Rolando Cavazos-Cadena

Weak conditions for the existence of optimal stationary policies in average Markov decision chains with unbounded costs

Kybernetika, Vol. 25 (1989), No. 3, 145--156

Persistent URL: http://dml.cz/dmlcz/125085

KYBERNETIKA — VOLUME 25 (1989), NUMBER 3

WEAK CONDITIONS FOR THE EXISTENCE OF OPTIMAL STATIONARY POLICIES IN AVERAGE MARKOV DECISION CHAINS WITH UNBOUNDED COSTS

ROLANDO CAVAZOS-CADENA

Average cost Markov decision chains with discrete time parameter are considered. The cost function is unbounded and satisfies an additional condition which frequently holds in applications. Also, we assume that there exists a *single* stationary policy for which the corresponding Markov chain is irreducible and ergodic with finite average cost. Within this framework, the existence of an average cost optimal stationary policy is proved.

1. INTRODUCTION

We are concerned with average cost Markov decision processes (MDP's) with denumerable state space and discrete time parameter. For these models, the problem of determining an optimal (stationary) policy has received considerable attention in the MDP literature; see, for instance, Thomas [4] and the references therein. Usually, searching for optimal policies is based on a *bounded* solution to the (average cost) optimality equation (*OE*). Indeed, under standard continuity-compactness conditions such a solution to the *OE* yields immediately an optimal stationary policy. Many *sufficient* conditions for the existence of a bounded solution to the *OE* have been given [4]. However, they impose strong restrictions on the model since (typically) the state space must be an irreducible ergodic class under *any* stationary policy, a restriction that is not satisfied in most interesting queuing/inventory problems; such is the case in Nain and Ross [9] and the inner references. Moreover, it was proved in Cavazos-Cadena [10, 11] that those restrictive conditions are *necessary* for the existence of a bounded solution of the *OE*. Thus, an approach to obtain optimal stationary policies relying on a bounded solution to the *OE* has, naturally, a limited scope.

On the other hand, most interesting models with infinite state space arising in applications have an unbounded (nonnegative) cost function. At this point, the so-called Lyapounov function condition could be used to obtain optimal stationary

policies (Hordijk [1]). However, under this condition, the Markov chain associated with *any* stationary policy must be ergodic, and again this is too restrictive for practical cases.

Recently, Sennott [5, 6, 7] has obtained optimal stationary policies under very general conditions. To obtain her results, she follows the usual approach of examining the average case as limit of discounted cases, but she does it more efficiently. This note goes on the way traced by Sennott.

Our main result can be (roughly) described as follows; We prove the existence of an average cost optimal stationary policy under the following assumptions: (i) The cost function is nonnegative and, for each $r \geqq 0$, the cost function is greater than r outside a finite set; see Assumption 1.2 below, and (ii) There exists a *single* stationary policy inducing an irreducible Markov chain with finite average cost.

Now, we describe formally the model we will be dealing with.

The Model. Let (S, A, C, p) be the usual MDP where the state space S is an infinite denumerable set, while the *metric space* A is the action set. With each $x \in S$, a nonempty set $A(x) \subset A$ is associated; $A(x)$ is the set of admissible actions at state x.

We assume that each $A(x)$ is a *compact* subspace of A. On the other hand, C is the cost function and p is the transition law. The system evolves as follows: At each time $t \in \mathbb{N} := \{0, 1, 2, \ldots\}$ the state of the system is observed, say $x \in S$, and an action $a \in A(x)$ is selected. Then, a cost $C(x, a)$ is incurred and the state of the system at time $t + 1$ will be $y \in S$ with probability $p_{xy}(a)$.

Assumption 1.1 For each $x, y \in S$, the mappings $a \to p_{xy}(a)$, $a \in A(x)$, and $a \to$ $\to C(x, a)$, $a \in A(x)$ are *lower semicontinuous*.

A *policy* π is a rule for choosing actions which may be randomized and may depend on the current state as well as on the past history. The class of all policies is denoted by P. A *stationary* policy is characterized by the following: For each $x \in S$, when the system is at x, the policy chooses the same action regardless the past history. Thus, the class of stationary policies is naturally identified with $F := \mathop{\mathsf{X}}_{x \in S} A(x)$, which the class of all functions $f: S \to A$ satisfying $f(x) \in A(x)$, $x \in S$. Notice that F is is a *compact* metric space in the product topology; Dugundji [3, p. 224]. The state and action processes are denoted by $\{X_n\}$ and $\{A_n\}$ respectively. Given the initial state x and the policy π that is being employed the distribution of the joint process $\{X_n, A_n\}$ is uniquely determined (Ash [12, p. 109]) and is denoted by $P_\pi[\,.\,|X_0 = x]$, while $\mathsf{E}_\pi[\,.\,|X_0 = x]$ is the corresponding expectation operator. Also, under the action of any stationary policy the state process is a Markov chain with stationary transition mechanism.

Throughout the remainder we adopt the (lim sup) *average cost criterion*: For $x \in S$ and $\pi \in P$, the average cost at state x under policy π is defined by

$$J(x, \pi) := \limsup_n \mathsf{E}_\pi\Big[\sum_{t=0}^{n} C(X_t, A_t) \,\Big|\, X_0 = x\Big]/(n + 1)\,, \tag{1}$$

while

$$J(x) := \inf_{\pi \in P} J(x, \pi)$$

is the optimal average cost at state x. A policy π is (average cost) *optimal* if $J(x, \pi) = J(x)$ for all $x \in S$.

Assumption 1.2. (i) The cost function is nonnegative. Moreover, (ii) For each $r \geqq 0$, there exists a *finite* set $G(r) \subset S$ such that

$$C(x, a) \geqq r \quad \text{whenever} \quad a \in A(x) \quad \text{and} \quad x \in S - G(r).$$

Under this restriction the expectation in (1) is well defined (its value may be ∞). This assumption is satisfied, for instance, in queuing models with $K \geqq 1$ classes of customers, where the state space is $\mathbb{N}^K$ and the cost function depends polynomially on the state; usually C is a linear or quadratic function [9]. Finally, to obtain optimal stationary policies we need the following communicating/recurrence condition.

Assumption 1.3. (cf. [5, 6, 7]). There exists $f^* \in F$ such that the corresponding Markov chain is *irreducible and ergodic*. Moreover,

$$g^* := \sum_x \pi_{f^*}(x)\, C(x, f^*(x)) < \infty, \tag{2}$$

where $\{\pi_{f^*}(x) \mid x \in S\}$ is the unique *invariant distribution* of the Markov chain determined by f^*; Loève [8, p. 39–42].

In (2), $\sum_x$ indicates summation over $x \in S$. This convention will be used consistently.

The main result of this note is Theorem 3.1 which can be summarized as follows: Under Assumptions 1.1–1.3 *there exists an average cost optimal stationary policy*. To prove this we parallel the development in [5, 6, 7], that is, we examine the average cost case as limit of discounted cases. In doing so, we obtain an inequality that yields our existence result; see Theorem 3.1 below. Also, under an additional condition we obtain a solution to the average cost *OE*.

The organization of the paper is as follows: Section 2 contains all the preliminaries we need from the discounted case, with Theorem 2.2 being the backbone of the argumentation, while our main result is presented in Section 3. We conclude in Section 4 with additional results on the optimality equation and some brief comments.

Remark 1.1 Even without explicit reference we suppose that Assumptions 1.1–1.3 hold true. For an event W, its indicator function is denoted by $I[W]$. Finally, the notation in Assumptions 1.2 and 1.3 will be maintained.

2. PRELIMINARIES FROM THE DISCOUNTED CASE

For each $x \in S$, $\pi \in P$ and $\alpha \in [0, 1)$, the *α-discounted cost* at state x under policy π is defined by

$$V_\alpha(x, \pi) := \mathsf{E}_\pi\Big[\sum_{t=0}^{\infty} \alpha^t\, C(X_t, A_t) \,\Big|\, X_0 = x\Big], \quad \text{and}$$

$$V_\alpha(x) := \inf_{\pi \in P} V_\alpha(x, \pi)$$

is the *optimal* α-discounted cost at state x. A policy π is α-discounted optimal if $V_\alpha(x) = V_\alpha(x, \pi)$ for all $x \in S$.

The results we need from the discounted cost case are given in Theorems 2.1 and 2.2 below. First, we introduce some useful notation together with a Lemma.

Definition 2.1. (i) For each $G \subset S$, the *stopping time* T_G is defined by

$$T_G := \min\{n \geqq 1 \mid X_n \in G\},$$

where, by convention, the minimum of the empty set is ∞. When $G = \{x\}$ is a singleton, the simpler notation T_x is used.

(ii) For each $x, z \in S$, $h(x, z)$ is defined by

$$h(x, z) := \mathsf{E}_{f^*}\Big[\sum_{t=0}^{T_z - 1} C(X_t, A_t) \,\Big|\, X_0 = x\Big];$$

see Assumption 1.3.

Lemma 2.1. (i) For each $x, z \in S$ we have that $h(x, z) < \infty$, and

$$\lim_n \mathsf{E}_{f^*}\Big[\sum_{t=0}^{n} C(X_t, A_t) \,\Big|\, X_0 = x\Big]/(n+1) = g^* \tag{3}$$

(ii) For each $x \in S$, $\lim_{\alpha \to 1} (1 - \alpha)\, V_\alpha(x, f^*) = g^*$.

The first part is a (well known) consequence of Assumption 1.3 and can be proved, for instance, using the results on (delayed) renewal processes in Ross [13, ch. 3]. Part (ii) follows from (3) together with Proposition 4–7 in Heyman and Sobel [2, p. 173].

We begin with the following basic results.

Theorem 2.1. (i) For each $x \in S$ and $\alpha \in [0, 1)$ we have that $0 \leqq V_\alpha(x) < \infty$. Moreover,

(a) $\limsup_{\alpha \to 1} (1 - \alpha)\, V_\alpha(x) \leqq g^*$ for all $x \in S$, and

(b) The α-discounted optimality equation holds:

$$V_\alpha(x) = \inf_{a \in A(x)} \big[C(x, a) + \alpha \textstyle\sum_y p_{xy}(a)\, V_\alpha(y)\big], \quad x \in S. \tag{4}$$

(ii) For each $x \in S$ and $\alpha \in [0, 1)$, the mapping

$$a \to C(x, a) + \alpha \textstyle\sum_y p_{xy}(a)\, V_\alpha(y), \quad a \in A(x) \tag{5}$$

is *lower semicontinuous*. Then, this mapping has a minimizer $f_\alpha(x) \in A(x)$, and the policy $f_\alpha \in F$ is *α-discounted optimal*.
(iii) For all $x, z \in S$, and $\alpha \in [0, 1)$, $V_\alpha(x) - V_\alpha(z) \leqq h(x, z)$.

Proof. (i) Let $x \in S$ be fixed. Using Lemma 2.1 (ii), $(1 - \alpha)\, V_\alpha(x) \leqq (1 - \alpha) \cdot V_\alpha(x, f^*)$ yields part (a). Then $V_\alpha(x)$ is finite for α near to 1, and since $C \geqq 0$, we see that $\alpha \to V_\alpha(x) \geqq 0$ is nondecreasing in $\alpha \in [0, 1)$. Thus, we conclude that $0 \leqq V_\alpha(x) < \infty$ for all $\alpha \in [0, 1)$. Now, part (b) follows in the usual way; cf. Theorem 2.1 in Ross [14, p. 31].

(ii) Let $\alpha \in [0, 1)$, $x \in S$ and $a' \in A(x)$ be arbitrary but fixed. Then, Assumption 1.1 together with Fatou's Lemma yield that

$$\liminf_{a \to a'} \left[C(x, a) + \alpha \sum_y p_{xy}(a)\, V_\alpha(y)\right] \geqq C(x, a') + \alpha \sum_y p_{xy}(a')\, V_\alpha(y)\,.$$

Thus, the mapping (5) is lower semicontinuous. Since $A(x)$ is compact, this mapping has a minimizer $f_\alpha(x) \in A(x)$ [3, p. 227], and the α-discounted optimality of f_α is obtained as in the proof of Theorem 1.2 in [14, p. 50].

(iii) Let $z \in S$ and $\alpha \in [0, 1)$ be fixed and define $\pi \in P$ as follows: In $[0, T_z)\, \pi$ chooses same actions as f^*, while π coincides with f_α in $[T_z, \infty)$. For each $x \in S$, it is clear that the distributions of T_z with respect to $P_{f^*}[\,.\,|X_0 = x]$ and $P_\pi[\,.\,|X_0 = x]$ are the same. Hence, $P_\pi[T_z < \infty \mid X_0 = x] = 1$. Also, using the definition of π together with the Markov property, it follows that, for each $x \in S$

$$\begin{aligned} V_\alpha(x) \leqq V_\alpha(x, \pi) &= \mathsf{E}_\pi\Big[\sum_{t=0}^{T_z - 1} \alpha^t\, C(X_t, A_t) + \alpha^{T_z}\, V_\alpha(z) \,\Big|\, X_0 = x\Big] \\ &= \mathsf{E}_{f^*}\Big[\sum_{t=0}^{T_z - 1} \alpha^t\, C(X_t, A_t) + \alpha^{T_z}\, V_\alpha(z) \,\Big|\, X_0 = x\Big] \end{aligned}$$

and, since $V_\alpha(z) \geqq 0$, it implies that $V_\alpha(x) \leqq h(x, z) + V_\alpha(z)$. □

The above Theorem yields an upper bound for $V_\alpha(x) - V_\alpha(z)$. Now, we need a minimizer for $V_\alpha(\cdot)$ when α is near to 1. This minimizer is obtained below.

Theorem 2.2. There exist a *finite* set G and $\beta \in [0, 1)$ with the following property: For each $\alpha \in [\beta, 1)$ we can find $x_\alpha \in G$ such that

$$V_\alpha(x) \geqq V_\alpha(x_\alpha) \quad \text{for all} \quad x \in S\,, \tag{6}$$

that is, if $1 > \alpha \geqq \beta$, $V_\alpha(\cdot)$ attains its minimum in the set G.

Proof. Define the finite set G by

$$G := G(g^* + 1)\,;$$

see Assumption 1.2. Then, $C(x, a) \geqq g^* + 1$ for $a \in A(x)$ and $x \in S - G$. Now, take $\beta \in [0, 1)$ such that

$$(1 - \alpha)\, V_\alpha(x) \leqq g^* + 1 \quad \text{for} \quad 1 > \alpha \geqq \beta \quad \text{and} \quad x \in G\,; \tag{7}$$

this is possible by the finiteness of G and Theorem 2.1 (ia). Finally, for $1 > \alpha \geqq \beta$

select $x_\alpha \in G$ satisfying $V_\alpha(x) \geqq V_\alpha(x_\alpha)$, $x \in G$ where, again, we are using that G is finite. To complete the proof, we only need to show that x_α satisfies (6) for $x \in S - G$. Thus, let $x \in S - G$ and $1 > \alpha \geqq \beta$ be arbitrary but fixed. To simplify the notation the stopping time T_G is simply denoted by T. Now, observe that

$$\mathsf{E}_{f_\alpha}[I[T = \infty] \sum_{t=0}^{\infty} \alpha^t C(X_t, A_t) \mid X_0 = x] \geqq (g^* + 1) P_{f_\alpha}[T = \infty \mid X_0 = x]/(1 - \alpha)$$

and then, since $x_\alpha \in G$, the above inequality and (7) imply

$$\text{(8)} \quad \mathsf{E}_{f_\alpha}[I[T = \infty] \sum_{t=0}^{\infty} \alpha^t C(X_t, A_t) \mid X_0 = x] \geqq V_\alpha(x_\alpha) P_{f_\alpha}[T = \infty \mid X_0 = x] .$$

Also, the Markov property yields

$$\text{(9)} \quad \mathsf{E}_{f_\alpha}[I[T < \infty] \sum_{t=0}^{\infty} \alpha^t C(X_t, A_t) \mid X_0 = x] =$$

$$= \mathsf{E}_{f_\alpha}[I[T < \infty] (\sum_{t=0}^{T-1} \alpha^t C(X_t, A_t) + \alpha^T V_\alpha(x_T)) \mid X_0 = x] \geqq$$

$$\geqq \mathsf{E}_{f_\alpha}[I[T < \infty] (\sum_{t=0}^{T-1} \alpha^t (1 - \alpha) V_\alpha(x_\alpha) + \alpha^T V_\alpha(x_\alpha)) \mid X_0 = x] =$$

$$= V_\alpha(x_\alpha) P_{f_\alpha}[T < \infty \mid X_0 = x]$$

where we have used that (i) for $t \in [0, T)$, $C(X_t, A_t) \geqq g^* + 1 \geqq (1 - \alpha) V_\alpha(x_\alpha)$, and (ii) $V_\alpha(X_T) \geqq V_\alpha(x_\alpha)$ (since $X_T \in G$). Then, (8) and (9) together imply

$$V_\alpha(x) = V_\alpha(x, f_\alpha) \geqq V_\alpha(x_\alpha) . \qquad \square$$

3. THE MAIN THEOREM

We are now ready to prove our existence result. For $\alpha \in [\beta, 1)$, define

$$h_\alpha(x) := V_\alpha(x) - V_\alpha(x_\alpha) , \quad x \in S ;$$

observe that $h_\alpha(\cdot) \geqq 0$. We begin with the following.

Lemma 3.1. Let $\{\alpha_n\} \subset [\beta, 1)$ be a sequence converging to 1. There exists a subsequence, denoted by $\{\beta_k\}$, such that

(i) The following limits exist:

$$\text{(10)} \quad \lim_k f_{\beta_k}(x) =: f(x) , \quad x \in S ;$$

$$\text{(11)} \quad \lim_k h_{\beta_k}(x) =: h(x) , \quad x \in S ;$$

$$\text{(12)} \quad \lim_k (1 - \beta_k) V_{\beta_k}(x) =: g ,$$

where this limit does not depend on $x \in S$, and

(ii) The sequence $\{x_{\beta_k}\}$ is constant.

Proof. Let $z \in S$ be fixed. Without loss of generality we can assume – taking

a subsequence if necessary – that $\lim_n (1 - a_n) V_{\alpha_n}(z)$ exists. Now observe that, by Theorem 2.1 (iii), $|V_{\alpha_n}(x) - V_{\alpha_n}(z)| \leqq \max\{h(x, z), h(z, x)\}$, $x \in S$ and then $\lim_n (1 - \alpha_n) V_{\alpha_n}(x)$ exists and is independent of x. On the other hand, for $x \in S$ and $n \in \mathbb{N}$, we have that $0 \leqq h_{\alpha_n}(x) = V_{\alpha_n}(x) - V_{\alpha_n}(x_{\alpha_n}) \leqq h(x, x_{\alpha_n}) \leqq$ $\leqq \max\{h(x, y) \mid y \in G\} =: H(x)$. Then, for $n \in \mathbb{N}$,

$$w_n := (x_{\alpha_n}; f_{\alpha_n}(x), h_{\alpha_n}(x) \mid x \in S) \in G \times \{\mathsf{X}_{x \in S}(A(x) \times [0, H(x)])\}.$$

Now, let G be endowed with the discrete topology and observe that the right hand side is a (sequentially) compact space in the product topology. Then, there exist a convergent subsequence $\{w_{n_k}\}$ and it follows that $\{\beta_k\} = \{\alpha_{n_k}\}$ satisfies the desired conclusion. □

Using the notation of (10), (11) and (12) our main result follows.

Theorem 3.1. Under Assumptions 1.1–1.3 there exists an average cost optimal stationary policy. More precisely, let $\{\alpha_n\} \subset [\beta, 1)$ be any sequence converging to 1, and take a subsequence $\{\beta_k\}$ as in Lemma 3.1. Then,

(i) f is average cost optimal and $g = J(x, f)$ for all $x \in S$. Also

$$g + h(x) \geqq C(x, f(x)) + \sum_y p_{xy}(f(x))\, h(y), \quad x \in S. \tag{13}$$

Moreover,

(ii) *Any* $f \in F$ satisfying (13) is average cost optimal.

Proof. Using Proposition 4-7 in [2, p. 173], we see that for $x \in S$ and $\pi \in P$, $g = \lim_n (1 - \beta_n) V_{\beta_n}(x) \leqq \limsup_n (1 - \beta_n) V_{\beta_n}(x, \pi) \leqq J(x, \pi)$, and then $g \leqq J(x)$, $x \in S$. Now, from the definition of f_β and (4), we see that $V_{\beta_n}(x) = C(x, f_{\beta_n}(x)) +$ $+ \beta_n \sum_y p_{xy}(f_{\beta_n}(x))\, V_{\beta_n}(y)$, $x \in S$ and, since $\{x_{\beta_n}\}$ is a constant sequence, this is equivalent to

$$(1 - \beta_n) V_{\beta_n}(x_{\beta_0}) + h_{\beta_n}(x) = C(x, f_{\beta_n}(x)) + \beta_n \sum_y p_{xy}(f_{\beta_n}(x))\, h_{\beta_n}(y), \quad x \in S.$$

Now, take lim inf as $n \to \infty$ in both sides of this equality. In this case, (10)–(12) together with Assumption 1.1 and Fatou's Lemma imply (13). To complete the proof let $f \in F$ be *any* policy satisfying (13). Then, by a simple induction argument we see that for $x \in S$ and $n \in \mathbb{N}$,

$$(n + 1)\, g + h(x) \geqq \mathsf{E}_f\Big[\sum_{t=0}^{n} C(X_t, A_t) + h(X_n + 1) \mid X_0 = x\Big].$$

Since $h \geqq 0$, this inequality immediately yields that for all $x \in S$, $g \geqq J(x, f)$, and since $J(x, f) \geqq J(x) \geqq g$, f is average cost optimal with average cost g. □

Corollary 3.1. $\lim_{\alpha \to 1} (1 - \alpha) V_\alpha(x)$ exists and is independent of $x \in S$. Moreover, the common value is the optimal average cost.

Proof. For an arbitrary sequence $\{\alpha_n\} \subset [\beta, 1)$ converging to 1, we have proved

the existence of a subsequence $\{\beta_n\}$ satisfying the conclusions in Lemma 3.1. Moreover, Theorem 3.1 states that g given by (12) is *the* optimal average cost which is *uniquely* determined. The conclusion follows from the arbitrariness of $\{\alpha_n\}$. □

Under an additional restriction we can prove that the average cost *OE* holds:

Assumption 3.1. For some state $z \in S$ the following holds:

$$\sum_y p_{xy}(a)\, h(y, z) < \infty \quad \text{for all} \quad x \in S \quad \text{and} \quad a \in A(x). \tag{14}$$

Theorem 3.2. For a sequence $\{\alpha_n\} \subset [\beta, 1)$ select a subsequence $\{\beta_n\}$ as in Lemma 3.1. Then, under Assumptions 1.1–1.3 and 3.1, g and $h(\cdot)$ satisfy the average cost *OE*:

$$g + h(x) = \inf_{a \in A(x)} \left[C(x, a) + \sum_y p_{xy}(a)\, h(y)\right], \quad x \in S. \tag{15}$$

Proof. We will prove that our assumptions imply that (14) holds if we replace z by *any* other element of S, that is,

$$\sum_y p_{xy}(a)\, h(y, w) < \infty \quad \text{for all} \quad x, w \in S \quad \text{and} \quad a \in A(x). \tag{16}$$

The result follows immediately from this. Indeed, using that $\{x_{\beta_n}\}$ is a constant sequence, the α-discounted optimality equation (4) yields that for $n \in \mathbb{N}$, $x \in S$ and $a \in A(x)$,

$$(1 - \beta_n)\, V_{\beta_n}(x_{\beta_0}) + h_{\beta_n}(x) \leqq C(x, a) + \sum_y p_{xy}(a)\, h_{\beta_n}(y).$$

Since $0 \leqq h_{\beta_n}(y) = V_{\beta_n}(y) - V_{\beta_n}(x_{\beta_0}) \leqq h(y, x_{\beta_0})$, the above inequality together with (16) and the dominated convergence theorem imply that $g + h(x) \leqq C(x, a) + \sum_y p_{xy}(a)\, h(x)$. Thus,

$$g + h(x) \leqq \inf_{a \in A(x)} \left[C(x, a) + \sum_y p_{xy}(a)\, h(y)\right],$$

and then (15) follows using (13). Thus, we only need to prove (16). Let z be as in (14) and take $w \in S$ arbitrary but *fixed*. Define $T'_w := \min\{n \geqq 1 \,|\, X_{T_z + n} = w\}$ if $T_z < \infty$ and this set is nonempty; $T'_w := \infty$ otherwise. Then $T_z + T'_w$ is the first time the system is at w *after* a visit to z. It follows that $T_z + T'_w \geqq T_w$, and using that $C \geqq 0$, we see that, for each $x \in S$

$$h(x, w) = \mathsf{E}_{f^*}\Big[\sum_{t=0}^{T_w - 1} C(X_t, A_t) \,\Big|\, X_0 = x\Big] \leqq$$

$$\leqq \mathsf{E}_{f^*}\Big[\sum_{t=0}^{T_z - 1} C(X_t, A_t) + \sum_{t=T_z}^{T_z + T'_w - 1} C(X_t, A_t) \,\Big|\, X_0 = x\Big] =$$

$$= h(x, z) + \mathsf{E}_{f^*}\Big[\sum_{t=0}^{T_w - 1} C(X_t, A_t) \,\Big|\, X_0 = z\Big],$$

where we have used that $P_{f^*}[T_z < \infty \,|\, X_0 = x] = 1$ together with the Markov property. Thus, we conclude that $h(x, w) \leqq h(x, z) + h(z, w)$, and then (14) implies (16) since $h(z, w)$ is a finite constant. □

4. CONCLUDING REMARKS AND RESULTS

We have given sufficient conditions for the existence of average cost optimal stationary policies. Our approach parallels the one followed by Sennott in [5, 6, 7]. However, the main *difference* between Sennott's results and those in this note, is that we do *not* assume the existence of a fixed state for which the α-discounted cost is "minimal" for all $\alpha \in [0, 1)$. Rather, using our assumptions, we have "constructed" a minimizer of $V_\alpha(\cdot)$ for α near to 1, and the key point is that this minimizer can be selected in a *finite* set. As already mentioned, our conditions are satisfied in interesting queuing models. Also, it is clear that our results hold when the cost function is just *bounded below* and the other conditions remain valid.

On the other hand, there are – at least – two interesting problems to be considered: The first one refers to the optimality criterion. We have used the lim sup average cost criterion (1). Replacing lim sup by lim inf in (1) yields the lim inf average cost criterion. Now, the lim inf (lim sup) criterion represents the smallest (largest) limit point of the expected average costs over finite horizons, and then, since minimizing the smallest average cost is "more appealing" than minimizing the largest one, it is natural to ask if results similar to Theorem 3.1 and Theorem 3.2 can be obtained with respect to the lim inf average cost criterion. The second problem refers to the *OE*: Is it possible to prove that g and $h(\cdot)$ in (11) and (12) satisfy the average cost *OE* without Assumption 3.1?

Presently, we just have a *partial* (affirmative) answer to the second problem which is stated in Theorem 4.1 below. Assume throughout the following that Assumptions 1.1 – 1.3 hold true. Within this framework, we have proved that (13) holds, and we see that, for all $x \in S$,

$$g + h(x) \geqq \inf_{a \in A(x)} [C(x, a) + \sum_z p_{xz}(a)\, h(z)] . \tag{17}$$

Define the *discrepancy function* $\Phi: S \to \mathbb{R}$ by

$$\Phi(x) := g + h(x) - \inf_{a \in A(x)} [C(x, a) + \sum_z p_{xy}(a)\, h(z)] , \quad x \in S . \tag{18}$$

Clearly, $\Phi \geqq 0$ and the equality holds in (17) exactly when $\Phi(x) = 0$. Now, for each $x \in S$ the term in brackets in (17) is a lower semicontinuous function of $a \in A(x)$, and then, it has a minimizer $\ell(x) \in A(x)$; see the proof of Theorem 2.1 (ii). Thus, we can write

$$\inf_{a \in A(x)} [C(x, a) + \sum_z p_{xz}(a)\, h(z)] = c(x, \ell(x)) + \sum_z p_{xz}(\ell(x))\, h(z) , \quad x \in S$$

and using (17) we see that (13) holds with $f = \ell$. Then ℓ is optimal by Theorem 3.1. Also (18) is equivalent to

$$g + h(x) = C(x, \ell(x)) + \Phi(x) + \sum_z p_{xz}(\ell(x))\, h(z) , \quad x \in S ,$$

and, as in the proof of Theorem 3.1 (ii) we obtain,

$$g \geqq \limsup_n \mathsf{E}_\ell[\sum_{t=0}^{n} C(X_t, A_t) + \Phi(X_t) \mid X_0 = x]/(n+1) \geqq$$

$$\geqq \limsup_n \mathsf{E}_\ell[\sum_{t=0}^{n} C(X_t, A_t) \mid X_0 = x]/(n+1) = g\,,$$

since ℓ is optimal. We conclude that, for all $x \in S$,

$$g = \limsup_n \mathsf{E}_\ell[\sum_{t=0}^{n} C(X_t, A_t) + \Phi(X_t) \mid X_0 = x]/(n+1) = \tag{19}$$

$$= \limsup_n \mathsf{E}_\ell[\sum_{t=0}^{n} C(X_t, A_t) \mid X_0 = x]/(n+1)\,.$$

In what follows, a state that is positive recurrent (null recurrent, transient) with respect to the Markov chain induced by ℓ is referred, simply, as an ℓ-positive recurrent (-null recurrent, -transient) state. Also, for the properties of Markov chains used here see, for instance, Loève [8, p. 39–42].

Theorem 4.1. (i) The class of ℓ-positive recurrent states is nonempty.

(ii) If y is an ℓ-positive recurrent state we have

$$g + h(y) = \inf_{a \in A(y)} [C(y, a) + \sum_z p_{yz}(a)\, h(z)]\,. \tag{20}$$

Proof. Let $G := G(g+1)$ (see Assumption 1.2) and assume that G does *not* contain any ℓ-positive recurrent state. In this case we have (since G is *finite*) that, for $x \in S$,

$$\sum_{t=0}^{n} P_\ell[X_t \in G \mid X_0 = x]/(n+1) \to 0 \quad \text{as} \quad n \to \infty\,. \tag{21}$$

On the other hand, since $C(X_t, A_t) \geqq g + 1$ for $X_t \in S - G$ we see that

$$\mathsf{E}_\ell[\sum_{t=0}^{n} C(X_t, A_t) \mid X_0 = x] \geqq \mathsf{E}_\ell[\sum_{t=0}^{n} C(X_t, A_t)\, I[X_t \in S - G] \mid X_0 = x] \geqq$$

$$\geqq (g+1)\, \mathsf{E}_\ell[\sum_{t=0}^{n} I[X_t \in S - G] \mid X_0 = x] = (g+1) \sum_{t=0}^{n} P_\ell[X_t \in S - G \mid X_0 = x] =$$

$$= (g+1)\,[(n+1) - \sum_{t=0}^{n} P_\ell[X_t \in G \mid X_0 = x]]\,.$$

From this we get easily, using (21), that

$$\limsup_n \mathsf{E}_\ell[\sum_{t=0}^{n} C(X_t, A_t) \mid X_0 = x]/(n+1) \geqq g + 1$$

which contradicts the optimality of ℓ. We conclude that G contains – at least – one ℓ-positive recurrent state and (i) follows. To prove (ii) let $y \in S$ be ℓ-positive recurrent and let R be the ℓ-recurrence class containing y. In this case, for any

$Q: S \to [0, \infty)$ we have that, as $n \to \infty$

$$\mathsf{E}_\ell[\sum_{t=0}^{n} Q(X_t) \mid X_0 = y]/(n+1) \to \sum_z \pi(R, z)\, Q(z); \tag{22}$$

here, $\{\pi(R, z) \mid z \in S\}$ is the *unique* invariant distribution (of the Markov chain determined by ℓ) which is *concentrated* on R. We recall that $\pi(R, z) > 0$ if and only if $z \in R$. From (22) we obtain that, as $n \to \infty$,

$$\mathsf{E}_\ell[\sum_{t=0}^{n} C(X_t, A_t) + \Phi(X_t) \mid X_0 = y]/(n+1) \to \sum_z \pi(R, z)\,[C(z, \ell(z)) + \Phi(z)]$$

and

$$\mathsf{E}_\ell[\sum_{t=0}^{n} C(X_t, A_t) \mid X_0 = y]/(n+1) \to \sum_z \pi(R, z)\, C(z, \ell(z))\,.$$

Combining these two convergences with (19) we get

$$g = \sum_z \pi(R, z)\,[C(z, \ell(z)) + \Phi(z)] = \sum_z \pi(R, z)\, C(z, \ell(z))\,.$$

This implies, since g is finite and Φ is nonnegative, that $0 = \sum_z \pi(R, z)\, \Phi(z) \geqq$ $\geqq \pi(R, y)\, \Phi(y) \geqq 0$. Since $y \in R$, $\pi(R, y) > 0$ and we see that $\Phi(y) = 0$ which is equivalent to (20). This completes the proof. □

According to Theorem 4.1 the average cost *OE* holds at all ℓ-positive recurrent states. However, our second question is still open for ℓ-null recurrent or ℓ-transient states. Also notice that if the state space S is irreducible when policy ℓ is employed, *all* the states are ℓ-positive recurrent and then, (20) holds for all $y \in S$. Active research to provide a complete answer to both of the problems posed above is presently in progress.

(Received August 30, 1988.)

REFERENCES

[1] A. Hordijk: Dynamic Programming and Potential Theory. (Mathematical Centre Tract 51.) Mathematisch Centrum, Amsterdam 1974.

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

[3] J. Dugundji: Topology. Allyn and Bacon, Boston 1966.

[4] L. C. Thomas: Connectedness conditions for denumerable state Markov decision processes. In: Recent Developments in Markov Decision Processes (Hartley, Thomas, White, eds.), Academic Press, New York 1981, pp. 181—204.

[5] L. I. Sennott: A new condition for the existence of optimal stationary policies in average cost Markov decision processes. Oper. Res. Lett. 5 (1986), 17—23.

[6] L. I. Sennott: A new condition for the existence of optimum stationary policies in average cost Markov decision processes — unbounded cost case. Proceedings of the 25th IEEE Conf. on Dec. and Control, Athens, Greece 1986, pp. 1719—1721.

[7] L. I. Sennott: Average cost optimal stationary policies in infinite state Markov decision processes — Existence and an algorithm. Submitted (1987).

[8] M. Loève: Probability Theory I. Springer-Verlag, New York—Berlin—Heidelberg 1977.

[9] P. Nain and K. W. Ross: Optimal priority assignement with hard constraints. Submitted to IEEE Trans. Automat. Control (1986).
[10] R. Cavazos-Cadena: Necessary conditions for the optimality equation in average-reward Markov decision processes. Appl. Math. Optim. *19* (1989), 1, 97—112.
[11] R. Cavazos-Cadena: Necessary and sufficient conditions for a bounded solution to the optimality equation in average reward Markov decision chains. System Control Lett. *10* (1988), 71—78.
[12] R. B. Ash: Real Analysis and Probability. Academic Press, New York 1972.
[13] S. M. Ross: Applied Probability Models with Optimization Applications. Holden-Day, San Francisco 1970.
[14] S. M. Ross: Introduction to Stochastic Dynamic Programming. Academic Press, New York 1983.

Dr. Rolando Cavazos-Cadena, Departamento de Estadística y Cálculo, Universidad Autónoma Agraria Antonio Narro, Buenavista 25315, Saltillo, Coahuila, Mexico and *Department of Mathematics, Texas Tech University, Lubbock, TX 79409, U.S.A.*