Kybernetika

Rosa María Flores-Hernández

Monotone optimal policies in discounted Markov decision processes with transition probabilities independent of the current state: existence and approximation

Kybernetika, Vol. 49 (2013), No. 5, 705--719

Persistent URL: http://dml.cz/dmlcz/143520

KYBERNETIKA — VOLUME 49 (2013), NUMBER 5, PAGES 705–719

MONOTONE OPTIMAL POLICIES IN DISCOUNTED MARKOV DECISION PROCESSES WITH TRANSITION PROBABILITIES INDEPENDENT OF THE CURRENT STATE: EXISTENCE AND APPROXIMATION

Rosa M. Flores-Hernández

In this paper there are considered Markov decision processes (MDPs) that have the discounted cost as the objective function, state and decision spaces that are subsets of the real line but are not necessarily finite or denumerable. The considered MDPs have a cost function that is possibly unbounded, and dynamic independent of the current state. The considered decision sets are possibly non-compact. In the context described, conditions to obtain either an increasing or decreasing optimal stationary policy are provided; these conditions do not require assumptions of convexity. Versions of the policy iteration algorithm (PIA) to approximate increasing or decreasing optimal stationary policies are detailed. An illustrative example is presented. Finally, comments on the monotonicity conditions and the monotone versions of the PIA that are applied to discounted MDPs with rewards are given.

Keywords: Markov decision process, total discounted cost, total discounted reward, increasing optimal policy, decreasing optimal policy, policy iteration algorithm

Classification: 90C40, 93E20

1. INTRODUCTION

In this paper, there are considered Markov decision processes (MDPs) with the discounted cost as the objective function (see [3, 8, 9, 13]). Conditions to ensure the existence of monotone optimal stationary policies are described, when the existence of an optimal stationary policy is assumed. Such conditions are given in terms of the elements of the Markov decision model (for other kinds of conditions see [2]); in particular, these conditions allow the state space X and the decision space A to be subsets of $\mathbb{R}$ that can be finite, denumerable or non-denumerable.

Specifically, there are considered discrete-time MDPs on real spaces that have an infinite time horizon and discounted cost as the objective function. The MDPs considered have a dynamic of the system that is independent of the current state, a cost function that is not necessarily bounded, and compact or non-compact decision sets.

The dynamic described here permits the construction of conditions that do not require the properties of stochastic order, superadditivity or subadditivity in the transition law, and monotonicity in the reward function (as in [5, 13, 14]). In previous work the dynamic

of the system is more general; however, the state spaces are finite ([5]) or denumerable ([13]), the decision spaces are finite ([5, 14]), or the reward function is bounded ([14]). Contrary to the conditions presented in [7], the conditions presented here do not contain assumptions of convexity.

Models that have dynamic independent of the current state do not embrace all control structures, but is natural for certain types of problems. This dynamic is observed in models of capital accumulation, fisheries management, and reservoir operation ([6, 9, 12]). Other examples concern single-product dynamic inventory models and models regarding replacement and maintenance ([10]). MDPs with finite decision sets, where the transition probabilities depend on the decisions taken and not on the current state were named "invariant MDPs" by Assaf (see [1]). In this paper, this terminology will not be used. In [1] (Section 5), it is shown that (modulo measurability technicalities) any problem may be transformed into an invariant one. More examples that possess this special structure or those that can be converted into such problems can be found in economic literature (see [10]).

Policy iteration, which is also known as the approximation in policy space (see [3, 8, 13]), is a method for solving optimal decision problems (ODPs). It directly refers to the particular structure of the ODPs.

In the same context described four paragraphs above, in this paper, modified versions of the policy iteration algorithm (PIA) are provided and implemented in $\mathbb{R}$. These new algorithms allow one to approximate monotone (increasing or decreasing) optimal stationary policies. Contrary to the existing algorithm (see [8]), in the versions presented here, when the initial stationary policy is increasing (or decreasing), the policies obtained at each iteration are also increasing (or decreasing) and stationary. [13] considers an increasing version of the PIA for MDPs with rewards, but there both the state and decision spaces are finite subsets of $\mathbb{R}$.

The main and novel contribution of the paper is the presentation of a detailed study of the MDPs on real spaces with costs and dynamic independent of the current state, by the following: a) Non-convex conditions which guarantee the existence of monotone optimal policies are provided. These conditions extend the previous ones given for discrete MDPs in [13], and complete the study of the existence of monotone optimal policies given in Section 4 of [7], because in this reference only convex conditions are provided. b) Suitable versions of the PIA in order to approximate the (monotone) optimal policies are given. These versions are refinements of the PIA proposed for MDPs on Borel spaces in [8]. In fact, the versions of the PIA provided permit to take, as initial condition, monotone stationary policies instead of measurable ones, as the version presented in [8] require. c) There is included an illustration of how the theory presented works by means of an elaborated example.

The paper is organized as follows. Section 2, provides basic concepts and results in $\mathbb{R}$ and on MDPs with costs. In Section 3, the non-convex conditions under which it is possible to guarantee the existence of monotone optimal stationary policies are provided. In Section 4, the monotone versions of the PIA are detailed. In Section 5, an example is detailed to illustrate the existence of an increasing optimal stationary policy and to illustrate the corresponding version of the PIA. Section 6, presents remarks about the conditions that ensure the existence of monotone optimal stationary policies for

MDPs with rewards and the corresponding modified versions of the PIA. Finally, the conclusions are supplied in Section 7.

2. PRELIMINARIES

2.1. Terminology and some results in $\mathbb{R}$

This subsection contains concepts and results in $\mathbb{R}$ (see [7] and [14] for the more general context in lattice theory). For such a space, $x \wedge y := \inf\{x, y\}$ and $x \vee y := \sup\{x, y\}$.

Let Γ be a fixed subset of $\mathbb{R}$.

Let Θ and Υ be subsets of Γ. Θ is *lower than* Υ, and is denoted as $\Theta \sqsubseteq \Upsilon$, if $\theta \wedge \upsilon \in \Theta$ and $\theta \vee \upsilon \in \Upsilon$ for all $\theta \in \Theta$ and $\upsilon \in \Upsilon$.

Let Z be a nonempty subset of $\mathbb{R}$. For $x \in Z$, let $\Gamma(x)$ be a nonempty subset of Γ. It is said that the multifunction $x \to \Gamma(x)$ is *ascending* if $x \to \Gamma(x)$ is increasing with respect to the relation $\sqsubseteq$, i. e., $\Gamma(x) \sqsubseteq \Gamma(y)$ for $x \leq y$ in Z. It is said that $x \to \Gamma(x)$ is *descending* if $x \to \Gamma(x)$ is decreasing with respect to the relation $\sqsubseteq$.

Let X and A be nonempty Borel subsets of $\mathbb{R}$ with Borel $\sigma-$ algebras $\mathcal{B}(X)$ and $\mathcal{B}(A)$, respectively. For each $x \in X$, let $A(x)$ be a nonempty (measurable) subset of A (i. e., $x \to A(x)$ is a multifunction from X to A). Suppose that $\mathbb{K} := \{(x, a) : x \in X,\ a \in A(x)\}$ is a measurable subset of $X \times A$.

A function $T : \mathbb{K} \to \mathbb{R}$ is *subadditive* (it has antitone or decreasing differences) on $\mathbb{K}$ if $T(y, b) + T(x, a) \leq T(y, a) + T(x, b)$ for all $x \leq y$ in X and $a \leq b$, with $a, b \in A(x) \cap A(y)$. T is called *superadditive* (it has isotone or increasing differences) on $\mathbb{K}$ if $-T$ is subadditive on $\mathbb{K}$.

2.2. Markov decision processes

This subsection contains concepts and results of Markov decision processes (MDPs) as in [7].

Let $\{X, A, \{A(x) : x \in X\}, Q, c\}$ be a discrete-time, stationary *Markov decision model* and consists of the *state space* X, the *decision set* A, the *admissible decision sets* $A(x)$, $x \in X$, the *transition law* Q, and the *one-stage cost* c. The cost function $c : \mathbb{K} \to \mathbb{R}$ is measurable.

In this paper, the sets X, A, $A(x)$ for each $x \in X$, and $\mathbb{K}$ are considered as in Subsection 2.1. The transition law is independent of the current state, and therefore $Q(B|a)$, $B \in \mathcal{B}(X)$ and $a \in A(x)$, is a stochastic kernel on X, given A. Specifically, the evolution of the MDPs considered is given by the transition probability law Q induced by the difference equation of the type:

$$x_{t+1} = U(a_t, \xi_t), \tag{1}$$

$t = 0, 1, ...$, where $x_0 = x \in X$ is the initial state, $\{\xi_t\}$ is a sequence of independent and identically distributed random variables that, take on values in some real Borel space S with density Δ. Let ξ denote a generic element of the sequence $\{\xi_t\}$ (ξ will be used in this paper to specify the assumptions related to the sequence $\{\xi_t\}$), and let $U : A \times S \to X$ be a measurable function.

Let $\mathbb{F}$ be the set of *decision functions* or *measurable selectors*, i. e. the set of all measurable functions $\varrho : X \to A$, such that $\varrho(x) \in A(x)$ for all $x \in X$. A sequence $\pi = \{\varrho_t\}$, such that for each t, $\varrho_t \in \mathbb{F}$ is called a *Markov policy*. A *stationary policy* is a Markov policy π, such that $\varrho_t = \varrho$, for all $t = 0, 1, ...$ and $\varrho \in \mathbb{F}$. The stationary policy $(\varrho, \varrho, \ldots)$ will be identified with the element of the sequence, i. e., ϱ. The set of all policies will be denoted by Π.

The *expected total discounted cost* is taken into account as the objective function and is given by:

$$V(\pi, x) := E_x^\pi \left[\sum_{t=0}^{\infty} \alpha^t c(x_t, a_t) \right], \tag{2}$$

where $x_0 = x$ is the initial state, π is the policy that drives the system, and the discount factor is given by $\alpha \in (0, 1)$.

A policy π^* is called *optimal* if

$$V(\pi^*, x) = \inf_{\pi \in \Pi} V(\pi, x), \tag{3}$$

for all $x \in X$, and the minimum cost $V^*(x) := V(\pi^*, x)$, $x \in X$, is referred to as the *optimal value function*.

For $\{X, A, \{A(x), x \in X\}, Q, c\}$ a fixed Markov decision model, as the one specified above, the following results are considered and are adapted to the dynamic used in this paper:

Assumption 2.1. (Assumptions 4.2.1 and 4.2.2 in [8])

a) The one-stage cost $c : \mathbb{K} \to \mathbb{R}$ is nonnegative, lower semi-continuous (l.s.c.) and inf-compact on $\mathbb{K}$, that is for every $x \in X$ and $s \in \mathbb{R}$, the set $A_s(x) := \{a \in A(x) : c(x, a) \leq s\}$ is compact.

b) The transition law Q is strongly continuous, i. e.,

$$u'(a) := \int u(z)\, Q(\mathrm{d}z|a)$$

is continuous and bounded on A for each function $u : X \to \mathbb{R}$ measurable and bounded.

c) There is a policy π such that $V(\pi, x) < \infty$ for all $x \in X$.

Π^0 denotes the family of policies for which Assumption 2.1 c) holds.

Lemma 2.2. (Theorem 4.2.3 parts a, b and c in [8]) Suppose that Assumption 2.1 holds, then:

a) The discounted cost optimal value function V^* satisfies the *discounted cost optimality equation* (DCOE), i. e. for all $x \in X$,

$$V^*(x) = \min_{a \in A(x)} \left[c(x, a) + \alpha \int V^*(z)\, Q(\mathrm{d}z|a) \right]. \tag{4}$$

b) There is $f^* \in \mathbb{F}$, such that

$$V^*(x) = c(x, f^*(x)) + \alpha \int V^*(z)\, Q(\mathrm{d}z|f^*(x)), \tag{5}$$

where $x \in X$, and f^* is optimal. Conversely, if f^* is a stationary optimal policy, then it satisfies (5).

c) If π^* is a policy such that $V(\pi^*, \cdot)$ is a solution to the DCOE and satisfies

$$\lim_{n\to\infty} \alpha^n E_x^{\pi}\, V(\pi^*, x_n) = 0$$

for all $\pi \in \Pi^0$ and $x \in X$, then $V(\pi^*, \cdot) = V^*(\cdot)$; hence π^* is a discounted optimal policy.

Remark 2.3. (see [8], p. 51) For any deterministic stationary policy f, the discounted cost $V(f, \cdot)$ satisfies for all $x \in X$,

$$V(f, x) = c(x, f(x)) + \alpha \int V(f, y)\, Q(\mathrm{d}y|f(x)).$$

3. MONOTONICITY CONDITIONS

Define the function $LV : \mathbb{K} \to \mathbb{R}$ as

$$LV(x, a) := c(x, a) + \alpha \int V^*(z)\, Q(\mathrm{d}z|a), \tag{6}$$

which corresponds to the function that is minimized in (4).

For each $x \in X$, define $A^*(x)$ by

$$A^*(x) := \left\{ a \in A(x) : LV(x, a) = \min_{a^* \in A(x)} LV(x, a^*) \right\}$$

where $A^*(x)$ represents the set of minimizers of the DCOE.

Lemma 3.1. (Lemma 6.1 in [7]) Assumption 2.1 implies that $A^*(x)$ is a nonempty compact set for every $x \in X$.

Let $f : X \to A$ be the greatest admissible decision for each $x \in X$ that satisfies the DCOE, i. e.

$$f(x) = \sup A^*(x), \tag{7}$$

which is well defined and $f(x) \in A^*(x) \subset A(x)$, by Lemma 3.1.

Assumption 3.2.

a) X is discrete (i. e., finite or denumerable).

b) $X \subset \mathbb{R}$ is an interval and f is monotone.

c) There exists a unique optimal policy for the discounted MDP taken into account.

Remark 3.3. a) It is not difficult to observe that each condition in Assumption 3.2 implies that f, defined in (7), is measurable.

b) In [4] the authors give conditions under which Assumption 3.2 c) is satisfied.

3.1. Increasing optimal policies

Condition 3.4.

a) $x \to A(x)$ is ascending.

b) $c(\cdot,\cdot)$ is subadditive on $\mathbb{K}$.

Lemma 3.5. Under Condition 3.4 b), the function LV that is defined in (6) is subadditive.

P r o o f . Since the second term of (6) does not depend on x given that $\int V^*(z)Q(\mathrm{d}z|a) = \int V^*(U(a,s))\,\Delta(s)\,\mathrm{d}s$ by (1), the subadditivity of LV is a consequence of the fact that c is subadditive. □

Theorem 3.6. Suppose that Assumption 2.1 holds and one of the conditions of Assumption 3.2 results, then there exists an increasing optimal stationary policy under Condition 3.4.

P r o o f . The proof proceeds by contradiction. From Lemma 2.2 b), let f be defined as in (7). Suppose that for $x, y \in X$ with $x \leq y$, $f(y) < f(x)$, then, as $x \to A(x)$ is ascending, and by Lemma 3.1, $f(x) \in A^*(x) \subset A(x)$ and $f(y) \in A^*(y) \subset A(y)$, it follows that $f(y) = f(x) \wedge f(y) \in A(x)$ and $f(x) = f(x) \vee f(y) \in A(y)$. Thus, $f(x), f(y) \in A(x) \cap A(y)$. Since $x \leq y$, $f(y) < f(x)$, $f(x), f(y) \in A(x) \cap A(y)$, and LV is subadditive (Lemma 3.5), it is obtained that

$$LV(y, f(x)) + LV(x, f(y)) \leq LV(y, f(y)) + LV(x, f(x)).$$

According to the DCOE and since $f(y) \in A(x)$, it follows that

$$LV(x, f(x)) \leq LV(x, f(y)).$$

Thus,

$$0 \leq LV(x, f(y)) - LV(x, f(x)) \leq LV(y, f(y)) - LV(y, f(x)),$$

i. e.

$$LV(y, f(x)) \leq LV(y, f(y)).$$

This contradicts the definition of $f(y)$. Therefore, f is an increasing optimal policy. It is measurable, according to Remark 3.3 a). □

3.2. Decreasing optimal policies

Condition 3.7.

a) $x \to A(x)$ is descending.

b) $c(\cdot,\cdot)$ is superadditive on $\mathbb{K}$.

Lemma 3.8. Under Condition 3.7 b), the function LV that is defined in (6) is superadditive.

P r o o f. As in the proof of Lemma 3.5, the superadditivity of LV is a consequence of the fact that c is superadditive, given that the second term of (6) does not depend on the state x. □

Theorem 3.9. Suppose that Assumption 2.1 holds and suppose that one of the conditions of Assumption 3.2 results, then there exists a decreasing optimal stationary policy under Condition 3.7.

P r o o f. This proof is similar to the proof of Theorem 3.6. Now, suppose that $f(x) < f(y)$ for $x \leq y$ in X. Using Lemma 3.8 and $LV(y, f(y)) \leq LV(y, f(x))$, it follows that $LV(x, f(y)) \leq LV(x, f(x))$. This is a contradiction given that $f(x) \in A^*(x)$ and $f(y) \in A(x)$ (because $A(\cdot)$ is descending), for $x \in X$. Therefore, f is a decreasing optimal policy and is measurable by Remark 3.3 a). □

Remark 3.10.

a) Conditions 3.4 and 3.7 do not require the state and decision sets to be convex, and thus it is possible to consider discrete models, i. e. Markov decision models for which X and/or A are finite or denumerable sets.

b) In Heyman and Sobel [9] (Section 8.3) and Mendelssohn and Sobel [12], there are several examples in resources management that are presented, where the dynamic of the system is similar to (1). For example, there are models of capital accumulation, fisheries management, and reservoir operation. In individual optimal consumption and savings models, consider time t, where x_t denotes the capital on hand in units of dollars or physical quantities as the context dictates, a_t represents the amount of x_t that is reinvested, and $z_t = x_t - a_t$ is the amount consumed. Equation (1) represents the connection between the reinvestment decision and accumulated capital.

c) Jaśkiewicz [10] mentions a single-product dynamic inventory model, where the dynamic of the system is similar to equation (1). In such a model an action a is the stock after ordering, and ξ is a random variable denoting the demand. The transition probability function $Q(\cdot|a)$ is a distribution function of $a - \xi$, which describes the movement of the system from the current stock x to a new one y.

4. MONOTONE VERSIONS OF THE POLICY ITERATION ALGORITHM (PIA)

In this section, suitable versions of the PIA in order to approximate the (monotone) optimal policies are presented. For it, let $V(g, x)$ be the expected discounted total cost when the policy g is used and take the initial state to be $x_0 = x$.

4.1. PIA implemented to approximate increasing optimal policies

In this new version of the PIA, the initial condition is assumed to be an increasing stationary policy and the policies that are obtained at each iteration of the algorithm are also increasing and stationary.

Let $\mathcal{I}$ be the set of increasing stationary policies.

The increasing version of the PIA, which is an improved version, in $\mathbb{R}$, of the PIA proposed for MDPs on Borel spaces in [8], is shown below.

Algorithm 4.1. Set $n = 0$ and select $g_0 \in \mathcal{I}$.

1) (Policy evaluation) Given that $g_n \in \mathcal{I}$, calculate the corresponding cost v_n by solving the equation

$$v_n(x) = c(x, g_n(x)) + \alpha \int v_n(y)\, Q(\mathrm{d}y|g_n(x)) \tag{8}$$

for all $x \in X$. By Remark 2.3 $v_n(\cdot) = V(g_n, \cdot)$.

2) (Policy improvement) Determine $g_{n+1} \in \mathcal{I}$ such that for all $x \in X$,

$$\begin{aligned} &c(x, g_{n+1}(x)) + \alpha \int v_n(y)\, Q(\mathrm{d}y|g_{n+1}(x)) \\ &= \min_{a \in A(x)} \left[c(x,a) + \alpha \int v_n(y)\, Q(\mathrm{d}y|a) \right] \end{aligned} \tag{9}$$

and calculate v_{n+1} according to (8).

If $v_{n+1}(x) = v_n(x)$ for all $x \in X$, then take $v = v_n$ and stop; g_n could be the increasing optimal stationary policy. Otherwise, substitute g_n by g_{n+1}, increment n by 1 and return to step 2.

Theorem 4.2. Suppose that Assumption 2.1, one of the conditions of Assumption 3.2, and Condition 3.4 hold. Algorithm 4.1 yields a sequence $\{g_n(\cdot)\}$ of increasing stationary policies. If there exists an n for which $v_{n+1}(x) = v_n(x)$ for all $x \in X$, then $v = v_n$ is a solution to the DCOE

$$v(x) = \min_{a \in A(x)} \left[c(x,a) + \alpha \int v(z)\, Q(\mathrm{d}z|a) \right]. \tag{10}$$

In addition, if v satisfies

$$\lim_{t \to \infty} \alpha^t E_x^\pi v(x_t) = 0 \tag{11}$$

for all $\pi \in \Pi^0$ and $x \in X$, then $V^* = v$ and g_n is a discounted optimal policy.

P r o o f . The policy $g_{n+1}(\cdot)$ for $n \in \mathbb{N}$, which is obtained in step 2 of Algorithm 4.1, is increasing and stationary by Theorem 3.6 and the fact that

$$LV_n(x,a) := c(x,a) + \alpha \int v_n(y)\, Q(\mathrm{d}y|a),$$

for $(x,a) \in \mathbb{K}$ and $n = 0, \ldots$, is subadditive (now consider $LV_n(x,a)$, for $n = 0, \ldots$ instead of $LV(x,a)$). The last statement is a consequence of the fact that $c(\cdot,\cdot)$ is subadditive on $\mathbb{K}$ (see Condition 3.4 b)).

If there exist an n for which $v_{n+1}(x) = v_n(x)$ for all $x \in X$, then it follows from (9) and Remark 2.3 that $v = v_n$ satisfies the DCOE (10). If (11) holds, the desired conclusion follows from Lemma 2.2 c). □

4.2. PIA implemented to approximate decreasing optimal policies

In this new version of the PIA, the initial condition is a decreasing stationary policy and the policies obtained at each iteration are also decreasing and stationary.

Let $\mathcal{D}$ be the set of decreasing stationary policies.

The decreasing version of the PIA, which is an improved version, in $\mathbb{R}$, of the PIA presented in [8], is similar to Algorithm 4.1 with the difference being $g_n \in \mathcal{D}$, $n = 0, 1, \ldots$.

Algorithm 4.3. Set $n = 0$ and select $g_0 \in \mathcal{D}$.

1) (Policy evaluation) Given $g_n \in \mathcal{D}$, calculate the corresponding cost v_n by solving equation (8) for all $x \in X$. By Remark 2.3 $v_n(\cdot) = V(g_n, \cdot)$.

2) (Policy improvement) Determine $g_{n+1} \in \mathcal{D}$ such that for all $x \in X$, (9) holds, and calculate v_{n+1} according to (8).

If $v_{n+1}(x) = v_n(x)$ for all $x \in X$, then take $v = v_n$ and stop; g_n could be the decreasing optimal stationary policy. Otherwise, substitute g_n with g_{n+1}, increment n by 1 and return to step 2.

Theorem 4.4. Suppose that Assumption 2.1, one of the conditions of Assumption 3.2, and Condition 3.7 hold. Algorithm 4.3 yields a sequence $\{g_n(\cdot)\}$ of decreasing stationary policies. If there exists an n for which $v_{n+1}(x) = v_n(x)$ for all $x \in X$, then $v = v_n$ is a solution to the DCOE (10). If in addition, v satisfies (11) for all $\pi \in \Pi^0$ and $x \in X$, then $V^* = v$ and g_n is a discounted optimal policy.

P r o o f . The proof is similar to the proof of Theorem 4.2 by changing the words "increasing and subadditive" to "decreasing and superadditive", respectively, and using Theorem 3.9 and Condition 3.7 b) instead of Theorem 3.6 and Condition 3.4 b). □

Remark 4.5. If there exist positive numbers m and k, with $1 \leq k \leq \frac{1}{\alpha}$ and w as a nonnegative measurable function on X such that for all $(x, a) \in \mathbb{K}$,

a) $c(x, a) \leq mw(x)$,

b) $\int w(y)\, Q(\mathrm{d}y|a) \leq kw(x)$,

then (11) holds (see remark in p. 58 and Proposition 4.3.1 b) in [8]).

5. AN EXAMPLE

Example 5.1. Consider $X = A = [0, \infty)$, and for each $x \in X$, $A(x) = \left[\frac{x}{2}, \infty\right)$. The dynamic of the system is given as in (1) by

$$x_{t+1} = a_t + \xi_t$$

for $t = 0, 1, \ldots$, $S = [0, \infty)$, $E[\xi] = \mu$, and there is a continuous $\Delta(\cdot)$. Let $c(x, a) = x + a$, for $(x, a) \in \mathbb{K}$.

Lemma 5.2. Example 5.1 satisfies Assumption 2.1, Assumption 3.2 b), and Condition 3.4. Therefore, for this example there exists an increasing optimal stationary policy by Theorem 3.6.

P r o o f . In Example 5.1, $c(\cdot,\cdot)$ is nonnegative, continuous (particularly l.s.c.) and inf-compact, given that for $x \in X$ and $s \in \mathbb{R}$, $\{a \geq \frac{x}{2} : c(x,a) \leq s\} = [\frac{x}{2}, s-x]$ if $s \geq \frac{3}{2}x$, and $\{a \geq \frac{x}{2} : c(x,a) \leq s\} = \emptyset$ if $s < \frac{3}{2}x$; hence $\{a \geq \frac{x}{2} : c(x,a) \leq s\}$ is compact for all s. $Q(B|a) = \int I_B(a+s)\Delta(s)\,\mathrm{d}s$, $a \in [\frac{x}{2}, \infty)$, $B \in \mathcal{B}(X)$, and using the Change of Variable Theorem, one can obtain that

$$Q(B|a) = \int_B \Delta(u-a)\,\mathrm{d}u,$$

i. e., $\Delta(\cdot - a)$ is a density for $Q(\cdot|a)$ with respect to the Lebesgue measure on $\mathbb{R}$. Since $\Delta(\cdot)$ is continuous and taking into account Lemma 2.3 in [4], Q is strongly continuous. Then, considering $g(x) = x$ for $x \in X$, it is possible to verify that $V(g,x) < \infty$ (see the proof of Lemma 5.3 below). Thus, Assumption 2.1 holds.

To prove Condition 3.4 consider the following: take $x, y \in X$ with $x \leq y$, $a \in A(x)$ and $b \in A(y)$. To prove that $A(x) \sqsubseteq A(y)$, it is sufficient to consider the following three cases: $a \in \left[\frac{x}{2}, \frac{y}{2}\right)$, $a \in [\frac{y}{2}, b)$ or $a \in [b, \infty)$. If $a \in [\frac{y}{2}, b)$ or $a \in [b, \infty)$, then $a \wedge b \in A(y) \subset A(x)$ and $a \vee b \in A(y)$; if $a \in \left[\frac{x}{2}, \frac{y}{2}\right)$, then $a \wedge b = a \in A(x)$ and $a \vee b = b \in A(y)$. Since $A(x) \subset \mathbb{R}$ for $x \in X$, it follows that $x \to A(x)$ is ascending. One can also verify that $c(\cdot,\cdot)$ is subadditive.

Thus, there exists an increasing optimal policy for Example 5.1, and as $X = [0, \infty)$ is an interval in $\mathbb{R}$, Assumption 3.2 b) is true. □

Lemma 5.3. For Example 5.1, $v_{n+1}(x) = v_n(x)$ for all x, when $n = 1$ in Algorithm 4.1. Also, g_1 is an increasing optimal stationary policy because (11) is satisfied.

P r o o f . Set $n = 0$ and select $g_0(x) = x$, which is an increasing stationary policy. Now calculating the corresponding discounted cost:

$$v_0(x) := V(g_0, x) = E_x^{g_0}\left[\sum_{t=0}^{\infty} \alpha^t c(x_t, a_t)\right] = \sum_{t=0}^{\infty} \alpha^t\, E_x^{g_0}\left[c(x_t, a_t)\right],$$

using the last expression and the Change of Variable Theorem, it is obtained that:
for $t = 0$, $E_x^{g_0}[c(x_0, a_0)] = 2x$;
for $t = 1$,

$$\begin{aligned} E_x^{g_0}[c(x_1, a_1)] &= \int_{[0,\infty)} c(y, g_0(y))\, Q(\mathrm{d}y|g_0(x)) = \int_{[0,\infty)} 2y\, Q(\mathrm{d}y|x) \\ &= \int_{[0,\infty)} 2(x+s)\Delta(s)\,\mathrm{d}s = 2x + 2\mu; \end{aligned}$$

for $t = 2$,

$$E_x^{g_0}\left[c(x_2, a_2)\right] = \int_{[0,\infty)} \left(\int_{[0,\infty)} c(y, g_0(y))\, Q(\mathrm{d}y|g_0(z)) \right) Q(\mathrm{d}z|g_0(x))$$
$$= \int_{[0,\infty)} (2z + 2\mu)\, Q(\mathrm{d}z|x) = 2x + 4\mu.$$

In general, it is possible to prove that $E_x^{g_0}\left[c(x_k, a_k)\right] = 2x + 2k\mu$, for all $k = 0, 1, \ldots$. Thus,

$$v_0(x) = \sum_{t=0}^{\infty} \alpha^t\, E_x^{g_0}\left[c(x_t, a_t)\right] = \sum_{t=0}^{\infty} \alpha^t (2x + 2t\mu) = 2x\frac{1}{1-\alpha} + 2\mu \sum_{t=0}^{\infty} t\alpha^t$$
$$= 2x\frac{1}{1-\alpha} + 2\frac{\alpha}{(1-\alpha)^2}\mu.$$

Then, the next decision function g_1 must satisfy:

$$c(x, g_1(x)) + \alpha \int v_0(y)\, Q(\mathrm{d}y|g_1(x)) = \min_{a \in A(x)} \left[c(x, a) + \alpha \int v_0(y)\, Q(\mathrm{d}y|a) \right]$$

for all $x \in X$. Solving the second member of this equality, it is obtained that:

$$\min_{a \in A(x)} \left[c(x, a) + \alpha \int v_0(y)\, Q(\mathrm{d}y|a) \right]$$
$$= \min_{a \in [\frac{x}{2}, \infty)} \left[x + a + \alpha \int \left(2y\frac{1}{1-\alpha} + 2\mu\frac{\alpha}{(1-\alpha)^2} \right) Q(\mathrm{d}y|a) \right]$$
$$= \min_{a \in [\frac{x}{2}, \infty)} \left[x + a + 2\frac{\alpha^2}{(1-\alpha)^2}\mu + \frac{2\alpha}{1-\alpha} \int (a + s)\Delta(s)\, \mathrm{d}s \right]$$
$$= \min_{a \in [\frac{x}{2}, \infty)} \left[x + 2\frac{\alpha^2}{(1-\alpha)^2}\mu + \frac{2\alpha}{1-\alpha}\mu + \left(1 + \frac{2\alpha}{1-\alpha} \right) a \right].$$

Hence $g_1(x) = \frac{x}{2}$, $x \in X$. Calculating the corresponding discounted cost:

$$v_1(x) := V(g_1, x) = \sum_{t=0}^{\infty} \alpha^t\, E_x^{g_1}\left[c(x_t, a_t)\right],$$

it is obtained that:
for $t = 0$, $E_x^{g_1}[c(x_0, a_0)] = x + \frac{x}{2} = \frac{3}{2}x$;
for $t = 1$,

$$E_x^{g_1}[c(x_1, a_1)] = \int_{[0,\infty)} c(y, g_1(y))\, Q(\mathrm{d}y|g_1(x)) = \int_{[0,\infty)} \frac{3}{2} y\, Q\left(\mathrm{d}y|\frac{x}{2}\right)$$
$$= \frac{3}{2} \int_{[0,\infty)} \left(\frac{x}{2} + s \right) \Delta(s)\, \mathrm{d}s = \frac{3}{2^2}x + \frac{3}{2}\mu;$$

for $t = 2$,

$$E_x^{g_1}\left[c(x_2, a_2)\right] = \int_{[0,\infty)} \left(\int_{[0,\infty)} c(y, g_1(y))\, Q(\mathrm{d}y|g_1(z)) \right) Q(\mathrm{d}z|g_1(x))$$
$$= \int_{[0,\infty)} \left(\frac{3}{2^2} z + \frac{3}{2}\mu \right) Q\left(\mathrm{d}z|\frac{x}{2}\right) = \frac{3}{2^3}x + \left(\frac{3}{2^2} + \frac{3}{2}\right)\mu.$$

Using induction, it can be proved that

$$E_x^{g_1}\left[c(x_k, a_k)\right] = \frac{3}{2^{k+1}}x + 3\mu \sum_{i=1}^{k} \frac{1}{2^i} = \frac{3}{2^{k+1}}x + 3\mu\left(1 - \left(\frac{1}{2}\right)^k\right),$$

$k = 1, \ldots$. Thus,

$$v_1(x) = \sum_{t=0}^{\infty} \alpha^t\, E_x^{g_1}\left[c(x_t, a_t)\right] = \frac{3}{2}x\alpha^0 + \sum_{t=1}^{\infty} \alpha^t \left[\frac{3}{2^{t+1}}x + 3\mu\left(1 - \left(\frac{1}{2}\right)^t\right)\right]$$
$$= \frac{3}{2}x + 3\mu\left(\frac{1}{1-\alpha} - 1\right) + \left(\frac{3}{2}x - 3\mu\right) \sum_{t=1}^{\infty} \left(\frac{\alpha}{2}\right)^t$$
$$= \frac{3}{2}x + 3\mu\frac{\alpha}{1-\alpha} + \left(\frac{3}{2}x - 3\mu\right)\left(\frac{1}{1-\frac{\alpha}{2}} - 1\right)$$
$$= \frac{3}{2}\left(1 + \frac{\alpha}{2-\alpha}\right)x + 3\mu\alpha\left(\frac{1}{1-\alpha} - \frac{1}{2-\alpha}\right).$$

As $v_1(\cdot) \neq v_0(\cdot)$ substitute g_0 by g_1 and set $n = 1$. The next decision function g_2 must satisfy:

$$c(x, g_2(x)) + \alpha \int v_1(y)\, Q(\mathrm{d}y|g_2(x)) = \min_{a \in A(x)} \left[c(x, a) + \alpha \int v_1(y)\, Q(\mathrm{d}y|a) \right]$$

for all $x \in X$. Solving the second member of this equality, it is obtained that:

$$\min_{a \in A(x)} \left[c(x, a) + \alpha \int v_1(y)\, Q(\mathrm{d}y|a) \right]$$
$$= \min_{a \in [\frac{x}{2}, \infty)} \left[x + a + \alpha \int \left(\frac{3}{2-\alpha} y + 3\mu\alpha \left(\frac{1}{1-\alpha} - \frac{1}{2-\alpha} \right) \right) Q(\mathrm{d}y|a) \right]$$
$$= \min_{a \in [\frac{x}{2}, \infty)} \left[\left(1 + \frac{3\alpha}{2-\alpha}\right) a + x + \frac{3\alpha}{2-\alpha}\mu + 3\mu\alpha^2 \left(\frac{1}{1-\alpha} - \frac{1}{2-\alpha} \right) \right].$$

Hence $g_2(\cdot) = \frac{x}{2} = g_1(\cdot)$. Therefore, $v_2(\cdot) = v_1(\cdot)$.

Also, (11) holds by Remark 4.5 given that:

a) $c(x, a) = x + a \leq 2a + a = 3aw(x)$, where $w(x) = 1$ and $m = 3a$, and

b) $\int w(y)\, Q(\mathrm{d}y|a) = 1 = kw(x)$, for $k = 1 < \frac{1}{\alpha}$.

The proof is concluded by Lemma 5.2 and Theorem 4.2. □

6. REMARKS ON THE MONOTONICITY CONDITIONS AND MONOTONE VERSIONS OF THE PIA FOR MDPS WITH REWARDS

Analogously to what was done in the case of MDPs with costs, the conditions to ensure the existence of monotone optimal stationary policies and the monotone versions of the PIA for discounted MDPs with rewards, can be provided. This is discussed briefly in this section.

To do this, it is necessary to change c by r in the context of Subsection 2.2 until (2), and in (3) to consider the supremum instead of the infimum. Assumptions W and D in [11] are used instead of Assumption 2.1 to guarantee the existence of optimal stationary policies (Theorem 2 in [11]) for the corresponding discounted optimality equation.

Remark 6.1. a) Similar to Theorem 3.6 and using Condition 3.4 with superadditivity instead of subadditivity, it can be proved that there exists an increasing optimal stationary policy.

b) Similar to Theorem 3.9 and using Condition 3.7 with subadditivity instead of superadditivity, it can be proved that there exists a decreasing optimal stationary policy.

For obtaining monotone versions of the PIA, to consider the maximum instead of the minimum together with the changes proposed in the second paragraph of this section.

Remark 6.2. a) Similar to Algorithm 4.1, an increasing version of the PIA for MDPs with rewards can be established. Similar to Theorem 4.2 using Condition 3.4 and superadditivity instead of subadditivity, it is possible to prove that this algorithm yields a sequence of increasing stationary policies. Under additional assumptions (as in Theorem 4.2), an increasing optimal policy is obtained.

b) Similar to Algorithm 4.3, a decreasing version of the PIA for MDPs with rewards can be established. Similar to Theorem 4.4 using Condition 3.7 with subadditivity instead of superadditivity, it is possible to prove that this algorithm yields a sequence of decreasing stationary policies. Under additional assumptions (as in Theorem 4.4), a decreasing optimal policy is obtained.

c) Following the proof of Proposition 4.3.1 in [8] and observing that $\lim_{n\to\infty} \alpha^n E_x^\pi |V(\pi', x_n)| = 0$ implies that $\lim_{n\to\infty} \alpha^n E_x^\pi V(\pi', x_n) = 0$ for all $\pi, \pi' \in \Pi^0$ and $x \in X$, it is possible to show that Remark 4.5 is also valid for discounted MDPs with rewards by considering $\sup_{a\in A(x)} |r(x,a)|$ instead of $c(x,a)$ in a) of such remark.

d) It can be proved that for the elementary consumption-investment problem given in [8], pp. 37–38 (or for the problem of fisheries management in [9], Section 8.3), there exists an increasing optimal stationary policy. If $\Delta(\cdot)$ is continuous, then in the corresponding PIA, $v_2(x) = v_1(x)$ for all x, and g_1 is an increasing optimal stationary policy, by Remark 6.2 c).

In such example, $X = A = [0, \infty)$, and for each $x \in X$, $A(x) = [0, x]$. The dynamic of the system is given as in (1) by $x_{t+1} = a_t \cdot \xi_t$, for $t = 0, 1, \ldots$, with $S = [0, \infty)$,

$\mu = E[\xi] > 1$, and $0 < \alpha\mu < 1$. $r(x,a) = p(x-a)$, for $(x,a) \in \mathbb{K}$ and constant $p > 0$. See Remark 3.10 b) for a brief interpretation of this model.

7. CONCLUSIONS

The results explained in the previous sections permit to consider MDPs with costs and rewards that have finite, denumerable or non-denumerable state and decision spaces due to the conditions stated that guarantee the existence of a monotone (increasing or decreasing) optimal stationary policy do not require convexity assumptions. Regarding the monotone versions of the PIA given in this paper, it can be observed that when the initial stationary policy has a certain monotonicity, the policies obtained at each iteration have the same monotonicity. Thus, knowing that the optimal stationary policy has a certain monotonicity, the algorithms presented here permit one to search among the same types of policies instead of searching the largest set of all measurable policies.

ACKNOWLEDGEMENT

The author wishes to thank two anonymous referees for their helpful comments and suggestions, which were used to aid the improvement of this paper.

This work was supported in part by CONACYT (México) and AS CR (Czech Republic) under Grant No. 171396.

(Received June 8, 2012)

REFERENCES

[1] D. Assaf: Invariant problems in discounted dynamic programming. Adv. in Appl. Probab. *10* (1978), 472–490.

[2] N. Bäuerle and U. Rieder: Markov Decision Processes with Applications to Finance. Springer–Verlag, Berlin – Heidelberg 2011.

[3] D. P. Bertsekas: Dynamic Programming: Deterministic and Stochastic Models. Prentice Hall, New Jersey 1987.

[4] D. Cruz-Suárez, R. Montes-de-Oca and F. Salem-Silva: Conditions for the uniqueness of optimal policies of discounted Markov decision processes. Math. Methods Oper. Res. *60* (2004), 415–436.

[5] A. Dragut: Structured optimal policies for Markov decision processes: lattice programming techniques. In: Wiley Encyclopedia of Operations Research and Management Science (J. J. Cochran, ed.), John Wiley and Sons, 2010, pp. 1–25.

[6] D. Duffie: Security Markets. Academic Press, San Diego 1988.

[7] R. M. Flores-Hernández and R. Montes-de-Oca: Monotonicity of minimizers in optimization problems with applications to Markov control processes. Kybernetika *43* (2007), 347–368.

[8] O. Hernández-Lerma and J. B. Lasserre: Discrete-Time Markov Control Processes: Basic Optimality Criteria. Springer–Verlag, New York 1996.

[9] D. P. Heyman and M. J. Sobel: Stochastic Models in Operations Research, Vol. II. Stochastic Optimization. McGraw–Hill, New York 1984.

[10] A. Jaśkiewicz: A note on risk-sensitive control of invariant models. Syst. Control Lett. *56* (2007), 663–668.

[11] A. Jaśkiewicz and A. S. Nowak: Discounted dynamic programming with unbounded returns: application to economic models. J. Math. Anal. Appl. *378* (2011), 450–462.

[12] R. Mendelssohn and M. J. Sobel: Capital accumulation and the optimization of renewable resource models. J. Econom. Theory *23* (1980), 243–260.

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

[14] D. M. Topkis: Supermodularity and Complementarity. Princeton University Press, Princeton, New Jersey 1998.

Rosa María Flores-Hernández, Universidad Autónoma de Tlaxcala, Facultad de Ciencias Básicas, Ingeniería y Tecnología. Calz. Apizaquito s/n, Km. 1.5, Apizaco, Tlaxcala 90300. México.
e-mail: rosam@xanum.uam.mx