Rolando Cavazos-Cadena

Solution to the optimality equation in a class of Markov decision chains with the average cost criterion

Kybernetika, Vol. 27 (1991), No. 1, 23--37

Persistent URL: http://dml.cz/dmlcz/125657

KYBERNETIKA — VOLUME 27 (1991), NUMBER 1

SOLUTION TO THE OPTIMALITY EQUATION IN A CLASS OF MARKOV DECISION CHAINS WITH THE AVERAGE COST CRITERION*

ROLANDO CAVAZOS-CADENA

We consider *average* cost Markov Decision Chains with denumerable state space and discrete time parameter. The existence of a *single* stationary policy under which the state space is an irreducible and ergodic class is assumed. The cost function is *unbounded* and satisfies the following *structural condition:* Given a real number, the cost always exceeds that number except at a finite set of states. Within this framework it is proved that the optimality equation has a solution that yields optimal stationary policies, and that the lim sup and lim inf average criteria are equivalent.

1. INTRODUCTION

We are concerned with Markov Decision Processes (MDP's) with denumerable state space and discrete time parameter. Besides standard continuity — compactness conditions, the class of models we consider in this note is determined by the following assumptions: (i) There exists a *single* stationary policy under which (a) the state space is an irreducible and ergodic class *and* (b) the average cost is finite, and (ii) The cost function has a "penalized structure" in the following sense: Given a real number r, there exists a finite set of states, say $G = G(r)$, such that, outside G the cost is larger than r; see Assumption 3.2. Within this framework (which is the same as in [5]) we obtain the following results: (1) The Average Cost Optimality Equation (ACOE) admits a solution that yields optimal stationary policies in "the standard way" ([14] or [11, Th. 6.17]); (2) we prove that the lim sup and lim inf average criteria are *equivalent* in the sense that they yield the *same* optimal value function; see Section 3 for a more precise description. These results are a solution to the problems posed in [5, Section 4].

The organization of the paper is as follows: Section 2 contains some notation

* This research was supported in part by the Consejo Nacional de Ciencia y Technología (CONACYT) under Grant PCEXCNA 050156 and by SEMAC under Grant 89-2/00ifn$.

and terminology while in Section 3 the model and the problems under consideration are formally described. Section 4 contains the preliminaries we need to prove that the ACOE admits a solution in Section 5. The equivalence of the lim sup and lim inf optimality criteria is proved in Section 6 and we conclude in Section 7 with some brief comments.

2. NOTATION AND TERMINOLOGY

As usual, $\mathbb{R}$ stands for the set of real numbers and $\mathbb{N} := \{0, 1, 2, \ldots\}$. Given the sets A and B, their cartesian product is denoted by $A \times B$; when A and B are topological spaces, $A \times B$ is endowed with the product topology. If A is a metric space, $\mathscr{B}(A)$ is the corresponding Borel σ-field, and $\mathbb{P}(A)$ is the class of all probability measures on $(A, \mathscr{B}(A))$. For metric spaces A and H, the class of all transition probabilities on A given H is denoted by $\mathbb{P}(A \mid H)$, i.e., $\pi = \pi(\cdot \mid \cdot) \in \mathbb{P}(A \mid H)$ if and only if (i) For each $h \in H$, $\pi(\cdot \mid h)$ is a probability measure on $(A, \mathscr{B}(A))$, and (ii) For each $A' \in \mathscr{B}(A)$, $\pi(A' \mid \cdot)$: $H \to [0, 1]$ is measurable. On the other hand, if Y is a random vector, $\sigma(Y)$ is the σ-field generated by Y and, for an event W, the corresponding indicator function is denoted by $I[W]$. Finally, (in)equalities involving conditional expectations are understood to hold true almost surely (a.s.) with respect to the underlying probability measure *without* including this statement explicitly.

3. THE MODEL

Let (S, A, C, p) be a Markov decision process, where the *state space* S is a nonempty denumerable set endowed with the discrete topology and the *action set* A is assumed to be a *compact metric* space. The *transition law* is p and C: $S \times A \to \mathbb{R}$ is the *cost function*. This model represents a dynamical system evolving as follows: At each time $t \in \mathbb{N}$ the state of the system is observed, say $X_t = x \in S$. Then, an action $A_t = a \in A$ is chosen and, consequently (i) a cost $C(x, a)$ is incurred and (ii) regardless of the states observed and actions applied prior to t, the state of the system at time $t + 1$ will be $y \in S$ with probability $p_{xy}(a)$; of course, $\sum_{y \in S} p_{xy}(a) =: \sum_{y} p_{xy}(a) = 1$.

Assumption 3.1. (i) The cost function C is nonnegative.
(ii) For each x, $y \in S$, the mappings

$$a \to p_{xy}(a), \quad a \in A \tag{3.1}$$

and

$$a \to C(x, a), \quad a \in A$$

are *lower semicontinuous*.

Remark 3.1. (i) In the usual description of an MDP it is supposed that (in general) not all the actions in A are available at each state x. Rather, the set of *admissible*

actions at the state $x \in S$ is assumed to be a nonempty set $A(x) \subset A$ and the MDP can be represented as $(S, A, \{A(x)\}, C, p)$; here, we are assuming that $A(x) = A$ for all $x \in S$. However, as noted by Borkar [4], this condition does *not* imply any loss of generality; every MDP with denumerable state space for which the sets $A(x)$ are *not* (necessarily) equal, can be transformed into an MDP for which *all* the sets $A(x)$ are the *same*.

(ii) Let $x, y \in S$ be arbitrary but *fixed*. Using Theorems 2 and 4 in [1, Appendix 6], it is not difficult to see that Assumption 3.1 implies lower semicontinuity of the mapping $a \to \sum_{z \neq y} p_{xz}(a)$, $a \in A$ and then $a \to p_{xy}(a) = 1 - \sum_{z \neq y} p_{xz}(a)$, $a \in A$ is *upper semicontinuous*. Hence: The mappings in (3.1) are, in fact, *continuous*.

Policies. For $t \in \mathbb{N}$, let H_t be the set of *possible histories* up to time n, that is, $H_0 := S$, and $H_t := (S \times A)^t \times S$ for $t \geqq 1$. A typical element of H_t is denoted by $h_t = (x_0, a_0, \ldots, x_{t-1}, a_{t-1}, x_t)$. A *policy* $\pi = \{\pi_t\}_0^\infty$ is a sequence of conditional probabilities such that $\pi_t \in \mathcal{P}(A \mid H_t)$, $t \in \mathbb{N}$; the class of all policies is denoted by $\mathcal{P}$. A policy π is *randomized stationary* if there exists $\gamma \in \Gamma := \mathcal{P}(A \mid S)$ such that the equality $\pi_t(\cdot \mid h_t) = \gamma(\cdot \mid x_t)$ always holds. We naturally identify Γ with the class of all randomized stationary policies. On the other hand, let $\mathbb{F} := \prod_{x \in S} A$ i.e., $\mathbb{F}$ is the class of all functions $f: S \to A$; notice that $\mathbb{F}$ is *compact metric* in the product topology [9]. A policy π is (deterministic) *stationary* if there exists $f \in \mathbb{F}$ satisfying the following: For all $t \in \mathbb{N}$ and $h_t \in H_t$, $\pi_t(\cdot \mid h_t)$ is the unit of mass concentrated at $f(x_t)$; the class of stationary policies is identified with $\mathbb{F}$. With the above conventions, $\mathbb{F} \subset \Gamma \subset \mathcal{P}$.

The state-action process $\{(X_t, A_t)\}$ can be thought of as the coordinate sequence in $\Omega := (S \times A)^\infty$. Given the initial state $X_0 = x$ and the policy π being used, a *unique* probability measure is determined on $(\Omega, \mathcal{B}(\Omega))$ [1, p. 109] or [7, p. 80]. This measure is denoted by P_x^π and E_x^π stands for the corresponding expectation operator; of course, we always have $P_x^\pi[X_0 = x] = 1$. Finally we observe that, under the action of a policy $\gamma \in \Gamma$, the state process $\{X_t\}$ is a Markov chain with stationary transition mechanism [7, 11]; the corresponding one-step transition probability matrix is denoted by P^γ, i.e.,

$$P_{xy}^\gamma := \int p_{xy}(a)\, \gamma(\mathrm{d}a \mid x), \quad x, y \in S.$$

As usual, we (sometimes) identify P^γ with the Markov chain determined by γ.

Optimality Criteria. To evaluate the performance of a policy we now introduce the lim sup and the lim inf average cost criteria.

Definition 3.1. Let $x \in S$ and $\pi \in \mathcal{P}$ be arbitrary but *fixed*.

(i) The lim sup *average cost* at state x under policy π is

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

and

$$J_+(x) := \inf_{\pi \in \mathscr{P}} J_+(x, \pi)$$

is the *optimal* lim sup *average cost* at state x.

(ii) the lim inf *average cost* at state x when policy π is used is defined by

$$J_-(x, \pi) := \liminf_n \mathsf{E}_x^\pi\Big[\sum_{t=0}^{n} C(X_t, A_t)\Big]/(n+1)\,;$$

$$J_-(x) := \inf_{\pi \in \mathscr{P}} J_-(x, \pi)$$

is the *optimal* lim inf *average cost* at state x.

(iii) A policy π is lim sup (resp. lim inf) *optimal* if, for all $x \in S$,

$$J_+(x, \pi) = J_+(x) \quad (\text{resp. } J_-(x, \pi) = J_-(x))\,.$$

Notice that, since $C \geqq 0$, the above expectations are well defined (their value may be ∞).

Assumption 3.2. (i) (Penalized structure of the cost function). For each real number r, there exists a *finite* set $G = G(r) \subset S$ such that

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

(ii) (Communication/Finite average cost). There exist a (single) policy $f^* \in \mathbb{F}$ satisfying (a) and (b) below.

(a) Under f^*, the state process is an *irreducible and ergodic* Markov chain, and

(b) If $\{q_{f^*}(x) \mid x \in S\}$ is the *unique* invariant distribution of the Markov chain induced by f^* [8, 11],

$$g^* := \sum_x q_{f^*}(x)\, C(x, f^*(x)) < \infty\,.$$

Remark 3.2. (i) It is well known that, with f^* and g^* as in Assumption 3.2, $g^* = J_+(x, f^*) = J_-(x, f^*)$ for all $x \in S$ [8, 11]. Hence, from the definition of the optimal value functions J_+ and J_- it follows that

$$\infty > g^* \geqq J_+(x) \geqq J_-(x) \geqq 0\,, \quad x \in S\,.$$

(ii) Assumption 3.2 is satisfied in interesting queuing models [2, 10] for which $S = \mathbb{N}^k (k = 1, 2, \ldots)$ and C is a polynomial function (usually linear or quadratic) of the state.

The problems. Under our assumptions, the existence of a lim sup optimal stationary policy was proved in [5]. Also, it was proved there that $J_+(\cdot)$ is a *constant function*. However to obtain a solution to the ACOE (see (3.2) below) additional conditions were imposed. Our first problem concerns the existence of (appropriate) solutions to the ACOE under Assumptions 3.1 and 3.2 *alone*.

Problem 1. Let $g = J_+(x)$, $x \in S$. Does $h: S \to \mathbb{R}$ exist such that (a) and (b) below are satisfied?

(a) h is bounded below.

(b) For all $x \in S$,

$$g + h(x) = \min_{a \in A} [C(x, a) + \sum_y p_{xy}(a)\, h(y)] . \tag{3.2}$$

Our second problem concerns the "equivalence" of the lim sup and lim inf average cost criteria: Given $\pi \in \mathscr{P}$, the lim sup (lim inf) criterion represents the largest (smallest) limit points of the expected average costs over finite horizons under policy π. Since minimizing the smallest average *cost* is "more appealing" than minimizing the largest one, it is natural to ask if the lim sup and lim inf criteria are equivalent in the sense that they yield the *same* optimal value function.

Problem 2. Is it true that $J_+(x) = J_-(x)$ for all $x \in S$?

Our solution to problem 1 is along the ideas developed in [5], while problem 2 is solved (essentiallly) by adapting results from [4] to our present framework.

4. PRELIMINARIES

In this section we recall the results from [5] which are needed to solve the first of the problems posed above; see Lemmas 4.1–4.3 below. We begin by introducing the *discounted* criterion: For $\alpha \in [0, 1)$, $x \in S$ and $\pi \in \mathscr{P}$, the *α-discounted cost* at state x under policy π is defined by

$$V_\alpha(x, \pi) := \mathsf{E}_x^\pi[\sum_{t=0}^{\infty} \alpha^t\, C(X_t, A_t)] ,$$

while

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

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 $x \in S$.

Lamma 4.1 (cf. [5].) (i) Let $\alpha \in [0, 1)$. For each $x \in S$, $V_\alpha(x) < \infty$, and there exists an α-discounted optimal *stationary* policy $f_\alpha \in \mathbb{F}$.

(ii) $\lim_{\alpha \to 1-} (1 - \alpha)\, V_\alpha(x) =: g$ exists and does *not* depend on $x \in S$. Moreover,

(iii) $J_+(x) = g$ for all $x \in S$, and

(iv) There exist $\beta \in [0, 1)$ and a *finite* set $\bar{G} \subset S$ such that, for $\alpha \in [\beta, 1)$, $V_\alpha(\cdot)$ attains its minimum in $\bar{G}$. In other words, given $\alpha \in [\beta, 1)$, there exists $x_\alpha \in \bar{G}$ such that

$$V_\alpha(x) \geqq V_\alpha(x_\alpha) , \quad x \in S . \tag{4.1}$$

Now, let G be a *nonempty* subset of S. The stopping time T_G is defined by

$$T_G := \min\{n > 0 \mid X_n \in G\} \tag{4.2}$$

where, by convention, the minimum of the empty set is ∞. If $G = \{z\}$ is a singleton we simply write T_z instead of T_G. Define $h^*: S \times S \to \mathbb{R}$ by

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

Lemma 4.2. (cf. [5].) Let $x, z \in S$ be arbitrary. Then
(i) $h^*(x, z)$ is *finite*, and
(ii) $V_\alpha(x) - V_\alpha(z) \leqq h^*(x, z)$.

Throughout the paper $x_0 \in S$ is *arbitrary but fixed* and for $\alpha \in [0, 1)$, $h_\alpha: S \to \mathbb{R}$ is defined by

$$h_\alpha(x) := V_\alpha(x) - V_\alpha(x_0), \quad x \in S. \tag{4.3}$$

By Lemma 4.2 we have

$$h_\alpha(x) \leqq M(x) := h^*(x, x_0) < \infty, \quad x \in S, \quad \alpha \in [0, 1). \tag{4.4}$$

Now, let $\beta \in [0, 1)$ and the *finite* set $\bar{G}$ be as in Lemma 4.1 (iv) and define N by

$$N := \max\{h^*(x_0, x) \mid x \in \bar{G}\}; \tag{4.5}$$

notice that, since $\bar{G}$ is a *finite* set, Lemma 4.2 (i) implies that N is a *finite* number. Let $\alpha \in [\beta, 1)$ and take $x_\alpha \in \bar{G}$ satisfying (4.1). Then $V_\alpha(x_0) - V_\alpha(x_\alpha) \leqq N$ (see Lemma 4.2 (ii) and (4.5)) which implies $h_\alpha(x) = [V_\alpha(x) - V_\alpha(x_\alpha)] + [V_\alpha(x_\alpha) - V_\alpha(x_0)] \geqq -N$ for all $x \in S$. Combining this with (4.4) we obtain:

$$-N \leqq h_\alpha(x) \leqq M(x), \quad x \in S, \quad \alpha \in [\beta, 1). \tag{4.6}$$

Lemma 4.3. For $\alpha \in [0, 1)$, let $f_\alpha \in \mathbb{F}$ be α-discounted optimal.

(i) For each sequence $\{\alpha_n\} \subset [\beta, 1)$ converging to 1, there exists a subsequence $\{\beta_n\}$ such that the following limits exist for all $x \in S$:

$$\lim_n f_{\beta_n}(x) =: f(x) \in A \tag{4.7}$$

and

$$\lim_n h_{\beta_n}(x) =: h(x) \in [-N, M(x)]. \tag{4.8}$$

(ii) Let $f \in \mathbb{F}$ and $h: S \to \mathbb{R}$ be determined by (4.7) and (4.8). Then, f is lim sup *optimal* and

$$g + h(x) \geqq C(x, f(x)) + \sum_y p_{xy}(f(x))\, h(y), \quad x \in S. \tag{4.9}$$

(iii) *Every* policy $f \in \mathbb{F}$ satisfying (4.9) is lim sup *optimal*.

This is a slight variant of the results in [5, Section 3].

5. THE OPTIMALITY EQUATION

We now solve problem 1. Let $\{\beta_n\} \subset [\beta, 1)$ be a *fixed* sequence converging to 1 so that the convergences in (4.7) and (4.8) hold true for all $x \in S$. Taking a subsequence if necessary we can assume that

$$\beta_n \text{ increases to } 1 \text{ as } n \to \infty . \tag{5.1}$$

Our solution to problem 1 is the following.

Theorem 5.1. (i) The *optimal* lim sup *average cost* g and the function h in (4.8) satisfy the ACOE, that is

$$g + h(x) = \min_{a \in A} \left[C(x, a) + \sum_y p_{xy}(a)\, h(y)\right], \quad x \in S . \tag{5.2}$$

Moreover,

(ii) The equality holds in (4.9) for all $x \in S$, and

(iii) Let $\bar{f} \in \mathbb{F}$ be such that, for each $x \in S$, $\bar{f}(x)$ is a minimizer of the term in brackets in (5.2). Then $\bar{f}$ is lim sup optimal.

Part (ii) of this Theorem follows immediately from part (i), and part (iii) can be obtained from Lemma 4.3 (iii). Thus, only part (i) needs to be proved. Since the proof is somewhat technical we first establish some preliminaries given in the form of Lemmas 5.1 and 5.2 below.

Let $G_1 := G(g + 1)$ be the *finite* set guaranteed by Assumption 3.2 and, by notational convenience, write

$$T_1 := T_{G_1} ; \tag{5.3}$$

see (4.2) for the definition of T_{G_1}. We recall that $C(x, a) \geqq g + 1$ whenever $x \notin G_1$. On the other hand, it is clear that $[T_1 > t] \subset [X_t \notin G_1]$, $t \in \mathbb{N} - \{0\}$. Therefore,

$$C(X_t, A_t) \geqq g + 1 \quad \text{if} \quad 1 \leqq t < T_1 . \tag{5.4}$$

Lemma 5.2. Let the policy $l \in \mathbb{F}$ and $\Phi: S \to [0, \infty)$ satisfy

$$g + h(x) = C(x, l(x)) + \Phi(x) + \sum_y p_{xy}(l(x))\, h(y), \quad x \in S . \tag{5.5}$$

Then, for each $x \in S$,

(i) $h(x) \geqq \mathsf{E}_x^l\left[\sum_{t=0}^{T_1 - 1} (C(X_t, A_t) + \Phi(X_t) - g) + h(X_{T_1})\right]$, and

(ii) $\mathsf{E}_x^l[T_1] < \infty$.

Proof. First, we observe that a simple induction argument yields that

$$(n + 1)\, g + h(x) = \mathsf{E}_x^l\left[\sum_{t=0}^{n} (C(X_t, A_t) + \Phi(X_t)) + h(X_{n+1})\right], \quad n \in \mathbb{N}, \quad x \in S .$$

Since the left hand side is finite and $C + \Phi \geqq 0$, this immediately implies that $\mathsf{E}_x^l[h(X_n)] < \infty$ for all $n \in \mathbb{N}$ and $x \in S$; recall that $\mathsf{E}_x^l[h(X_0)] = h(x)$. Using that h

is bounded below we conclude that

$$\mathsf{E}_x^I[|h(X_n)|] < \infty\,, \quad x \in S\,, \quad n \in \mathbb{N}\,. \tag{5.6}$$

Let $x \in S$ be arbitrary but *fixed*. For $t \in \mathbb{N}$, (5.5) implies

$$h(X_t) = (C(X_t, A_t) + \Phi(X_t) - g) + \mathsf{E}_x^I[h(X_{t+1})|\,\mathscr{F}_t]$$

where $\mathscr{F}_t := \sigma(X_0, \ldots, X_t)$; since $[T_1 > t] \in \mathscr{F}_t$. the above equation implies that

$$h(X_t)\,I[T_1 > t] = (C(X_t, A_t) + \Phi(X_t) - g)\,I[T_1 > t] +$$
$$+ \mathsf{E}_x^I[h(X_{t+1})\,I[T_1 > t]\,|\,\mathscr{F}_t]\,,$$

and, after writing $I[T_1 > t] = I[T_1 > t+1] + I[T_1 = t+1]$ in the conditional expectation, a transposition yields

$$h(X_t)\,I[T_1 > t] - \mathsf{E}_x^I[h(X_{t+1})\,I[T_1 > t+1]\,|\,\mathscr{F}_t] =$$
$$= (C(X_t, A_t) + \Phi(X_t) - g)\,I[T_1 > t] + \mathsf{E}_x^I[h(X_{t+1})\,I[T_1 = t+1]\,|\,\mathscr{F}_t]\,.$$

Take expectations in both sides of this equality to obtain

$$\mathsf{E}_x^I[h(X_t)\,I[T_1 > t]] - \mathsf{E}_x^I[h(X_{t+1})\,I[T_1 > t+1]] =$$
$$= \mathsf{E}_x^I(h(X_t)\,I[T_1 > t] - \mathsf{E}_x^I[h(X_{t+1})\,I[T > t+1]\,|\,\mathscr{F}_t]) =$$
$$= \mathsf{E}_x^I[(C(X_t, A_t) + \Phi(X_t) - g)\,I[T_1 > t]] + \mathsf{E}_x^I[h(X_{t+1})\,I[T_1 = t+1]]\,,$$

and summing up from $t = 0$ to n we conclude: For all $n \in \mathbb{N}$,

$$h(x) - \mathsf{E}_x^I[h(X_{n+1})\,I[T_1 > n+1]] =$$
$$= \mathsf{E}_x^I[\sum_{t=0}^{n}(C(X_t, A_t) + \Phi(X_t) - g)\,I[T_1 > t]] +$$
$$+ \mathsf{E}_x^I[\sum_{t=0}^{n} h(X_t)\,I[T_1 = t+1]]\,; \tag{5.7}$$

notice that $\mathsf{E}_x^I[h(X_0)\,I[T_1 > 0]] = \mathsf{E}_x^I[h(X_0)] = h(x)$. Now observe that, since C and Φ are nonnegative, (5.4) implies that $-g \leqq C(X_0, A_0) + \Phi(X_0) - g \leqq$ $\leqq \sum_{t=0}^{n}(C(X_t, A_t) + \Phi(X_t) - g)\,I[T_1 > t] \nearrow \sum_{t=0}^{\infty}(C(X_t, A_t) + \Phi(X_t) - g)\,I[T_1 > t] =$ $= \sum_{t=0}^{T_1-1}(C(X_t, A_t) + \Phi(X_t) - g)$. Therefore, the monotone convergence theorem yields

$$\lim_n \mathsf{E}_x^I[\sum_{t=0}^{n}(C(X_t, A_t) + \Phi(X_t) - g)\,I[T_1 > t]] =$$
$$= \mathsf{E}_x^I[\sum_{t=0}^{T_1-1}(C(X_t, A_t) + \Phi(X_t) - g)]\,. \tag{5.8}$$

On the other hand, since $h \geqq -N$ (see (4.8)), Fatou's Lemma implies

$$\liminf_n \mathsf{E}_x^I[\sum_{t=0}^{n} h(X_{t+1})\,I[T_1 = t+1]] = \liminf_n \mathsf{E}_x^I[h(X_{T_1})\,I[T_1 \leqq n+1]]$$
$$\geqq \mathsf{E}_x^I[h(X_{T_1})\,I[T_1 < \infty]] \geqq -N P_x^I[T_1 < \infty]\,. \tag{5.9}$$

Also, it is clear that

$$NP_x^l[T_1 = \infty] = \liminf_n \mathsf{E}_x^l[NI[T_1 > n+1]] \geqq$$
$$\geqq \liminf_n \mathsf{E}_x^l[-h(X_{n+1})\, I[T_1 < n+1]] \,. \tag{5.10}$$

Taking lim inf as $n \to \infty$ in both sides of (5.7) and using (5.8)–(5.10) we immediately obtain

$$h(x) + NP_x^l[T_1 = \infty] \geqq \mathsf{E}_x^l[\sum_{t=0}^{T_1-1} (C(X_t, A_t) + \Phi(X_t) - g)] +$$
$$+ \mathsf{E}_x^l[h(X_{T_1})\, I[T_1 < \infty]] \,. \tag{5.11}$$

Hence:

$$h(x) + NP_x^l[T_1 = \infty] \geqq C(x, l(x)) + \Phi(x) - g +$$
$$+ \mathsf{E}_x^l[\sum_{t=1}^{T_1-1} (C(X_t, A_t) + \Phi(X_t) - g)] - NP_x^l[T_1 < \infty] \geqq$$
$$\geqq C(x, l(x)) + \Phi(x) - g + \mathsf{E}_x^l[\sum_{t=1}^{T_1-1} 1] - NP_x^l[T_1 < \infty] \,;$$

here, we have used that (a) $\Phi \geqq 0$ together with (5.4), and (b) the last inequality in (5.9). From this we conclude that $\mathsf{E}_x^l[T_1] \leqq 1 + g + h(x) + N - \Phi(x) - C(x, l(x)) < \infty$ which proves part (ii) of the Lemma. It also yields $P_x^l[T_1 < \infty] = 1$ and then, part (i) follows from (5.11). □

Remark 5.1. The inequality (5.6) played an important role in the above proof, namely, it "legitimates" the arguments involving conditional expectations.

Lemma 5.2. Let $x \in S$ be arbitrary and assume that the policy $l \in \mathbb{F}$ satisfies

$$\mathsf{E}_x^l[T_1] < \infty \,. \tag{5.12}$$

Then

$$h(x) \leqq \mathsf{E}_x^l[\sum_{t=0}^{T_1-1} (C(X_t, A_t) - g) + h(X_{T_1})] \,. \tag{5.13}$$

Proof. Let $\alpha \in [0, 1)$ and define the policy π as follows: For $t < T_1$, $\pi_t := l$ and $\pi_t := f_\alpha$ for $t \geqq T_1$. This definition of π and the Markov property together imply [11]

$$V_\alpha(x, \pi) = \mathsf{E}_x^l[\sum_{t=0}^{T_1-1} \alpha^t\, C(X_t, A_t) + \alpha^{T_1}\, V_\alpha(X_{T_1})] \,. \tag{5.14}$$

Now, define $g_\alpha := (1 - \alpha)\, V_\alpha(x_0)$. Straightforward computations yield

$$V_\alpha(x_0) = \mathsf{E}_x^l[\sum_{t=0}^{T_1-1} \alpha^t g_\alpha + \alpha^{T_1}\, V_\alpha(x_0)] \,;$$

combining this equality with (5.14) we obtain

$$h_\alpha(x) = V_\alpha(x) - V_\alpha(x_0) \leqq V_\alpha(x, \pi) - V_\alpha(x_0) =$$
$$= \mathsf{E}_x^l[\sum_{t=0}^{T_1-1} \alpha^t (C(X_t, A_t) - g_\alpha) + \alpha^{T_1}\, h_\alpha(X_{T_1})] \,,$$

and then,

$$h_\alpha(x) \leqq \mathsf{E}_x^l\Big[\sum_{t=0}^{T_1-1} \alpha^t(C(X_t, A_t) - g) + \alpha^{T_1} h_\alpha(X_{T_1})\Big] + (g - g_\alpha)\,\mathsf{E}_x^l[T_1]. \quad (5.15)$$

Now, since $\beta_n \nearrow 1$ as $n \to \infty$, Lemma 4.2 (ii) and (5.12) together imply that

$$(g - g_{\beta_n})\,\mathsf{E}_x^l[T_1] \to 0 \quad \text{as} \quad n \to \infty\,. \quad (5.16)$$

On the other hand, for all $n \in \mathbb{N}$,

$$\left|\beta_n^{T_1} h_{\beta_n}(X_{T_1})\right| \leqq M < \infty$$

where $M := \max\{|N|, |M(x)|,\ x \in G_1\}$, see (4.6) and recall that X_{T_1} belongs to the *finite* set G_1. Using the bounded convergence theorem we obtain

$$\lim_n \mathsf{E}_x^l[\beta_n^{T_1} h_{\beta_n}(X_{T_1})] = \mathsf{E}_x^l[h(X_{T_1})]\,. \quad (5.17)$$

Finally, it is clear that, as $n \to \infty$

$$-g \leqq C(X_0, A_0) - g \leqq \sum_{t=0}^{T_1-1} \beta_n^t(C(X_t, A_t) - g) \nearrow \sum_{t=0}^{T_1-1} (C(X_t, A_t) - g)$$

(see (5.1) and (5.4)), and the monotone convergence theorem yields

$$\lim_n \mathsf{E}_x^l\Big[\sum_{t=0}^{T_1-1} \beta_n^t(C(X_t, A_t) - g)\Big] = \mathsf{E}_x^l\Big[\sum_{t=0}^{T_1-1} (C(X_t, A_t) - g)\Big]\,. \quad (5.18)$$

To conclude, replace α by β_n in (5.15) and take limit as $n \to \infty$ in both sides of the resulting inequality. Using (5.16)–(5.18) the desired result follows immediately. □

After these lemmas we are ready to prove that g and h satisfy the ACOE.

Proof of Theorem 5.1. As already noted, only part (i) needs to be proved. Using that h is bounded below, Assumption 3.1 implies that, for each $x \in S$, the mapping $a \to C(x, a) + \sum_y p_{xy}(a)\, h(y)$, $a \in A$ has a minimizer $l(x) \in A$; see the proof of Theorem 2.1 (ii) in [5]. Then,

$$h(x) + g \geqq \min_{a \in A} \big[C(x, a) + \sum_y p_{xy}\, h(y)\big] = C(x, l(x)) + \sum_y p_{xy}(l(x))\, h(y)$$

where the inequality is due to (4.9). Define $\Phi\colon S \to \mathbb{R}$ by

$$\Phi(x) := h(x) + g - C(x, l(x)) - \sum_y p_{xy}(l(x))\, h(y)\,, \quad x \in S\,.$$

Notice that (5.2) is the same as $\Phi(x) = 0$ for all $x \in S$ and that $\Phi \geqq 0$. Since the relation defining Φ is equivalent to (5.5), Lemma 5.1 yields that for *arbitrary* $x \in S$,

$$h(x) \geqq \mathsf{E}_x^l\Big[\sum_{t=0}^{T_1-1} (C(X_t, A_t) + \Phi(X_t) - g) + h(X_{T_1})\Big] \quad (5.19)$$

and $\mathsf{E}_x^l[T_1] < \infty$. Then Lemma 5.2 implies

$$h(x) \leqq \mathsf{E}_x^l\Big[\sum_{t=0}^{T_1-1} (C(X_t, A_t) - g) + h(X_{T_1})\Big]\,,$$

and combining this inequality and (5.19) we immediately obtain

$$0 \geqq \mathsf{E}_x^f\Big[\sum_{t=0}^{T_1-1} \Phi(X_t)\Big] \geqq \mathsf{E}_x^f[\Phi(X_0)] = \Phi(x) \geqq 0\,,$$

i.e., $\Phi(x) = 0$, and the result follows since $x \in S$ is arbitrary. □

6. EQUIVALENCE OF AVERAGE OPTIMALITY CRITERIA

We now prove that the optimal value functions corresponding to the lim sup and lim inf average criteria coincide. This result is a straightforward consequence of the following theorem which was proved in [4] under conditions on the transition law and the cost function which are substantially stronger than those we are assuming here.

Theorem 6.1. For each $x \in S$ and $\pi \in \mathscr{P}$,

$$\liminf_n \sum_{t=0}^{n} C(X_t, A_t)/(n+1) \geqq g \quad P_x^\pi\text{-a.s.}$$

Before giving a proof of this theorem, we point out the following.

Corollary 6.1. For all $x \in S$

$$J_+(x) = J_-(x)$$

Proof. Let $x \in S$ be *fixed*. For arbitrary $\pi \in \mathscr{P}$, we have

$$J_-(x,\pi) = \liminf_n \mathsf{E}_x^\pi\Big[\sum_{t=0}^{n} C(X_t, A_t)/(n+1)\Big] \geqq$$

$$\geqq \mathsf{E}_x^\pi\Big[\liminf_n \sum_{t=0}^{n} C(X_t, A_t)/(n+1)\Big] \geqq \mathsf{E}_x^\pi[g] = g$$

where we have used Fatou's Lemma and Theorem 6.1. The above inequality yields $J_-(x) \geqq g$ (since π is arbitrary) and then the conclusion follows from $g = J_+(x) \geqq$ $\geqq J_-(x)$; see Lemma 4.1 (iii) and Remark 3.2. □

The proof of Theorem 6.1 is based on Lemmas 6.1 and 6.2 below. Let $(X_t, A_t)_{t=0}^\infty \in$ $\in \Omega\ (=(S \times A)^\infty)$ be a sample path and define $\nu_t \in \mathbb{P}(S \times A)$, $t \in \mathbb{N}$ by

$$\nu_t(L) := \sum_{n=0}^{t} I[(X_t, A_t) \in L]/(t+1), \quad L \in \mathscr{B}(S \times A),\ t \in \mathbb{N}\,. \tag{6.1}$$

Throughout the remainder, $x \in S$ and $\pi \in \mathscr{P}$ are arbitrary but *fixed*.

Lemma 6.1. There exists $\Omega' \subset \Omega$ such that

(i) $P_x^\pi[\Omega'] = 1$, and

(ii) On Ω' the following holds: Let $\{t_k\} \subset \mathbb{N}$ be a sequence such that $t_k \to \infty$

as $k \to \infty$ and assume that $\{\nu_{t_k}\}$ converges weakly to $\nu \in \mathbb{P}(S \times A)$ [3]. Then,

$$\mu(y) = \sum_x \mu(x)\, P^{\gamma}_{xy}\,, \quad y \in S\,, \tag{6.2}$$

where $\mu \in \mathbb{P}(S)$ and $\gamma \in \Gamma$ are defined by

$$\mu(y) := \nu(\{y\} \times A)\,, \quad y \in S \tag{6.3}$$

and, for $A' \in \mathscr{B}(A)$,

$$\begin{aligned} \gamma(A' \mid x) &:= \nu(\{x\} \times A')/\mu(x) \quad \text{if} \quad \mu(x) \neq 0 \\ &:= \Delta(A') \qquad\qquad\quad \text{if} \quad \mu(x) = 0 \end{aligned} \tag{6.4}$$

where $\Delta \in \mathbb{P}(A)$ is arbitrary but *fixed*.

This result can be proved along the lines in [4, Lemma 3.6]; in fact, just minor changes are need to adapt the discussion in [4] to our present framework.

Lemma 6.2. Suppose that $\mu \in \mathbb{P}(S)$ and $\gamma \in \Gamma$ satisfy (6.2) and let $C^{\gamma}\colon S \to \mathbb{R}$ be given by

$$C^{\gamma}(x) := \int C(x, a)\, \gamma(\mathrm{d}a|x)\,. \tag{6.5}$$

(Notice that, under the action of the policy $\gamma \in \Gamma$, $C^{\gamma}(x)$ is the conditional expectation of $C(X_t, A_t)$ given $X_t = x$, where $t \in \mathbb{N}$ is arbitrary.)

Then,

$$\sum_x \mu(x)\, C^{\gamma}(x) \geqq g\,.$$

Proof. To begin with, recall that if $x \in S$ satisfies $\mu(x) > 0$ then x is P^{γ}-positive recurrent [8, p. 41]. Let $R(\subset S)$ be a P^{γ}-positive recurrence class and denote by $Q_R \in \mathbb{P}(S)$ the corresponding *invariant* distribution; this means that (6.2) is satisfied with Q_R instead of μ and $Q_R(x) > 0$ if and only if $x \in R$. It is well known that, for $y \in R$ [8, 11]

$$\begin{aligned} &\sum_x Q_R(x)\, C^{\gamma}(x) = \lim_n \mathsf{E}^{\gamma}_y[\sum_{t=0}^{n} C^{\gamma}(X_t)]/(n+1) = \\ &= \lim_n \mathsf{E}^{\gamma}_y[\sum_{t=0}^{n} C(X_t, A_t)]/(n+1) = J_+(y, \gamma) \geqq J_+(y)\,, \end{aligned}$$

and then,

$$\sum_x Q_R(x)\, C^{\gamma}(x) \geqq g\,; \tag{6.6}$$

see Lemma 4.1 (iii). To conclude recall that any μ satisfying (6.2) can be written as [8, p. 41]

$$\mu = \sum_{R \in \mathscr{R}} c_R Q_R \tag{6.7}$$

where $\mathscr{R}$ consists of all P^{γ}-positive recurrence classes and $c_R := \sum_{x \in R} \mu(x)$, $R \in \mathscr{R}$. Using that $\sum_{R \in \mathscr{R}} c_R = 1$, the conclusion, immediately follows from (6.6) and (6.7). □

Proof of Theorem 6.1. Let Ω' be as in Lemma 6.1. Take a sample path $(X_t, A_t)_{t=0}^{\infty} \in \Omega'$ and define $\mathscr{L}$ by

$$\mathscr{L} := \liminf_n \sum_{t=0}^{n} C(X_t, A_t)/(n+1) = \liminf_n \int C(s,a)\, \nu_n(\mathrm{d}(s,a))\,;$$

see (6.1). Now, pick a subsequence $\{n_k\} \subset \mathbb{N}$ such that

$$\mathscr{L} = \lim_k \int C(s,a)\, \nu_{n_k}(\mathrm{d}(s,a))$$

and consider the following cases.

Case 1: $\{\nu_{n_k}\}_{k=0}^{\infty} \subset \mathbb{P}(S \times A)$ is *not* tight [3, p. 37].

Since A is compact, there exists $\varepsilon > 0$ and a subsequence $\{n_k'\}$ of $\{n_k\}$ such that, for each $G \subset S$, G finite,

$$\liminf_k \nu_{n_k'}[(S-G) \times A] \geqq \varepsilon \tag{6.8}$$

Then, for *finite* sets $G \subset S$,

$$\mathscr{L} = \lim_k \int C(s,a)\, \nu_{n_k'}(\mathrm{d}(s,a)) \geqq \liminf_k \int_{(S-G)\times A} C(s,a)\, \nu_{n_k'}(\mathrm{d}(s,a))$$

$(C \geqq 0)$

and (6.8) immediately implies

$$\mathscr{L} \geqq \varepsilon m(S-G)\,. \tag{6.9}$$

where $m(S-G) := \inf\{C(x,a) \mid x \in S-G,\ a \in A\}$. Letting G increase to S in (6.9) and using Assumption 3.2 (i) we conclude that $\mathscr{L} = \infty$.

Case 2: $\{\nu_{n_k}\}$ *is* tight [3].

In this case there exists a subsequence $\{n_k'\}$ of $\{n_k\}$ such that $\{\nu_{n_k'}\}$ converges weakly to $\nu \in \mathbb{P}(S \times A)$ [3, Ch. 1, Section 6]. For $N \in \mathbb{N}$ define $C_N : S \times A \to \mathbb{R}$ by $C_N(x,a) := \min\{C(x,a), N\}$, $(x,a) \in S \times A$. Clearly, $0 \leqq C_N \leqq C$ and C_N is lower semicontinuous [1, Appendix 6, Theorem 4]. Then,

$$\mathscr{L} = \lim_k \int C(s,a)\, \nu_{n_k'}(\mathrm{d}(s,a)) \geqq \liminf_k \int C_N(s,a)\, \nu_{n_k'}(\mathrm{d}(s,a)) \geqq$$

$$\geqq \int C_N(s,a)\, \nu(\mathrm{d}(s,a))$$

(see [3, p. 17]). Using that $C_N \nearrow C$ as $N \to \infty$ the monotone convergence theorem together with the last inequality imply

$$\mathscr{L} \geqq \int C(s,a)\, \nu(\mathrm{d}(s,a)) = \int C(s,a)\, \gamma(\mathrm{d}a|s)\, \mu(\mathrm{d}s) =$$

$$= \int C^{\gamma}(s)\, \mu(\mathrm{d}s) = \sum_s \mu(s)\, C^{\gamma}(s)$$

where μ and γ and C^{γ} are as in (6.3), (6.4) and (6.5) respectively. Since the sample path $(X_t, A_t)_{t=0}^{\infty}$ belongs to Ω', Lemma 6.1 (ii) impiies that (6.2) is satisfied and then, Lemma 6.2 yields that $\mathscr{L} \geqq \sum_s \mu(s)\, C^{\gamma}(x) \geqq g$.

In short: We have seen that for arbitrary $(X_t, A_t)_{t=0}^{\infty} \in \Omega'$, the inequality $\mathscr{L} = \liminf_n \sum_{t=0}^{N} C(X_t, A_t)/(n+1) \geq g$ holds true. This completes the proof since $P_x^{\pi}[\Omega'] = 1$. □

7. CONCLUDING REMARKS

We have given sufficient conditions for the equivalence of average cost optimality criteria and for the existence of solutions to the ACOE yielding optimal stationary policies. Our conditions on the recurrence structure of the model (Assumption 3.2 (ii)) are substantially weaker than those usually found in the MDP literature, e.g. the simultaneous Doeblin or the Lyapounov function condition; see [14] and the inner references. In contrast, we imposed a *very special* (and hence, *restrictive*) structure on the cost function, (Assumption 3.2 (i)) which has played a central role in our argumentation. However, it should be noted that our assumptions hold true for importants models in applications [2, 10].

On the other hand, Sennott has proposed in [12, 13] a set of conditions guaranteeing the existence of lim sup optimal stationary policies. Indeed, among the (sufficient) conditions presently available to obtain that existence result, the conditions in [13] are, in certain sense, *the weakest*; see the example in [6]. It was proved in [5] that the conditions in [13] are valid under our Assumption 3.2. Thus, the following seems to be an *interesting problem*: What conditions (if any) should be added to the assumptions in [13] so that results similar to Theorem 5.1 and Corrolary 6.1 hold true? Research in this direction is presently in progress.

(Received May 12, 1989.)

REFERENCES

[1] R. B. Ash: Real Analysis and Probability. Academic Press, New York 1972.

[2] J. S. Baras, D.-J. Ma and A. M. Makowski: K competing queues with geometric service requirements and linear costs: The μc rule is always optimal. Systems Control Lett. *6* (1985), 3, 173–180.

[3] P. Billingsley: Convergence of Probability Measures. Wiley, New York 1968.

[4] V. S. Borkar: On minimum cost per unit of time control of Markov chains. SIAM J. Control Optim. *22* (1984), 6, 965–978.

[5] R. Cavazos-Cadena: Weak conditions for the existence of optimal stationary policies in average Markov decision chains with unbounded costs. Kybernetika *25* (1989), 3, 145–156.

[6] R. Cavazos-Cadena and L. I. Sennott: Comparing recent assumptions for the existence of average optimal stationary policies (submitted for publication).

[7] K. Hinderer: Foundations of Non-Stationary Dynamic Programming with Discrete Time Parameter. Springer-Verlag, Berlin–Heidelberg–New York 1970.

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

[9] J. Munkres: Topology, a First Course. Prentice-Hall, Englewood Cliffs, New Jersey 1975.

[10] P. Nain and K. W. Ross: Optimal priority assignment with hard constraints. IEEE Trans. Automat. Control *31* (1986), 10, 883–888.

[11] S. M. Ross: Applied Probability Models with Optimization Applications. Holden-Day, San Francisco, California 1970.
[12] L. I. Sennot: A new condition for the existence of optimal stationary policies in average cost Markov decision processes. Oper. Res. Lett. 5 (1986), 17—23.
[13] L. I. Sennot: A new condition for the existence of optimum stationary policies in average cost Markov decision processes — unbounded cost case. Proceedings of the 25th IEEE Conference on Decision and Control, Athens, Greece 1986, pp. 1719—1721.
[14] L. C. Thomas: Connectedness conditions for denumerable state Markov decision processes. In: Recent Developments in Markov Decision Processes (R. Hartley, L. C. Thomas and D. J. White, eds.), Academic Press, New York 1980, pp. 181—204.

Dr. Rolando Cavazos-Cadena, Departmento de Estadística y Cálculo, Universidad Autónoma Agraria Antonio Narro, Buenavista 25315, Saltillo Coah, México and Department of Mathematics, Texas Tech University, Lubbock, Texas 79409, U.S.A.