Kybernetika

Beatris A. Escobedo-Trujillo; Carmen G. Higuera-Chan

Time-varying Markov decision processes with state-action-dependent discount factors and unbounded costs

Kybernetika, Vol. 55 (2019), No. 1, 166–182

Persistent URL: `http://dml.cz/dmlcz/147711`

KYBERNETIKA — VOLUME *55* (2019), NUMBER 1, PAGES 166–182

TIME-VARYING MARKOV DECISION PROCESSES WITH STATE-ACTION-DEPENDENT DISCOUNT FACTORS AND UNBOUNDED COSTS

Beatris A. Escobedo-Trujillo and Carmen G. Higuera-Chan

In this paper we are concerned with a class of time-varying discounted Markov decision models $\mathcal{M}_n$ with unbounded costs c_n and state-action dependent discount factors. Specifically we study controlled systems whose state process evolves according to the equation $x_{n+1} = G_n(x_n, a_n, \xi_n)$, $n = 0, 1, \ldots$, with state-action dependent discount factors of the form $\alpha_n(x_n, a_n)$, where a_n and ξ_n are the control and the random disturbance at time n, respectively. Assuming that the sequences of functions $\{\alpha_n\}$,$\{c_n\}$ and $\{G_n\}$ converge, in certain sense, to α_∞, c_∞ and G_∞, our objective is to introduce a suitable control model for this class of systems and then, to show the existence of optimal policies for the limit system $\mathcal{M}_\infty$ corresponding to α_∞, c_∞ and G_∞. Finally, we illustrate our results and their applicability in a class of semi-Markov control models.

Keywords: discounted optimality, non-constant discount factor, time-varying Markov decision processes

Classification: 93E20, 90C40

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 with common distribution θ. In addition, $\{G_n\}$ is a sequence of given functions such that

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

where $1_B(\cdot)$ denotes the indicator function of the set B. Assuming possible unbounded time-varying costs c_n and time-varying state-action dependent discount factors $\alpha_n(x_n, a_n)$, where $\{\alpha_n\}$ and $\{c_n\}$ are sequences of functions such that $\alpha_n(x, a) \to \alpha_\infty(x, a)$ and

DOI: 10.14736/kyb-2019-1-0166

$c_n(x,a) \to c_\infty(x,a)$, for each (x,a), one of our main objectives is to show the existence of an $\alpha_\infty-$discount optimal policy for the general limiting system

$$x_{t+1} = G_\infty(x_t, a_t, \xi_t), \ \ t = 0, 1, \ldots. \tag{3}$$

Our approach consists of analyzing, on first step, the time-invariant system

$$x_{t+1} = G_n(x_t, a_t, \xi_t), \ \ t = 0, 1, \ldots$$

with discount rate and cost functions $\alpha_n(\cdot,\cdot)$ and $c_n(\cdot,\cdot)$ respectively, for each fixed $n \in \mathbb{N}_0$. In this case we show the existence of a stationary optimal policy f_n^*. Then, letting $n \to \infty$ we obtain the corresponding results for the limiting system (3). In particular we prove that an accumulation point f_∞^* of $\{f_n^*\}$ defines an optimal policy for the limiting model $\mathcal{M}_\infty$ corresponding to G_∞, α_∞ and c_∞.

A similar problem but with a constant discount factor has been studied in [8], which could be restrictive in some situations. In our case, for each $n \in \mathbb{N}_0 \cup \{\infty\}$, the role of the discount factor rate during the evolution of the system is as follows: at initial state x_0, the controller chooses an action a_0. Then a cost $c_n(x_0, a_0)$ is incurred, and the system moves to a new state x_1 according to a transition probability determined by G_n and the distribution of the noise ξ_0. Once the system is in state x_1 the controller selects an action a_1 and incurs a discounted cost $\alpha_n(x_0, a_0)c_n(x_1, a_1)$ and the process is repeated. For stage $m \geq 1$ the discounted cost is

$$\alpha_n(x_0, a_0)\alpha_n(x_1, a_1) \cdots \alpha_n(x_{m-1}, a_{m-1})c_n(x_m, a_m).$$

So that, the optimality of the control policies will be analyzed according to the following performance index

$$E\left[\sum_{t=0}^{\infty} \prod_{k=0}^{t-1} \alpha_n(x_k, a_k)c_n(x_t, a_t)\right] \tag{4}$$

which defines the total expected discount cost.

As an additional result, we analyze the behavior of the non-stationary control policy $\pi = \{f_n^*\}$ in the model $\mathcal{M}_\infty$, that is, the policy giving the decision $a_n = f_n(x_n)$ at each stage $n \in \mathbb{N}_0$. In this case, due to the nature of the discounted criterion, we can prove that π is asymptotically optimal in the model $\mathcal{M}_\infty$. From this perspective, the problem can be seen as an adaptive control problem in $\mathcal{M}_\infty$, when the dynamic G_∞, the discount factor and cost function, α_∞ and c_∞, are unknown. Then, $\{G_n\}, \{\alpha_n\}$ and $\{c_n\}$ can be considered as sequences of estimators (approximators) of G_∞, α_∞ and c_∞ respectively.

A similar class of adaptive control problem has been analyzed in [5, 10, 17] for systems modeled as a time-invariant stochastic difference equation as (3), where $\{\xi_t\}$ is a sequence of i.i.d. random variable (r.v.) with unknown distribution. On the other hand, control problems with nonconstant time-invariant discount factor have been studied, for instance, in [4, 16]. Indeed, in [4] is analyzed the case of randomized discounted rates, while in [16] discount factors depending on state-action and a random noise are considered. Moreover, in [9] are studied time-varying systems as (2) where $\{\xi_t\}$ are i.i.d. r.v. with unknown density, by imposing restrictive conditions on the control model and applying a complicated density estimation method.

The importance of studying this kind of systems is its application to model the real time evolution of the concentration $x_n = (x_n^{(1)}, x_n^{(2)})$ of a biomass $x_n^{(1)}$ and a substrate $x_n^{(2)}$ in a bioreaction, controlled by an action a_n. Such reaction appears, for instance, in depollution and the agrofood industry (see, e. g., [1, 11]) where $\{x_n\}$ evolve as

$$x_{n+1} = H(x_n)g_n(x_n) + K(x_n, a_n) + \xi_n, \tag{5}$$

where H, g_n and K are given functions and $\{\xi_n\}$ is a sequence of random noises. In particular $g_n(\cdot)$ is a time-varying component in the model which represents the microbial growth rate. Under appropriate condition related with temperature, pH, etc., $g_n(\cdot)$ tends to stabilize at some growth rate $g_\infty(\cdot)$. In this case the time-varying system converges to a limit system

$$x_{t+1} = H(x_t)g_\infty(x_t) + K(x_t, a_t) + \xi_t.$$

Although the main applications of the system (2) focus on this biotechnological process, in this paper we present as part of our main contributions, an interesting example of a class of semi-Markov control processes with time varying sojourn time distribution. In this case, we prove that the semi-Markov performance index can be written as a index with state-action dependent discount factor as (4). That is, the results obtained in our main model can be applied.

This paper is the first part of a work focused on Markov decision processes (MDPs) with time-varying state-action-dependent discount factors, as described above, from several aspects. In this first part we provide the theoretical foundations which include the modeling, guaranteeing the existence of optimal policy, and showing its applicability.

The rest of the paper is organized as follows. In Section 2, we introduce the Markov control models we are concern with. Next, in Section 3, it is defined the optimality criterion and presented the general assumptions. The existence of an optimal stationary policy and the construction of a policy pointwise asymptotically optimal for the control model $\mathcal{M}_\infty$ are given in Section 4 (Theorem 4.3 and Theorem 4.5). In order to illustrate our results, in Section 5, it is presented an example of a class of semi-Markov control processes whose performance index is written as a index with discount factor dependent on the state-action. Finally, we conclude in Section 6, presenting some direction on future works.

Notation. We denote $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$, $\mathbb{N}_\infty = \mathbb{N}_0 \cup \{\infty\}$. A Borel space X is a Borel subset of a complete separable metric space and we denote by $\mathcal{B}(X)$ its Borel $\sigma-$algebra. Examples of Borel spaces are countable sets with the discrete topology; Euclidean spaces with the usual topology; a compact metric space; the product (finite or countable) of Borel spaces.

Given a Borel space, we denote by $M(X)$ the family of measurable and bounded functions on X, and $L(X)$ denotes the subclass of lower semi-continuous functions in

$M(X)$. Let X and Y be Borel spaces. A stochastic kernel $Q(\cdot|\cdot)$ on X given Y is a function such that $Q(\cdot|y)$ is a probability measure on X for each $y \in Y$ and $Q(B|\cdot)$ is a measurable function on Y for each $B \in \mathcal{B}(X)$.

2. MARKOV CONTROL MODELS

For each $n \in \mathbb{N}_\infty$, we consider the time-varying control model with state-action-dependent discount factors

$$\mathcal{M}_n := (X, A, \{A(x) \mid x \in X\}, Q_n, \alpha_n, c_n) \tag{6}$$

associated to the system

$$x_{t+1} = G_n(x_t, a_t, \xi_t), \ \ t \in \mathbb{N}_0. \tag{7}$$

The state space X and the action space A are Borel subsets, while $\{A(x) \mid x \in X\}$ is a family of non-empty Borel subsets of A such that for each state $x \in X$, $A(x)$ represents the set of admissible controls when the system is in state x. We define the set

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

of admissible state-action pairs which is assumed to be a Borel subset of the Cartesian product of X and A. Further, $Q_n\,(\cdot|\cdot)$ is a stochastic kernel which denotes the transition law among the states corresponding to (7), and $\alpha_n : \mathbb{K} \to (0,1)$, $n \in \mathbb{N}_\infty$, is a measurable function representing the discount factor. Finally, the one-stage cost $c_n : \mathbb{K} \to \mathbb{R}$, $n \in \mathbb{N}_\infty$, is a non-negative measurable real-valued function, possibly unbounded.

Observe that for each $t \in \mathbb{N}_0$, $n \in \mathbb{N}_\infty$, $(x,a) \in \mathbb{K}$, and $B \in \mathcal{B}(X)$, the transition law Q_n takes the form

$$\begin{aligned} Q_n(B \mid x,a) :&= \operatorname{Prob}\left[G_n(x_t,a_t,\xi_t) \in B \mid x_t = x, a_t = a\right] \\ &= E1_B[G_n(x,a,\xi_t)] \\ &= \int_S 1_B[G_n(x,a,s)]\theta(\mathrm{d}s), \end{aligned} \tag{8}$$

where $\{\xi_t\}$ is a sequence of i.i.d. random vectors (r.v.'s) on a probability space $(\Omega, \mathcal{F}, P)$, with values in a Borel set S and a common distribution θ, that is

$$\theta(D) := P\left[\xi_t \in D\right], \ \ t \in \mathbb{N}_0, D \in \mathcal{B}(S).$$

Control policies. We define the spaces of admissible histories up to time t by $\mathbb{H}_0 := X$ and

$$\mathbb{H}_t := (\mathbb{K} \times S)^t \times X, \ t \geq 1.$$

An element $h_t \in \mathbb{H}_t$ is a vector or history, of the form $h_t = (x_0, a_0, \ldots, x_{t-1}, a_{t-1}, x_t)$, where $(x_n, a_n) \in \mathbb{K}$ for $n = 0, \ldots, t-1$ and $x_t \in X$.

Definition 2.1. A *control policy*, is a sequence $\pi = \{\pi_t\}$ of stochastic kernels π_t on A given $\mathbb{H}_t$, satisfying the constraint $\pi_t(A(x_t)|h_t) = 1 \ \ \forall h_t \in \mathbb{H}_t, \ t = 0, 1, \ldots.$

Definition 2.2. A control policy $\pi = \{\pi_t\}$ is said to be *Markovian*, if each stochastic kernel π_t satisfies $\pi_t(\cdot|h_t) = \pi_t(\cdot|x_t)$. A policy $\pi = \{\pi_t\}$ is said to be *stationary*, if there exists a measurable function f such that $\pi_t(\cdot|h_t)$ is concentrated at $f(x_t)$, for all $h_t \in \mathbb{H}_t$, $t = 0, 1, \ldots$.

We denote the set of all control policies by Π, the set of all Markov policies by Π_M, and we refer to $\mathbb{F}$ as the set of stationary policies. Finally, we identify a stationary policy π with its corresponding function f.

3. OPTIMALITY CRITERION

For a fixed $n \in \mathbb{N}_\infty$, and for $\pi \in \Pi$ and an initial state $x_0 = x \in X$, we define the total expected discounted cost with state-action-dependent discount factor for the control model $\mathcal{M}_n$ as

$$V_n(\pi, x) := E_x^{(n)\pi}\left[\sum_{t=0}^{\infty} \Gamma_t^{(n)} c_n(x_t, a_t)\right], \tag{9}$$

where

$$\Gamma_t^{(n)} := \prod\nolimits_{k=0}^{t-1} \alpha_n(x_k, a_k) \ \text{ if } \ t \geq 1, \quad \Gamma_0 = 1,$$

and $E_x^{(n),\pi}$ is the expectation operator respect to a probability measure $P_x^{(n)\pi}$ (see [3]). Hence, the optimal value function for the control model $\mathcal{M}_n, n \in \mathbb{N}_\infty$, is

$$V_n(x) := \inf_{\pi \in \Pi} V_n(\pi, x), \ x \in X. \tag{10}$$

Therefore, a policy $\pi^* \in \Pi$ is optimal for the control model $\mathcal{M}_n$, $n \in \mathbb{N}_\infty$, if

$$V_n(x) = V_n(\pi^*, x) \text{ for all } \ x \in X. \tag{11}$$

Next, we will impose the following assumption to our model.

Assumption 3.1. (a) For each $n \in \mathbb{N}_0$, the function $G_n : \mathbb{K} \times S \to X$ is continuous, and furthermore, there exists a continuous function $G_\infty : \mathbb{K} \times S \to X$ such that 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 $n \to \infty$, for each $B \in \mathcal{B}(X)$.

(b) For each $x \in X$, the set $A(x)$ is compact. Moreover, the multifunction $x \to A(x)$ is upper semi-continuous. That is, for each open set $A' \subset A$, the set $\{x \in X : A(x) \subset A'\}$ is open in X.

(c) For each $n \in \mathbb{N}_0$, the one-stage cost c_n is a nonnegative function belonging to $L(\mathbb{K})$, and there exists a function c_∞ in $L(\mathbb{K})$ such that c_n converges to c_∞. Moreover, there exists a continuous function $W : X \to [1, \infty)$ and positive constants $\bar{c} > 0$, $\beta \in (0, 1)$ and $b < \infty$ such that

$$c_n(x, a) \leq \bar{c}W(x), \quad (x, a) \in \mathbb{K},$$

and

$$\int_X W(y) Q_n(\mathrm{d}y|x, a) \leq \beta W(x) + b, \quad (x, a) \in \mathbb{K}, \ n \in \mathbb{N}_0. \tag{12}$$

(d) For each $n \in \mathbb{N}_0$, the discount factor function α_n is continuous on $\mathbb{K}$ and

$$\alpha^* := \sup_{(x,a)\in\mathbb{K}} \alpha_n(x,a) < 1. \tag{13}$$

Moreover, there exists a continuous function α_∞ such that, for each $(x,a) \in \mathbb{K}$,

$$\alpha_n(x,a) \to \alpha_\infty(x,a), \quad \text{as} \quad n \to \infty.$$

Assumption 3.1 is crucial in the development of our work since it guarantees that both the total expected discounted cost (9) and the optimal value function (10) are well defined, and moreover, the existence of minimizers. In particular, Assumption 3.1(a) is satisfied by the additive-noise systems of the form

$$G_n(x_t, a_t, \xi_t) := \bar{G}_n(x_t, a_t) + \xi_t.$$

Observe that the biotechnological processes given in (5) fall in this class.

Let $\mathbb{B}_W$ be the Banach space of all measurable functions $v : X \to \mathbb{R}$ with finite weighted norm

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

We denote by $\mathbb{L}_W$ the subspace of nonnegative lower semi-continuous (l.s.c.) functions in $\mathbb{B}_W$.

Remark 3.2. Following standard calculations we can prove (see [7]) that Assumption 3.1 yields, for each $\pi \in \Pi$, $x \in X$ and $t, n \in \mathbb{N}_0$,

(a)

$$E_x^{(n)\pi} W(x_t) \leq \beta^t W(x) + \bar{b} \leq (1+\bar{b})W(x), \quad \text{where } \bar{b} = \frac{b}{1-\beta},$$

(b)

$$V_n(\pi, x) \leq \frac{(1+\bar{b})W(x)}{1-\alpha^*},$$

(c)

$$V_n(x) \leq \frac{(1+\bar{b})W(x)}{1-\alpha^*}.$$

4. MAIN RESULTS

We summarize our main results as follows. In Theorem 4.3 we prove that under Assumption 3.1, the value function for the Markov model $\mathcal{M}_n$ converges to the value function of the Markov model $\mathcal{M}_\infty$. Moreover, there exists an optimal policy for the control model $\mathcal{M}_\infty$ which is an accumulation point of a sequence of optimal policies for the control model $\mathcal{M}_n$. Next, Theorem 4.5 establishes that the sequence $\{f_n\}$ formed by the minimizers corresponding to the model $\mathcal{M}_n$, $n \in \mathbb{N}_0$, defines a control policy which is pointwise asymptotically optimal for the control model $\mathcal{M}_\infty$. In order to state these facts precisely, we introduce some preliminary results.

For each measurable function u on X, $(x, a) \in \mathbb{K}$, and $n \in \mathbb{N}_\infty$, we define the operators

$$T_a^{(n)} u(x) := c_n(x, a) + \alpha_n(x, a) \int_X u(y) Q_n(\mathrm{d}y|x, a),$$

$$T^{(n)} u(x) := \inf_{a \in A(x)} T_a^{(n)} u(x)$$

and

$$T_f^{(n)} u(x) := c_n(x, f) + \alpha_n(x, f) \int_X u(y) Q_n(\mathrm{d}y|x, f), \ f \in \mathbb{F}.$$

Observe that T is monotone in the sense that if $v \geq u$ then $Tv \geq Tu$. In addition, from Assumption 3.1 and [20, Corollary 4.3], we have that T maps $\mathbb{L}(X)$ into itself.

The following theorem gives a characterization of the optimal cost and the existence of an optimal policy for each $n \in \mathbb{N}_0$ fixed. We can see the proof of this result in [6, Theorem 4.2.3] and [7].

Theorem 4.1. Suppose that Assumption 3.1 holds. Then, for each $n \in \mathbb{N}_0$,

(a) $V_n \in \mathbb{L}_W$ is a solution of the Optimality Equation for the control model $\mathcal{M}_n$, i. e., $T^{(n)} V_n = V_n$.

(b) There exists a stationary policy $f_n^* \in \mathbb{F}$ such that, for all $x \in X$,

$$V_n(x) = T_{f_n^*}^{(n)} V_n(x), \tag{14}$$

and f_n^* is an optimal stationary policy for the control model $\mathcal{M}_n$.

Remark 4.2. From [23], there exists $f_\infty^* \in \mathbb{F}$ such that, for each $x \in X$, $f_\infty^*(x) \in A(x)$ is an accumulation point of $\{f_n^*(x)\}$. Hence, for each $x \in X$, there exists a subsequence $\{n_i(x) = n_i\}$ of $\{n\}$ such that

$$f_{n_i}^*(x) \to f_\infty^*(x) \quad as \quad i \to \infty.$$

Theorem 4.3. Suppose that Assumption 3.1 holds. Then

(a) $V_\infty \in \mathbb{L}_W$ is a solution in $\mathbb{L}_W$ of the Optimality Equation for the control model $\mathcal{M}_\infty$, i. e., $T^{(\infty)} V_\infty = V_\infty$.

(b) As $n \to \infty$, $V_n(x) \to V_\infty(x)$, $x \in X$.

(c) For all $x \in X$,

$$V_\infty(x) = T_{f_\infty^*}^{(\infty)} V_\infty(x), \tag{15}$$

and f_∞^* is an optimal stationary policy for the control model $\mathcal{M}_\infty$.

Proof. (a) Let $\{u_k\}$ be an increasing sequence of continuous and bounded functions such that $u_k(x) \nearrow W(x)$, $x \in X$, which exists from the continuity of W. Then, for each $(x,a) \in \mathbb{K}$, $k, n \in \mathbb{N}_0$,

$$\int_X tu_k(y)Q_n(\mathrm{d}y|x,a) \leq \int_X W(y)Q_n(\mathrm{d}y|x,a) \leq \beta W(x) + b.$$

Letting $n \to \infty$, from Assumption 3.1 (a) we get

$$\int_X u_k(y)Q_\infty(\mathrm{d}y|x,a) \leq \beta W(x) + b, \ k \in \mathbb{N}_0.$$

Now, letting $k \to \infty$ we prove that inequality (12) holds for $n = \infty$, that is

$$\int_X W(y)Q_\infty(\mathrm{d}y|x,a) \leq \beta W(x) + b, \ (x,a) \in \mathbb{K}. \tag{16}$$

From the continuity of the functions G_∞ and α_∞, the proof of this part follows by applying the argument of the proof of Theorem 4.1 (a).

(b) Let

$$V_L(x) := \liminf_{n\to\infty} V_n(x) \quad \text{and} \quad V_U(x) := \limsup_{n\to\infty} V_n(x).$$

Observe that from Remark 3.2(c), $V_L, V_U \in \mathbb{B}_W$. From Theorem 4.1(a) and applying [6, Lemma 4.2.4] and [22, p. 231] we have

$$\begin{aligned} V_L(x) &= \min_{a\in A(x)} \{\liminf_{n\to\infty} c_n(x,a) + \liminf_{n\to\infty} \alpha_n(x,a) \int_X V_n(y)Q_n(\mathrm{d}y|x,a)\} \\ &\geq \min_{a\in A(x)} \{c_\infty(x,a) + \alpha_\infty(x,a) \int_X V_L(y)Q_\infty(\mathrm{d}y|x,a)\}, \ x \in X. \end{aligned}$$

Since Q_∞ is a stochastic kernel and V_L is nonnegative, we have that for every $r \in \mathbb{R}$, the set

$$\{(x,a) \in \mathbb{K} : \int_X V_L(y)Q_\infty(\mathrm{d}y|x,a) \leq r\} \in \mathcal{B}(\mathbb{K}). \tag{17}$$

This fact holds because the function $V_L'(\cdot,\cdot) := \int_X V_L(y)Q(\mathrm{d}y|\cdot,\cdot)$ is lower semicontinuous (l.s.c.). Indeed, let (x^l, a^l) be a sequence in $\mathbb{K}$ such that $(x^l, a^l) \to (x,a) \in \mathbb{K}$. Additionally, let $\{v_t\}$ be a sequence of bounded functions such that $v_t(x) \uparrow V_L(x)$ for all $x \in X$, which is possible due to V_L is a nonnegative function in $\mathbb{B}_W$. Then, for each t

$$\liminf_{l\to\infty} \int_X V_L(y)Q_\infty(\mathrm{d}y|x^l,a^l) \geq \liminf_{l\to\infty} \int_X v_t(y)Q_\infty(\mathrm{d}y|x^l,a^l) = \int_X v_t(y)Q_\infty(\mathrm{d}y|x,a).$$

Now, letting $t \to \infty$ by the Monotone Convergence Theorem we obtain that V_L' is l.s.c., so the set in (17) is closed, and so a measurable set.

Hence, from [20, Cor. 4.3], for every arbitrary $\epsilon > 0$, there exists $f_\epsilon \in \mathbb{F}$ such that

$$c_\infty(x,f_\epsilon) + \alpha_\infty(x,f_\epsilon) \int_X V_L(y)Q_\infty(\mathrm{d}y|x,f_\epsilon) \leq V_L(x) + \epsilon.$$

Iterating this inequality we obtain

$$\begin{aligned} V_L(x) &\geq E_x^{(\infty)f_\epsilon}[c(x_0,f_\epsilon)+\sum_{t=1}^{m-1}\prod_{k=0}^{t-1}\alpha_\infty(x_k,a_k)c_\infty(x_t,f_\epsilon)] \\ &+E_x^{(\infty)f_\epsilon}[\prod_{k=0}^{m-1}\alpha_\infty(x_k,f_\epsilon)V_L(x_m)]-\epsilon\sum_{t=1}^{m-1}\prod_{k=0}^{t-1}\alpha_\infty(x_k,a_k) \\ &=E_x^{(\infty)f_\epsilon}\sum_{t=0}^{m-1}\Gamma_t^{(\infty)}c_\infty(x_t,f_\epsilon)+E_x^{(\infty)f_\epsilon}[\Gamma_m^{(\infty)}V_L(x_m)]-\epsilon E_x^{(\infty)f_\epsilon}\sum_{t=0}^{m-1}\Gamma_t^{(\infty)} \\ &\geq E_x^{(\infty)f_\epsilon}\sum_{t=0}^{m-1}\Gamma_t^{(\infty)}c_\infty(x_t,f_\epsilon)-\epsilon E_x^{(\infty)f_\epsilon}\sum_{t=0}^{m-1}\Gamma_t^{(\infty)}. \end{aligned} \tag{18}$$

Observe that

$$E_x^{(\infty)f_\epsilon}\sum_{t=0}^{m-1}\Gamma_t^{(\infty)}\leq E_x^{(\infty)f_\epsilon}\sum_{t=0}^{m-1}(\alpha^*)^t\leq\frac{1}{1-\alpha^*}.$$

Then, letting $m\to\infty$ in (18) we get

$$V_L(x)\geq V_\infty(f_\epsilon,x)-\frac{\epsilon}{1-\alpha^*}.$$

As $\epsilon>0$ was arbitrary, we conclude that, for each $x\in X$ $V_L(x)\geq V_\infty(f_\epsilon,x)$, and therefore

$$V_L(x)\geq V_\infty(x),\ x\in X. \tag{19}$$

Now we proceed to prove the inequality

$$V_U(x)\leq V_\infty(x),\ x\in X.$$

First observe that for all $(x,a)\in\mathbb{K}$,

$$V_n(x)\leq c_n(x,a)+\alpha_n(x,a)\int_X V_n(y)Q_n(\mathrm{d}y|x,a),$$

which implies that

$$V_U(x)\leq c_\infty(x,a)+\alpha_\infty(x,a)\int_X V_U(y)Q_\infty(\mathrm{d}y|x,a).$$

Iterating this inequality for any arbitrary police $\pi\in\Pi$ we obtain

$$V_U(x)\leq E_x^{(\infty)\pi}\sum_{t=0}^{m-1}\Gamma_t^{(\infty)}c_\infty(x_t,a_t)+E_x^{(\infty)\pi}[\Gamma_m^{(\infty)}V_U(x_m)]. \tag{20}$$

On the other hand, from (16) and Remark 3.2(a) we have, for $\pi\in\Pi$, $x\in X$ and $t\in\mathbb{N}_0$,

$$E_x^{(\infty)\pi}W(x_t)\leq(1+\bar{b})W(x).$$

In addition, from Remark 3.2(c)

$$V_U(x) \leq \frac{(1+\bar{b})}{1-\alpha^*} W(x), \ x \in X.$$

Thus, from Assumption 3.1(d)

$$\begin{aligned} 0 &\leq E_x^{(\infty)} \Gamma_m^{(\infty)} V_U(x_m) \leq (\alpha^*)^m E_x^{(\infty)\pi} V_U(x_m) \\ &\leq \frac{(\alpha^*)^m (1+\bar{b})}{1-\alpha^*} E_x^{(\infty)\pi} W(x_m) \\ &\leq \frac{(\alpha^*)^m (1+\bar{b})^2}{1-\alpha^*} W(x). \end{aligned}$$

Therefore,

$$\lim_{m\to\infty} E_x^{(\infty)} \Gamma_m V_U(x_m) = 0, \ \pi \in \Pi, x \in X. \tag{21}$$

Hence, letting $m \to \infty$ in (20), from (21) we obtain

$$V_U(x) \leq V_\infty(\pi, x), \ \pi \in \Pi, x \in X,$$

which, in turn implies

$$V_U(x) \leq V_\infty(x), \ x \in X. \tag{22}$$

Finally, combining (19) and (22) we get

$$V_\infty(x) = V_L(x) = V_U(x),$$

that is,

$$V_n(x) \to V_\infty(x), x \in X, \text{ as } n \to \infty.$$

(c) For a fixed and arbitrary $x \in X$, from (14) and letting $n_i = i$, we have

$$V_i(x) = c_i(x, f_i) + \alpha_i(x, f_i) \int_X V_i(y) Q_i(\mathrm{d}y | x, f_i).$$

Letting $i \to \infty$ we obtain

$$V_\infty(x) \geq c_\infty(x, f_\infty) + \alpha_\infty(x, f_\infty) \int_X V_\infty(y) Q_\infty(\mathrm{d}y | x, f_\infty), \ x \in X.$$

As x is arbitrary, from part (a) of this Theorem, for every $x \in X$,

$$V_\infty(x) = c_\infty(x, f_\infty) + \alpha_\infty(x, f_\infty) \int_X V_\infty(y) Q_\infty(\mathrm{d}y | x, f_\infty).$$

Hence, standard arguments prove that the policy f_∞ is optimal for the control model $\mathcal{M}_\infty$. □

4.1. Asymptotic discount optimality

In this section we are interested in to analyse the optimality of the control policy $\hat{\pi} = \{f_n^*\}$ in the control model $\mathcal{M}_\infty$. That is, the policy giving the action $a_t = f_t(x_t)$ at each stage t for the limit control model $\mathcal{M}_\infty$. It is worth observing that $\hat{\pi} = \{f_n^*\}$ is a non-stationary policy in $\mathcal{M}_\infty$, where f_n^* satisfies $V_n = T_{f_n^*}^{(n)} V_n$, for each finite $n \in \mathbb{N}_0$. Furthermore, considering the nature of the discounted index and the convergence $V_n \to V_\infty$, on one hand we have that the discounted criterion depends on the actions taken in the early stages, and on the other hand the best information provided by such an approximation scheme is in the distant stages. These opposite facts imply that this approach does not yield optimality of $\hat{\pi}$ in the model $\mathcal{M}_\infty$, and therefore its optimality will be studied in an asymptotic sense. To ease computations, we assume that $c_n = c$ for a function $c : \mathbb{K} \to \mathbb{R}$ satisfying Assumption 3.1(c). Let $\Phi^{(\infty)} : \mathbb{K} \to \mathbb{R}$ be the function defined as

$$\Phi^{(\infty)}(x, a) = c(x, a) + \alpha_\infty(x, a) \int_X V_\infty(y) Q_\infty(\mathrm{d}y|x, a) - V_\infty(x).$$

Observe that the optimality equation for the model $\mathcal{M}_\infty$ is equivalent to

$$\min_{a \in A(x)} \Phi^{(\infty)}(x, a) = 0.$$

Furthermore, a control policy $f_\infty^* \in \mathbb{F}$ is optimal if $\Phi^{(\infty)}(x, f_\infty^*) = 0,\ x \in X$. This fact gives rise to the following definition.

Definition 4.4. A Markov policy $\pi = \{f_n\}$ is pointwise asymptotically optimal for the control model $\mathcal{M}_\infty$ if, for each $x \in X$,

$$\lim_{n \to \infty} \Phi^{(\infty)}(x, f_n(x)) = 0.$$

Hence, we can state our result as follows.

Theorem 4.5. Under Assumption 3.1, the control policy $\hat{\pi} = \{f_n^*\}$ is pointwise asymptotically optimal in the control model $\mathcal{M}_\infty$.

Proof. Let $\Phi^{(n)} : \mathbb{K} \to \mathbb{R}$ be the function defined as

$$\Phi^{(n)}(x, a) := c(x, a) + \alpha_n(x, a) \int_X V_n(y) Q_n(\mathrm{d}y|x, a) - V_n(x).$$

Observe that

$$\Phi^{(n)}(x, f_n^*) = 0.$$

Then, from Theorem 4.1(b) for each $x \in X$,

$$\begin{aligned}\Phi^{(\infty)}(x, f_n^*(x)) &= |\Phi^{(\infty)}(x, f_n^*(x)) - \Phi^{(n)}(x, f_n^*(x))| \\ &\le \sup_{a \in A(x)} |\Phi^{(\infty)}(x, a) - \Phi^{(n)}(x, a)|. \end{aligned} \tag{23}$$

Hence, to prove the theorem it is sufficient to show that, for each $x \in X$,

$$\sup_{a\in A(x)} |\Phi^{(\infty)}(x,a) - \Phi^{(n)}(x,a)| \to 0 \text{ as } n \to \infty. \tag{24}$$

Observe that

$$\begin{aligned}
\sup_{a\in A(x)} |\Phi^{(\infty)}(x,a) - \Phi^{(n)}(x,a)| = {} & \sup_{a\in A(x)} |c(x,a) + \alpha_\infty(x,a)\int_X V_\infty(y)Q_\infty(\mathrm{d}y|x,a) \\
& - V_\infty(x) - c(x,a) - \alpha_n(x,a)\int_X V_n(y)Q_n(\mathrm{d}y|x,a) + V_n(x)| \\
\le {} & \sup_{a\in A(x)} |\alpha_\infty(x,a)\int_X V_\infty(y)Q_\infty(\mathrm{d}y|x,a) - \alpha_n(x,a)\int_X V_\infty(y)Q_\infty(\mathrm{d}y|x,a)| \\
& + \sup_{a\in A(x)} |\alpha_n(x,a)\int_X V_\infty(y)Q_\infty(\mathrm{d}y|x,a) - \alpha_n(x,a)\int_X V_n(y)Q_\infty(\mathrm{d}y|x,a)| \\
& + \sup_{a\in A(x)} |\alpha_n(x,a)\int_X V_n(y)Q_\infty(\mathrm{d}y|x,a) - \alpha_n(x,a)\int_X V_n(y)Q_n(\mathrm{d}y|x,a)| \\
& + \sup_{a\in A(x)} |V_n(x) - V_\infty(x)|. \\
\le {} & \sup_{a\in A(x)} |\alpha_\infty(x,a) - \alpha_n(x,a)|\int_X V_\infty(y)Q_\infty(\mathrm{d}y|x,a) && (25) \\
& + \sup_{a\in A(x)} \alpha_n(x,a)\int_X |V_\infty(y) - V_n(y)|Q_\infty(\mathrm{d}y|x,a) && (26) \\
& + \sup_{a\in A(x)} \alpha_n(x,a)\int_X |V_n(y)||Q_\infty(\mathrm{d}y|x,a) - Q_n(dy|x,a)| && (27) \\
& + \sup_{a\in A(x)} |V_n(x) - V_\infty(x)|. && (28)
\end{aligned}$$

Thus, by Assumption 3.1 (a),(d) and Theorem 4.3(b) the terms (25-28) goes to zero as n increase, that is

$$\lim_{n\to\infty} \sup_{a\in A(x)} |\Phi^{(\infty)}(x,a) - \Phi^{(n)}(x,a)| = 0,$$

so (24) is demonstrated, and therefore the policy $\hat{\pi}$ is pointwise asymptotically optimal. □

The previous approach can be thought of as an adaptive control approach for the model $\mathcal{M}_\infty$ in the following sense. We suppose that the transition kernel $Q_\infty(\cdot|\cdot)$ and discount factor function $\alpha_\infty(\cdot,\cdot)$ from control model $\mathcal{M}_\infty$ are unknown and the sequences $\{Q_n(\cdot|\cdot)\}$ and $\{\alpha_n(\cdot,\cdot)\}$ are sequences of approximators converging to Q_∞ and α_∞, respectively, in the sense of the Assumption 3.1. Then, $\hat{\pi} = \{f_n^*\}$ is an adaptive policy which is pointwise asymptotically optimal in the model $\mathcal{M}_\infty$

A similar class of adaptive control problem has been analyzed in [5, 9, 10, 17] for systems modeled as a stochastic difference equations as (3), where $\{\xi_t\}$ is a sequence of i.i.d. r.v. with unknown distribution. In this case, by applying suitable statistical estimation techniques together with control schemes, are constructed asymptotically discounted optimal policies.

5. AN APPLICATION TO SEMI-MARKOV CONTROL MODELS

We consider a standard semi-Markov control model

$$\mathcal{M}_{SM} = (X, A, Q, H, D, d) \tag{29}$$

with the following interpretation. At time of the $n-$th decision epoch T_n, the system is in the state $x_n = x$ and the controller chooses a control $a_n = a \in A(x)$. Then the system remains in state x during a nonnegative random time δ_{n+1} with distribution $H(\cdot|x,a)$, and the following happen: 1) an immediate cost $D(x,a)$ is incurred; 2) the system jumps to a new state $x_{n+1} = y$ according to the transition law $Q(\cdot|x,a)$; and 3) a cost rate $d(x,a)$ is imposed until the transition occurs. Once the transition to state y occurs, the process is repeated.

Observe that the decision epochs are $T_n := T_{n-1} + \delta_n$ for $n \in \mathbb{N}$, and $T_0 = 0$. The random variables $\delta_{n+1} = T_{n+1} - T_n$ are called the sojourn or holding times at state x_n.

As usual the costs are continuously discounted, and therefore the one-stage cost takes the form:

$$c(x,a) := D(x,a) + d(x,a)\int_0^\infty \int_0^t \exp(-\alpha s)\mathrm{d}s H(\mathrm{d}t|x,a),\ (x,a) \in \mathbb{K}.$$

Further, it is possible to prove (see [14, 15]) that for each $(x,a) \in \mathbb{K}$,

$$c(x,a) = D(x,a) + \tau(x,a)d(x,a), \tag{30}$$

where,

$$\tau(x,a) := \frac{1 - \Delta(x,a)}{\alpha} \tag{31}$$

and

$$\Delta(x,a) := \int_0^\infty e^{-\alpha t} H(\mathrm{d}t|x,a). \tag{32}$$

The standard performance index is defined as follows. For each $\pi \in \Pi$ and $x_0 = x \in X$,

$$V(\pi,x) = E_x^\pi \left[\sum_{n=0}^\infty e^{-\alpha T_n} c(x_n, a_n) \right].$$

In the spirit of our work, we are interested in to study the following time-varying semi-Markov decision model

$$\mathcal{M}_{SM}^{(n)} = (X, A, Q_n, H_n, D, d)$$

where, $\{Q_n\}$ is a sequence of stochastic kernels on X given $\mathbb{K}$, converging (setwise) to a stochastic kernel Q_∞ on X given $\mathbb{K}$, and $\{H_n\}$ is a sequence of distribution functions converging weakly to H_∞.

Similarly to (30) – (32), we define

$$c_n(x,a) = D(x,a) + \tau_n(x,a)d(x,a)$$

where

$$\tau_n(x,a) := \frac{1-\Delta_n(x,a)}{\alpha}$$

and

$$\Delta_n(x,a) := \int_0^\infty e^{-\alpha t} H_n(\mathrm{d}t|x,a) < 1.$$

Observe that, as $n \to \infty$,

$$\Delta_n(\cdot,\cdot) \to \Delta_\infty(\cdot,\cdot)$$

and

$$\tau_n(\cdot,\cdot) \to \tau_\infty(\cdot,\cdot),$$

where

$$\Delta_\infty(x,a) = \int_0^\infty e^{-\alpha t} H_\infty(\mathrm{d}t|x,a).$$

For $n \in \mathbb{N}_\infty$, we define the total expected cost

$$V^{(n)}(\pi,x) = E_x^{(n)\pi}\left[\sum_{m=0}^{\infty} e^{\alpha T_m} c_n(x_m,a_m)\right]. \tag{33}$$

It is well-known that to analyze semi-Markov control models it is sufficient to consider the Markov policies Π_M (see, e. g., [19]). Hence, if we restrict ourselves to the Markovian policies, we will prove that the semi-Markov index (33) can be expressed as a discounted index with state-action dependent discount factor (see (9)). Hence, provided that Assumption 3.1 holds, the results in Theorem 4.3 are applicable. Specifically we have the following result:

Proposition 5.1. For each $n \in \mathbb{N}_\infty$ and $\pi \in \Pi_M$,

$$V^{(n)}(\pi,x) = E_x^{(n)\pi} \sum_{m=0}^{\infty} \Gamma_m^{(n)} c_n(x_m,a_m), \tag{34}$$

where $\Gamma_m^{(n)} := \prod_{k=0}^{m-1} \Delta_n(x_k,a_k)$ and $\Gamma_0^{(n)} = 1$.

Proof. Let $\pi = \{\pi_n\}$ be a Markov policy. We proceed to analyze each term of the sum on the right side of equation (33). The first one is for $m = 0$,

$$\begin{aligned} E_x^{(n)\pi} c_n(x_0, a_0) &= \int_A c_n(x, a)\pi_0(\mathrm{d}a|x_0) \\ &= \int_A \Gamma_0^{(n)} c_n(x_0, a)\pi_0(\mathrm{d}a|x_0) \\ &= E_x^{(n)\pi}\Gamma_0^{(n)} c_n(x_0, a_0). \end{aligned}$$

Now, for $m = 1$, using that $\pi \in \Pi_M$ we have

$$\begin{aligned} &E_x^{(n)\pi} e^{-\alpha T_1} c_n(x_1, a_1) \\ &= \int_A \int_X \int_A \int_0^\infty e^{-\alpha\delta_1} c_n(x_1, a_1) H_n(\mathrm{d}\delta_1|x_0, a_0)\pi_1(\mathrm{d}a_1|x_0, a_0, \delta_1, x_1) Q_n(\mathrm{d}x_1|x_0, a_0)\pi_0(\mathrm{d}a_0|x_0) \\ &= \int_A \int_0^\infty e^{-\alpha\delta_1} H_n(\mathrm{d}\delta_1|x_0, a_0) \int_X \int_A c_n(x_1, a_1)\pi_1(\mathrm{d}a_1|x_1, a_0, x_0) Q_n(\mathrm{d}x_1|x_0, a_0)\pi_0(\mathrm{d}a_0|x_0) \\ &= \int_A \Delta_n(x_0, a_0) \int_X \int_A c_n(x_1, a_1)\pi_1(\mathrm{d}a_1|x_1, a_0, x_0) Q_n(\mathrm{d}x_1|x_0, a_0)\pi_0(\mathrm{d}a_0|x_0) \\ &= \int_A \int_X \int_A \Delta_n(x_0, a_0) c_n(x_1, a_1)\pi_1(\mathrm{d}a_1|x_1) Q_n(\mathrm{d}x_1|x_0, a_0)\pi_0(\mathrm{d}a_0|x_0) \\ &= E_x^{(n)\pi} \Delta_n(x_0, a_0) c_n(x_1, a_1) \\ &= E_x^{(n)\pi} \Gamma_1^{(n)} c_n(x_1, a_1). \end{aligned}$$

Similarly we can prove for $m = k$,

$$E_x^{(n)\pi} e^{-\alpha(T_k)} c_n(x_k, a_k) = E_x^{(n)\pi}\Gamma_k^{(n)} c_n(x_k, a_k).$$

Thus using an inductive process over m, each term of $V^{(n)}$ can be written in terms of $\Gamma_m^{(n)}$, which proves (34). □

Remark 5.2. Taking into account the previous section, and Proposition 5.1, we can analyze the case of semi-Markov control models with unknown holding time distribution. Indeed, if we assume that the holding time distribution H_∞ is unknown, $\{H_n\}$ can be considered as a sequence of estimators of H_∞. This class of problems has been studied in [13, 14] considering that the distribution H_∞ has a density independent of the state-action pair.

6. CONCLUDING REMARKS

The paper is the first part of a project whose objective is to study time-varying MDPs with state-action dependent discount factors. Specifically, in this paper we have introduced the main elements to analyze this class of control systems from a theoretical point of view. Indeed, under suitable assumptions on the control model $\mathcal{M}_n$, $n < \infty$, we have proved the existence of optimal policies for the limit model $\mathcal{M}_\infty$. Such assumptions are essentially continuity and $W-$boundedness conditions on the cost functions together with a convergence property of the dynamics. For the sake of generality, our model is defined on Borel spaces and it is assumed possibly unbounded one-stage costs. This fact allows us to consider examples as general as the LQ-systems (linear systems/quadratic costs) where the one-stage cost is unbounded, some class of inventory and queueing models, biotechnological processes as in (5), as well as the corresponding partially observable systems whose analysis, under the standard approach, leads us to define an equivalent control problem where the state space is a set of probability measures, which in turn is a Borel space. Furthermore, as discussed in Sections 4 and 5, our results have important applications to two widely studied problems in this general context, namely, the adaptive control problem and the control problem for semi-Markov processes.

The second part of the project, which is a future work in progress of the authors, consists of developing approximation algorithms for optimal policies and value functions. There are well-known and powerful methods that address this problem but in the scenario of constant discount factors. Among these are, for instance, approximate dynamic programming (see, e. g., [2, 18, 21] and references therein) and the analysis by means of Turnpike Theorems (see, e. g., [12, 24]). In the context of our paper, that is assuming time-varying state-actions dependent discount factors, the problem remains open.

ACKNOWLEDGEMENT

Work supported partially by Consejo Nacional de Ciencia y Tecnología (CONACyT, México) under grant CB2015/254306.

(Received March 13, 2018)

REFERENCES

[1] G. Bastin and D. Dochain: On-line Estimation and Adaptive Control of Bioreactors. Elsevier, Amsterdam 2014.

[2] D. P. Bertsekas: Approximate policy iteration: a survey and some new methods. J. Control Theory Appl. *9* (2011), 310–335. DOI:10.1007/s11768-011-1005-3

[3] E. B. Dynkin and A. A. Yushkevich: Controlled Markov Processes. Springer-Verlag, New York 1979. DOI:10.1007/978-1-4615-6746-2

[4] J. González-Hernández, R. R. López-Martínez, and J. A. Minjárez-Sosa: Approximation, estimation and control of stochastic systems under a randomized discounted cost criterion. Kybernetika *45* (2009), 737–754.

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

[6] O. Hernández-Lerma and J. B. Lasseerre: Discrete-Time Markov Control Processes: Basic Optimality Criteria. Springer, New York 1996. DOI:10.1007/978-1-4612-0729-0

[7] O. Hernández-Lerma and J. B. Lasserre: Further Topics on Discrete-time Markov Control Processes. Springer-Verlag, New York 1999. DOI:10.1007/978-1-4612-0561-6

[8] O. Hernández-Lerma and N. Hilgert: Limiting optimal discounted-cost control of a class of time-varying stochastic systems. Syst. Control Lett. *40* (2000), 1, 37–42. DOI:10.1016/s0167-6911(99)00121-8

[9] N. Hilgert and J. A. Minjárez-Sosa: Adaptive policies for time-varying stochastic systems under discounted criterion. Math. Meth. Oper. Res. *54* (2001), 3, 491–505. DOI:10.1007/s001860100170

[10] N. Hilgert and J. A. Minjárez-Sosa: Adaptive control of stochastic systems with unknown disturbance distribution: discounted criteria. Math. Meth. Oper. Res. *63* (2006), 443–460. DOI:10.1007/s00186-005-0024-6

[11] N. Hilgert, R. Senoussi, and J. P. Vila: Nonparametric estimation of time-varying autoregressive nonlinear processes. C. R. Acad. Sci. Paris Série *1* 1996), 232, 1085–1090. DOI:10.1109/.2001.980647

[12] M. E. Lewis and A. Paul: Uniform turnpike theorems for finite Markov decision processes. Math. Oper. Res.. DOI:10.1287/moor.2017.0912

[13] F. Luque-Vásquez and J. A. Minjárez-Sosa: Semi-Markov control processes with unknown holding times distribution under a discounted criterion. Math. Meth. Oper. Res. *61* (2005), 455–468. DOI:10.1007/s001860400406

[14] F. Luque-Vásquez, J. A. Minjárez-Sosa, and L. C. Rosas-Rosas: Semi-Markov control processes with partially known holding times distribution: Discounted and average criteria. Acta Appl. Math. *114* (2011), 3, 135–156. DOI:10.1007/s10440-011-9605-y

[15] F. Luque-Vásquez, J. A. Minjárez-Sosa, and L. C. Rosas-Rosas: Semi-Markov control processes with unknown holding times distribution under an average criterion cost. Appl. Math. Optim. Theory Appl. *61* (2010), 3, 317–336. DOI:10.1007/s00245-009-9086-9

[16] J. A. Minjárez-Sosa: Markov control models with unknown random state-action-dependent discount factors. TOP *23* (2015), 743–772. DOI:10.1007/s11750-015-0360-5

[17] J. A. Minjárez-Sosa: Approximation and estimation in Markov control processes under discounted criterion. Kybernetika *40* (2004), 6, 681–690.

[18] W. B. Powell: Approximate Dynamic Programming. Solving the Curse of Dimensionality John Wiley and Sons Inc, 2007. DOI:10.1002/9780470182963

[19] M. L. Puterman: Markov Decision Processes. Discrete Stochastic Dynamic Programming. John Wiley and Sons 1994. DOI:10.1002/9780470316887

[20] U. Rieder: Measurable selection theorems for optimization problems. Manuscripta Math. *24* (1978), 115–131. DOI:10.1007/bf01168566

[21] M. T. Robles-Alcaráz, O. Vega-Amaya, and J. A. Minjárez-Sosa: Estimate and approximate policy iteration algorithm for discounted Markov decision models with bounded costs and Borel spaces. Risk Decision Analysis *6* (2017), 2, 79-95. DOI:10.3233/rda-160116

[22] H. L. Royden: Real Analysis. Prentice Hall 1968.

[23] M. Schäl: Conditions for optimality and for the limit on n-stage optimal policies to be optimal. Z. Wahrs. Verw. Gerb. *32* (1975), 179–196. DOI:10.1007/bf00532612

[24] J. F. Shapiro: Turnpike planning horizon for a markovian decision model. Magnament Sci. *14* (1968), 292–300. DOI:10.1287/mnsc.14.5.292

Beatris A. Escobedo-Trujillo, Engineering Faculty, Universidad Veracruzana, Coatzacoalcos, Ver. 96538. México.
e-mail: bescobedo@uv.mx

Carmen G. Higuera-Chan, Departamento de Matemáticas, Universidad de Sonora. Rosales s/n, Col. Centro, 83000 Hermosillo, Sonora. México.
e-mail: carhiguera@mat.uson.mx