Kybernetika

Rocio Ilhuicatzi-Roldán; Hugo Cruz-Suárez; Selene Chávez-Rodríguez

Markov decision processes with time-varying discount factors and random horizon

Kybernetika, Vol. 53 (2017), No. 1, 82–98

Persistent URL: `http://dml.cz/dmlcz/146709`

KYBERNETIKA — VOLUME *53* (2017), NUMBER 1, PAGES 82–98

MARKOV DECISION PROCESSES WITH TIME-VARYING DISCOUNT FACTORS AND RANDOM HORIZON

Rocio Ilhuicatzi-Roldán, Hugo Cruz-Suárez
and Selene Chávez-Rodríguez

This paper is related to Markov Decision Processes. The optimal control problem is to minimize the expected total discounted cost, with a non-constant discount factor. The discount factor is time-varying and it could depend on the state and the action. Furthermore, it is considered that the horizon of the optimization problem is given by a discrete random variable, that is, a random horizon is assumed. Under general conditions on Markov control model, using the dynamic programming approach, an optimality equation for both cases is obtained, namely, finite support and infinite support of the random horizon. The obtained results are illustrated by two examples, one of them related to optimal replacement.

Keywords: Markov decision process, dynamic programming, varying discount factor, random horizon

Classification: 93E20, 90C40, 90C39

1. INTRODUCTION

This paper deals with Markov Decision Processes (MDPs). MDPs are used to model dynamic systems that change over time under the presence of uncertainty. The systems are observed by a controller at discrete time stages, thus a sequence of actions is determined, which is known as a policy. To evaluate the quality of each policy a performance criterion or objective function is considered. In this document the expected total discounted cost is considered, where the discount factor is time-varying, which could depend on the state and the action. Furthermore, it is assumed that the development of the process can be interrupted by external factors to the system, that is, a horizon with uncertainty is considered.

The motivation to study discounted criterion comes from financial and economic aspects. The discount factor is applied to model depreciation of money with respect to time, which depends on real circumstances of the interest rate. In these cases, it is necessary to adjust the value of the discount factor according to the market situations. Consequently, considering a fixed discount factor is not realistic. In fact, attention has been paid to the discounted models with non-constant discount factor. In MDPs literature there are several works in this direction which present generalizations of the

DOI: 10.14736/kyb-2017-1-0082

discount factor and they are studied under different methodologies. To mention some of them, with multiple discount factors: [1, 5] and [6]; dependent on the state: [13, 20, 21, 22, 23, 24, 25, 26] and [27]; dependent on the state and the action: [15, 17] and [19], with a varying discount factor: [4, 7] and [13], and with a randomized discount factor: [8, 9, 10, 11] and [12].

This document presents a study of the optimal control problem with factors varying in time, which are dependent on the state and the action. As mentioned earlier in the previous paragraph, there are several works with these considerations, however, they do not consider a random horizon in the performance criteria. In this work, a discrete random variable is used to model the horizon of the process. Furthermore, it is assumed that the random horizon is independent of the sequence of state-action pairs generated in each stage (see Assumption 3.1). Under this hypothesis in [3], we study the optimal control problem with total expected cost as performance criterion, which has been applied to optimal replacement problems ([16]). Considering an independent random horizon of the control process can be used to represent various situations, for example, in finance, where there may be a drastic change of the investor's plan in the future with certain probability (see [2]).

The objective of this work is to study the optimal control problem on Borel spaces with a varying discount factor and a random planning horizon. Furthermore, one of the objectives of this paper is to present examples with an explicit solution, which illustrate the theory of this work. In the reviewed literature only two examples with an explicit solution were found (see [20] and [24]). In short, under certain conditions in the control model, the main objectives of this paper are:

a) Establish a functional equation that characterizes the optimal solution of the control problem using the dynamic programming approach.

b) Guarantee the existence of stationary optimal policies.

c) Provide examples in which a) and b) are illustrated.

The document is organized as follows: MDPs basic theory is presented in the second section. Later, in the third section the problem with a varying discount factor and a random horizon is described in detail. Afterward, in the fourth section, an analysis of the control problem is presented via dynamic programming approach. Finally, in the fifth section, the developed theory is illustrated by two examples, one of them relative to optimal replacement.

Notation and Terminology: The following notation is used throughout this document. Let Z be a Borel space, that is, a Borel subset of a complete and separable metric space, its Borel σ-algebra is denoted by $\mathcal{B}(Z)$. The indicator function of a set C is denoted by I_C. The set of real numbers is indicated with the letter $\mathbb{R}$.

2. PRELIMINARIES

Consider a Markov Decision Model conformed by the following components:

$$\mathcal{M} := (X, A, \{A(x) \mid x \in X\}, Q, \{\alpha_t(\cdot) \mid t \in \{0, 1, \ldots\}\}, c(\cdot), \tau).$$

The first component, X, is called the *state space.* The second component, A, is denominated the *action space.* In this document X and A are Borel spaces. $\{A(x) \mid x \in X\}$ is a family of nonempty subsets $A(x)$ of A, $A(x)$ denotes the set of *feasible actions* (controls) in the state $x \in X$. Then, the set of feasible state-action pairs defined as $\mathbb{K} := \{(x,a) \mid x \in X, a \in A(x)\}$, is assumed to be a measurable subset of $X \times A$. The following component is the *transition law* Q, it is a stochastic kernel on X given $\mathbb{K}$. For each $t \in \{0,1,\ldots\}$, $\alpha_t : \mathbb{K} \to (0,1]$ is a measurable function, which represents a *discount factor* applied at the cost c in the epoch t ($\alpha_0 := 1$). The measurable function $c : \mathbb{K} \to \mathbb{R}$ denotes the *cost per stage* (or one-stage cost) function. Finally, τ is a random variable defined on the probability space $(\Omega', \mathcal{G}, P)$, which represents a *random horizon* of the problem. It is assumed that τ is a discrete random variable with probability mass function given by

$$\rho_t := P(\tau = t), \quad t = 1, 2, 3, \ldots, T,$$

where T is a positive integer or $T = \infty$.

A Markov Decision Process (MDP) evolves as follows: at the initial decision epoch, the system occupies state $x_0 = x \in X$ and a decision maker (or controller) chooses an action $a_0 = a \in A(x)$. Then, a cost $c(x_0, a_0)$ is incurred and the system jumps to a state x_1 according to the transition law $Q(\cdot \mid x, a)$. Immediately after the jump occurs, the controller selects an action $a_1 \in A(x_1)$ and incurs a discounted cost $\alpha_0(x_0, a_0)\alpha_1(x_1, a_1)$ $c(x_1, a_1)$. Afterward, the system moves to a state x_2 and the process is repeated. Thus, for each $n \geq 1$ an admissible history h_n of a MDP up to the nth transition is obtained, $h_n = (x_0, a_0, \ldots, x_{n-1}, a_{n-1}, x_n)$, with $(x_k, a_k) \in \mathbb{K}$ for $k = 0, 1, \ldots, n-1$, and $x_n \in X$. Let $\mathbb{H}_n, n = 0, 1, \ldots,$ denote the set of all admissible histories of the system up to the nth transition. Thus, a *control policy* $\pi = \{\pi_n\}$ is a sequence of stochastic kernels π_n on A given $\mathbb{H}_n$, satisfying the constraint: $\pi_n(A\,(x_n)\,|h_n) = 1$, for each $h_n \in \mathbb{H}_n, n = 0, 1 \ldots$. The collection of all policies is denoted by Π. Define $\mathbb{F}$ as the set of all measurable functions $f : X \to A$ such that $f\,(x) \in A\,(x)$ for each $x \in X$. Then, a *Markov policy* is a sequence $\{f_t\}$ such that $f_t \in \mathbb{F}$, for $t = 0, 1, \ldots$. In particular, a Markov policy $\pi = \{f_t\}$ is said to be stationary if f_t is independent of t, i.e. $f_t = f \in \mathbb{F}$, for all $t = 0, 1, \ldots$, in this case, f_t is denoted by f and refers to $\mathbb{F}$ as the set of *stationary policies.*

In many cases, the evolution of a Markov control process is specified by a discrete time or difference equation of the form

$$x_{t+1} = F(x_t, a_t, \xi_t), \quad t = 0, 1, 2, \ldots,$$

with $x_0 \in X$ given, where $\{\xi_t\}$ is a sequence of independent and identically distributed random variables with values in a Borel space S and a common distribution μ, independent of the initial state x_0. In this case, the transition law Q is given by

$$Q(B|x,a) = \int_S I_B(F(x,a,s))\mu(\mathrm{d}s), B \in \mathcal{B}(X),$$

$(x, a) \in \mathbb{K}$.

Let $(\Omega, \mathcal{F})$ be the measurable space consisting of the canonical sample space $\Omega := (X \times A)^\infty$ and the corresponding product σ-algebra $\mathcal{F}$. The elements of Ω are sequences of the form $\omega = (x_0, a_0, x_1, a_1, \ldots)$ with $x_t \in X$ and $a_t \in A$ for all $t = 0, 1, 2, \ldots$. Let

$\pi = \{\pi_t\}$ be an arbitrary policy and δ be an arbitrary probability measure on X called the initial distribution. Then, by the theorem of Ionescu–Tulcea (see [14]), there is a unique probability measure P_δ^π on $(\Omega, \mathcal{F})$ which is supported on $\mathbb{H}_\infty$, i. e., $P_\delta^\pi(\mathbb{H}_\infty) = 1$. The stochastic process $(\Omega, \mathcal{F}, P_\delta^\pi, \{x_t\})$ is called a discrete-time *Markov control process* or a *Markov decision process.*

The expectation operator with respect to P_δ^π is denoted by E_δ^π. If δ is concentrated at the initial state $x \in X$, then P_δ^π and E_δ^π are written as P_x^π and E_x^π, respectively.

3. STATEMENT OF THE PROBLEM WITH VARYING DISCOUNT FACTOR AND RANDOM HORIZON

The objective in this section is to introduce the optimal control problem associated to the Markov decision model $\mathcal{M}$. Consider for $\pi \in \Pi$ and $x \in X$ the following performance criterion:

$$v^\tau(\pi, x) := E\left[c(x_0, a_0) + \sum_{t=1}^{\tau} \prod_{k=0}^{t-1} \alpha_k(x_k, a_k) c(x_t, a_t)\right], \tag{1}$$

where E denotes the expected value with respect to the joint distribution of the process $\{(x_t, a_t) : t \geq 0\}$ and the random variable τ. In this document, the performance criterion (1) will be called *total expected discounted cost with time-varying discount factors and random horizon*, in short, DRH.

Assumption 3.1. Throughout the paper it is assumed that for each $x \in X$ and $\pi \in \Pi$, the induced process $\{(x_t, a_t)\}$ is independent of the random variable τ.

Consider the performance criterion (1), then the *optimal control problem* consists of determining a policy π^*, such that

$$v^\tau(\pi^*, x) = \inf_{\pi \in \Pi} v^\tau(\pi, x),$$

$x \in X$, and π^* will be called an *optimal policy.* The function V defined by

$$V(x) = \inf_{\pi \in \Pi} v^\tau(\pi, x),$$

$x \in X$, will be called the *optimal value function.*

Now, some changes will be applied in the objective function (1), in order to have a suitable version that allows us to use the technique of dynamic programming. Then, using basic properties of conditional expectation and independence (see Assumption 3.1), it yields

$$\begin{aligned}
v^{\tau}(\pi,x) &= E\left[E\left[c(x_0,a_0)+\sum_{t=1}^{\tau}\prod_{k=0}^{t-1}\alpha_k(x_k,a_k)c(x_t,a_t)\middle|\tau\right]\right]\\
&= \sum_{n=1}^{T}E_x^{\pi}\left[c(x_0,a_0)+\sum_{t=1}^{n}\prod_{k=0}^{t-1}\alpha_k(x_k,a_k)c(x_t,a_t)\right]\rho_n\\
&= c(x_0,a_0)+\sum_{t=1}^{T}\sum_{n=t}^{T}E_x^{\pi}\left[\prod_{k=0}^{t-1}\alpha_k(x_k,a_k)c(x_t,a_t)\right]\rho_n\\
&= E_x^{\pi}\left[c(x_0,a_0)+\sum_{t=1}^{T}\prod_{k=0}^{t-1}\alpha_k(x_k,a_k)c(x_t,a_t)P_t\right], \qquad (2)
\end{aligned}$$

where $P_t := P(\tau \geq t), \quad t=1,\ldots,T.$

Remark 3.2. a) Note that if the distribution of the random horizon τ has a finite or infinite support, the optimization problem with random horizon is considered as a problem with a finite or infinite horizon, respectively.

b) If τ is concentrated on T, the objective function (2) is simplified to the following expression:

$$v^{\tau}(\pi,x)=E_x^{\pi}\left[c(x_0,a_0)+\sum_{t=1}^{T}\prod_{k=0}^{t-1}\alpha_k(x_k,a_k)c(x_t,a_t)\right]. \qquad (3)$$

Furthermore, if the discount factor $\alpha_k(x,a)=\alpha\in(0,1)$, for each $(x,a)\in\mathbb{K}$ in (3), then the objective function is the usual discounted cost criteria, see for instance [14] and [18].

Let $\widehat{\alpha}_0 := P_0 = 1$ and $\widehat{\alpha}_k := \frac{P_k}{P_{k-1}}$, for $k=1,2,\ldots,T$. The factors $\{\widehat{\alpha}_k\}$ can be considered as the following conditional probability: $\widehat{\alpha}_k = P(\tau \geq k+1 \mid \tau \geq k)$. Furthermore, for each $t \geq 1$, P_t can be written in the following way:

$$P_t=\prod_{k=0}^{t-1}\widehat{\alpha}_k. \qquad (4)$$

Then for each $x\in X$ and $\pi\in\Pi$, by (4), it is verified that

$$\begin{aligned}
v^{\tau}(\pi,x) &= E_x^{\pi}\left[c(x_0,a_0)+\sum_{t=1}^{T}\prod_{k=0}^{t-1}\widehat{\alpha}_k\alpha_k(x_k,a_k)c(x_t,a_t)\right]\\
&= E_x^{\pi}\left[c(x_0,a_0)+\sum_{t=1}^{T}\prod_{k=0}^{t-1}\widetilde{\alpha}_k(x_k,a_k)c(x_t,a_t)\right], \qquad (5)
\end{aligned}$$

where $\widetilde{\alpha}_k(x_k,a_k) := \widehat{\alpha}_k\alpha_k(x_k,a_k), k=0,1,2,\ldots.$

The following assumptions will be applied in Section 4 to validate the dynamic programming approach for DRH. The first block of assumptions will be used to ensure the existence of minimizers of the dynamic programming equation.

Assumption 3.3. (a) The set-valued mapping $x \mapsto A(x)$ is upper semicontinuous.

(b) The one-stage cost c is lower semicontinuous (l.s.c.), non-negative and inf-compact on $\mathbb{K}$.

(c) Q is strongly continuous.

(d) The discount functions $\widetilde{\alpha}_t$, $t = 0, 1, 2, \ldots$ are l.s.c.

The following assumption is necessary to guarantee the finiteness property of the optimal value function when $T = +\infty$.

Assumption 3.4. There exists a policy $\pi \in \Pi$ such that $v^\tau(\pi, x) < \infty$ for each $x \in X$.

Remark 3.5. In the literature, there are several works using an analogous criterion to (5), see for instance [17, 20, 22, 24] and [25]. However, in these references, it is assumed that the discount factor is uniformly bounded in (0,1), i. e. there exists $\beta \in (0, 1)$ such that

$$\sup_{(x,a)\in\mathbb{K}} \alpha(x, a) \le \beta. \tag{6}$$

Observe that in Assumption 3.3 and Assumption 3.4 this condition is not considered. Then, it is possible to consider the undiscounted case. This case is important in real situations, for example, suppose that the random horizon represents the working life of a machine (or electric equipment) and the cost function is equal to one. In this case the objective function consists in minimizing the cost function over all up-time of the machine.

4. DYNAMIC PROGRAMMING APPROACH

In this section, it will be presented an analysis of the optimal control problem via the dynamic programming approach. It is important to clarify that the procedure applied in this section is motivated by the semicontinuous MDPs dynamic programming approach (see, for instance, [14]). The novelty in this document is that the discount factor is time-varying and it could depend on the state and the action (see (5)), then it is necessary to present an adequate version of the dynamic programming equation and validate it. Firstly, it will be presented the case $T < \infty$.

Theorem 4.1. Suppose that Assumption 3.3 holds and T is a positive integer. Define for each $x \in X$ and $t = T, T-1, \ldots, 0$, the following measurable functions:

$$J_t(x) := \min_{a\in A(x)} \left[c(x, a) + \widetilde{\alpha}_t(x, a) \int_X J_{t+1}(y) Q(\mathrm{d}y \mid x, a) \right], \tag{7}$$

and $J_{T+1}(x) := 0$, $x \in X$. Then for each $t = 0, 1, \ldots, T$, there exists $f_t \in \mathbb{F}$ such that f_t attains the minimum in (7) for all $x \in X$ and $\pi^* = \{f_t\}$ is the optimal policy. Furthermore, the optimal value function is given by $V(x) = v^\tau(\pi^*, x) = J_0(x)$, $x \in X$.

Proof. Firstly observe that, under Assumption 3.3, for each $t = 0, 1, \ldots, T$, there exists $f_t \in \mathbb{F}$ such that f_t attains the minimum in (7) due to Theorem 3.3.5 in [14]. Then, it is simply necessary to prove that the optimal value function is J_0. To this end, define the cost from time t onwards when the policy π is used and $x_t = x$, as follows:

$$C_t(\pi, x) := E^{\pi}\left[c(x_t, a_t) + \sum_{j=t+1}^{T} \prod_{k=t}^{j-1} \widetilde{\alpha}_k(x_k, a_k) c(x_j, a_j) \middle| x_t = x \right],$$

for $t = 0, 1, \ldots, T$, and $C_{T+1}(\pi, x) := 0$.

It will be proved that for each $\pi \in \Pi$ and $x \in X$ the following inequality holds:

$$C_t(\pi, x) \geq J_t(x), \tag{8}$$

for $t = 0, 1, \ldots, T$. Observe that if $\pi = \pi^*$, (8) holds with equality. Furthermore, if $t = 0$, the following identities are hold

$$\begin{aligned} J_0(x) &= C_0(\pi^*, x) \\ &= v^{\tau}(\pi^*, x). \end{aligned}$$

The proof of (8) is for backward induction. Suppose that for some $t = T, T-1, \ldots, 0$,

$$C_{t+1}(\pi, x) \geq J_{t+1}(x), x \in X. \tag{9}$$

Then

$$C_t(\pi, x) = E^{\pi}\left[c(x_t, a_t) + \sum_{j=t+1}^{T} \prod_{k=t}^{j-1} \widetilde{\alpha}_k(x_k, a_k) c(x_j, a_j) \middle| x_t = x \right]$$

$$\begin{aligned} &= E^{\pi}\left[c(x_t, a_t) + \widetilde{\alpha}_t(x_t, a_t) \left[c(x_{t+1}, a_{t+1}) + \sum_{j=t+2}^{T} \prod_{k=t+1}^{j-1} \widetilde{\alpha}_k(x_k, a_k) c(x_j, a_j) \middle| x_t = x \right]\right] \\ &= \int_A \left[c(x, a) + \widetilde{\alpha}_t(x, a) \int_X C_{t+1}(\pi, y) Q(\mathrm{d}y \mid x, a) \right] \pi_t(\mathrm{d}a \mid x). \end{aligned}$$

Now, using the induction hypothesis

$$\begin{aligned} C_t(\pi, x) &\geq \int_A \left[c(x, a) + \widetilde{\alpha}_t(x, a) \int_X J_{t+1}(y) Q(\mathrm{d}y \mid x, a) \right] \pi_t(\mathrm{d}a \mid x) \\ &\geq \min_{a \in A(x)} \left[c(x, a) + \widetilde{\alpha}_t(x, a) \int_X J_{t+1}(y) Q(\mathrm{d}y \mid x, a) \right], \end{aligned}$$

hence $C_t(\pi, x) \geq J_t(x)$, $x \in X$ and $t = 0, 1, \ldots, T+1$.
On the other hand, if $C_{t+1}(\pi, x) = J_{t+1}(x)$ for all $x \in X$ with $\pi = \pi^*$, $\pi_t(\cdot \mid h_t)$ is the measure of Dirac concentrated at $f_t(x_t)$, then the equality holds throughout the previous calculations obtaining $C_t(\pi^*, x) = J_t(x)$. Then, if $C_t(\pi, x) \geq J_t(x)$, in particular for $t = 0$, $v^{\tau}(\pi, x) \geq J_0(x)$ and for $\pi = \pi^*$, $v^{\tau}(\pi^*, x) = J_0(x)$. □

Now we show the case where $T=+\infty$. First, define the expected total cost from time n onwards applied to (5) given the initial condition $x_n=x$ and $\pi\in\Pi$, as follows:

$$v_n(\pi,x):=E_x^\pi\left[c(x_n,a_n)+\sum_{t=n+1}^{\infty}\prod_{k=n}^{t-1}\widetilde{\alpha}_k(x_k,a_k)c(x_t,a_t)\right], \tag{10}$$

and for $x\in X$ define

$$V_n(x):=\inf_{\pi\in\Pi}v_n(\pi,x). \tag{11}$$

Furthermore, for $N>n\geq 0$, define

$$v_{n,N}(\pi,x):=E_x^\pi\left[c(x_n,a_n)+\sum_{t=n+1}^{N}\prod_{k=n}^{t-1}\widetilde{\alpha}_k(x_k,a_k)c(x_t,a_t)\right], \tag{12}$$

with $\pi\in\Pi$, $x\in X$, and

$$V_{n,N}(x):=\inf_{\pi\in\Pi}v_{n,N}(\pi,x), \tag{13}$$

$x\in X$.

Define for $u\in L(X):=\{u:X\to R\mid u \text{ is non-negative and l.s.c.}\}$ and $n=0,1,\ldots$, the following operator defined on X as

$$T_nu(x)=\min_{a\in A(x)}\left[c(x,a)+\widetilde{\alpha}_n(x,a)\int_X u(y)Q(\mathrm{d}y\mid x,a)\right],$$

$x\in X$.

Remark 4.2. Under Assumption 3.3, it is straightforward to see that the following statements are hold:

a) $u\in L(X)\implies T_nu(x)\in L(X)$, $n=0,1,\ldots$.

b) Let $u\in L(X)$ and define

$$G_n(x,a):=c(x,a)+\widetilde{\alpha}_n(x,a)\int_X u(y)Q(\mathrm{d}y\mid x,a),\quad (x,a)\in\mathbb{K}.$$

Observe that for each $n\in\{0,1,\ldots\}$, G_n is a *l.s.c.* function on $\mathbb{K}$, due to Assumption 3.3. In consequence, since the multifunction $x\mapsto A(x)$ is *l.s.c.*, for each $n\geq 0$, there exists $f_n\in\mathbb{F}$ such that

$$T_nu(x)=G_n(x,f_n(x)),$$

$x\in X$. This fact follows as an application of the measurable selection theorems, see for instance [14].

c) Furthermore, observe that for each $n=0,1,2,\ldots$, $\lambda\in\mathbb{R}$ and $x\in X$, it holds that

$$M:=\{a\in A(x)\mid G_n(x,a)\leq\lambda\}\subset N:=\{a\in A(x)\mid c_n(x,a)\leq\lambda\},$$

since the cost function is a non-negative function. In consequence, since M is a closed set and N is a compact set, it follows that G_n is an inf-compact function on $\mathbb{K}$, for each $n=0,1,2,\ldots$.

The following lemmas will be very useful in the proof of the main result of this section.

Lemma 4.3. Suppose that Assumption 3.3 holds and let $\{u_n\}$ be a sequence in $L(X)$. If $u_n \geq T_n u_{n+1}$, $n = 0, 1, 2, \ldots$, then $u_n \geq V_n$, $n = 0, 1, 2, \ldots$.

Proof. Let $\{u_n\}$ be a sequence in $L(X)$ and suppose that

$$u_n \geq T_n u_{n+1} \quad n = 0, 1, 2, \ldots.$$

Then, by Remark 4.2 b), for each $n \geq 0$ there exists $f_n \in \mathbb{F}$ such that

$$u_n(x) \geq c(x, f_n(x)) + \widetilde{\alpha}_n(x, f_n(x)) \int_X u_{n+1}(y) Q(\mathrm{d}y \mid x, f_n(x)),$$

$x \in X$. Iterating this inequality, it is obtained that

$$\begin{aligned} u_n(x) \quad &\geq \quad E_x^{\pi}\left[c(x_n, f_n(x_n)) + \sum_{t=n+1}^{N-1} \prod_{j=n}^{t-1} \widetilde{\alpha}_j(x_j, f_j(x_j)) c(x_t, f_t(x_t))\right] \\ &\quad + \prod_{j=n}^{N-1} \widetilde{\alpha}_j(x_j, f_j(x_j)) E_x^{\pi}\big[u(x_N)\big], \end{aligned} \tag{14}$$

$x \in X$, where

$$E_x^{\pi}\left[u(x_N)\right] = \int_X u(y) Q^N(\mathrm{d}y \mid x_n, f_n(x_n)),$$

and $Q^N(\cdot \mid x_n, f_n(x_n))$ denotes the N-step transition kernel of the Markov control process $\{x_t\}$, when the policy $\pi = \{f_k\}$ is used, beginning at a stage n.
Since u is non-negative and $x_n = x$, (14) imply that

$$u_n(x) \geq E_x^{\pi}\left[c(x_n, f_n(x_n)) + \sum_{t=n+1}^{N-1} \prod_{j=n}^{t-1} \widetilde{\alpha}_j(x_j, f_j(x_j)) c(x_t, f_t(x_t))\right].$$

Hence, letting $N \to \infty$, it yields

$$u_n(x) \geq v_n(\pi, x) \geq V_n(x),$$

$x \in X$. □

Lemma 4.4. Suppose that Assumption 3.3 holds. Then, for every $n \geq 0$ and $x \in X$ the sequence $\{V_{n,N} \mid N \geq 0\}$ is non-decreasing and converges to V_n, that is,

$$V_{n,N}(x) \uparrow V_n(x) \quad \textit{as} \quad N \to \infty.$$

Proof. Let $x \in X$ be arbitrary but fixed. Observe that by Theorem 4.1 the functions defined as

$$U_t(x) = \min_{a\in A(x)} \left[c(x,a) + \widetilde{\alpha}_t(x,a) \int_X U_{t+1}(y) Q(\mathrm{d}y \mid x,a) \right], \tag{15}$$

for $t = N-1, N-2 \ldots, n$, with $U_N(x) = 0$, are *l.s.c.* (see Remark 4.2) and if $t = n$,

$$V_{n,N}(x) = \min_{a\in A(x)} \left[c(x,a) + \widetilde{\alpha}_n(x,a) \int_X V_{n+1,N}(y) Q(\mathrm{d}y \mid x,a) \right], \tag{16}$$

due to U_n is the optimal value of an optimal control problem of $N-n$ stages, i. e. $U_n(x) = V_{n,N}(x)$. In consequence, by the non-negativity of the cost function c, $\{V_{n,N} : N\} \subset L(x)$ is a non-decreasing sequence and

$$V_{n,N}(x) \le V_n(x), N > n.$$

Then, for each $n \ge 0$ there exists an unique function $u_n \in L(X)$, such that

$$V_{n,N}(x) \uparrow u_n(x) := \sup_{N>n} V_{n,N}(x).$$

It will now be proved that u_n coincides with V_n, for all $n \ge 0$. To this end, observe that

$$V_{n,N}(x) \le v_{n,N}(\pi, x) \le v_n(\pi, x), \pi \in \Pi.$$

Hence, $V_{n,N}(x) \le V_n(x), N > n$, then

$$u_n(x) \le V_n(x), n \ge 0. \tag{17}$$

On the other hand, from (16), when $N \to \infty$, it is obtained that

$$u_n(x) = \min_{a\in A(x)} \left[c(x,a) + \widetilde{\alpha}_n(x,a) \int_X u_{n+1}(y) Q(\mathrm{d}y \mid x,a) \right], \tag{18}$$

$n = 0, 1, 2, \ldots$, the interchange between limit and minimum is guaranteed by Lemma 4.2.4 in [14] (see Remark 4.2). Then, using the previous lemma, it follows that,

$$u_n(x) \ge V_n(x), n \ge 0. \tag{19}$$

Finally, since the state $x \in X$ is arbitrary, from (17) and (19) the result follows. □

Theorem 4.5. Suppose that Assumptions 3.3 and 3.4 hold, then

(a) the optimal value function V_n, $n = 0, 1, 2, \ldots$, satisfies the optimality equation

$$V_n(x) = \min_{a\in A(x)} \left[c(x,a) + \widetilde{\alpha}_n(x,a) \int_X V_{n+1}(y) Q(\mathrm{d}y \mid x,a) \right], \tag{20}$$

$x \in X$, and if $\{u_n\}$ is another sequence that satisfies the optimality equations in (20), then $u_n \ge V_n$.

(b) There exists a policy $\pi^* = \{f_n \in \mathbb{F} \mid n \geq 0\}$ such that, for each $n = 0, 1, 2, \ldots$, the control $f_n(x) \in A(x)$ attains the minimum in (20), i.e.

$$V_n(x) = c(x, f_n(x)) + \widetilde{\alpha}_n(x, f_n(x)) \int_X V_{n+1}(y) Q(\mathrm{d}y \mid x, f_n(x)), \tag{21}$$

$x \in X$, and the policy π^* is optimal.

Proof.

a) The proof of Lemma 4.4 guarantees that the sequence $\{V_n\}$ satisfies the optimality equations in (20), and by Lemma 4.3, if $\{u_n\}$ satisfies $u_n = T_n u_{n+1}$, it is concluded that $u_n \geq V_n$.

b) The existence of $f_n \in \mathbb{F}$ that satisfies (21) is ensured by Remark 4.2. Now, iterating (21) with $x_n = x \in X$, it is obtained that

$$\begin{aligned} V_n(x) &= E_x^{\pi}\left[c(x_n, f_n(x_n)) + \sum_{t=n+1}^{N-1} \prod_{j=n}^{t-1} \widetilde{\alpha}_j(x_j, f_j(x_j)) c(x_t, f_t(x_t)) \right] \\ &\quad + \prod_{j=n}^{N-1} \widetilde{\alpha}_j(x_j, f_j(x_j)) E_x^{\pi}\left[u(x_N)\right] \\ &\geq E_x^{\pi}\left[c(x_n, f_n(x_n)) + \sum_{t=n+1}^{N-1} \prod_{j=n}^{t-1} \widetilde{\alpha}_j(x_j, f_j(x_j)) c(x_t, f_t(x_t)) \right], \end{aligned}$$

$n \geq 0$ and $N > n$. This implies, that, letting $N \to \infty$, $V_n(x) \geq v_n(\pi^*, x)$, $x \in X$ and $\pi^* = \{f_k\} \subseteq \mathbb{F}$. Moreover, in particular for π^*, $V_n(x) \leq v_n(\pi^*, x)$, $x \in X$. Therefore, $V_n(x) = v_n(\pi^*, x), x \in X$ and $\pi^* = \{f_n\}$ is optimal.

□

5. EXAMPLES

Example 5.1. In this example Theorem 4.1 is applied to the following machine (equipment) replacement model. This class of models has been studied, for instance, in [18]. However, in the present case, we present a non-constant discount factor. The state of the system $x \in X = \{1, 2, \ldots, D\}$, D is a positive integer, represents the condition of the machine at each decision epoch. The higher the value of x, the worse the condition of the machine. Suppose that at the beginning of each period, the state of the machine is noted and an action upon whether or not to replace the machine is made. If the decision to replace is made, then it is assumed that the machine is instantaneously replaced by a new machine whose state is 1. Then the action space, which coincides with the space of admissible actions, is given by $A = \{0, 1\}$, action 0 corresponds to operating the machine for an additional period, while action 1 corresponds to replacing it and pay a cost $R > 0$. Let $P = (p_{i,j})_{D \times D}$ be the matrix of transition probabilities for going from level i to level j. Because no machine can move to a better level of deterioration, $p_{i,j} = 0$ if $j < i$. Let $g : \{1, 2, 3, \dots, D\} \to \mathbb{R}$ be a known function, which will measure the cost of operation of the machine. Suppose that g is non-decreasing, i. e. $g(1) \le g(2) \le \dots \le g(D)$.

The problem consists on determining an optimal replacement policy that minimizes the expected total discounted cost of operation, considering a varying discount factor. Furthermore, a random horizon τ with an uniform probability distribution is considered, that is, $P(\tau = k) = 1/T$.

Let $q(y|x, a)$ be the probability that the machine moves from level x to level y given the action a. Then

$$q(y|x, a = 0) = p_{x,y}$$

and

$$q(y|x, a = 1) = \begin{cases} 1 & \text{if } y = 1 \\ 0 & \text{otherwise.} \end{cases}$$

The cost-per-stage function is given by:

$$c(x, a) = \begin{cases} g(x) & \text{if } a = 0 \\ g(1) + R & \text{if } a = 1. \end{cases}$$

Hence, the dynamic programming equation (7) given in Theorem 4.1 for the replacement problem is written in the following way:

$$\begin{aligned} J_{T+1}(x) &= 0, \\ J_t(x) &= \min\{R + g(0) + \widehat{\alpha}_t \alpha_t(x, 1) J_{t+1}(1), g(x) + \widehat{\alpha}_t \alpha_t(x, 0) \sum_{y=x}^{D} q(y|x, 0) J_{t+1}(y)\}, \end{aligned}$$

$t = T, T-1, T-2, \dots, 0$, $x \in X$ and $\widehat{\alpha}_t = P(\tau \ge t+1)/P(\tau \ge t)$. Consider the following discount factors: $\alpha_0(x, a) = 1$ and

$$\alpha_t(x, a) = \begin{cases} \frac{x(1-a)+\beta a}{D(1+\beta^t)}, & x = 0, 1, 2, \dots, D-1, \\ \beta, & x = D, \end{cases}$$

$t = 1, 2, 3, \dots, T$ with $\beta \in (0, 1)$.

Suppose that $\beta = 0.8$, $T = 14$, $D = 6$,

$$P = \begin{pmatrix} 0.20 & 0.25 & 0.20 & 0.15 & 0.15 & 0.05 \\ 0 & 0.10 & 0.25 & 0.15 & 0.10 & 0.40 \\ 0 & 0 & 0.10 & 0.30 & 0.40 & 0.20 \\ 0 & 0 & 0 & 0.20 & 0.30 & 0.50 \\ 0 & 0 & 0 & 0 & 0.30 & 0.70 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$g(1) = 5$, $g(2) = 7$, $g(3) = 29$, $g(4) = 34$, $g(5) = 42$, $g(6) = 55$ and $R = 45$.

In Table 1 and Table 2, the optimal policy and the optimal value function are presented, respectively.

Stage	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
State															
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	0	0	0	0	0	0	0	1	1	1	1	1	0	0
4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
6	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table 1. Optimal policies.

Initial State x	1	2	3	4	5	6
Optimal Value	7.29953	13.83538	40.12755	50.44435	50.44435	55.33231

Table 2. Optimal value.

Example 5.2. Consider a Markov decision model with state space $X = [0, \infty)$ and action space $A = A(x) = [1, \infty), x \in X$. The dynamic of this system is given by the following difference equation:

$$x_{t+1} = \beta x_t a_t + \xi_t, \tag{22}$$

$t = 0, 1, 2, \ldots$, where $\beta \in (0, 1)$ and with $x_0 = x$ known. Suppose than $\{\xi_t\}$ is a sequence of independent and identically distributed random variables, with $E[\xi_t] = 0$ and $Var[\xi_t] = E[\xi_t^2] = 1$. The cost function is given by the following quadratic cost:

$$c(x, a) = x^2 + a^2,$$

$(x, a) \in \mathbb{K}$. For this example consider that the varying discount factor is given by

$$\alpha_t(x, a) = \frac{1}{a^2},$$

$(x, a) \in \mathbb{K}$. Finally, it will be assumed a random horizon τ, which is concentrated on T with $T = \infty$, i.e. it will be considered as an objective function as (3) with an infinite horizon.

Firstly, observe that Assumption 3.3, trivially holds. Then it simply necessary to verify Assumption 3.4. To this end, consider the stationary policy $f(x) = 1/\beta, x \in X$. Then, using (22), it is obtained that

$$x_n = x + \sum_{k=0}^{n-1} \xi_k, \tag{23}$$

for $n \geq 1$. In consequence,

$$\begin{aligned}
v^{\tau}(f,x) &= E_x^f\left[c(x_0,a_0)+\sum_{t=1}^{\infty}\prod_{k=0}^{t-1}\alpha_k(x_k,a_k)\,c(x_t,a_t)\right]\\
&= E_x^f\left[x^2+\frac{1}{\beta^2}+\sum_{t=1}^{\infty}\beta^{2t}\left(\left(x+\sum_{k=0}^{t-1}\xi_k\right)^2+\frac{1}{\beta^2}\right)\right]\\
&= E_x^f\left[x^2+\frac{1}{\beta^2}+\sum_{t=1}^{\infty}\beta^{2t}\left(x^2+2x\sum_{k=0}^{t-1}\xi_k+\sum_{k=0}^{t-1}\xi_k^2+\sum_{k,j=0,k\neq j}^{t-1}\xi_k\xi_j+\frac{1}{\beta^2}\right)\right]
\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{\beta^2}+x^2\sum_{t=1}^{\infty}\beta^{2t}+\sum_{t=1}^{\infty}t\beta^{2t}+\sum_{t=1}^{\infty}\beta^{2(t-1)}\\
&= \frac{1}{\beta^2}+\frac{x^2+1}{1-\beta^2}+\frac{\beta^2}{(1-\beta^2)}<\infty.
\end{aligned}$$

In this problem, the value iteration functions are given by

$$\begin{aligned}
V_{0,0}(x) &= 0\\
V_{0,N}(x) &= \min_{a\in[1,\infty)}\left[x^2+a^2+\frac{1}{a^2}E\left[V_{0,N-1}(\beta xa+\xi)\right]\right], \qquad N=1,2,3\ldots
\end{aligned}$$

and $x \in X$, where ξ is a generic element of $\{\xi_t\}$.
Iterating, it is obtained that

$$\begin{aligned}
V_{0,1}(x) &= \min_{a\in A(x)}\left[x^2+a^2\right]\\
&= C_1x^2+D_1
\end{aligned}$$

where $C_1 = D_1 = 1$ and $f_1(x) = 1$.
Next,

$$\begin{aligned}
V_{0,2}(x) &= \min_{a\in[1,\infty)}\left[x^2+a^2+\frac{1}{a^2}E\left[C_1(\beta xa+\xi)^2+D_1\right]\right]\\
&= \min_{a\in[1,\infty)}\left[x^2+a^2+C_1\beta^2x^2+\frac{C_1+D_1}{a^2}\right]\\
&= \min_{a\in[1,\infty)}\left[\frac{a^4+C_1+D_1}{a^2}+(1+C_1\beta^2)x^2\right]\\
&= (1+C_1\beta^2)x^2+2\sqrt{C_1+D_1}\\
&= C_2x^2+D_2,
\end{aligned}$$

where $C_2 = (1 + C_1\beta^2)$ and $D_2 = 2\sqrt{C_1 + D_1}$ with $f_2(x) = \sqrt[4]{C_1 + D_1}$.

Then,

$$\begin{aligned} V_{0,3}(x) &= \min_{a\in[1,\infty)}\left[x^2+a^2+\frac{1}{a^2}E\left[C_2(\beta xa+\xi)^2+D_2\right]\right] \\ &= \min_{a\in[1,\infty)}\left[\frac{a^4+C_2+D_2}{a^2}+(1+C_2\beta^2)x^2\right] \\ &= (1+C_2\beta^2)x^2+2\sqrt{C_2+D_2} \\ &= C_3x^2+D_3, \end{aligned}$$

where $C_3=(1+C_2\beta^2)$ and $D_3=2\sqrt{C_2+D_2}$ with $f_3(x)=\sqrt[4]{C_2+D_2}$.

By induction, it is obtained that

$$V_{0,N}(x)=C_Nx^2+D_N$$

and

$$f_N(x)=2\sqrt[4]{D_{N-1}+C_{N-1}},$$

where the constants C_N and D_N satisfies the following recurrence equations:

$$\begin{aligned} D_1 &= 1 \\ D_N &= 1+D_{n-1}\beta^2, \qquad N=2,3,4,\ldots \end{aligned}$$

and

$$\begin{aligned} C_1 &= 1 \\ C_N &= 2\sqrt{C_{N-1}+D_{N-1}} \qquad N=2,3,4,\ldots. \end{aligned}$$

Now, taking limit when N goes to ∞, it follows that

$$\lim_{N\to\infty}D_n=\lim_{N\to\infty}\sum_{t=0}^{N-1}\beta^{2t}=\frac{1}{1-\beta^2},$$

and

$$\lim_{N\to\infty}C_n=2\left(1+\sqrt{\frac{2-\beta^2}{1-\beta^2}}\right).$$

Then, by Lemma 4.4, the optimal value function is given by

$$\begin{aligned} V(x) &= \lim_{N\to\infty}V_{0,N}(x) \\ &= \frac{x^2}{1-\beta^2}+2\left(1+\sqrt{\frac{2-\beta^2}{1-\beta^2}}\right), \end{aligned}$$

$x\in X$, and the optimal policy is

$$f(x)=\sqrt[4]{2\left(1+\sqrt{\frac{2-\beta^2}{1-\beta^2}}\right)+\frac{1}{1-\beta^2}}$$

$x\in X$.

(Received March 26, 2016)

REFERENCES

[1] Y. Carmon and A. Shwartz: Markov decision processes with exponentially representable discounting. Oper. Res. Lett. *37* (2009), 51–55. DOI:10.1016/j.orl.2008.10.005

[2] X. Chen and X. Yang: Optimal consumption and investment problem with random horizon in a BMAP model. Insurance Math. Econom. *61* (2015), 197–205. DOI:10.1016/j.insmatheco.2015.01.004

[3] H. Cruz-Suárez, R. Ilhuicatzi-Roldán, and R. Montes-de-Oca: Markov decision processes on Borel spaces with total cost and random horizon. J. Optim. Theory Appl. *162* (2014), 329–346. DOI:10.1007/s10957-012-0262-8

[4] E. Della Vecchia, S. Di Marco, and F. Vidal: Dynamic programming for variable discounted Markov decision problems. In: Jornadas Argentinas de Informática e Investigación Operativa (43JAIIO) XII Simposio Argentino de Investigación Operativa (SIO), Buenos Aires 2014, pp. 50–62.

[5] E. Feinberg and A. Shwartz: Constrained dynamic programming with two discount factors: applications and an algorithm. IEEE Trans. Automat. Control *44* (1999), 628–631. DOI:10.1109/9.751365

[6] E. Feinberg and A. Shwartz: Markov decision models with weighted discounted criteria. Math. Oper. Res. *19* (1994), 152–168. DOI:10.1287/moor.19.1.152

[7] Y. H. García and J. González-Hernández: Discrete-time Markov control process with recursive discounted rates. Kybernetika *52* (2016), 403–426. DOI:10.14736/kyb-2016-3-0403

[8] J. González-Hernández, R. R. López-Martínez, and J. A. Minjarez-Sosa: Adaptive policies for stochastic systems under a randomized discounted criterion. Bol. Soc. Mat. Mex. *14* (2008), 149–163.

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

[10] J. González-Hernández, R. R. López-Martínez, J. A. Minjarez-Sosa, and J. A. Gabriel-Arguelles: Constrained Markov control processes with randomized discounted cost criteria: occupation measures and external points. Risk and Decision Analysis *4* (2013), 163–176.

[11] J. González-Hernández, R. R. López-Martínez, J. A. Minjarez-Sosa, and J. A. Gabriel-Arguelles: Constrained Markov control processes with randomized discounted rate: infinite linear programming approach. Optimal Control Appl. Methods *35* (2014), 575–591. DOI:10.1002/oca.2089

[12] J. González-Hernández, R. R. López-Martínez, and J. R. Pérez-Hernández: Markov control processes with randomized discounted cost. Math. Methods Oper. Res. *65* (2007), 27–44. DOI:10.1007/s00186-006-0092-2

[13] X. Guo, A. Hernández-del-Valle, and O. Hernández-Lerma: First passage problems for nonstationary discrete-time stochastic control systems. Eur. J. Control *18* (2012), 528–538. DOI:10.3166/ejc.18.528-538

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

[15] K. Hinderer: Foundations of non-stationary dynamic programming with discrete time parameter. In: Lectures Notes Operations Research (M. Bechmann and H. Künzi, eds.), Springer-Verlag *33*, Zürich 1970. DOI:10.1007/978-3-642-46229-0

[16] R. Ilhuicatzi-Roldán and H. Cruz-Suárez: Optimal replacement in a system of n-machines with random horizon. Proyecciones *31* (2012), 219–233. DOI:10.4067/s0716-09172012000300003

[17] J. A. Minjares-Sosa: Markov Control Models with unknown random state-action-dependent discounted factors. TOP *23* (2015), 743–772. DOI:10.1007/s11750-015-0360-5

[18] M. L. Puterman: Markov Decision Process: Discrete Stochastic Dynamic Programming. John Wiley and Sons, New York 1994.

[19] M. Schäl: Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal. Probab. Theory Related Fields *32* (1975), 179–196. DOI:10.1007/bf00532612

[20] Q. Wei and X. Guo: Markov decision processes with state-dependent discounted factors and unbounded rewards/costs. Oper. Res. Lett. *39* (2011), 369–374. DOI:10.1016/j.orl.2011.06.014

[21] Q. Wei and X. Guo: Semi-Markov decision processes with variance minimization criterion. 4OR, *13* (2015), 59-79. DOI:10.1007/s10288-014-0267-2

[22] X. Wu and X. Guo: First passage optimality and variance minimisation of Markov decision processes with varying discounted factors. J. Appl. Probab. *52* (2015), 441–456. DOI:10.1017/s0021900200012560

[23] X. Wu, X. Zou, and X. Guo: First passage Markov decision processes with constraints and varying discount factors. Front. Math. China *10* (2015), 1005–1023. DOI:10.1007/s11464-015-0479-6

[24] X. Wu and J. Zhang: An application to the finite approximation of the first passage models for discrete-time Markov decision processes with varying discount factors. In: Proc. 11th World Congress on Intelligent Control and Automation 2015, pp. 1745–1748. DOI:10.1109/wcica.2014.7052984

[25] X. Wu and J. Zhang: Finite approximation of the first passage models for discrete-time Markov decision processes with varying discounted factors. Discrete Event Dyn. Syst. *26* (2016), 669–683. DOI:10.1007/s10626-014-0209-3

[26] L. Ye and X. Guo: Continuous-time Markov decision processes with state-dependent discount factors. Acta Appl. Math. *121* (2012), 5–27. DOI:10.1007/s10440-012-9669-3

[27] Y. Zhang: Convex analytic approach to constrained discounted Markov decision processes with non-constant discount factors. TOP *21* (2013), 378–408. DOI:10.1007/s11750-011-0186-8

Rocio Ilhuicatzi-Roldán, Universidad Autónoma de Tlaxcala, Facultad de Ciencias Básicas, Ingeniería y Tecnología, Av. Ángel Solana s/n, San Luis Apizaquito, Tlaxcala. México.
e-mail: rocioil@hotmail.com

Hugo Cruz-Suárez, Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias Físico Matemáticas, Av. San Claudio y 18 Sur, Puebla, México.
e-mail: hcs@fcfm.buap.mx

Selene Chávez-Rodríguez, Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias Físico Matemáticas, Av. San Claudio y 18 Sur, Puebla. México.
e-mail: nagiroge@hotmail.com