

Limiting Average Cost Adaptive Control Problem for Time-Varying Stochastic Systems

Nadine Hilgert* and J. Adolfo Minjárez-Sosa†

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 $x_{n+1} = G_n(x_n, a_n, \xi_n)$, $n = 0, 1, \dots$, where the ξ_n are i.i.d. \mathfrak{R}^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 $n \rightarrow \infty$. Assuming observability of ξ_n , we construct an adaptive policy which is average 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; average and discounted cost criteria; optimal adaptive policy.

1 Introduction

We are concerned with a discrete-time, time-varying stochastic control system of the form

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

*Laboratoire d'Analyse des Systèmes et de Biométrie, INRA-ENSA.M, 2 place Viala, 34060 Montpellier Cedex 1, France. (hilgert@ensam.inra.fr).

†*Corresponding author.* Departamento de Matemáticas, Universidad de Sonora, Rosales s/n, Col. Centro, 83000, Hermosillo, Sonora, México. (aminjare@gauss.mat.uson.mx)

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 \mathfrak{R}^k having an unknown density ρ . In addition, $\{G_n\}$ is a sequence of given functions converging to some function G_∞ in the following way:

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

where $1_B(\cdot)$ denotes the indicator function of the set B (See Assumption 2.2 for more details on this condition). This type of systems appears, for instance, in some time-varying controlled biotechnological processes (Ref. 1,2). We will illustrate the main results of this paper with a generic model of bioreaction.

Assuming that the realizations of the processes $\{\xi_t\}$ and $\{x_t\}$ are completely observable, our main objective is to introduce average cost 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. This work is motivated by a previous paper (Ref. 3) which deals with the construction of asymptotically discounted optimal adaptive policies for (3). We take advantage of these results for studying the average optimality via the average cost optimality inequality, and using a variant of the well-known vanishing discount factor approach (see (Ref. 4) in the case of a time-invariant system). It consists in fixing an appropriate sequence $\{\alpha_t\}$, $\alpha_t \nearrow 1$, of discount factors, and in exploiting the corresponding α_t -discounted optimality equation and its limit as $t \rightarrow \infty$. The adaptation to the time-varying system is realized by considering the average cost problem for the time-invariant system

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

for each fixed $n \in \mathbb{N}_0$, and then letting $n \rightarrow \infty$ to obtain the corresponding result for the limiting system (3). See also the case of an additive-noise model with known density in (Ref. 5). Put in this form, our main result, Theorem 3.1, can also be seen as a further result of (Ref. 4) on system (3) where the function G_∞ is unknown and estimated by some consistent functional estimator G_n .

The remainder of the paper is organized as follows. In Section 2 we introduce the Markov control models we are concerned with and the assumptions

required. The adaptive policies are introduced in Section 3 together with the main result, Theorem 3.1, and some preliminary facts. A generic example of a biotechnological process satisfying all the hypotheses of the paper is described in Section 4. Finally Section 5 is devoted to the proof of Theorem 3.1.

2 Markov control models

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

$$M_n := (X, A, \{A(x) \mid x \in X\}, Q_n, c) \quad (5)$$

associated to the system (4), in which 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 $x \in X$, the set $A(x) \in \mathbb{B}(A)$ denotes the set of admissible controls when the system is in state x , and is supposed nonempty. 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, the transition law Q_n corresponding to (4) is a stochastic kernel on X given \mathbb{K} , that is, for all $t \in \mathbb{N}_0$, $(x, a) \in \mathbb{K}$ and $B \in \mathbb{B}(X)$,

$$\begin{aligned} Q_n(B \mid x, a) &:= \text{Prob}[G_n(x_t, a_t, \xi_t) \in B \mid x_t = x, a_t = a] \\ &= \int_{\mathfrak{R}^k} 1_B[G_n(x, a, s)]\rho(s)ds, \end{aligned} \quad (6)$$

where $1_B(\cdot)$ denotes the indicator function of the set B , and $\{\xi_t\}$ is a sequence of i.i.d. random vectors (r.v.'s) on a probability space (Ω, \mathcal{F}, P) , with values in \mathfrak{R}^k and common distribution with an unknown density ρ . Finally, the cost-per-stage $c(x, a)$ is a nonnegative measurable real-valued function on \mathbb{K} .

Two assumptions are made on some components of the control models and on the dynamics of system (1).

Assumption 2.1 (*Bounds and semicontinuity.*)

a) For all $x \in X$ the function $a \rightarrow c(x, a)$ is lower semicontinuous (l.s.c.) on $A(x)$. Moreover, there exists a measurable function $W : X \rightarrow [\bar{W}, \infty)$ such that $\sup_{A(x)} c(x, a) \leq W(x)$, $x \in X$, where \bar{W} is a positive constant.

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

Note that, by Assumption 2.1(a), $c(x, a)$ can be an unbounded cost-per-stage function, provided that is upper bounded by some function $W(x)$.

Assumption 2.2 (*On the dynamics of the system.*) For each $n \in \mathbb{N}_0$, the function $G_n : \mathbb{K} \times \mathfrak{R}^k \rightarrow X$ is continuous, and furthermore, there exists a continuous function $G_\infty : \mathbb{K} \times \mathfrak{R}^k \rightarrow X$ such that (2) holds, that is: the transition law $Q_n(B \mid x, a) = E1_B[G_n(x, a, \xi_t)]$ converges (setwise) to $Q_\infty(B \mid x, a) = E1_B[G_\infty(x, a, \xi_t)]$ as $n \rightarrow \infty$, for each $B \in \mathcal{B}(X)$.

One particular case which satisfies Assumption 2.2 is the following: 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 ρ is bounded and continuous, then (2) trivially holds if G_n converges pointwise to G_∞ . See (Ref. 5,6).

Let Π be the set of all control policies and $\mathbb{F} \subset \Pi$ be the subset of all deterministic stationary policies (Ref. 7). As usual, every stationary policy $\pi \in \mathbb{F}$ is identified with a measurable function $f : X \rightarrow A$ such that $f(x) \in A(x)$ for every $x \in X$, so that π is of the form $\pi = \{f, f, f, \dots\}$. 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 $x \in X$, $s \in \mathfrak{R}^k$ and $n = 0, 1, \dots, \infty$.

For a fixed $n = 0, 1, \dots, \infty$, when using a policy $\pi \in \Pi$, given the initial state $x_0 = x$, we define the total expected α -discounted cost when the control model is M_n [see (5)] as

$$V_{\alpha, n}(\pi, x) := E_x^{(n)\pi} \left[\sum_{t=0}^{\infty} \alpha^t c(x_t, a_t) \right], \quad (7)$$

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., Ref. 7). We also define the long-run expected average cost when the control model is M_n as

$$J_n(\pi, x) := \limsup_{m \rightarrow \infty} m^{-1} E_x^{(n)\pi} \left[\sum_{t=0}^{m-1} c(x_t, a_t) \right]. \quad (8)$$

The functions

$$V_{\alpha,n}(x) := \inf_{\pi \in \Pi} V_{\alpha,n}(\pi, x) \quad \text{and} \quad J_n(x) := \inf_{\pi \in \Pi} J_n(\pi, x) \quad \text{for } x \in X \quad (9)$$

are the optimal α -discounted cost and the optimal average cost respectively, corresponding to the control model M_n .

A policy $\pi^* \in \Pi$ is said to be α -discounted optimal for the control model M_n ($n = 0, 1, \dots, \infty$) if $V_{\alpha,n}(x) = V_{\alpha,n}(\pi^*, x)$ for all $x \in X$. Similarly, a policy $\pi^* \in \Pi$ is said to be average cost optimal (or simply AC-optimal) for the control model M_n ($n = 0, 1, \dots, \infty$) if $J_n(x) = J_n(\pi^*, x)$ for all $x \in X$.

Our main problem is precisely to introduce AC-optimal adaptive policies for the limiting system (3) when the density ρ of the r.v.'s ξ_t is unknown. To that aim, we shall require two sets of hypotheses in addition to Assumptions 2.1 and 2.2. Assumption 2.3 is made of technical requirements on the unknown density ρ and on the function W introduced in Assumption 2.1(a), whereas Assumption 2.4 refers to the transition law and will be necessary to ensure a solution to the average cost optimality inequality (ACOI).

Assumption 2.3 (On the density ρ and the function W .) Fix an arbitrary $\varepsilon \in (0, 1/2)$ and denote L_q the space $L_q(\mathfrak{R}^k)$ where $q := 1 + 2\varepsilon$.

a) $\rho \in L_q$;

b) there exists a constant L such that for each $z \in \mathfrak{R}^k$

$$\|\Delta_z \rho\|_{L_q} \leq L |z|^{1/q}, \quad (10)$$

where $\Delta_z \rho(s) := \rho(s+z) - \rho(s)$, $s \in \mathfrak{R}^k$ and $|\cdot|$ is the Euclidean norm in \mathfrak{R}^k ;

c) there exists a nonnegative and measurable function $\bar{\rho} : \mathfrak{R}^k \rightarrow \mathfrak{R}$ such that $\rho(s) \leq \bar{\rho}(s)$ almost everywhere with respect to the Lebesgue measure;

d) for all $s \in \mathfrak{R}^k$, the function φ defined by

$$\varphi(s) := \sup_X [W(x)]^{-1} \sup_{a \in A(x), n \in \mathbb{N}_0} \bar{W}[G_n(x, a, s)] \quad (11)$$

is finite, and moreover

$$e) \int_{\mathfrak{R}^k} \varphi^2(s) |\bar{\rho}(s)|^{1-2\varepsilon} ds < \infty.$$

The function φ in (11) might be nonmeasurable. In such a case we suppose the existence of a measurable upper bound $\bar{\varphi}$ of φ for which Assumption 2.3(e) holds. Besides, from (11), note that, for each $n = 0, 1, \dots, \infty$, 2.3(d), (e) holds with φ_n instead of φ , where

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

Assumption 2.4 (*Bounds for the transition law.*) *There exists a probability measure ν on $(X, \mathcal{B}(X))$ and two constants $\bar{\psi} > 0$ and $\beta_0 \in (0, 1)$ for which the following holds: for each $f \in \mathcal{IF}$ and $n \in \mathbb{N}_0$, there is a measurable function $0 \leq \psi_{f,n}(\cdot) \leq 1$ such that for any $x \in X$ and $B \in \mathcal{B}(X)$,*

$$i) Q_n(B | x, f) \geq \psi_{f,n}(x)\nu(B);$$

$$ii) \int_X \psi_{f,n}(x)\nu(dx) \geq \bar{\psi};$$

$$iii) \int_{\mathfrak{R}^k} W^p[G_n(x, f, s)]\rho(s)ds \leq \beta_0 W^p(x) + \psi_{f,n}(x)b_0 \text{ for some } p > 1, \text{ with } b_0 := \int_X W^p(y)\nu(dy) < \infty.$$

Remark 2.1 *The set of Assumptions 2.1 to 2.4 is more restrictive than the assumption used in (Ref. 3) for the discounted criterion because of additional difficulties in the asymptotic analysis of the average cost. In fact, in (Ref. 3), it was only necessary to impose Assumptions 2.1 to 2.3 together with*

$$\int_{\mathfrak{R}^k} W^p[G_n(x, a, s)]\rho(s)ds \leq \beta_0 W^p(x) + b_0, \quad (12)$$

for all $x \in X$, $a \in A(x)$, $n \in \mathbb{N}_0$, and for some $p > 1$, $\beta_0 < 1$, $b_0 < \infty$. Now, it is easy to see that (12) follows from Assumptions 2.4(i), (iii) taking the same p , β_0 and b_0 , and by the fact that for each $x \in X$ and $a \in A(x)$, there is a stationary policy $f \in \mathcal{F}$ with $f(x) = a$ (see Example 2.6 in Ref. 8). Moreover, in Lemma 5.1(a) below, it is stated that

$$\int_{\mathfrak{R}^k} W[G_n(x, a, s)]\rho(s)ds \leq \beta W(x) + b, \text{ for all } x \in X, a \in A(x), n \in \mathbb{N}_0, \quad (13)$$

where $\beta = \beta_0^{1/p}$ and $b = b_0^{1/p}$.

3 Construction of the adaptive policy and main results

The average optimality of the adaptive policy is studied via a variant of the so-called vanishing discount factor approach. This approach consists in first choosing an appropriate sequence $\{\alpha_t\}$, $\alpha_t \nearrow 1$, of discount factors, then replacing the unknown density ρ by suitable estimators ρ_t , and finally exploiting the corresponding α_t -discounted optimality equations for the model M_n , $n = 0, 1, \dots, \infty$, taking limit as $t \rightarrow \infty$. To state our main result, we first present two lemmas that summarize important facts on discounted and average criteria, next we introduce a sequence $\{\rho_t\}$ of estimators of ρ which has been used in (Ref. 9,10,4).

Results on discounted and average criteria. Let W be the function introduced in Assumption 2.1. We denote by L_W^∞ the normed linear space of all measurable functions $u : X \rightarrow \mathfrak{R}$ with a finite norm $\|u\|_W$ defined as

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

Lemma 3.1 (Ref. 6,3). *Let $\alpha \in (0, 1)$ be an arbitrary but fixed discount factor. If ρ satisfies the condition (12) (or (13)), then under Assumptions 2.1 and 2.2,*

a) $V_{\alpha,n}(x) \rightarrow V_{\alpha,\infty}(x)$, as $n \rightarrow \infty$, for all $x \in X$;

b) for $n = 0, 1, \dots, \infty$, the value function $V_{\alpha, n}$ satisfies the α -discounted cost optimality equation

$$V_{\alpha, n}(x) = \inf_{a \in A(x)} \left[c(x, a) + \alpha \int_{\mathfrak{R}^k} V_{\alpha, n}[G_n(x, a, s)] \rho(s) ds \right], \quad x \in X; \quad (15)$$

c) for any $\delta > 0$ and $n = 0, 1, \dots, \infty$, there exists a policy $f_{\delta, n} \in \mathcal{IF}$ such that

$$c(x, f_{\delta, n}) + \alpha \int_{\mathfrak{R}^k} V_{\alpha, n}[G_n(x, f_{\delta, n}, s)] \rho(s) ds \leq V_{\alpha, n}(x) + \delta \quad \forall x \in X. \quad (16)$$

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

It is easy to see, from results in (Ref. 11), that if ρ satisfies the condition (13), under Assumption 2.1(a), we have

$$V_{\alpha, n}(x) \leq \frac{CW(x)}{1 - \alpha} \quad \text{for all } x \in X, \quad \alpha \in (0, 1), \quad n \in \mathbb{N}_0. \quad (17)$$

Thus, from Lemma 5.1(c) below, we have that

$$V_{\alpha, \infty}(x) \leq \frac{CW(x)}{1 - \alpha} \quad \text{for all } x \in X, \quad \alpha \in (0, 1). \quad (18)$$

Lemma 3.2 (Ref. 5) *Suppose that Assumption 2.1, 2.2 and 2.4 hold. Then:*

a) *For each $n = 0, 1, \dots, \infty$, there exist a constant j_n^* and a nonnegative function $h_n \in L_W^\infty$ such that, for all state $x \in X$, the average cost optimality inequality (ACOI) holds, i.e.,*

$$j_n^* + h_n(x) \geq \inf_{A(x)} \left[c(x, a) + \int_{\mathfrak{R}^k} h_n[G_n(x, a, s)] \rho(s) ds \right]. \quad (19)$$

Moreover, j_n^ is the optimal average cost for the model M_n , i.e., $j_n^* = \inf_{\pi \in \Pi} J_n(\pi, x)$.*

b) *The sequence $\{j_n^*\}$ converges to j_∞^* as $n \rightarrow \infty$.*

Remark 3.1 a) *The relation between the discounted and average criteria is given as follows. Fix an arbitrary state $z \in X$, and let $j_{\alpha,n} := (1 - \alpha)V_{\alpha,n}(z)$ and $h_{\alpha,n}(x) := V_{\alpha,n}(x) - V_{\alpha,n}(z)$, for $\alpha \in (0, 1)$, $n = 0, 1, \dots, \infty$. Then, following standard arguments in the literature on average cost MCPs (see, e.g., Ref. 12,13), it is possible to prove that, for each finite $n \in \mathbb{N}_0$,*

$$\lim_{t \rightarrow \infty} j_{\alpha_t, n} = j_n^*, \quad (20)$$

for any sequence $\{\alpha_t\}$ of discount factor such that $\alpha_t \nearrow 1$. In addition (see Ref. 12), the pair $(j_n^*, h_n(x))$, with $h_n(x) := \liminf_{t \rightarrow \infty} h_{\alpha_t, n}(x)$ for $n \in \mathbb{N}_0$, satisfies the ACOI, and we have

$$\sup_{\alpha \in (0,1)} \|h_{\alpha,n}\|_W < \infty, \quad n \in \mathbb{N}_0. \quad (21)$$

b) From Lemmas 3.1(a) and 3.2(b) it is easy to see that (20) and (21) also hold for $n = \infty$, i.e.,

$$\lim_{t \rightarrow \infty} j_{\alpha_t, \infty} = j_\infty^* \quad (22)$$

and

$$\sup_{\alpha \in (0,1)} \|h_{\alpha, \infty}\|_W < \infty. \quad (23)$$

Density estimation. Let $\xi_0, \xi_1, \dots, \xi_{t-1}$ be independent realizations (observed up to time $t - 1$), of r.v.'s with the unknown density ρ . We suppose that ρ satisfies Assumptions 2.3 and relation (12) or (13).

We estimate the density ρ applying a method of statistical estimation that was originally proposed in (Ref. 9) to obtain an asymptotically discounted optimal (ADO) adaptive policy for time-invariant control model (see also Ref. 10,4). This estimation scheme was slightly modified in (Ref. 3) to construct ADO adaptive policies for time-varying model M_n , $n = 0, 1, \dots, \infty$. In that paper, it is shown the existence of estimators $\rho_t(s) := \rho_t(s; \xi_0, \xi_1, \dots, \xi_{t-1})$, $s \in \mathfrak{R}^k$, satisfying the following properties: For each $t \in \mathbb{N}_0$

- i) $\rho_t \in L_q$;
- ii) ρ_t is a density function on \mathfrak{R}^k ;
- iii) $\rho_t(s) \leq \bar{\rho}(s)$ a.e.;

iv) for all $(x, a) \in \mathbb{K}$, $n = 0, 1, \dots, \infty$,

$$\int_{\mathbb{R}^k} W[G_n(x, a, s)]\rho_t(s)ds \leq \beta W(x) + b. \quad (24)$$

Furthermore, from (Ref. 9,3) we know the following:

Lemma 3.3 *Suppose that Assumption 2.3 holds. Then for some $\gamma > 0$,*

$$E \|\rho_t - \rho\|^{p'} = \mathbf{O}(t^{-\gamma}) \quad \text{as } t \rightarrow \infty, \quad (25)$$

where $1/p + 1/p' = 1$ and $\|\cdot\|$ is the pseudo-norm on the space of all densities μ on \mathbb{R}^k defined as:

$$\|\mu\| := \sup_X [W(x)]^{-1} \sup_{A(x)} \int_{\mathbb{R}^k} W[G_\infty(x, a, s)]\mu(s)ds. \quad (26)$$

For arbitrary density μ in \mathbb{R}^k , the pseudo-norm $\|\mu\|$ may be infinite. However, $\|\mu\| < \infty$ for any density function satisfying (13).

Construction of the adaptive policy. Having the estimators ρ_t of ρ , we now define an adaptive control policy as follows.

We fix an arbitrary nondecreasing sequence of discount factors $\{\hat{\alpha}_t\}$ such that $1 - \hat{\alpha}_t = \mathbf{O}(t^{-\tau})$ as $t \rightarrow \infty$, and

$$\lim_{m \rightarrow \infty} \frac{\kappa(m)}{m} = 0, \quad (27)$$

where $0 < \tau < \gamma/(3p')$ (with γ and p' as in Lemma 3.3) and $\kappa(m)$ is the number of changes of value of $\{\hat{\alpha}_t\}$ for $t = 0, 1, \dots, m$.

For each fixed t , let $V_{\hat{\alpha}_t, n}^{(\rho_t)}(\pi, x) := E_x^{(n), \pi, \rho_t} [\sum_{k=0}^{\infty} \hat{\alpha}_t^k c(x_k, a_k)]$ be the total expected $\hat{\alpha}_t$ -discounted cost for the model M_n and process (4) in which the r.v.'s ξ_0, ξ_1, \dots , have the common density ρ_t , and let $V_{\hat{\alpha}_t, n}^{(\rho_t)}(x) := \inf_{\pi \in \Pi} V_{\hat{\alpha}_t, n}^{(\rho_t)}(\pi, x)$, $x \in X$, be the corresponding value function. The sequences $h_{\hat{\alpha}_t, n}^{(\rho_t)}(\cdot)$ and $j_{\hat{\alpha}_t, n}^{(\rho_t)}$ are defined accordingly [see Remark 3.1].

Remark 3.2 Taking into account the property (24) of the estimators ρ_t , it follows from Lemma 3.1(c) that for each $t \geq 1$, $n = 0, 1, \dots, \infty$, and for any $\hat{\delta}_t > 0$, there exists a stationary policy $\hat{f}_{t,n} \in \mathcal{IF}$ such that

$$c(x, \hat{f}_{t,n}) + \hat{\alpha}_t \int_{\mathfrak{R}^k} V_{\hat{\alpha}_t, n}^{(\rho_t)}[G_n(x, \hat{f}_{t,n}, s)]\rho(s)ds \leq V_{\hat{\alpha}_t, n}^{(\rho_t)}(x) + \hat{\delta}_t, \quad \forall x \in X. \quad (28)$$

Moreover, from (17) and (18), for each $t \geq 1$ and $n = 0, 1, \dots, \infty$

$$V_{\hat{\alpha}_t, n}^{(\rho_t)}(x) \leq \frac{CW(x)}{1 - \hat{\alpha}_t} \text{ for all } x \in X. \quad (29)$$

For each $t \in \mathbb{N}$, we set $h_t := (x_0, a_0, \xi_0, \dots, x_{t-1}, a_{t-1}, \xi_{t-1}, x_t)$, the history up to time t , where $(x_m, a_m) \in \mathbb{IK}$, $\xi_m \in \mathfrak{R}^k$, $m = 0, 1, \dots, t-1$ and $x_t \in X$.

Definition 3.1 Let $\{\hat{\delta}_t\}$ be an arbitrary convergent sequence of positive numbers, and let $\hat{\delta} := \lim_{t \rightarrow \infty} \hat{\delta}_t$. In addition, for each $n = 0, 1, \dots, \infty$, let $\{\hat{f}_{t,n}\}$ be a sequence of functions in \mathcal{IF} satisfying (28). The adaptive policy $\hat{\pi}_n = \{\hat{\pi}_{t,n}\}$ is defined as $\hat{\pi}_{t,n}(h_t) = \hat{\pi}_t(h_t; \rho_t) := \hat{f}_{t,n}(x_t)$ for each $t \in \mathbb{N}$, where $\hat{\pi}_{0,n}(x)$ is any fixed action in $A(x)$.

Using similar arguments as in (Ref. 3), it is easy to see that the policy $\hat{\pi}_\infty$ is the sequence $\{\hat{\pi}_{t,\infty}\}$, where each component $\hat{\pi}_{t,\infty}$, $t \in \mathbb{N}$, can be obtained as an accumulation point of the sequence $\{\hat{f}_{t,n}(x_t)\}$ indexed by n .

Now, our main result (proved in §5) is the following:

Theorem 3.1 Suppose that Assumptions 2.1 to 2.4 hold. Then the adaptive policy $\hat{\pi}_\infty$ is $\hat{\delta}$ -average cost optimal for the model M_∞ , that is, $J_\infty(\hat{\pi}_\infty, x) \leq j_\infty^* + \hat{\delta}$ for all $x \in X$, where j_∞^* is the optimal average cost in Lemma 3.2 for the model M_∞ . In particular, if $\hat{\delta} = 0$, then the policy $\hat{\pi}_\infty$ is average cost optimal for the model M_∞ .

Remark 3.3 a) Since Assumptions 2.3 and 2.4 are stated for each finite $n \in \mathbb{N}_0$, it is well known that the adaptive policy $\hat{\pi}_n$ is $\hat{\delta}$ -average cost optimal for the model M_n (see, e.g., Ref. 4). The interesting point of Theorem 3.1 is that this result also holds for $n = \infty$.

b) It is well-known that average cost optimal stationary policies exist for the control model M_n , $n = 0, 1, \dots, \infty$, if the minimum on the right-hand side of (19) is attained for each $x \in X$. However, to ensure the existence of such minimum, typically we require rather restrictive continuity and compactness conditions on the control model M_n (see, for instance, (Ref. 14,12,15,16, 11, 13,17) in the case $n \in \mathbb{N}_0$; and see (Ref. 5) for $n = \infty$). Hence, it can happen that under the assumptions made in this paper, stationary policies for the average criterion do not exist for the control model M_n , $n = 0, 1, \dots, \infty$, with a known density ρ of the r.v.'s ξ_t , while our main result, Theorem 3.1, guarantees the existence of average cost optimal adaptive policies.

4 Example

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

$$x_{n+1} = \left(H(x_n)g_n(x_n) + G(x_n, a_n) + \xi_n \right)^+ \quad (n \in \mathbb{N}_0), \quad (30)$$

$x_0 = x$ given, with state space $X = [0, \infty)$ and actions sets $A(x) = A$ for all $x \in X$, where A is a compact subset of \mathfrak{R} . The functions H , g_n and G are continuous and such that $H(0) = 0$ and $G(0, \cdot) \equiv 0$. The sequence of r.v.'s $\{\xi_n\}$ is i.i.d.

This model represents, for example, the real time evolution of the concentration x_n of a biomass in a bioreaction, directed by a control action a_n . Such reactions are very common in depollution and in the agro-food industry (Ref. 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 (30) “tends” to a time-homogeneous system such as (3).

To assure that the system (30) has a nice stable behavior, we make the following assumption on its dynamic:

Assumption 4.1 *There exists a positive constant $\eta < 1/2$ such that*

$$\limsup_{|x| \rightarrow \infty} \sup_{i \in \mathbb{N}_0} \sup_{a \in A(x)} \frac{|H(x)g_i(x) + G(x, a)|}{|x|} = \eta.$$

See for example (Ref. 18) 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

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

The r.v.'s ξ_0, ξ_1, \dots are supposed to be i.i.d. with unknown density $\rho \in L_q(\mathfrak{R})$, bounded and continuous, and satisfying the inequality

$$\|\Delta_z \rho\|_q \leq L |z|^{1/q}, \quad (31)$$

for some given constants $L < \infty$ and $q > 1$. A sufficient condition (see Ref. 9) for (31) is the following: there exist a finite set $B \subset \mathfrak{R}$ (possibly empty) and a constant $L' \geq 0$ such that:

- i) ρ has a bounded derivative ρ' on $\mathfrak{R} \setminus B$ which belongs to $L_q(\mathfrak{R})$;
- ii) the function $|\rho'(x)|$ is nonincreasing for $x \geq L'$ and nondecreasing for $x \leq -L'$.

In addition, we assume that $E(\xi_0) \leq \frac{1}{4} - (\eta - 1)^2$ and that there exists a constant $M < \infty$ such that $\rho(s) \leq M \min\{1, 1/|s|^{1+r}\}$, for all $s \in \mathfrak{R}$. Clearly, Assumptions 2.1 and 2.2 are satisfied defining, for $x \in X$, $W(x) := C(x + \delta)^{1/2}$, where $\delta = \frac{1}{2} - \eta$ and $C = \max(1, \frac{x^*}{\delta})$.

On the other hand, it is easy to check that

$$\varphi(s) \leq 1 + \delta^{1/2} + C^{-1} |s|^{1/2} / \inf_X W(x) < \infty, \quad s \in \mathfrak{R}.$$

Thus, Assumption 2.3 is verified by taking $\bar{\rho}(s) := M \min\{1, 1/|s|^{1+r}\}$, $\forall s \in \mathfrak{R}$, with appropriate $r > 0$.

Finally, denote $\beta_0 := \eta + \frac{1}{2} < 1$, $\nu(\cdot)$ the Dirac measure concentrated at $x = 0$ and $\psi_{f,n}(x) := P(H(x)g_n(x) + G(x, f) + \xi_0 \leq 0)$ for all $f \in \mathbb{F}$, $n \in \mathbb{N}_0$ and any $x \in X$. Assumption 2.4 then holds with $p = 2$ and $\bar{\psi} := P(\xi_0 \leq 0) > 0$ (by the fact that $E(\xi_0) \leq \frac{1}{4} - (\eta - 1)^2$), and Theorem 3.1 applies.

5 Proof of Theorem 3.1

The proof consists in the adaptation of the results obtained in (Ref. 4) to the limiting model M_∞ . To that aim, we need the two following lemmas.

Lemma 5.1 *Suppose that Assumption 2.1(a) holds. Then, for every $n \in \mathbb{N}_0$ the condition (12) implies that*

a) (Ref. 9) for every $(x, a) \in \mathbb{K}$,

$$\int_{\mathbb{R}^k} W[G_n(x, a, s)]\rho(s)ds \leq \beta W(x) + b, \quad (32)$$

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

b) (Ref. 9) $\sup_{t \geq 1} E_x^{(n)\pi}[W^p(x_t)] < \infty$ and $\sup_{t \geq 1} E_x^{(n)\pi}[W(x_t)] < \infty$.

If moreover Assumption 2.2 holds, then

c) (Ref. 6) for all $(x, a) \in \mathbb{K}$,

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

and

$$\int_{\mathbb{R}^k} W[G_\infty(x, a, s)]\rho(s)ds \leq \beta W(x) + b, \quad (33)$$

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 $\pi \in \Pi$, $x \in \bar{X}$.

In the remainder, we will repeatedly use the following inequalities. For any $u \in L_W^\infty$ and any density μ that satisfies (33) (see (24)), we have

$$|u(x)| \leq \|u\|_W W(x) \quad (34)$$

and

$$\int_{\mathbb{R}^k} u[G_\infty(x, a, s)]\mu(s)ds \leq \|u\|_W [\beta W(x) + b], \quad (35)$$

for all $x \in X$ and $a \in A(x)$. The relation (34) is a consequence of the definition (14), and (35) holds because of (33) and (34).

Lemma 5.2 *Under Assumptions 2.1 to 2.4, for each $x \in X$ and $\pi \in \Pi$, we have*

$$\lim_{t \rightarrow \infty} E_x^{(\infty)\pi} \left\| V_{\hat{\alpha}_t, \infty} - V_{\hat{\alpha}_t, \infty}^{(\rho_t)} \right\|_W^{p'} = 0. \quad (36)$$

Proof. Let D be the set of densities μ on \mathfrak{R}^k satisfying the relation

$$\int_{\mathfrak{R}^k} W[G_\infty(x, a, s)] \mu(s) ds \leq \beta W(x) + b, \quad (37)$$

for each $(x, a) \in \mathbb{K}$. Observe that, from (33), $\rho \in D$ and from (24), $\rho_t \in D$, $t \in \mathbb{N}_0$.

For each $t \in \mathbb{N}_0$ and any density $\mu \in D$, we define the operator $T_{\mu, \hat{\alpha}_t}^{(\infty)} \equiv T_\mu : L_W^\infty \rightarrow L_W^\infty$ as

$$T_\mu u(x) := \inf_{A(x)} \left\{ c(x, a) + \hat{\alpha}_t \int_{\mathfrak{R}^k} u[G_\infty(x, a, s)] \mu(s) ds \right\}, \quad x \in X, \quad u \in L_W^\infty. \quad (38)$$

Note that under Assumptions 2.1 and 2.2, from Lemma 3.1(b),

$$T_\rho V_{\hat{\alpha}_t, \infty} = V_{\hat{\alpha}_t, \infty} \quad \text{and} \quad T_{\rho_t} V_{\hat{\alpha}_t, \infty}^{(\rho_t)} = V_{\hat{\alpha}_t, \infty}^{(\rho_t)}, \quad \text{for each } t \in \mathbb{N}_0. \quad (39)$$

On the other hand, for each $t \in \mathbb{N}_0$, define $\theta_t \in (\hat{\alpha}_t, 1)$ as $\theta_t := (1 + \hat{\alpha}_t)/2$, and let $W_t(x) := W(x) + d_t$ for $x \in X$, where $d_t := b(\theta_t/\hat{\alpha}_t - 1)^{-1}$. Let $L_{W_t}^\infty$ be the space of measurable functions $u : X \rightarrow \mathfrak{R}$ with the norm $\|u\|_{W_t}$, $t \in \mathbb{N}_0$, which is defined as

$$\|u\|_{W_t} := \sup_{x \in X} \frac{|u(x)|}{W_t(x)}.$$

Observe that $d_t \leq 2b/(1 - \hat{\alpha}_t)$, $t \in \mathbb{N}_0$. Hence,

$$\|u\|_{W_t} \leq \|u\|_W \leq l_t \|u\|_{W_t}, \quad t \in \mathbb{N}_0,$$

where $l_t := 1 + 2b / [(1 - \hat{\alpha}_t) \inf_{x \in X} W(x)]$. Thus, (36) will be proved if we show that for each $x \in X$ and $\pi \in \Pi$

$$l_t^{p'} E_x^{(\infty)\pi} \left\| V_{\hat{\alpha}_t, \infty} - V_{\hat{\alpha}_t, \infty}^{(\rho_t)} \right\|_{W_t}^{p'} \rightarrow 0, \text{ as } t \rightarrow \infty. \quad (40)$$

A consequence of Lemma 2 in (Ref. 19) is that, for each $t \in \mathbb{N}_0$ and $\mu \in D$, the inequality $\int_{\mathbb{R}^k} W[G_\infty(x, a, s)] \mu(s) ds \leq W(x) + b$ implies that the operator T_μ (see (38)) is a contraction with respect to the norm $\|\cdot\|_{W_t}$, with modulus θ_t , i.e.,

$$\|T_\mu v - T_\mu u\|_{W_t} \leq \theta_t \|v - u\|_{W_t} \quad \forall v, u \in L_W^\infty, t \in \mathbb{N}_0. \quad (41)$$

Now, from (41) and (39) we have, for each $t \in \mathbb{N}_0$,

$$\left\| V_{\hat{\alpha}_t, \infty} - V_{\hat{\alpha}_t, \infty}^{(\rho_t)} \right\|_{W_t} \leq \|T_\rho V_{\hat{\alpha}_t, \infty} - T_{\rho_t} V_{\hat{\alpha}_t, \infty}\|_{W_t} + \theta_t \left\| V_{\hat{\alpha}_t, \infty} - V_{\hat{\alpha}_t, \infty}^{(\rho_t)} \right\|_{W_t},$$

which implies that

$$l_t \left\| V_{\hat{\alpha}_t, \infty} - V_{\hat{\alpha}_t, \infty}^{(\rho_t)} \right\|_{W_t} \leq \frac{l_t}{1 - \theta_t} \|T_\rho V_{\hat{\alpha}_t, \infty} - T_{\rho_t} V_{\hat{\alpha}_t, \infty}\|_{W_t}. \quad (42)$$

On the other hand, from definition (26), (18) and (24), and the fact that $[W_t(\cdot)]^{-1} < [W(\cdot)]^{-1}$ for all $t \in \mathbb{N}_0$, we obtain

$$\begin{aligned} & \|T_\rho V_{\hat{\alpha}_t, \infty} - T_{\rho_t} V_{\hat{\alpha}_t, \infty}\|_{W_t} \\ & \leq \hat{\alpha}_t \sup_X [W_t(x)]^{-1} \sup_{A(x)} \int_{\mathbb{R}^k} V_{\hat{\alpha}_t, \infty}[G_\infty(x, a, s)] |\rho(s) - \rho_t(s)| ds \\ & \leq \frac{C \hat{\alpha}_t}{1 - \hat{\alpha}_t} \sup_X [W(x)]^{-1} \sup_{A(x)} \int_{\mathbb{R}^k} W[G_\infty(x, a, s)] |\rho(s) - \rho_t(s)| ds \\ & \leq \frac{C}{1 - \hat{\alpha}_t} \|\rho - \rho_t\|. \end{aligned} \quad (43)$$

Note that by the definition of $\hat{\alpha}_t$ and θ_t ,

$$\frac{1}{(1 - \theta_t)(1 - \hat{\alpha}_t)^2} = \mathbf{O}(t^{3\tau}) \text{ as } t \rightarrow \infty. \quad (44)$$

Now, combining (42), (43), (44), and using the definition of l_t , we get

$$\begin{aligned}
& l_t^{p'} \left\| V_{\hat{\alpha}_t, \infty} - V_{\hat{\alpha}_t, \infty}^{(\rho_t)} \right\|_{W_t}^{p'} \\
& \leq C^{p'} \left[\frac{1}{(1 - \theta_t)(1 - \hat{\alpha}_t)} + \frac{2b}{(1 - \theta_t)(1 - \hat{\alpha}_t)^2 \inf_X W(x)} \right]^{p'} \|\rho - \rho_t\|^{p'} \\
& = C^{p'} \mathbf{O}(t^{3p'\tau}) \|\rho - \rho_t\|^{p'} \text{ as } t \rightarrow \infty. \tag{45}
\end{aligned}$$

To conclude, we take the expectation $E_x^{(\infty)\pi}$ on both sides of (45), and observing that $E_x^{(\infty)\pi} \|\rho - \rho_t\|^{p'} = E \|\rho - \rho_t\|^{p'}$ (since ρ_t doesn't depend on x and π), we obtain (40) by virtue of Lemma 3.3 and the fact that $3\tau p' < \gamma$ [see the definition of $\hat{\alpha}_t$]. This proves Lemma 5.2. \diamond

Remark 5.1 *It is easy to prove that*

$$\lim_{t \rightarrow \infty} E_x^{(\infty)\pi} \left\| V_{\hat{\alpha}_t, \infty} - V_{\hat{\alpha}_t, \infty}^{(\rho_t)} \right\|_W W(x_t) = 0 \text{ for } x \in X, \pi \in \Pi. \tag{46}$$

Indeed, denoting $\bar{C} := \left(E_x^{(\infty)\pi} [W^p(x_t)] \right)^{1/p} < \infty$ [see Lemma 5.1(c)] and applying Hölder's inequality, we obtain

$$E_x^{(\infty)\pi} \left\| V_{\hat{\alpha}_t, \infty} - V_{\hat{\alpha}_t, \infty}^{(\rho_t)} \right\|_W W(x_t) \leq \bar{C} \left(E_x^{(\infty)\pi} \left[\left\| V_{\hat{\alpha}_t, \infty} - V_{\hat{\alpha}_t, \infty}^{(\rho_t)} \right\|_{W_t}^{p'} \right] \right)^{1/p'}.$$

Thus, letting $t \rightarrow \infty$, Lemma 5.2 yields (46).

Proof of Theorem 3.1.

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

$$\begin{aligned}
\Phi_t & := c(k_t) + \hat{\alpha}_t \int_{\mathfrak{R}^k} V_{\hat{\alpha}_t, \infty} [G_\infty(k_t, s)] \rho(s) ds - V_{\hat{\alpha}_t, \infty}(x_t) \\
& = c(k_t) + \hat{\alpha}_t E_x^{(\infty)\hat{\pi}_\infty} [V_{\hat{\alpha}_t, \infty}(x_{t+1}) \mid k_t] - V_{\hat{\alpha}_t, \infty}(x_t).
\end{aligned} \tag{47}$$

From definition of $h_{\alpha, \infty}$ and $j_{\alpha, \infty}$ (see Remark 3.1(a)), it is easy to see that

$$\Phi_t = c(k_t) + \hat{\alpha}_t E_x^{(\infty)\hat{\pi}_\infty} [h_{\hat{\alpha}_t, \infty}(x_{t+1}) \mid k_t] - j_{\hat{\alpha}_t, \infty} - h_{\hat{\alpha}_t, \infty}(x_t).$$

Now, for $m \geq k \geq 1$,

$$\begin{aligned} m^{-1} E_x^{(\infty)\hat{\pi}_\infty} \left[\sum_{t=k}^m c(k_t) - j_{\hat{\alpha}_t, \infty} \right] &= m^{-1} E_x^{(\infty)\hat{\pi}_\infty} \left[\sum_{t=k}^m \Phi_t \right] \\ &+ m^{-1} E_x^{(\infty)\hat{\pi}_\infty} \left[\sum_{t=k}^m (h_{\hat{\alpha}_t, \infty}(x_t) - \hat{\alpha}_t h_{\hat{\alpha}_t, \infty}(x_{t+1})) \right]. \end{aligned} \quad (48)$$

On the other hand, from (21), (34) and Lemma 5.1(c), we have, for $\alpha \in (0, 1)$ and a constant $C' < \infty$, $E_x^{(\infty)\hat{\pi}_\infty} [h_{\alpha, \infty}(x_t)] < C'$. Denoting $\alpha_1^*, \alpha_2^*, \dots, \alpha_{\kappa(m)}^*$, $m \geq 1$, the different values of $\hat{\alpha}_t$ for $t \leq m$ (see condition (27)), and using that $\{\hat{\alpha}_t\}$ is a nondecreasing sequence we have

$$\begin{aligned} m^{-1} E_x^{(\infty)\hat{\pi}_\infty} \left[\sum_{t=k}^m (h_{\hat{\alpha}_t, \infty}(x_t) - \hat{\alpha}_t h_{\hat{\alpha}_t, \infty}(x_{t+1})) \right] \\ &= m^{-1} E_x^{(\infty)\hat{\pi}_\infty} \left[\sum_{t=k}^m (h_{\hat{\alpha}_t, \infty}(x_t) - \hat{\alpha}_t h_{\hat{\alpha}_t, \infty}(x_t)) \right] \\ &+ m^{-1} E_x^{(\infty)\hat{\pi}_\infty} \left[\sum_{t=k}^m \hat{\alpha}_t (h_{\hat{\alpha}_t, \infty}(x_t) - h_{\hat{\alpha}_t, \infty}(x_{t+1})) \right] \\ &\leq (1 - \hat{\alpha}_k) C' + m^{-1} 2C' \sum_{i=1}^{\kappa(m)} \alpha_i^* \\ &\leq (1 - \hat{\alpha}_k) C' + 2C' \kappa(m) m^{-1} \\ &\leq (1 - \hat{\alpha}_k) C' + \mathbf{o}(1), \quad x \in X. \end{aligned} \quad (49)$$

Now, from (15) and (47) (see also (39)) we have

$$\begin{aligned} \Phi_t &= c(k_t) + \hat{\alpha}_t \int_{\mathfrak{R}^k} V_{\hat{\alpha}_t, \infty} [G_\infty(k_t, s)] \rho(s) ds \\ &- \inf_{A(x_t)} \left[c(x_t, a) + \hat{\alpha}_t \int_{\mathfrak{R}^k} V_{\hat{\alpha}_t, \infty} [G_\infty(x_t, a, s)] \rho(s) ds \right] \\ &\leq |I_1(t)| + |I_2(t)| + |I_3(t)|, \end{aligned}$$

where

$$\begin{aligned}
I_1(t) &:= \hat{\alpha}_t \int_{\mathfrak{R}^k} V_{\hat{\alpha}_t, \infty}[G_\infty(k_t, s)] \rho(s) ds - \hat{\alpha}_t \int_{\mathfrak{R}^k} V_{\hat{\alpha}_t, \infty}^{(\rho_t)}[G_\infty(k_t, s)] \rho(s) ds, \\
I_2(t) &:= \hat{\alpha}_t \int_{\mathfrak{R}^k} V_{\hat{\alpha}_t, \infty}^{(\rho_t)}[G_\infty(k_t, s)] \rho(s) ds - \hat{\alpha}_t \int_{\mathfrak{R}^k} V_{\hat{\alpha}_t, \infty}^{(\rho_t)}[G_\infty(k_t, s)] \rho_t(s) ds, \\
I_3(t) &:= c(k_t) + \hat{\alpha}_t \int_{\mathfrak{R}^k} V_{\hat{\alpha}_t, \infty}^{(\rho_t)}[G_\infty(k_t, s)] \rho_t(s) ds \\
&\quad - \inf_{A(x_t)} \left[c(x_t, a) + \hat{\alpha}_t \int_{\mathfrak{R}^k} V_{\hat{\alpha}_t, \infty}[G_\infty(x_t, a, s)] \rho(s) ds \right].
\end{aligned}$$

Using (34) and (35)

$$\begin{aligned}
|I_1(t)| &\leq \hat{\alpha}_t \int_{\mathfrak{R}^k} \left| V_{\hat{\alpha}_t, \infty}[G_\infty(k_t, s)] - V_{\hat{\alpha}_t, \infty}^{(\rho_t)}[G_\infty(k_t, s)] \right| \rho(s) ds \\
&\leq \hat{\alpha}_t \left\| V_{\hat{\alpha}_t, \infty} - V_{\hat{\alpha}_t, \infty}^{(\rho_t)} \right\|_W [\beta W(x_t) + b].
\end{aligned} \tag{50}$$

Taking the expectation $E_x^{(\infty)\hat{\pi}_\infty}$ on both sides of (50), the Lemma 5.2 and (46) yield

$$E_x^{\hat{\pi}} |I_1(t)| \rightarrow 0, \text{ as } t \rightarrow \infty. \tag{51}$$

Now, from the definition of $\hat{\alpha}_t$ and (29), $\left\| V_{\hat{\alpha}_t, \infty}^{(\rho_t)} \right\|_W = \mathbf{O}(t^\tau)$. Thus, from definition (26),

$$\begin{aligned}
|I_2(t)| &\leq \hat{\alpha}_t \int_{\mathfrak{R}^k} V_{\hat{\alpha}_t, \infty}^{(\rho_t)}[G_\infty(k_t, s)] |\rho(s) - \rho_t(s)| ds \\
&\leq \hat{\alpha}_t W(x_t) \left\| V_{\hat{\alpha}_t, \infty}^{(\rho_t)} \right\|_W \|\rho - \rho_t\|.
\end{aligned} \tag{52}$$

Hence, taking expectation and applying Hölder's inequality in (52) we get

$$\begin{aligned}
E_x^{(\infty)\hat{\pi}_\infty} |I_2(t)| &\leq \left([\mathbf{O}(t^\tau)]^{p'} E_x^{(\infty)\hat{\pi}_\infty} \|\rho - \rho_t\|^{p'} \right)^{1/p'} \\
&= \left[\mathbf{O}(t^{\tau p' - \gamma}) \right]^{1/p'} \rightarrow 0 \text{ as } t \rightarrow \infty,
\end{aligned} \tag{53}$$

due to the fact that $\tau < \gamma/3p'$ (see the definition of $\hat{\alpha}_t$).

For the term $|I_3(t)|$, from the definition of the policy $\hat{\pi}$ and combining (15) and (28) (see also (39)), adding and subtracting the term

$$\inf_{A(x_t)} \left\{ c(x_t, a) + \hat{\alpha}_t \int_{\mathfrak{R}^k} V_{\hat{\alpha}_t, \infty}^{(\rho_t)} [G_\infty(x_t, a, s)] \rho_t(s) ds \right\}$$

in $I_3(t)$, we get

$$|I_3(t)| \leq \hat{\delta}_t + \hat{\alpha}_t \sup_{A(x_t)} \left| \int_{\mathfrak{R}^k} V_{\hat{\alpha}_t, \infty}^{(\rho_t)} [G_\infty(x_t, a, s)] \rho_t(s) ds - \int_{\mathfrak{R}^k} V_{\hat{\alpha}_t, \infty} [G_\infty(x_t, a, s)] \rho(s) ds \right|.$$

The latter inequality yields

$$\begin{aligned} |I_3(t)| &\leq \hat{\delta}_t + \hat{\alpha}_t \sup_{A(x_t)} \int_{\mathfrak{R}^k} V_{\hat{\alpha}_t, \infty}^{(\rho_t)} [G_\infty(x_t, a, s)] |\rho(s) - \rho_t(s)| ds \\ &\quad + \hat{\alpha}_t \sup_{A(x_t)} \int_{\mathfrak{R}^k} \left| V_{\hat{\alpha}_t, \infty}^{(\rho_t)} [G_\infty(x_t, a, s)] - V_{\hat{\alpha}_t, \infty} [G_\infty(x_t, a, s)] \right| \rho(s) ds. \end{aligned}$$

Thus, from (26),

$$\begin{aligned} |I_3(t)| &\leq \hat{\delta}_t + \hat{\alpha}_t W(x_t) \left\| V_{\hat{\alpha}_t, \infty}^{(\rho_t)} \right\|_W \|\rho - \rho_t\| \\ &\quad + \hat{\alpha}_t \left\| V_{\hat{\alpha}_t, \infty} - V_{\hat{\alpha}_t, \infty}^{(\rho_t)} \right\|_W [\beta W(x) + b]. \end{aligned}$$

Hence, from (50), (51), (52) and (53), we get $E_x^{(\infty)\hat{\pi}_\infty} |I_3(t)| \rightarrow \hat{\delta}$, as $t \rightarrow \infty$. Therefore

$$E_x^{(\infty)\hat{\pi}_\infty} [\Phi_t] \rightarrow \hat{\delta}, \text{ as } t \rightarrow \infty. \quad (54)$$

Finally, from (48), (49) and (54), we have for any $k \geq 1$ and $n \rightarrow \infty$,

$$m^{-1} E_x^{(\infty)\hat{\pi}_\infty} \left[\sum_{t=k}^m c(k_t) - j_{\hat{\alpha}_t, \infty} \right] = (1 - \hat{\alpha}_k) C' + \mathbf{o}(1) + \hat{\delta}, \quad x \in X.$$

It follows that [from (22), the fact that $\lim_{t \rightarrow \infty} \hat{\alpha}_t = 1$, and (8)]

$$J(\hat{\pi}_\infty, x) \leq j_\infty^* + \hat{\delta}, \quad x \in X.$$

This completes the proof of the theorem. \diamond

References

- 1.- Bastin G. and Dochain D. (1990) *On-line estimation and adaptive control of bioreactors*, Elsevier, Amsterdam.
- 2.- Hilgert N., Senoussi R. and Vila J.P. *Nonparametric identification of controlled nonlinear time varying processes*, SIAM Cont. Opt. (USA), 39,3 (2000), 950–960.
- 3.- Hilgert N. and Minjárez-Sosa J.A. *Adaptive policies for time-varying stochastic systems under discounted criterion*, ZOR - Math. Methods of Oper. Res., 53,2, (2001), to appear.
- 4.- Minjárez-Sosa J.A. *Nonparametric adaptive control for discrete-time Markov processes with unbounded costs under average criterion*, Appl. Math. (Warsaw), 26,3 (1999), 267–280.
- 5.- Hilgert N. and Hernández-Lerma O. *Limiting average cost control problems in a class of time-varying stochastic systems*, Appl. Math. (Warsaw), to appear.
- 6.- Hilgert N. and Hernández-Lerma O. *Limiting optimal discounted-cost control of a class of time-varying stochastic systems*, Syst. Control Lett., 40,1 (2000), 37–42.
- 7.- Dynkin E.B. and Yushkevich A.A. (1979) *Controlled Markov Processes*, Springer-Verlag, New York.
- 8.- Rieder U. *Measurable selection theorems for optimization problems*, Manuscripta Math., 24 (1978), 115–131.
- 9.- Gordienko E.I. and Minjárez-Sosa J.A. *Adaptive control for discrete-time Markov processes with unbounded costs: discounted criterion*, Kybernetika, 34,2 (1998), 217–234.
- 10.- Gordienko E.I. and Minjárez-Sosa J.A. *Adaptive control for discrete-time Markov processes with unbounded costs: average criterion*, ZOR - Math. Methods of Oper. Res., 48,1, (1998), 37–55.
- 11.- Hernández-Lerma O. *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).

- 12.- Gordienko E.I. and Hernández-Lerma O. *Average cost Markov control processes with weighted norms: existence of canonical policies*, Appl. Math. (Warsaw), 23 (1995), 199–218.
- 13.- Hernández-Lerma O. and Lasserre J.B. (1996) *Discrete-Time Markov Control Processes: Basic Optimality Criteria*, Springer-Verlag, New York.
- 14.- Gordienko E.I. *Adaptive strategies for certain classes of controlled Markov processes*, Theory Probab. Appl., 29 (1985), 504–518.
- 15.- Gordienko E.I. and Hernández-Lerma O. *Average cost Markov control processes with weighted norms: value iteration*, Appl. Math. (Warsaw), 23 (1995), 219–237.
- 16.- Hernández-Lerma O. (1989) *Adaptive Markov Control Processes*, Springer-Verlag, New York.
- 17.- Hernández-Lerma O. and Lasserre J.B. (1999) *Further Topics on Discrete-Time Markov Control Processes*, Springer-Verlag, New York.
- 18.- Duflo M. (1997) *Random iterative models*, Springer-Verlag, Berlin.
- 19.- Van Nunen J.A.E.E. and Wessels J. *A note on dynamic programming with unbounded rewards*, Manag. Sci., 24 (1978), 576–580.