \to \infty$, as c is l.s.c., from Theorem 3.3 we obtain

$$c(x, f_{\delta,\infty}) + \alpha \int_{\Re^k} V_\infty[G_\infty(x, f_{\delta,\infty}, s)]\rho(s)\, ds \leq V_\infty(x) + \delta \quad \forall x \in X,$$

which implies that $f_{\delta,\infty}$ is a δ-minimizer of V_∞ thanks to (22).

4 Adaptive policies

To construct an adaptive policy, we first present a statistical method to estimate ρ. It is based on a density estimation scheme that was originally proposed in [4] to obtain an asymptotically discount optimal adaptive policy for the time-invariant model M_∞, see also [5]. We slightly modify this estimation scheme to make it independent of M_∞.

Let $\xi_0, \xi_1, \ldots, \xi_{t-1}$ be independent realizations (observed up to time $t-1$), of r.v.'s with the unknown density ρ. We suppose that Assumptions 2.4 and 2.6 hold.

Let $\hat{\rho}_t(s) := \hat{\rho}_t(s; \xi_0, \xi_1, \ldots, \xi_{t-1})$, for $s \in \Re^k$, be an arbitrary sequence of estimators of ρ belonging to L_q, and such that for some $\gamma > 0$

$$E\|\rho - \hat{\rho}_t\|_q^{qp'/2} = \mathbf{O}(t^{-\gamma}) \quad \text{as } t \to \infty, \tag{27}$$

where p' is given by the relation $1/p + 1/p' = 1$. Examples of estimators satisfying (27) are given in [6].

Then, we estimate ρ by the projection ρ_t of $\hat{\rho}_t$ on the set D of densities in L_q defined as follows:

$$D := \Big\{ \mu \in L_q : \mu \text{ is a density function on } \Re^k,\ \mu(s) \le \bar{\rho}(s)\ a.e. \text{ and}$$

$$\int W[G_n(x,a,s)]\mu(s)\,ds \le \beta W(x) + b,\ \forall n \in \mathbb{N}_0,\ (x,a) \in \mathbb{K} \Big\}.$$

See Lemma 3.1(a) for the constants β and b.

From Assumption 2.4 and Lemma 3.1(a), we have that $\rho \in \tilde{D} \subset D$, and so D is nonempty. Moreover, the existence (and uniqueness) of the estimator ρ_t is guaranteed because D is convex and closed in L_q [4]. Note also that if Assumption 2.2 holds, Lemma 3.1(c) yields that ρ belongs to the following set D_∞, used in [4]:

$$D_\infty := \Big\{ \mu \in L_q : \mu \text{ is a density function on } \Re^k,\ \mu(s) \le \bar{\rho}(s)\ a.e. \text{ and}$$

$$\int W[G_\infty(x,a,s)]\mu(s)\,ds \le \beta W(x) + b,\ \forall (x,a) \in \mathbb{K} \Big\}. \tag{28}$$

Hence, D is a subset of D_∞, which yields that the following Lemma 4.1 still holds.

Lemma 4.1. *[4, 5] Suppose that Assumptions 2.4 and 2.6 hold. Then*

$$E\|\rho_t - \rho\|^{p'} = \mathrm{O}(t^{-\gamma}) \quad \text{as } t \to \infty, \tag{29}$$

where $\|\cdot\|$ *is the pseudo-norm on the space of all densities* μ *on* $\Re^k$ *defined as:*

$$\|\mu\| := \sup_X [W(x)]^{-1} \sup_{A(x)} \int_{\Re^k} W[G_\infty(x,a,s)]\mu(s)\,ds. \tag{30}$$

For arbitrary density μ in $\Re^k$, the pseudo-norm $\|\mu\|$ may be infinite. However, by (28), $\|\mu\| < \infty$ for μ in D.

In the remainder of the paper, we fix an arbitrary discount factor $\alpha \in (0,1)$. The optimality of the adaptive policy is studied in the sense of the following definition.

Definition 4.2. a) [17] A policy $\pi \in \Pi$ is said to be asymptotically discount optimal for the control model M_n $(n = 0, 1, \ldots, \infty)$ if

$$|V_n^{(k)}(\pi, x) - E_x^{(n)\pi}[V_n(x_k)]| \to 0 \quad \text{as } k \to \infty, \quad \text{for all } x \in X,$$

where

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

is the expected total discounted cost for the control model M_n from stage k onward and $a_t = \pi_t(h_t)$.

b) Let $\delta \geq 0$. A policy π is δ-asymptotically discount optimal for the control model M_n ($n = 0, 1, \ldots, \infty$) if

$$\limsup_{k\to\infty} |V_n^{(k)}(\pi, x) - E_x^{(n)\pi}[V_n(x_k)]| \leq \delta, \quad x \in X.$$

From Definition 4.2(a) and (10) we have that discount optimality implies asymptotic discount optimality, and this one in turn implies δ-asymptotic discount optimality.

For any $\mu \in D$ and $n = 0, 1, \ldots, \infty$, let us define the operator $T_\mu^{(n)} : L_W^\infty \to L_W^\infty$ as

$$T_\mu^{(n)} u(x) := \inf_{A(x)} \left\{ c(x, a) + \alpha \int_{\Re^k} u[G_n(x, a, s)]\mu(s)\, ds \right\}, \qquad x \in X, \quad u \in L_W^\infty. \tag{31}$$

Observe in particular that, from (23), $T_\rho^{(n)} V_n = V_n$.

For the construction of the adaptive policy we replace the unknown density ρ by its estimate ρ_t and exploit the corresponding discounted optimality equation for the model M_∞ (see (22)), or more generally for model M_n, $n = 0, 1, \ldots, \infty$. As $\rho_t \in D$ for all $t \geq 1$, the following Proposition 4.3 is a direct consequence of Lemmas 3.1, 3.2, Theorem 3.3 and Remark 3.4.

Proposition 4.3.

a) *Suppose that Assumptions 2.1(a) and 2.2 hold. Then, for each $t \geq 1$ and $n = 0, 1, \ldots, \infty$, there exists a function $V_n^{(t)} \in L_W^\infty$ such that $T_{\rho_t}^{(n)} V_n^{(t)} = V_n^{(t)}$. Moreover,*

$$V_n^{(t)}(x) \leq \frac{\bar{c}}{1-\alpha} W(x), \quad x \in X. \tag{32}$$

b) *Under Assumptions 2.1 and 2.2, for each $t \geq 1$, $n \in \mathbb{N}$ and $\delta_t^* > 0$, there exists a stationary policy $f_{t,n}^* \in \mathbb{F}$ such that*

$$c(x, f_{t,n}^*) + \alpha \int_{\Re^k} V_n^{(t)}[G_n(x, f_{t,n}^*, s)]\rho_t(s)\, ds \leq V_n^{(t)}(x) + \delta_t^*, \quad x \in X. \tag{33}$$

c) *It follows from part (b) and Remark 3.4 that, for any $t \geq 1$, there exists a stationary policy $f_{t,\infty}^* \in \mathbb{F}$ such that, for all $x \in X$, $f_{t,\infty}^*(x) \in A(x)$ is an accumulation point of $\{f_{t,n}^*(x)\}$, and we have*

$$c(x, f_{t,\infty}^*) + \alpha \int_{\Re^k} V_\infty^{(t)}[G_\infty(x, f_{t,\infty}^*, s)]\rho_t(s)\, ds \leq V_\infty^{(t)}(x) + \delta_t^*, \quad x \in X. \tag{34}$$

The minimization in (31), with ρ_t instead of μ, 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 $\omega \in \Omega$.

Definition 4.4. For each fixed $n = 0, 1, \ldots, \infty$ and any arbitrary sequence $\{\delta_t^*\}$ of positive numbers, let $\{f_{t,n}^*\}$ be a sequence of functions satisfying (33) for each integer $t \in \mathbb{N}_0$. We define the adaptive policy $\pi_n^* = \{\pi_{t,n}^*\}$ as follows:

$$\pi_{t,n}^*(h_t) = \pi_{t,n}^*(h_t; \rho_t) := f_{t,n}^*(x_t), \quad h_t \in \mathbb{H}_t, \quad t = 1, 2, \ldots$$

while $\pi_{0,n}^*(x)$ is any fixed action in $A(x)$.

Note that, from Proposition 4.3(c), π_∞^* is the sequence $\{\pi_{t,\infty}^*\}$, where each component $\pi_{t,\infty}^*$, $t = 1, 2, \ldots$, can be obtained as an accumulation point of the sequence $\{f_{t,n}^*(x_t)\}$, indexed by n.

As $\{\delta_t^*\}$ is arbitrary, we choose it convergent and denote $\delta^* := \lim_{t\to\infty} \delta_t^*$. We are now ready to state our main result.

Theorem 4.5. *Suppose that Assumptions 2.1, 2.2, 2.4 and 2.6 hold. Then the adaptive policy π_∞^* is δ^*-asymptotically discount optimal for the model M_∞. In particular, if $\delta^* = 0$ then the policy π_∞^* is asymptotically discount optimal.*

Remark 4.6. (a) Since Assumptions 2.4 and 2.6 are stated for each *finite* $n \in \mathbb{N}_0$, we have (see [4]) that the adaptive policy π_n^* introduced in Definition 4.4 is δ^*-asymptotically discount optimal for the model M_n, for each *finite* n. The whole point of Theorem 4.5 is that this result also holds for $n = \infty$.

b) The notion of asymptotic optimality introduced in Definition 4.2 can be characterized in terms of the so-called discounted discrepancy function, defined for each $n = 0, 1, \ldots, \infty$, as:

$$\Phi_n(x, a) := c(x, a) + \alpha \int_{\Re^k} V_n[G_n(x, a, s)]\rho(s)\, ds - V_n(x), \quad (x, a) \in \mathbb{K}, \quad (35)$$

which is nonnegative in view of (22) and (23). That is (see e.g. [7, 10]), a policy $\pi \in \Pi$ is asymptotically discount optimal for the control model M_n ($n = 0, 1, \ldots, \infty$) if

$$E_x^{(n)\pi}[\Phi_n(x_t, a_t)] \to 0 \quad \textit{as } t \to \infty, \quad \text{for all } x \in X.$$

Moreover, for $\delta \geq 0$, it is easy to see that a policy $\pi \in \Pi$ is δ-asymptotically discount optimal for the control model M_n ($n = 0, 1, \ldots, \infty$) if

$$\limsup_{t\to\infty} E_x^{(n)\pi}[\Phi_n(x_t, a_t)] \leq \delta, \quad x \in X. \quad (36)$$

Thus, Theorem 4.5 will be proved if we show that the adaptive policy π_∞^* satisfies (36).

Proof of Theorem 4.5. Let us fix an arbitrary number $\theta \in (\alpha, 1)$ and define $\overline{W}(x) := W(x) + d$, $x \in X$, where $d := b(\theta/\alpha - 1)^{-1}$. Let $L_{\overline{W}}^\infty$ be the space 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 that

$$\|u\|_{\overline{W}} \leq \|u\|_W \leq \|u\|_{\overline{W}}(1+d'), \tag{37}$$

where $d' := d/\inf_X W(x)$. Hence $L_W^\infty = L_{\overline{W}}^\infty$ and the norms $\|\cdot\|_W$ and $\|\cdot\|_{\overline{W}}$ are equivalent.

On the other hand, a consequence of [18, Lemma 2] is that, for each $\mu \in D$, the inequality (16) in Lemma 3.1 implies that the operator $T_\mu := T_\mu^\infty$ defined in (31) is a contraction of modulus θ with respect to the norm $\|\cdot\|_{\overline{W}}$, that is,

$$\|T_\mu v - T_\mu u\|_{\overline{W}} \leq \theta \|v-u\|_{\overline{W}}, \quad \forall v,u \in L_{\overline{W}}^\infty. \tag{38}$$

Hence, from Proposition 4.3(a) we can see that

$$\|V_\infty - V_\infty^{(t)}\|_{\overline{W}} \leq \|T_\rho V_\infty - T_{\rho_t} V_\infty\|_{\overline{W}} + \theta \|V_\infty - V_\infty^{(t)}\|_{\overline{W}},$$

which implies

$$\|V_\infty - V_\infty^{(t)}\|_{\overline{W}} \leq \frac{1}{1-\theta} \|T_\rho V_\infty - T_{\rho_t} V_\infty\|_{\overline{W}}, \quad t \in \mathbb{N}_0. \tag{39}$$

Now, from (18), (30) and the fact that $[\overline{W}(\cdot)]^{-1} < [W(\cdot)]^{-1}$, we obtain

$$\begin{aligned} \|T_\rho V_\infty - T_{\rho_t} V_\infty\|_{\overline{W}} &\leq \alpha \sup_X [\overline{W}(x)]^{-1} \sup_{A(x)} \int_{\Re^k} V_\infty[G_\infty(x,a,s)]|\rho(s)-\rho_t(s)|\, ds \\ &\leq \frac{\alpha \bar{c}}{1-\alpha} \sup_X [W(x)]^{-1} \sup_{A(x)} \int_{\Re^k} W[G_\infty(x,a,s)] \\ &\quad \times |\rho(s)-\rho_t(s)|\, ds \leq \frac{\bar{c}}{1-\alpha}\|\rho-\rho_t\|, \quad t \in \mathbb{N}_0. \end{aligned} \tag{40}$$

Hence, from (37) and combining (39) and (40), we get

$$\begin{aligned} \|V_\infty - V_\infty^{(t)}\|_W &\leq (1+d')\|V_\infty - V_\infty^{(t)}\|_{\overline{W}} \\ &\leq \frac{\bar{c}(1+d')}{(1-\theta)(1-\alpha)} \|\rho - \rho_t\|, \quad t \in \mathbb{N}. \end{aligned} \tag{41}$$

On the other hand, for each $t \in \mathbb{N}_0$, we define the function $\Phi_\infty^{(t)} : \mathbb{K} \to \Re$ as:

$$\Phi_\infty^{(t)}(x,a) := c(x,a) + \alpha \int_{\Re^k} V_\infty^{(t)}[G_\infty(x,a,s)]\rho_t(s)\, ds - V_\infty^{(t)}(x), \quad (x,a) \in \mathbb{K}.$$

By the definition (35) of Φ_∞, we get (by adding and subtracting the term $\alpha \int_{\Re^k} V_\infty^{(t)}[G_\infty(x,a,s)]\rho(s)\, ds$)

$$
\begin{aligned}
&|\Phi_\infty^{(t)}(x,a) - \Phi_\infty(x,a)| \\
&\quad \le |V_\infty(x) - V_\infty^{(t)}(x)| + \alpha \int_{\Re^k} V_\infty^{(t)}[G_\infty(x,a,s)] |\rho_t(s) - \rho(s)|\, ds \\
&\qquad + \alpha \int_{\Re^k} |V_\infty^{(t)}[G_\infty(x,a,s)] - V_\infty[G_\infty(x,a,s)]| \rho(s)\, ds \\
&\quad \le \| V_\infty - V_\infty^{(t)} \|_W W(x) + \frac{\alpha \bar{c}}{1-\alpha} \int_{\Re^k} W[G_\infty(x,a,s)] |\rho_t(s) - \rho(s)|\, ds \\
&\qquad + \alpha[\beta W(x) + b] \| V_\infty^{(t)} - V_\infty \|_W,
\end{aligned}
$$

for each $(x,a) \in \mathbb{K}$, $t \in \mathbb{N}_0$ [see also (32)]. Hence, from (30) and (41), as $W(\cdot) \ge 1$ and $\alpha < 1$, it follows

$$
\sup_X [W(x)]^{-1} \sup_{A(x)} |\Phi_\infty^{(t)}(x,a) - \Phi_\infty(x,a)| \le C' \|\rho_t - \rho\|, \tag{42}
$$

where $C' = \dfrac{\bar{c}}{1-\alpha}\left[1 + \dfrac{(1+\beta+b)(1+d')}{1-\theta}\right]$. Moreover, by definition of the adaptive policy π_∞^* in Definition 4.4 and (34), we have $\Phi_\infty^{(t)}(\cdot, \pi_{t,\infty}^*(\cdot)) \le \delta_t^*$, $t \in \mathbb{N}_0$. Thus

$$
\begin{aligned}
\Phi_\infty(x_t, \pi_{t,\infty}^*(h_t)) &\le |\Phi_\infty(x_t, \pi_{t,\infty}^*(h_t)) - \Phi_\infty^{(t)}(x_t, \pi_{t,\infty}^*(h_t)) + \delta_t^*| \\
&\le \sup_{A(x_t)} |\Phi_\infty(x_t,a) - \Phi_\infty^{(t)}(x_t,a)| + \delta_t^* \\
&\le W(x_t) \sup_X [W(x)]^{-1} \sup_{A(x)} |\Phi_\infty(x_t,a) - \Phi_\infty^{(t)}(x_t,a)| + \delta_t^* \\
&\le C' W(x_t) \|\rho_t - \rho\| + \delta_t^*, \quad t \in \mathbb{N}_0.
\end{aligned} \tag{43}
$$

The latter inequality implies

$$
E_x^{(\infty)\pi_\infty^*}[\Phi_\infty(x_t,a_t)] \le C' E_x^{(\infty)\pi_\infty^*}[W(x_t)\|\rho_t - \rho\|] + \delta_t^*,
$$

and, therefore, to prove that π_∞^* is δ^*-asymptotically discount optimal [see (36)], it is enough to show that $E_x^{(\infty)\pi_\infty^*}[W(x_t)\|\rho_t - \rho\|] \to 0$ as $t \to \infty$. Define $\bar{C} := (E_x^{(\infty)\pi_\infty^*}[W^p(x_t)])^{1/p}$. By Lemma 3.1(c), $\bar{C} < \infty$. Applying Hölder's inequality, we deduce

$$
E_x^{(\infty)\pi_\infty^*}[W(x_t)\|\rho_t - \rho\|] \le \bar{C}(E_x^{(\infty)\pi_\infty^*}[\|\rho_t - \rho\|^{p'}])^{1/p'}.
$$

Then, observing that $E_x^{(\infty)\pi_\infty^*}[\|\rho_t - \rho\|^{p'}] = E[\|\rho_t - \rho\|^{p'}]$ (since ρ_t does not depend on x and π_∞^*), Lemma 4.1 yields the desired results. □

5 Example

We now discuss an example in biotechnological processes to illustrate how to verify our assumptions. Consider the following system

$$x_{n+1} = (H(x_n)g_n(x_n) + G(x_n, a_n) + \xi_n)^+ \quad (n \in \mathbb{N}_0), \tag{44}$$

$x_0 = x$ given, with state space $X = [0, \infty) \times [0, \infty)$ and actions sets $A(x) = A$ for all $x \in X$, where A is a compact subset of $\Re^2$. The functions H, g_n and G are continuous, and $\{\xi_n\}$ is an i.i.d. sequence of r.v.'s with bounded and continuous density ρ.

This model represents, for example, the real time evolution of the concentrations x_n of a biomass and a substrate in a bioreaction, directed by two control actions a_n. Such reactions are very common in depollution and in the agrofood industry [1]. The function $g_n(x)$ then characterizes the microbial growth rate, which is a time-varying quantity, influenced by many factors (biomass and substrate concentrations, temperature, pH, etc). However, under suitable conditions, the growth rate $g_n(x)$ tends to a "stable" growth rate $g_\infty(x)$ (in the sense of Assumption 2.2 for example), and so the time-varying system (44) "tends" to a time-homogeneous system such as (3).

To assure that the system (44) has a nice stable behavior, we make the following assumption on its dynamic:

Assumption 5.1. There exist a positive constant $v < 1$ and a norm $\|\cdot\|_{\Re^2}$ on X such that

$$\limsup_{\|x\|_{\Re^2} \to \infty} \sup_{i \in \mathbb{N}_0} \sup_{a \in A(x)} \frac{\|(H(x)g_i(x) + G(x,a))^+\|_{\Re^2}}{\|x\|_{\Re^2}} = v.$$

See for example [2] for further details on this kind of hypotheses.

The control objective is defined as the regulation of $\{x_n\}$ around a fixed reference point $x^* \in X$. To that aim, we choose the following cost function

$$c(x) := \|x - x^*\|_{\Re^2}^{1/2}, \quad x \in X.$$

The r.v.'s $\xi_0, \xi_1, \ldots$ are supposed to be i.i.d. with unknown density $\rho \in L_q$ satisfying the inequality

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

for some given constants $L < \infty$ and $q > 1$.

In addition, we assume that $E(\|\xi_0\|_{\Re^2}) < \infty$ and that there exists a constant $M < \infty$ such that $\rho(s) \le M \min\{1, 1/\|s\|_{\Re^2}^{1+r}\}$, for all $s \in \Re^2$.

Clearly, Assumptions 2.1, 2.2 and the conditions (a)–(c) in the definition of the set $\tilde{D}$ are satisfied defining, for $x \in X$ and $s \in \Re^2$, $W(x) := (\|x\|_{\Re^2} + \delta)^{1/2}$ and $\bar{\rho}(s) := M \min\{1, 1/\|s\|_{\Re^2}^{1+r}\}$, where $\delta \ge \max(1, \|x^*\|_{\Re^2})$.

On the other hand, a straightforward calculation shows that the density ρ satisfies the inequality (12) with $\beta_0 = v < 1$ and $b_0 = 2\delta + E\|\xi_0\|_{\Re^2} < \infty$. Therefore, Assumption 2.4 holds.

To conclude, it is easy to see that $\varphi(s) \leq 1 + \delta^{1/2} + \|s\|_{\Re^2}^{1/2} / \inf_X W(x) < \infty$, $s \in \Re^2$. Thus, choosing appropriate $r > 0$ in the definition of $\bar{\rho}$, Assumption 2.6 is satisfied and Theorem 4.5 holds.

References

[1] Bastin G, Dochain D (1990) On-line estimation and adaptive control of bioreactors. Elsevier, Amsterdam

[2] Duflo M (1997) Random iterative models. Springer-Verlag, Berlin

[3] Dynkin EB, Yushkevich AA (1979) Controlled Markov processes. Springer-Verlag, New York

[4] Gordienko EI, Minjárez-Sosa JA (1998) Adaptive control for discrete-time Markov processes with unbounded costs: discounted criterion. Kybernetika 34:217–234

[5] Gordienko EI, Minjárez-Sosa JA (1998) Adaptive control for discrete-time Markov processes with unbounded costs: average criterion. Math. Meth. of Oper. Res. 48:37–55

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

[7] Hernández-Lerma O (1989) Adaptive Markov control processes. Springer-Verlag, New York

[8] Hernández-Lerma O, Lasserre JB (1997) Policy iteration for average cost Markov control processes on Borel spaces. Acta Appl. Math. 47:125–154

[9] Hernández-Lerma O, Lasserre JB (1999) Further topics on discrete-time Markov control processes. Springer-Verlag, New York

[10] Hernández-Lerma O, Muñoz-de-Ozak M (1992) Discrete-time MCPs with discounted unbounded costs: optimality criteria. Kybernetika 28:191–212

[11] Hilgert N, Hernández-Lerma O (2000) Limiting optimal discounted-cost control of a class of time-varying stochastic systems. Syst. Control Lett. 40(1):37–42

[12] Hilgert N, Senoussi R, Vila JP (1996) Nonparametric estimation of time-varying autoregressive nonlinear processes. C. R. Acad. Sci. Paris Série 1, 323:1085–1090

[13] Kurano M (1972) Discrete-time markovian decision processes with an unknown parameter – average return criterion. J. Oper. Res. Soc. Japan 15:67–76

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

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

[16] Schäl M (1975) Conditions for optimality and for the limit on *n*-stage optimal policies to be optimal. Z. Wahrs. Verw. Gerb. 32:179–196

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

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