Boletín de la Sociedad Matemática Mexicana Vol. 35 No. 1, 1990

AVERAGE OPTIMAL STATIONARY POLICIES IN MARKOV DECISION PROCESSES UNDER WEAK STABILITY CONDITIONS*

BY ROLANDO CAVAZOS-CADENA

1. Introduction

We are concerned with Markov decision processes (MDP's) with denumerable state space and finite control sets. The reward function is bounded —but otherwise arbitrary— and the performance index of a control policy is a long-run expected average reward criterion. For these models, the existence of an optimal stationary policy has been established under several conditions on the transition law of the system [12]. A common approach to this problem consists in imposing a (stability) condition so that the *average reward optimality equation* (AROE) has a bounded solution, which in turn yields an optimal stationary policy in a standard way [9,10,12,...]. Under natural restrictions on the recurrence structure of the model, a bounded solution to the AROE exists if, and only if, the so called *simultanous Doeblin condition* (SDC) holds [2,3], a requirement that is quite restrictive and is not satisfied in interesting applications [1,2-4,11].

Our main objective in this note is to introduce a weak form of the SDC which is sufficient to guarantee that, for *arbitrary* bounded reward function, there exists an optimal stationary policy; se Assumption (2.2) and Theorem (4.1) below. To obtain this result we follow the approach of examining a model with the average criterion as a "limit" of discounted programs.

The organization of the paper is as follows: In Section 2 the decison model is formally described —including the key stability Assumption (2.2)— and some preliminaries from discounted dynamic programming are given in Section 3. Then, the main theorem is proved in Section 4 and we conclude with an example in Section 5.

Notation. As usual, $\mathbf{R}$ and $\mathbf{N}$ stand for the sets of real numbers and nonnegative integers, respectively. The indicator function of an event W is denoted by $I[W]$ and, if Y is a random vector, $\sigma(Y)$ denotes de σ-algebra generated by Y. Finally, for a real-valued function r,

$$\| r \| := \sup \{|r(x)| \,|x \text{ is the domain of } r\}.$$

2. The Decision Model

Throughout the remainder we follow closely (but not completely) the notation and terminology in Ross [9,10]. Let (S, U, p, r) be the usal MDP where the state space S and the control set U are nonempty *countable* sets. For each $x \in S$, $U(x) \subset U$ is the (nonempty) set of *admissible* controls at state x, and

* This research was partially supported by the Third World Academy of Sciences (TWAS) under Grant TWAS RG MP 898-152.

the set of *admissible pairs* is $\mathbf{K} := \{(x,u)|x \in S,\ u \in U(x)\}$. On the other hand, $r : \mathbf{K} \to \mathbf{R}$ and $p(\cdot|\cdot) : S \times \mathbf{K} \to [0,1]$ are the *reward function* and the *transition law*, respectively; of course, $\Sigma_y p(y|k) = 1,\ k \in \mathbf{K}$.

This model represents a dynamical system evolving as follows: At each decision time $t \in \mathbf{N}$, the state of the system is observed, say $X_t = x \in S$, and a control $U_t = u \in U(x)$ is chosen. Then (i) a reward $r(u,x)$ is earned, and (ii) regardless of the states observed and controls applied prior to t, the state of the system at time $t+1$ will be $y \in S$ with probability $p(y|x,u)$; this is the *Markov property* of the process.

Assumption (2.1): (i) The reward function is *bounded* i.e., $\| r \| < \infty$ (ii) For each $x \in S$, the set $U(x)$ is *finite*.

Control Policies. A *policy* is a (possibly randomized) rule for choosing controls wich may depend on the current state as well as on the record of previous states and controls; see [6,8] for a more detailed description. The class of all policies is denoted by $\mathcal{P}$. Let $\mathbf{F} := \times\{U(x)|x \in S\}$, i.e., $\mathbf{F}$ consists of all (choice) functions $f : S \to U$ satisfying $f(x) \in U(x),\ x \in S$. A policy π is *stationary* if there exists $f \in \mathbf{F}$ such that, under π, the control applied at time t is $U_t = f(X_t)$; as usual, the class of all stationary policies is (naturally) identified with $\mathbf{F}$. Given the initial state $X_0 = x$ and the policy $\pi \in \mathcal{P}$ being used, the distribution of the state-control process $\{(X_t, U_t)\}$ is uniquely determined [6,8]. This distribution is denoted by $P_\pi[\cdot|X_0 = x]$, and $E_\pi[\cdot|X_0 = x]$ stands for the corresponding expectation operator.

The Optimality Criterion. For any $x \in S$ and $\pi \in \mathcal{P}$, the (lim inf-expected) *average* reward at state x under policy π is defined by

$$J(x,\pi) := \liminf E_\pi[\sum_{t=0}^{n} r(X_t, U_t|X_0 = x]/(n+1), \quad \text{and} \tag{2.1}$$

$$J(x) := \sup\{J(x,\pi)|\pi \in \mathcal{P}\} \tag{2.2}$$

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

It is known that, under Assumption (2.1) alone, an optimal stationary policy does not necessarily exist [5,9,10]. We shall see, however, that an optimal policy does exist if, *additionally*, we suppose that Assumption (2.2) below holds true. First, we introduce some notation involving a state $z \in S$ which will be *fixed* throughout the remainder.

Definition (2.1): (i) The stopping time T is defined by

$$T := \min\{n > 0|X_n = z\}, \tag{2.3}$$

where, by convention, the minimum of the empty set is ∞.
(ii) Let $f \in \mathbf{F}$ be arbitrary. The functions M_f and M from S into $[0,\infty)$ are defined as follows: For each $x \in S$

$$M_f(x) := E_f[T|X_0 = x]; \tag{2.4}$$

$$M(x) := \sup\{M_\phi(x)|\phi \in \mathbf{F}\}. \tag{2.5}$$

Assumption (2.2): (i) $M_f(x) < \infty$, $x \in S$, $f \in \mathbf{F}$.
(ii) There exsits $\gamma \in \mathbf{R}$ such that, for all $f, \phi \in \mathbf{F}$ and $x \in S$,

$$M_\phi(x)/M_f(x) \leq \gamma.$$

Remark (2.1): Assumption (2.2) is a weakened version of the following form of the SDC [2,3,9,10,12]:
For some finite constant b, $M_f(x) \leq b$, $x \in S$, $f \in \mathbf{F}$.

Remark (2.2): Suppose that Assumption (2.2) holds. Then, for all ϕ, $f \in \mathbf{F}$ and $x \in S$, $M_\phi(x) \leq \gamma \cdot M_f(x)$, which implies (see (2.5))

$$M(x) \leq \gamma \cdot M_f(x) < \infty, \quad x \in S \quad f \in \mathbf{F}. \tag{2.6}$$

3. Preliminaries from the Discounted Case

Let $x \in S$, $\pi \in \mathcal{P}$ and $\alpha \in (0,1)$ be arbitrary. The (total expected) *α-discounted reward* at state x under policy π is

$$V_\alpha(x,\pi) := E_\pi[\sum_{t=0}^{\infty} \alpha^t r(X_t, U_t)|X_0 = x], \quad \text{and}$$

$$V_\alpha := \sup\{V_\alpha(x,\pi)|\pi \in \mathcal{P}\}$$

is the optimal α-discontinued reward at x. A policy π^* is a *α-optimal* if $V_\alpha(x,\pi^*) = V_\alpha(x)$ for all $x \in S$. It is well known [9,10] that (i) $\| V_\alpha(\cdot) \| \leq \| r \| /(1-\alpha)$, and (ii) $V_\alpha(\cdot)$ satisfies

$$V_\alpha(x) = \max_{u \in U(x)} [r(x,u) + \alpha \cdot \Sigma_y p(y|x,u) V_\alpha(y)], \quad x \in S, \tag{3.1}$$

which is the optimal α-discounted optimality equation. Moreover, a policy $f_\alpha \in \mathbf{F}$ is α-optimal if and only if, for any $x \in S$, $f_\alpha(x)$ is a maximizer of the function within brackets in (3.1). Throughout the remainder, f_α stands for an α-optimal stationary policy.

For any $x \in S$ and $\alpha \in (0,1)$ define

$$g_\alpha(x) := (1-\alpha)V_\alpha(x), \quad \text{and} \tag{3.2}$$

$$h_\alpha(x) := V_\alpha(x) - V_\alpha(z), \tag{3.3}$$

where $z \in S$ is the (fixed) state in Definition (2.1). We observe that $\| g_\alpha \| = (1-\alpha) \| V_\alpha \| \leq \| r \|$. Also, the arguments in the proof of Theorem (6.19) in [9] yield the following

LEMMA (3.1). *Suppose that Assumption* (2.2) *holds and let* $M(x)$ *be as in* (2.5). *Then, for any* $\alpha \in (0,1)$ *and* $x \in S$,

$$|h(x)| \leq 2 \cdot \| r \| \cdot M(x) < \infty. \tag{3.4}$$

LEMMA (3.2). *Suppose that Assumptions* (2.1) *and* (2.2) *hold and let* $\{\alpha_n\} \subset (0,1)$ *be a sequence converging to* 1. *Then, for a subsequence of* $\{\alpha_n\}$, *say* $\{\beta_m\}$, *the following limits exist*:

$$\lim h_{\beta_m}(x) =: h(x) \in [-2 \cdot \| r \| \cdot M(x), 2 \cdot \| r \| \cdot M(x)], \quad x \in S; \tag{3.5}$$

$$\lim f_{\beta_m}(x) =: f^*(x) \in U(x), \quad x \in S; \tag{3.6}$$

$$\lim g_{\beta_m}(z) =: g \in [- \| r \|, \| r \|]. \tag{3.7}$$

Moreover, with g *as in* (3.7) *we have*

$$\lim g_{\beta_m}(x) = g, \quad x \in S.$$

For a proof see, for instance, [4,11] or the proof of Theorem (6.17) in [9]. Notice that, for any $x \in S$, $|g_{\beta_m}(x) - g_{\beta_m}(z)| = (1 - \beta_m)h_\beta(x)$, and then, (3.5) and (3.7) together yield (3.8).

Remark (3.1). Throughout the remainder, the sequence $\{\beta_m\}$, $g \in \mathbf{R}$, $f^* \in \mathbf{F}$ and $h : S \to \mathbf{R}$, are as in Lemma (4.2).

We now prove that g and $h(\cdot)$ satisfy the AROE (3.9) below.

LEMMA (3.3). *Under Assumptions* (2.1) *and* (2.2) *the following holds.*

(i) *For any* $(x,u) \in \mathbf{K}$, $\Sigma_y p(y|x,u)M(y) < \infty$.

(ii) *For all* $x \in S$,

$$g + h(x) = \max_{u \in U(x)} [r(x,u) + \Sigma_y p(y|x,u)h(y)], \tag{3.9}$$

and

$$g + h(x) = r(x, f^*(x)) + \Sigma_y p(y|x, f^*(x))h(y). \tag{3.10}$$

(iii) *For any* $x \in S$ *and* $\pi \in \mathcal{P}$

(a) $g \geq J(x,\pi)$, *and then,*

(b) $g \geq J(x)$.

Proof. (i) Let $(x,u) \in \mathbf{K}$ be fixed. Select $f \in \mathbf{F}$ satisfying $f(x) = u$, and notice that (see (2.3))

$$\begin{aligned} \infty > M_f(x) &= E_f[T|X_0 = x] = 1 + E_f[TI[T > 1]|X_0 = x] \\ &= 1 + \Sigma_{y \neq z} p(y|x,u) E_f[T|X_0 = y] \quad \text{(by the Markov property)} \\ &= 1 + \Sigma_{y \neq z} p(y|x,u) M_f(y). \end{aligned} \tag{3.11}$$

Hence,

$$\begin{aligned}\Sigma_y p(y|x,u)M(y) &\leq \gamma \cdot \Sigma_y p(y|x,u)M_f(y) \quad \text{(by (2.6))}\\ &= \gamma \cdot [p(z|x,u)M_f(z) + \Sigma_{y\neq z} p(y|x,u)M_f(y)]\\ &\leq \gamma \cdot [M_f(z) + M_f(x) - 1] < \infty; \quad \text{see (3.11).}\end{aligned}$$

(ii) From part (i), Lemma (3.1) and the dominated convergence theorem, we obtain

$$\Sigma_y p(y|x,u)h_{\beta_m}(y) \to \Sigma_y p(y|x,u)h(y) \quad \text{as} \quad m \to \infty. \tag{3.12}$$

Using this convergence, (3.9) follows as in the proof of Theorem (6.18) in [9]. To conclude, observe that, since f_{β_m} is β_m-optimal,

$$V_{\beta_m}(x) = r(x, f_{\beta_m}(x)) + \beta_m \cdot \Sigma_y p(y|x, f_{\beta_m}(x))V_{\beta_m}(y)], \quad x \in S,$$

and then, simple rearragements using (3.2) and (3.3) yield [9,10]

$$g_{\beta_m}(z) + h_{\beta_m}(x) = r(x, f_{\beta_m}(x)) + \beta_m \cdot \Sigma_y p(y|x, f_{\beta_m}(x))h_{\beta_m}(y), \quad x \in S.$$

Let $x \in S$ be arbitrary but fixed. Since $U(x)$ is *finite*, (3.6) implies that $f_m(x) = f^*(x)$ for m large enough and, in this case,

$$g_{\beta_m}(z) + h_{\beta_m}(x) = r(x, f^*(x)) + \beta_m \cdot \Sigma_y p(y|x, f^*(x))h_{\beta_m}(y).$$

Then (3.10) follows by taking limit as m goes to ∞ in both sides of this equality and using (3.12).

(iii) Notice that, for any $x \in S$, $g = \lim\,(1-\beta_m)V_{\beta_m}(x) \geq \liminf_{\alpha \nearrow 1}(1-\alpha)$ $V_\beta(x) \geq J(x,\pi)$; see [7,p. 173] for the last inequality. This yields part (a) and, since π is arbitrary, part (b) follows immediately; see (2.1) and (2.2). □

4. Optimal Stationary Policies

We now establish the existence of average optimal stationary policies; se Remark (3.1) for notation.

THEOREM (4.1). *Under Assumptions* (2.1)-(2.2) *the following holds.*
(i) *let* $f \in \mathbf{F}$ *satisfy*

$$g + h(x) = r(x, f(x)) + \Sigma_y p(y|x, f(x))h(y), \quad x \in S. \tag{4.1}$$

Then, f is (average) optimal. In particular,
(ii) f^* *is optimal.*

The proof of this theorem relies on the following lemma.

LEMMA (4.1). *Suppose that Assumption* (2.2) *is valid. Then, for all $x \in S$ and $f \in \mathbf{F}$, we have that*
(i) *For all $n \in \mathbf{N}$, $E_f[M(X_n)|X_0 = x] < \infty$,*
and, as $n \to \infty$, the three convergences below hold true:

$$\begin{aligned} &\text{(ii)}\ E_f[M(X_n)I[T > n]|X_0 = x] \to 0; \\ &\text{(iii)}\ E_f[M(X_n)|X_0 = x]/(n+1) \to 0; \\ &\text{(iv)}\ E_f[|h(X_n)|\,|X_0 = x]/(n+1) \to 0. \end{aligned} \tag{4.2}$$

Proof. (i) Let $f \in \mathbf{F}$ be arbitrary and notice that for all $x \in S$,

$$\begin{aligned} M_f(z) + M_f(x) \geq & M_f(z)p(z|x, f(x)) + M_f(x) \\ = & 1 + \Sigma_y p(y|x, f(x))M_f(y); \end{aligned}$$

(see (3.11)). Then, a simple induction argument yields $\infty > (n+1)(M_f(z) - 1) + M_f(x) \geq E_f[M_f(X_n)|X_0 = x]$, $x \in S$, $n \in \mathbf{N}$, and the conclusion follows using (2.6).

(ii) Notice that $T > (T-n)I[T > n] \searrow 0$ on the event $[T < \infty]$.

Then, Assumption (2.2) and the dominated convergence theorem imply

$$E_f[(T-n)I[T > n]|X_0 = x] \searrow 0 \ \text{ as } \ n \to \infty. \tag{4.3}$$

On the other hand, a simple conditioning argument using the Markov property yields

$$E_f[M_f(X_n)I[T > n]|X_0 = x] = E_f[(T-n)I[T > n]|X_0 = x],$$

and the result follows using (2.6) and (4.3).

(iii) Let $f \in \mathbf{F}$ and $x \in S$. To begin with, observe that

$$1 = P_f[X_t = z \ \text{ for some } \ t \leq n|X_0 = x] + P_f[X_t \neq z, 0 \leq t \leq n|X_0 = x],$$

and then,

$$\begin{aligned} 1 = & \sum_{s=0}^{n} P_f[X_s = z, X_t \neq z \ \text{ for } \ s < t \leq n|X_0 = x] \\ & + P_f[X_t \neq z, 0 \leq t \leq n|X_0 = x], \end{aligned} \tag{4.4}$$

where we have used that,

$$[X_t = z \ \text{ for some } \ t \leq n] = \bigcup_{s=0}^{n} [X_s = z, X_t \neq z \ \text{ for } \ s < t \leq n];$$

this is the partition of the event on the left-hand side according to the last visit to state z up to time n. Using (4.4) we see that

$$\begin{aligned} &E_f[M(X_n)|X_0 = x] \\ &\quad = \sum_{s=0}^{n} E_f[M(X_n)I[X_s = z, X_t \neq z, s < t \leq n|X_0 = x] \\ &\qquad + E_f[M(X_n)I[X_t \neq z, 0 \leq t \leq n|X_0 = x] \end{aligned}$$

$$= \sum_{s=0}^{n} E_f[M(X_{n-s})I[T > n-s]|X_0 = z]P_f[X_s = z|X_0 = x]$$
$$+ E_f[M(X_n)IT > n]|X_0 = x] \quad \text{(by the Markov property)}$$
$$\leq \sum_{s=0}^{n} E_f[M(X_{n-s})I[T > n-s]|X_0 = z] + E_f[M(X_n)IT > n]|X_0 = x].$$

Now, the conclusion follows using this inequality and part (ii).

(iv) Combine part (iii) with $|h(\cdot)| \leq 2 \cdot \| r \| \cdot M(\cdot)$; see (3.5). □

Proof of Theorem (4.1). Notice that part (ii) is a consequence of (3.10) and part (i). To prove (i), let $f \in \mathbf{F}$ satisfy (4.1). In this case, a simple induction argument using (3.5) and Lemma (4.1)(i) yields that, for all $n \in \mathbf{N}$ and $x \in S$,

$$g + h(x)/(n+1) = E_f[\sum_{t=0}^{n} r(X_t, U_t)|X_0 = x]/(n+1)$$
$$+ E_f[h(X_n)|X_0 = x]/(n+1),$$

and from (4.2) we see that, $g = \lim E_f[\sum_{t=0}^{n} r(X_t, U_t)|X_0 = x]/(n+1)$, which implies $J(x, f) = g$; see (2.1). Now Lemma (3.3)(iiib) yields $J(x, f) \geq J(x)$, $x \in S$ and then f is optimal; see (2.2). □

5. An Example

We now apply Theorem (4.1) to establish the existence of optimal stationary policies in a queueing system.

Example (5.1). Let $S := \mathbf{N}$ and set $U(x) = U := \{1, 2, \ldots, k\}$, $x \in \mathbf{N}$, where k is a positive integer. For each $t \in \mathbf{N}$ and $u \in U$, let A_t and $D_t(u)$ be $\mathbf{N}$-valued random variables defined on a common probability space and suppose that the evolution of the system is determined by

$$X_{t+1} = \max\{X_t + A_t - D_t(U_t), 0\}. \tag{5.1}$$

This model can be interpreted as follows: We have a service station with infinite buffer capacity such that (i) for $t \in \mathbf{N}$, X_t is the number of customers waiting for service at time t, (ii) A_t is the number of arrivals in the slot $(t, t+1)$; then the total amount of customers requiring for service in $[t, t+1)$ is $X_t + A_t$, and (iii) if the control u is applied at time t, $D_t(u)$ is the number of service completions that can be provided in $[t, t+1)$; served customers leave the system. The following conditions are enforced:

C1: The vectors $(A_t, D_t(1,), \ldots, D_t(k))$, $t \in \mathbf{N}$, are independent and identically distributed; this condition guarantees that the system determined by (5.1) is a MDP.

C2: For a positive integer c, $D_t(u) \leq c$, $t \in \mathbf{N}$, $u \in U$.

Set $\lambda := E[A_t]$ and $\mu(u) := E[D_t(u)]$, $u \in U$.

C3: $0 < \lambda < \mu(u), \quad u \in U$.

This model is a variant of the system studied in [11, Section 3]; see also [1]. We shall see in Proposition (5.1) below that C1-C3 together imply that Assumption (2.2) is valid. Then, since U is finite, Theorem (4.1) yields that, for arbitrary *bounded* reward function, there exists an optimal stationary policy.

Set $z = 0$ so that T in (2.3) is the first time for which the system is empty.

PROPOSITION (5.1). *Under conditions* C1-C3, *Assumption* (22) *holds true.*

Proof. Let $X_0 = x \in S$ be arbitrary. Notice that the total number of customers asking for service in $[0, 1)$ is $x + A_0$; in particular, the number of services to be provided in $[0, T)$ is, at least, $x + A_0$. Since at most c customers can be served in a unit of time (by C2), it is clear that $T \geq (x + A_0)/c$. Then,

$$E_\pi[T|X_0 = x] \geq (x + \lambda)/c, \quad \pi \in \mathcal{P}. \tag{5.2}$$

On the other hand, for a positive integer n, define

$$Y_n := x + \sum_{t=0}^{n-1} (A_t - D_t(U_t)), \quad n = 1, 2, \ldots .$$

With this notation we have (see (2.3) and (5.1)) that

$$T = \min \{n > 0 | Y_n < 0\}. \tag{5.3}$$

This implies that $Y_n > 0$ for $1 \leq n < T$. Also, when $T > 1$, $Y_T = Y_{T-1} + A_{T-1} - D_{T-1}(U_{T-1}) \geq Y_{T-1} - c \geq -c$. Consequently:

$$n \leq T \Rightarrow Y_n \geq -c. \tag{5.4}$$

Now, observe that we always have $[T > t] \subset \sigma(X_s, 0 \leq s \leq t) \subset \sigma(X_0, A_s, D_s(U_s), 0 \leq s \leq t - 1)$; see (2.3) and (5.1). Using this and C1, we see that, under the action of a policy $f \in \mathbf{F}$, $I[T > t]$ is *independent* of A_t and $D_t(U_t)$. Hence:

(i) $E_f[A_t I[T > t]|X_0 = x] = \lambda \cdot P_f[T > t|X_0 = x]$.

Also, observing that

$$E_f[D_t(U_t)|X_0 = x] = E_f[D_t(f(X_t)|X_0 = x] = \mu(f(x)),$$

it follows that

(ii) $E_f[D_t(U_t) I[T > t]|X_0 = x] = P_f[T > t|X_0 = x] \cdot E_f[\mu(f(X_t))|X_0 = x] \geq P_f[T > t|X_0 = x] \cdot \mu_0$, where $\mu_0 := \min \{\mu(u)|u \in U\} > \lambda$; see C3.

Let $n \geq 1$ be arbitrary and, for convenience, set $T(n) := \min \{T, n\}$. Using (i) and (ii) above together with (5.4) we obtain:

$$-c \le E_f[Y_{T(n)}|X_0 = x] = E_f[x + \sum_{t=0}^{T(n)-1} (A_t - D(U_t))|X_0 = x]$$

$$= x + E_f[\sum_{t=0}^{n-1} (A_t - D(U_t)) I[T > t]|X_0 = x]$$

$$\le x + (\lambda - \mu_0) \sum_{t=0}^{n-1} P_f[[T > t]|X_0 = x],$$

and then

$$M_f(x) = E_f[T|X_0 = x] = \lim \sum_{t=0}^{n-1} P_f[[T > t]|X_0 = x]$$

$$\le (x + c)/(\mu_0 - \lambda) < \infty; \tag{5.5}$$

this proves that part (i) in Assumption (2.2) is valid. To complete the proof, use (5.2) and (5.5) to obtain: For all ϕ, $f \in \mathbf{F}$ and $x \in S$,

$$M_\phi(x)/M_f(x) \le [(x + c)/(\mu_0 - \lambda)] \cdot c/(x + \lambda)$$

$$\le c^2/[\lambda(\mu_0 - \lambda)] =: \gamma < \infty. \quad \square$$

DEPARTAMENTO DE ESTADÍSTICA Y CÁLCULO
UNIVERSIDAD AUTÓNOMA AGRARIA ANTONIO NARRO
BUENAVISTA, 25315, SALTILLO, COAH., MÉXICO

REFERENCES

[1] V. S. BORKAR, *On minimum cost per unit of time control of Markov chains*, SIAM J. Control and Optimization, **22** (1984), 965-978.

[2] R. CAVAZOS-CADENA *Necessary conditions for a bounded solution to the average reward optimality equation*, J. Applied Mathematics and Optimization, **19** (1989), 97-112.

[3] ———, *Necessary and sufficient conditions for a bounded solution to the optimality equation in average reward Markov decision chains*, Systems and Control Letters, **10** (1988), 71-78.

[4] ———, *Weak conditions for the existence of optimal stationary policies in average cost Markov decision chains with unbounded costs*, Kybernetika, **25** (1989), 145-156.

[5] Ll. FISHER AND S. M. ROSS, *An example in denumerable decision processes*, Annals of Mathematical Statistics, **39** (1968), 674-675.

[6] O. HERNÁNDEZ-LERMA, Adaptive Markov Control Processes, Springer-Verlag, New York, 1989.

[7] D. HEYMAN AND M. SOBEL, Stochastic Models in Operation Research, II, McGraw-Hill, New York, 1984.

[8] K. HINDERER, Foundations of Non-stationary Dymamic Programming with Discrete Time Parameter, Lecture Notes in Operations Research **33**, Springer Verlag, New York, 1970.

[9] S. M. ROSS, Applied Probability Models with Optimization Applications, Holden-Day, San Francisco, 1970.

[10] ———, Introduction to Stochastic Dymamic Programming, Academic Press, New York, 1983.

[11] L. I. SENNOTT, *A new condition for the existence of optimal stationary policies in average cost Markov decision chains*, Operations Research Letters, **5**, (1988), 17-23.

[12] L. C. THOMAS, *Conectedness conditions for denumerable state Markov decision processes*, in Recent Developments in Markov Decision Processes, R. Hartley, L. C. Thomas and D. J. White editors, Academic Press, New York, (1980), 181-204.