Raúl Montes-de-Oca; Enrique Lemus-Rodríguez; Daniel Cruz-Suárez

A stopping rule for discounted Markov decision processes with finite action sets

Kybernetika, Vol. 45 (2009), No. 5, 755--767

Persistent URL: `http://dml.cz/dmlcz/140043`

A STOPPING RULE FOR DISCOUNTED MARKOV DECISION PROCESSES WITH FINITE ACTION SETS

Raúl Montes-de-Oca,* Enrique Lemus-Rodríguez
and Daniel Cruz-Suárez

In a Discounted Markov Decision Process (DMDP) with finite action sets the Value Iteration Algorithm, under suitable conditions, leads to an optimal policy in a finite number of steps. Determining an upper bound on the necessary number of steps till gaining convergence is an issue of great theoretical and practical interest as it would provide a computationally feasible stopping rule for value iteration as an algorithm for finding an optimal policy. In this paper we find such a bound depending only on structural properties of the Markov Decision Process, under mild standard conditions and an additional "individuality" condition, which is of interest in its own. It should be mentioned that other authors find such kind of constants using non-structural information, i. e., information not immediately apparent from the Decision Process itself. The DMDP is required to fulfill an ergodicity condition and the corresponding ergodicity index plays a critical role in the upper bound.

Keywords: Markov decision process, ergodicity condition, value iteration, discounted cost, optimal policy, myopic policies

AMS Subject Classification: 90C40, 93E20

1. INTRODUCTION

In recent research it has been possible to establish that, under suitable conditions (in particular, uniqueness of the optimal policy), the value iteration procedure produces a sequence of policies that converges to the optimal policy uniformly over compact sets (see [1] for the uniform on compact sets convergence of the value iteration policies to the optimal policy, and [3] for the detection of a value iteration policy which is a uniform on compact sets ε-approximation to the optimal policy). This poses the problem of the rate of convergence, and an interesting twist to this problem is observed when in some applications (not at all uncommon) the action sets are finite (due to design decisions or to discretization procedures): in such a context convergence could be observed in a finite number of steps, and Value Iteration would then become a candidate for an algorithm that computes the optimal policy. In

[*] Correspondence Author

this case, the issue of when stopping the algorithm becomes relevant. Even if a stopping condition is available (in this case, the optimality equation, for instance), it is important to have upper bounds on how long it will take for the procedure to find the solution. Studying an "Individuality" Condition (IC) on the actions −to be precisely described later – in a sense enforces that different actions are "really" different from each other. So it is possible, not only to observe the convergence to the optimal policy, but to find the desired lower and upper bounds on the number of iterations under an Ergodicity Condition (in particular, there will be assumed the existence of an ergodicity index λ). Let us stress that if the ergodicity index λ is known, such a bound is determined directly by the structure of the Markov Decision Process (MDP), in particular, it is not necessary to express it depending on other constants indirectly determined by the MDP, as in some cases.

There are two bonuses in this context worth mentioning. Firstly, there exists an interesting relationship between the "individuality" condition and the uniqueness of optimal policies, which vouches for its appropriateness. Secondly, the knowledge of both lower and upper bounds on the number of steps till gaining convergence allows us to determine if myopic optimal policies (i. e. optimal policies which also are minimizers of the first step of the value iteration algorithm) are absent from particular decision processes. This is of interest, as many classical examples do admit myopic optimal policies.

As both the lower and the upper bounds are determining by the roots of some polynomials whose degrees depend on the structure of the Decision Process, it is apparent that these degrees somehow reflect the complexity of the optimal policy.

Now, the structure of the paper will be outlined. In Section 2 the Discounted Markov Decision Processes are defined. In Section 3 the conditions used in the paper are stated, devoting a little time to the "individuality" condition, commenting on its nice relationship with respect to the uniqueness of optimal policies. In Section 4 all the necessary technical lemmas are established. In Section 5 the stopping rule is found through the main theorem, that is stated and proved, providing the lower and upper bounds on number of steps necessary to convergence through Value Iteration in terms of the ergodicity index λ and structural information (i. e. information readily available from the Decision Process). In Section 6 some illustrative examples are presented, where the bounds are used as a tool to determine if a model admits myopic policies, or such policies are to be ruled out. In the last Section several open problems in this area are outlined.

2. DISCOUNTED MARKOV DECISION PROCESSES

In this paper a quite standard Markov Decision Process setting will be used, and referred to as the Decision Model henceforth:

Decision Model. Let $(X, A, \{A(x) : x \in X\}, Q, c)$ be the usual discrete-time Markov decision model (see [4] and [6]), where the state space X is a Borel space, and the control space A is a *finite* set. For each $x \in X$, $A(x) \subset A$ is the subset of admissible actions at a state x, the set $\mathbb{K} = \{(x, a) : x \in X, a \in A(x)\}$ is assumed to be a Borel subset of $X \times A$. Consider the transition probability law $Q(B|x, a)$, where $B \in \mathbb{B}(X)$ ($\mathbb{B}(X)$ denotes the Borel σ-algebra of X) and $(x, a) \in \mathbb{K}$ is a stochastic kernel on X, given $\mathbb{K}$ (i. e. $Q(\cdot|x, a)$ is a probability measure on X for every $(x, a) \in \mathbb{K}$, and $Q(B|\cdot)$ is a measurable function on $\mathbb{K}$ for every $B \in \mathbb{B}(X)$). Finally, the cost per stage c is a nonnegative, upper bounded (by a bound denoted by M), and measurable function on $\mathbb{K}$.

Policies. A *control policy* π is a (measurable, possibly randomized) rule for choosing actions, and at each time t ($t = 0, 1, \ldots$) the control prescribed by π may depend on the current state as well as on the history of the previous states and actions. The set of all policies will be denoted by Π. Given the initial state $x \in X$, any policy π defines the unique probability distribution of the state-action process (x_t, a_t) , for details see [4] and [6]. This distribution will be denoted by P_x^π, while E_x^π stands for the corresponding expectation operator, and the stochastic process $\{x_t\}$ will be called *Markov decision process* (MDP). $\mathbb{F}$ denotes the set of measurable functions $f : X \to A$ such that $f(x) \in A(x)$ for all $x \in X$. A policy π is *stationary* if there exists $f \in \mathbb{F}$ such that, under π, the action $f(x_t)$ is applied at each time t. The class of stationary policies is naturally considered as $\mathbb{F}$.

Optimality Criterion. Given $\pi \in \Pi$ and initial state $x_0 = x \in X$, let

$$V(\pi, x) = E_x^\pi \left(\sum_{t=0}^{\infty} \alpha^t c(x_t, a_t) \right) \tag{1}$$

be the *total expected discounted cost.* The number $\alpha \in (0, 1)$ is called the *discount factor.*

A policy π^* is said to be *optimal* if $V(\pi^*, x) = V^*(x)$ for all $x \in X$, where

$$V^*(x) = \inf_\pi V(\pi, x), \tag{2}$$

$x \in X$. V^* defined in (2) is called the *optimal value function.*

An MDP with the total expected discounted cost as the optimality criterion will be referred to as a *discounted* MDP.

3. ASSUMPTIONS AND CONDITIONS

In this section the conditions on the Decision Model are listed and discussed.

Let $\Phi(X)$ be the (Banach) space of the bounded measurable function endowed with the supremum norm $\|u\| = \sup_{x \in X} |u(x)|$. For every $u \in \Phi(X)$, Tu is the function on X defined as

$$Tu(x) = \min_{a \in A(x)} \left[c(x, a) + \alpha \int u(y)\, Q(\mathrm{d}y|x, a) \right], \quad x \in X. \tag{3}$$

Remark 3.1. As c is bounded, it is well-known that T, defined in (3), is a contraction, i. e.

$$\|Tu - Tu'\| \le \alpha \|u - u'\|, \tag{4}$$

for all $u, u' \in \Phi(X)$ (see Note 4 p. 52 in [5]).

Hence, by Banach's Fixed-Point Theorem, V^* is a fixed point of T, i. e. V^* satisfies the Optimality Equation (OE):

$$V^*(x) = \min_{a \in A(x)} \left[c(x,a) + \alpha \int V^*(y)\, Q(\mathrm{d}y|x,a) \right], \tag{5}$$

for all $x \in X$.

Also, this Theorem allows to obtain that the Value Iteration Algorithm is valid. That is the value iteration functions defined as

$$v_n(x) = \min_{a \in A(x)} \left[c(x,a) + \alpha \int v_{n-1}(y)\, Q(\mathrm{d}y|x,a) \right], \tag{6}$$

$x \in X$ and $n = 1, 2, \ldots$, with $v_0 = 0$, are well-defined, and for each $x \in X$, $v_n(x) \to V^*(x)$.

The Basic Assumptions are fulfilled on a wide variety of cases, see [4, 5, 6] and [8] and deserve no further comment.

Basic Assumptions (BA).

(**a**) There is $f^* \in \mathbb{F}$ such that

$$V^*(x) = c(x, f^*(x)) + \alpha \int V^*(y)\, Q(\mathrm{d}y|x, f^*(x)), \tag{7}$$

$x \in X$, and f^* is optimal.

(**b**) For each $n = 1, 2, \ldots$, there exists $f_n \in \mathbb{F}$ such that, for each $x \in X$,

$$v_n(x) = c(x, f_n(x)) + \alpha \int v_{n-1}(y)\, Q(\mathrm{d}y|x, f_n(x)). \tag{8}$$

Ergodicity Condition (EC). There exists a number $\lambda \in (0,1)$ such that

$$\sup_{k,k'} \|Q(\cdot|k) - Q(\cdot|k')\|_v \leq 2\lambda, \tag{9}$$

where the sup is over all $k, k' \in \mathbb{K}$, $\|\cdot\|_v$ denotes the variation norm for signed measures, and λ will be call in this paper the *ergodicity index*.

This Ergodicity Condition has been discussed extensively in [4] pp. 56–60. There the reader can find three alternative necessary conditions for *EC*, much easier to verify. For instance, the fact that there exists a state x^* which for any state and action has a probability greater than a positive constant is very useful in this paper, since it implies *EC*. It may be asked if this condition is relevant in the discounted case. As a matter of fact, it is: see [7], Section 6.8 where it is used to find an upper bound to a difference of integrals in terms of the span of a bounded function and the ergodicity index. By the way, inequality (6.8.4) and the fact that it uses non-structural information on the MDP partially inspired the present paper.

Before stating the next condition, some notation will be provided: for each $x \in X$ and $a^* \in A(x)$, let

$$K_x(a^*) = \min_{a \in A(x)-\{a^*\}} |c(x,a) - c(x,a^*)| - \frac{\alpha\lambda M}{1-\alpha}, \tag{10}$$

where λ is given in (9), and let

$$\begin{aligned} K_x^* &= \min_{a^* \in A(x)} K_x(a^*) \\ &= \min_{a^* \in A(x)} \left\{ \min_{a \in A(x)-\{a^*\}} |c(x,a) - c(x,a^*)| \right\} - \frac{\alpha\lambda M}{1-\alpha} \\ &= D_x - \frac{\alpha\lambda M}{1-\alpha}, \end{aligned} \tag{11}$$

where $D_x := \min_{a^* \in A(x)} \left\{ \min_{a \in A(x)-\{a^*\}} |c(x,a) - c(x,a^*)| \right\}$.

Individuality Condition (IC). For each $x \in X$, $K_x^* > 0$.

An ancestor of the Individuality Condition is discussed in [9], p. 63 (inside the proof of Lemma 3.7 when defining δ). Intuitively, this condition means that, given a state x, different actions have different costs to at least a minimum uniform degree. Lately it will be seen how naturally this condition is in relation to the uniqueness of the optimal policy(see [2] for some conditions which imply the uniqueness of optimal policies in Discounted MDPs).

Remark 3.2. Without further mention, throughout the rest of the paper MDPs which satisfy the definitions given in Section 2 and the Assumptions and Conditions given in Section 3 will be taken into account. In particular, these Assumptions and Conditions will not be mentioned in each Lemma, Theorem, or Example, but are supposed to hold.

4. PRELIMINARY LEMMAS

Before stating and proving the main result, some Lemmas will be dealt with.

Lemma 4.1.

$$\|V^* - v_n\| = \sup_{x \in X} |V^*(x) - v_n(x)| \leq \frac{\alpha^n M}{1-\alpha}, \tag{12}$$

for all $n \geq 1$.

Proof. Inequality (12) is a direct consequence of the fact that T, defined in (3), is a contraction (see Remark 3.1). □

Let

$$G(x,a) := c(x,a) + \alpha \int V^*(y)\, Q(\mathrm{d}y|x,a), \tag{13}$$

$(x,a) \in \mathbb{K}$ (observe that $G(x, f^*(x)) = V^*(x), x \in X$).

Lemma 4.2. For each $x \in X$ and $n = 1, 2, \dots,$

$$|G(x, f_n(x)) - G(x, f^*(x))| \le \frac{2M\alpha^n}{1-\alpha}. \tag{14}$$

Proof. Let $x \in X$ and $n \ge 1$ be fixed.
Then, using (8) and Lemma 4.1, it follows that

$$\begin{aligned} |G(x, f_n(x)) - G(x, f^*(x))| &= |G(x, f_n(x)) - V^*(x)| \\ &= |c(x, f_n(x)) + \alpha \int V^*(y)\, Q(\mathrm{d}y|x, f_n(x)) \\ &\quad + v_n(x) - v_n(x) - V^*(x)| \\ &\le |\alpha \int (V^*(y) - v_{n-1}(y))\, Q(\mathrm{d}y|x, f_n(x)) \\ &\quad + (v_n(x) - V^*(x))| \\ &\le \alpha \int |V^*(y) - v_{n-1}(y)| Q(\mathrm{d}y|x, f_n(x)) \\ &\quad + |v_n(x) - V^*(x)| \\ &\le \frac{M\alpha\alpha^{n-1}}{1-\alpha} + \frac{M\alpha^n}{1-\alpha} \\ &= \frac{2M\alpha^n}{1-\alpha}. \end{aligned}$$

This finalizes the proof of Lemma 4.2. □

The *span* of a function $\Psi \in \Phi(X)$ is defined by

$$\mathrm{sp}(\Psi) := \sup_{x \in X} \Psi(x) - \inf_{x \in X} \Psi(x). \tag{15}$$

The following lemma is crucial to the main argument of the paper and provides a tool to integral comparison through the ergodicity index.

Lemma 4.3. For any (x,a) and (x',a') in $\mathbb{K}$ and any $\Psi \in \Phi(X)$,

$$\int \Psi(y)\, Q(\mathrm{d}y|x,a) - \int \Psi(y)\, Q(\mathrm{d}y|x',a') \le \lambda\, \mathrm{sp}(\Psi). \tag{16}$$

Proof. The proof of this lemma is a part of the proof of Lemma 3.5, p. 59 in [4].

□

For each $x \in X$ and $a^* \in A(x)$, let $H_{x,a^*} : A(x) - \{a^*\} \to \mathbb{R}$, given by

$$H_{x,a^*}(a) = |G(x,a) - G(x,a^*)|, \tag{17}$$

$a \in A(x) - \{a^*\}$.

Lemma 4.4. For each $x \in X$,

$$\min_{a^* \in A(x)} \left\{ \min_{a \in A(x)-\{a^*\}} H_{x,a^*}(a) \right\} \geq K_x^*. \tag{18}$$

Proof. In the proof, the basic inequalities (19), (20), and (21) will be used:

$$|z - w| \geq |z| - |w|, \tag{19}$$

for all $z, w \in \mathbb{R}$,

$$-\mathrm{sp}(V^*) \geq \frac{-M}{1-\alpha}, \tag{20}$$

and

$$\left| \int V^*(y)\, Q(\mathrm{d}y|x,a) - \int V^*(y)\, Q(\mathrm{d}y|x',a') \right| \leq \lambda\, \mathrm{sp}(V^*), \tag{21}$$

$(x,a),(x',a') \in \mathbb{K}$. (Observe that (21) is a direct consequence of Lemma 4.3 with $\Psi = V^*$.)

Let $x \in X$ and $a^* \in A(x)$. Take $a \in A(x) - \{a^*\}$. Then, from (19), (20) and (21) it follows that

$$\begin{aligned} H_{x,a^*}(a) &= |G(x,a) - G(x,a^*)| \\ &\geq |c(x,a) - c(x,a^*)| - \alpha \left| \int V^*(y)\, Q(\mathrm{d}y|x,a^*) \right. \\ &\quad \left. - \int V^*(y)\, Q(\mathrm{d}y|x,a) \right| \\ &\geq |c(x,a) - c(x,a^*)| - \alpha\lambda\, \mathrm{sp}(V^*) \\ &\geq |c(x,a) - c(x,a^*)| - \frac{\alpha\lambda M}{1-\alpha} \\ &\geq \min_{a \in A(x)-\{a^*\}} |c(x,a) - c(x,a^*)| - \frac{\alpha\lambda M}{1-\alpha} \\ &= K_x(a^*). \end{aligned}$$

Therefore,

$$\min_{a\in A(x)-\{a^*\}} H_{x,a^*}(a) \geq K_x(a^*), \tag{22}$$

and

$$\min_{a^*\in A(x)} \left\{ \min_{a\in A(x)-\{a^*\}} H_{x,a^*}(a) \right\} \geq \min_{a^*\in A(x)} K_x(a^*) = K_x^*. \tag{23}$$

This is the end of the proof of Lemma 4.4. □

Remark 4.5. From Lemma 4.4 it results that for each $x \in X$,

$$\begin{aligned} & \min_{a\in A(x)-\{f^*(x)\}} H_{x,f^*(x)}(a) \\ \geq \quad & \min_{a^*\in A(x)} \left\{ \min_{a\in A(x)-\{a^*\}} H_{x,a^*}(a) \right\} \geq K_x^*, \end{aligned} \tag{24}$$

i. e., for each $x \in X$,

$$\min_{a\neq f^*(x)} |G(x,a) - G(x,f^*(x))| \geq K_x^*. \tag{25}$$

Moreover, notice that (7), (13), (25), and IC imply the uniqueness of f^*.

5. MAIN RESULT

The previous lemmas finally allow to upper bound the number $N(x)$ of steps needed to find the optimal rule $f^*(x)$, and hence the stopping rule for a state x. In particular, if $A(x)$ and $c(x,a)$ do not depend on x, or if X is finite, $N(x)$ will not depend on x and the stopping rule will be uniform on X.

Let d be the discrete metric on A , i. e. $d(a,a') = 0$ if $a = a'$, and $d(a,a') = 1$ if $a \neq a'$.

Theorem 5.1. For each $x \in X$ and the $N(x)$th step of the value iteration procedure, $f_{N(x)}(x) = f^*(x)$, where $N(x) = [(\ln((1-\alpha)K_x^*)/2M))/\ln\alpha] + 1$, and $[z]$ is the integer part of z.

Proof. Let x be a fixed state. Firstly, note that

$$d(f^*(x), f^*(x)) = 0 = \frac{1}{K_x^*}|G(x,f^*(x)) - G(x,f^*(x))|. \tag{26}$$

(Recall that K_x^* is assumed to be positive.) On the other hand, using (25), for each $a \in A(x)$ with $a \neq f^*(x)$,

$$\begin{aligned} d(a, f^*(x)) &= 1 \\ &= \frac{\min_{a\neq f^*(x)}|G(x,a) - G(x,f^*(x))|}{\min_{a\neq f^*(x)}|G(x,a) - G(x,f^*(x))|} \\ &\leq \frac{1}{K_x^*}|G(x,a) - G(x,f^*(x))|. \end{aligned}$$

Hence,

$$d(a, f^*(x)) \leq \frac{1}{K_x^*}|G(x,a) - G(x, f^*(x))|, \tag{27}$$

for all $a \in A(x)$.

Now, letting $a = f_n(x)$ in (27) and using Lemma 4.2 it results that

$$\begin{aligned} d(f_n(x), f^*(x)) &\leq \frac{1}{K_x^*}|G(x, f_n(x)) - G(x, f^*(x))| \\ &\leq \frac{1}{K_x^*}\frac{2M\alpha^n}{(1-\alpha)}, \end{aligned}$$

for each $n \geq 1$. Finally, solving the inequality

$$\frac{1}{K_x^*}\frac{2M\alpha^n}{(1-\alpha)} < 1, \tag{28}$$

it follows that

$$n > \frac{\ln\{(1-\alpha)K_x^*/2M\}}{\ln\alpha}. \tag{29}$$

Take

$$N(x) := \left[\frac{\ln\{(1-\alpha)K_x^*/2M\}}{\ln\alpha}\right] + 1. \tag{30}$$

Therefore, $f_{N(x)}(x) = f^*(x)$.

Since x is arbitrary, Theorem 5.1 follows. □

Remark 5.2. It is not clear that $N(x)$ is always positive: observe that

$$\begin{aligned} 0 < (1-\alpha)D_x - \alpha\lambda M &= (1-\alpha)\min_{a^*\in A(x)}\left\{\min_{a\in A(x)-\{a^*\}}|c(x,a) - c(x,a^*)|\right\} \\ &\quad -\alpha\lambda M \\ &< (1-\alpha)2M - \alpha\lambda M \\ &= 2M - 2\alpha M - \alpha\lambda M \\ &< 2M, \end{aligned}$$

hence

$$0 < \frac{(1-\alpha)D_x - \alpha\lambda M}{2M} < 1, \tag{31}$$

and, using (11), it results that

$$\ln\{(1-\alpha)K_x^*/2M\} = \ln\{((1-\alpha)D_{x^-}\alpha\lambda M)/2M\} < 0. \tag{32}$$

Then, as $\ln\alpha < 0$,

$$\frac{\ln\{(1-\alpha)K_x^*/2M\}}{\ln\alpha} > 0. \tag{33}$$

Remark 5.3. From (30) it is possible to find conditions which allow to obtain the optimal policy f^* in k steps, $k = 1, 2, \ldots$. For instance, from (11), (30), and (33),

$$\begin{aligned} 0 \quad &< \quad \frac{\ln\{(1-\alpha)K_x^*/2M\}}{\ln \alpha} \\ &= \quad \frac{\ln\{((1-\alpha)D_x \text{-} \alpha\lambda M)/2M\}}{\ln \alpha} < 1 \end{aligned}$$

if and only if

$$(D_x + \lambda M + 2M)\alpha - D_x < 0. \tag{34}$$

(Recall that $M > 0, \lambda, \alpha \in (0, 1)$, and, from IC, $D_x > 0$.)

Hence, $f_1(x) = f^*(x)$ if and only if $(D_x + \lambda M + 2M)\alpha - D_x < 0$. Similarly, $f_2(x) = f^*(x)$ if and only if

$$1 \leq \frac{\ln\{((1-\alpha)D_x \text{-} \alpha\lambda M)/2M\}}{\ln \alpha} < 2, \tag{35}$$

that is, if and only if

$$(D_x + \lambda M + 2M)\alpha - D_x \geq 0, \tag{36}$$

and

$$2M\alpha^2 + (D_x + \lambda M)\alpha - D_x < 0. \tag{37}$$

For the rest of the cases, i. e., $f_k(x) = f^*(x)$, $k \geq 3$ holds, if and only if

$$2M\alpha^{k-1} + (D_x + \lambda M)\alpha - D_x \geq 0, \tag{38}$$

and

$$2M\alpha^k + (D_x + \lambda M)\alpha - D_x < 0. \tag{39}$$

Now, three examples will be presented. For each of them, it is not difficult to verify that BA hold (for instance, in each case, it is possible to verify that the assumptions needed in [6], pp. 44–45 hold).

6. EXAMPLES

Example 6.1. Let $X = \{0, 1\}$, and for $x \in X$, $A(x) = \{1, 2\}$. The transition law is given by $Q(\{0\}|0, 1) = p$, $Q(\{1\}|0, 1) = q$, $Q(\{0\}|1, 1) = r$, $Q(\{1\}|1, 1) = s$, $Q(\{0\}|0, 2) = p_1$, $Q(\{1\}|0, 2) = q_1$, $Q(\{0\}|1, 2) = r_1$, $Q(\{1\}|1, 2) = s_1$. The cost function is given by $c(0, 1) = \theta_1$, $c(0, 2) = \theta_2$, $c(1, 1) = \theta_3$, $c(1, 2) = \theta_4$.

Assumption 6.2.

(a) $p, q, r, s, p_1, q_1, r_1$ and s_1 are positive, $p+q = 1, r+s = 1, p_1+q_1 = 1, r_1+s_1 = 1$, and $p = \min\{p, q, r, s, p_1, q_1, r_1, s_1\}$.

(b) $\theta_1, \theta_2, \theta_3, \theta_4, \theta_2 - \theta_1$ and $\theta_4 - \theta_3$ are positive, and $\theta_4 = \max\{\theta_1, \theta_2, \theta_3, \theta_4\}$.

Lemma 6.3. Under Assumption 6.2,

$$N(0) = [\{\ln((1-\alpha)(\theta_2-\theta_1)-\alpha((2-p)/2)\theta_4)/2\theta_4\}/\ln\alpha]+1, \tag{40}$$

and

$$N(1) = [\{\ln((1-\alpha)(\theta_4-\theta_3)-\alpha((2-p)/2)\theta_4)/2\theta_4\}/\ln\alpha]+1. \tag{41}$$

Proof. Note that from Assumption 6.2 (b), $M = \theta_4$, $D_0 = \theta_2 - \theta_1$, and $D_1 = \theta_4 - \theta_3$. Now, observe that from Assumption 6.2 (a),

$$Q(\{0\}|x,a) \geq p > 0, \tag{42}$$

for all $(x,a) \in \mathbb{K}$. Hence, Condition 3.1 (1) in [4] p. 56 holds, and for Lemma 3.3 in [4], the *EC* holds as well, with $\lambda = (2-p)/2$ (see also the proof of Lemma 3.3 in [4]). Therefore, (11) and (30) imply (40), and (41). □

Now, a specific example to illustrate Example 6.1 will be presented.

Example 6.4. Consider Example 6.1 with $p = q = r = s = p_1 = q_1 = r_1 = s_1 = 1/2,\ \theta_1 = \theta_3 = 1,\ \theta_2 = \theta_4 = 20$. Here, $D_x = 19, x \in X,\ M = \theta_4 = 20, \lambda = 3/4$, and

$$N(x) = [\{\ln\{19(1-\alpha)-15\alpha\}/40\}/\ln\alpha]+1, \tag{43}$$

$x \in X$. A straightforward computation implies by the IC that $\alpha < 0.55882\ldots$. Then the next array (constructed with the help of the inequalities provided in Remark 5.3 solved in terms of α) shows the range of values of α which obtain the optimal policy f^* in k steps, $k = 1, 2, \ldots, 10$. (For instance, using (36) and (37), $f^*(x) = f_2(x)$, $x \in X$ if $0.256757 \leq \alpha < 0.384707$; observe that the array is the same for both $x = 0$ and $x = 1$.)

k	α
1	$(0, 0.256757)$
2	$[0.256757, 0.384707)$
3	$[0.384707, 0.450943)$
4	$[0.450943, 0.490645)$
5	$[0.490645, 0.51585)$
6	$[0.51585, 0.532117)$
7	$[0.532117, 0.542544)$
8	$[0.542544, 0.549101)$
9	$[0.549101, 0.553122)$
10	$[0.553122, 0.55553)$

Example 6.5. Let η and n be a positive constant and a positive integer, respectively. Take $X = [0, \eta]$, $A = A(x) = \{\eta + 1, \eta + 2, \ldots, \eta + n\}$, for all $x \in X$. The transition law is given by $x_{t+1} = \min\{[x_t + a_t - \xi_t]^+, \eta\}$, $t = 0, 1, 2, \ldots$. Here $[z]^+ = \max\{0, z\}$, and $\xi_0, \xi_1, \ldots$ are i.i.d. random variables taking values in $S = [0, +\infty)$. Let ξ be a generic element of the sequence $\{\xi_t\}$. The cost function is given by $c(x, \eta + k) = g(x) + (\eta + k)$, $x \in X$, $k = 1, \ldots, n$, where $g : X \to \mathbb{R}$ is a nonnegative, upper bounded measurable function. Let φ be a bound for g.

Assumption 6.6.

$$P\{\xi \geq 2\eta + n\} := \beta > 0. \tag{44}$$

Lemma 6.7. Under Assumption 6.6, for each $x \in X$,

$$N(x) = [(\ln((1 - \alpha) - \alpha((2 - \beta)/2)\varphi)/2\varphi)/\ln \alpha] + 1. \tag{45}$$

Proof. It is easy to obtain that $M = \varphi$ and $D_x = 1, x \in X$. From Assumption 6.6, for each $(x, a) \in \mathbb{K}$,

$$\begin{aligned} Q(\{0\}|x, a) &= P\{x_{t+1} = 0 | x_t = x, a_t = a\} \\ &= P\{\xi \geq x + a\} \\ &\geq P\{\xi \geq 2\eta + n\} \\ &= \beta. \end{aligned}$$

Hence, as in the proof of Lemma 6.3, it results that $\lambda = (2 - \beta)/2$. Therefore, from (11) and (30), (45) follows. □

7. CONCLUSIONS

It seems apparent that the *BA*, *EC* and *IC* allow a clean analysis of the problem when the action sets are finite. The Value Iteration procedure plays a significant role all through the analysis. There is one immediate question: would Policy Iteration provide extra information to this problem?

The polynomials in α appearing in (38) and (39) raise another question: how do they describe the "complexity" of the optimal policy? It would be very interesting to determine if they could be turned into a useful tool to determine whether a DMDP admits myopic policies or not.

Finally, if we revert to non-finite action sets, could we apply this scheme to the number of steps to enter a ball of given radius centered on the optimal policy? Besides it is open to determine how similar bounds are to be found when dealing with MDPs not satisfying *EC*.

Hopefully, these questions will be addressed to in forthcoming papers.

ACKNOWLEDGEMENT

The authors heartily thank all the referees for their comments and encouragement to further research on "the difficult problem of identifying the optimal action in discounted Markov decision chain".

(Received June 25, 2008.))

REFERENCES

[1] D. Cruz-Suárez and R. Montes-de-Oca: Uniform convergence of the value iteration policies for discounted Markov decision processes. Bol. Soc. Mat. Mexicana *12* (2006), 133–148.

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

[3] D. Cruz-Suárez, R. Montes-de-Oca, and F. Salem-Silva: Uniform approximations of discounted Markov decision processes to optimal policies. In: Proc. Prague Stochastics 2006 (M. Hušková and M. Janžura, eds.), MATFYZPRESS, Prague 2006, pp. 278–287.

[4] O. Hernández-Lerma: Adaptive Markov Control Processes Springer-Verlag, New York 1989.

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

[6] O. Hernández-Lerma and J. B. Lasserre: Further Topics on Discrete–Time Markov Control Processes. Springer-Verlag, New York 1999.

[7] M. L. Puterman: Markov Decision Processes. Discrete Stochastic Dynamic Programming. Wiley, New York 1994.

[8] R. Ritt and L. Sennott: Optimal stationary policies in general state Markov decision chains with finite action sets. Math. Oper. Res. *17* (1992), 901–909.

[9] N. L. Stokey and R. E. Lucas: Recursive Methods in Economic Dynamics. Harvard University Press, USA 1989.

Raúl Montes-de-Oca, Departamento de Matemáticas, Universidad Autónoma Metropolitana-Iztapalapa, Av. San Rafael Atlixco 186, Col. Vicentina, 09340 México, D.F. México.
e-mail: momr@xanum.uam.mx

Enrique Lemus-Rodríguez, Universidad Anáhuac México-Norte, Av. Lomas Anáhuac s/n, 52786 Huixquilucan, Edo. de México. México.
e-mail: elemus@anahuac.mx

Daniel Cruz-Suárez, División Académica de Ciencias Básicas, Universidad Juárez Autónoma de Tabasco, Apdo. Postal 5, 86690 Cunduacán, Tab. México.
e-mail: daniel.cruz@basicas.ujat.mx