

7. R. Cavazos-Cadena and **E. Fernández-Gaucherand**,  
“Denumerable Controlled Markov Chains with Average Reward Criterion: Sample Path Optimality,”  
**ZOR: Mathematical Methods of Operations Research**  
**41** (1995) 89-108.



# Denumerable Controlled Markov Chains with Average Reward Criterion: Sample Path Optimality<sup>1</sup>

ROLANDO CAVAZOS-CADENA<sup>2</sup>

Departamento de Estadística y Cálculo, Universidad Autónoma Agraria Antonio Narro, Buenavista 25315, Saltillo, COAH. MEXICO

EMMANUEL FERNÁNDEZ-GAUCHERAND<sup>3</sup>

Systems & Industrial Engineering Department, The University of Arizona, Tucson, AZ 85721, USA

*Abstract:* We consider discrete-time nonlinear controlled stochastic systems, modeled by controlled Markov chains with denumerable state space and compact action space. The corresponding stochastic control problem of maximizing average rewards in the long-run is studied. Departing from the most common position which uses *expected* values of rewards, we focus on a sample path analysis of the stream of states/rewards. Under a Lyapunov function condition, we show that stationary policies obtained from the average reward optimality equation are not only average reward optimal, but indeed sample path average reward optimal, for almost all sample paths.

*Key Words:* Denumerable Controlled Markov Chains, Average Reward Criterion, Sample Path Optimality, Lyapunov Function Condition.

## 1 Introduction

There are numerous applications, in many different fields, of denumerable controlled Markov chain (CMC) models with an infinite planning horizon; see Bertsekas (1987), Ephremides and Verdú (1989), Ross (1983), Stidham and Weber (1993), Tijms (1986).

We consider the stochastic control problem of maximizing average rewards in the long-run, for denumerable CMC. Departing from the most common position which uses *expected* values of rewards, we focus on a sample path analysis of the stream of states/actions. Under a Lyapunov function condition, we show that

---

<sup>1</sup> Research supported by a U.S.-México Collaborative Research Program funded by the National Science Foundation under grant NSF-INT 9201430, and by CONACyT-MEXICO.

<sup>2</sup> Partially supported by the MAXTOR Foundation for applied Probability and Statistics, under grant No. 01-01-56/04-93.

<sup>3</sup> Research partially supported by the Engineering Foundation under grant RI-A-93-10, and by a grant from the AT&T Foundation.

stationary policies obtained from the average reward optimality equation are not only expected average reward optimal, but indeed sample path average reward optimal. For a summary of similar results, but under a different set of conditions as those used here, see Arapostathis et al. (1993), Section 5.3.

The paper is organized as follows. In section 2 we present the model. Section 3 defines the standard stochastic control problem, under an expected average reward criterion. Section 4 introduces the sample path optimality average reward criterion, and the statement of our main result. After some technical preliminaries in section 5, our main result is proved in section 6.

## 2 The Model

We study discrete-time controlled stochastic dynamical systems, modeled by CMC described by the triplet  $\langle S, A, P \rangle$ , where the state space  $S$  is a denumerable set, endowed with the discrete topology;  $A$  denotes the control or action set, a nonempty compact subset of a metric space. Let  $K := S \times A$  denote the space of state-action pairs, endowed with the product topology. The evolution of the system is governed by a collection of stochastic matrices  $\{P(a) = [P_{x,y}(a)]\}_{a \in A}$ , i.e.,  $P(a)$  is a state transition matrix, with elements  $P_{x,y}(a)$ .

In addition, to assess the performance of the system, a measurable<sup>4</sup> (and possibly unbounded) one-stage reward function  $r: K \rightarrow \mathbb{R}$  is chosen. Thus, at time  $t \in \mathbb{N}_0 := \{0, 1, 2, \dots\}$ , the system is observed to be in some state, say  $X_t = x \in S$ , and a decision  $A_t = a \in A$  is taken. Then a reward  $r(x, a)$  is obtained, and by the next decision epoch  $t + 1$ , the state of the system will have evolved to  $X_{t+1} = y$  with probability  $P_{x,y}(a)$ . Given a Borel space  $B$ , let  $\mathcal{K}(B)$  denote the set of all real-valued and continuous functions on  $B$ . The following continuity assumptions are standard.

*Assumption 2.1:* For each  $x, y \in S$ ,  $P_{x,y}(\cdot) \in \mathcal{K}(A)$ ; furthermore  $r(\cdot, \cdot) \in \mathcal{K}(K)$ .  $\square$

*Remark 2.1:* We are assuming that all actions in  $A$  are available to the decision-maker, when the system is at any given state  $x \in S$ ; this is done with no loss in generality; see Arapostathis et al. (1993), Section 5.3, and Borkar (1991).

The available information for decision-making at time  $t \in \mathbb{N}_0$  is given by the history of the process up to that time  $H_t := (X_0, A_0, X_1, A_1, \dots, A_{t-1}, X_t)$ , which is a random variable taking values in  $H_t$ , where

<sup>4</sup> Given a topological space  $W$ , its Borel  $\sigma$ -algebra will be denoted by  $\mathcal{B}(W)$ ; measurability will always be understood as Borel measurability henceforth.

$$H_0 := S, \quad H_t := H_{t-1} \times (A \times S), \quad H_\infty := (S \times A)^\infty,$$

are the history spaces, endowed with their product topologies.

An *admissible control policy* is a (possibly randomized) rule for choosing actions, which may depend on the entire history of the process up to the present time  $(H_t)$ ; see Arapostathis et al. (1993) and Hernández-Lerma (1989). Thus, a policy is specified by a sequence  $\pi = \{\pi_t\}_{t \in \mathbb{N}_0}$  of stochastic kernels  $\pi_t$  on  $A$  given  $H_t$ , that is: a) for each  $h_t \in H_t$ ,  $\pi_t(\cdot | h_t)$  is a probability measure on  $\mathcal{A}(A)$ ; and b) for each  $B \in \mathcal{B}(A)$ , the map  $h_t \mapsto \pi_t(B | h_t)$  is measurable. The set of all admissible policies will be denoted by  $\Pi$ . In our subsequent exposition, two classes of policies will be of particular interest: the stationary deterministic and the stationary randomized policies: A policy  $\pi \in \Pi$  is said to be stationary deterministic if there exists a decision function  $f: S \rightarrow A$  such that  $A_t = f(x)$  is the action prescribed by  $\pi$  at time  $t$ , if  $X_t = x$ . The set of all stationary deterministic policies is denoted as  $\Pi_{SD}$ . On the other hand, a policy  $\pi \in \Pi$  is said to be a stationary randomized policy if there exists a stochastic kernel  $\gamma$  on  $A$  given  $S$ , such that for each  $B \in \mathcal{B}(A)$ ,  $\gamma(B | X_t)$  is the probability of the event  $[A_t \in B]$ , given  $H_t = (H_{t-1}, A_{t-1}, X_t)$ . The class of all stationary randomized policies is denoted by  $\Pi_{SR}$ ;  $\pi \in \Pi_{SD}$  or  $\pi' \in \Pi_{SR}$  will be equivalently identified by the appropriate decision function  $f$  or stochastic kernel  $\gamma$ , respectively.

Given the initial state  $X_0 = x$ , and a policy  $\pi \in \Pi$ , the corresponding state, action and history processes,  $\{X_t\}$ ,  $\{A_t\}$  and  $\{H_t\}$  respectively, are random processes defined on the canonical probability space  $(H_\infty, \mathcal{B}(H_\infty), \mathcal{P}_x^x)$  via the projections  $X_t(h_\infty) := x_t$ ,  $A_t(h_\infty) := a_t$ , and  $H_t(h_\infty) := h_t$ , for each  $h_\infty = (x, a_0, \dots, x_t, a_t, \dots) \in H_\infty$ , where  $\mathcal{P}_x^x$  is uniquely determined; see Arapostathis et al. (1993), Bertsekas/Shreve (1978), Hinderer (1970), Hernández-Lerma (1989). The corresponding expectation operator is denoted by  $E_x^\pi$ . The following notation will also be used in the sequel: given Borel spaces  $B$  and  $D$  (see Arapostathis et al. (1993)), then a)  $\mathcal{P}(B)$  denotes the set of all probability measures on  $B$ ; b)  $\mathcal{P}(B|D)$  denotes the set of all stochastic kernels on  $B$  given  $D$ .

## 3 The Stochastic Control Problem

Our interest is in measuring the performance of the system in the long run, i.e., after a steady state regime has been reached. A commonly used criterion for this purpose is the *expected long-run average reward*, where the stochastic nature of the stream of rewards is itself averaged by the use of expected values; see Arapostathis et al. (1993). Thus, we have the following definition.

*Expected Average Reward (EAR):* The long-run expected average reward obtained by using  $\pi \in \Pi$ , when the initial state of the system is  $x \in S$ , is given by

$$J(x, \pi) := \liminf_{N \rightarrow \infty} \frac{1}{N+1} \mathbb{E}_x^* \left[ \sum_{t=0}^N r(X_t, A_t) \right], \quad (3.1)$$

and the optimal expected average reward is defined as

$$J^*(x) := \sup_{\pi \in \Pi} \{J(x, \pi)\}. \quad (3.2)$$

A policy  $\pi^* \in \Pi$  is said to be EAR optimal if  $J(x, \pi^*) = J^*(x)$ , for all  $x \in S$ .

*Remark 3.1:* In (3.1) above, the limit superior could also have been used, instead of the limit inferior. Although both alternatives are equivalent under some additional conditions, see Theorem 7.2 in Cavazos-Cadena (1991), in general the behavior of the inferior and superior limits is distinct, see pp. 181–183 in Dynkin and Yushkevich (1979).

*Remark 3.2:* In choosing the limit inferior, we are embracing a conservative philosophy, in that we choose to maximize the *worst expectations* for the gains. That is, the limit inferior has the economic interpretation of *guaranteed rewards*, even if the (long) horizon is not known in advance, or if the termination point is not determined by the decision-maker. On the other hand, the limit superior assumes a bolder, more optimistic position, in that what is being maximized is the *best expectation* of the gains. This criterion is rarely considered, since it makes sense only if the decision-maker has control on when to stop the process.

Our analysis will be carried out under the following assumption, which among other things guarantees that the expected value in (3.1) above is well defined, and that EAR optimal stationary policies exist.

*Assumption 3.1: Lyapunov Function Condition (LFC).* There exists a function  $\ell: S \rightarrow [0, \infty)$ , and a fixed state  $z^*$  such that:

(i) For each  $(x, a) \in K$ ,

$$1 + |r(x, a)| + \sum_{y \neq z^*} p_{x,y}(a) \ell(y) \leq \ell(x);$$

(ii) For each  $x \in S$ , the mapping  $f \mapsto \mathbb{E}_x^* \{\ell(X_1)\} = \sum_{y \in S} p_{x,y}(f(x)) \ell(y)$ , is continuous in  $f \in \Pi_{SD}$ ;

(iii) For each  $f \in \Pi_{SD}$  and  $x \in S$ ,

$$\mathbb{E}_x^* \{\ell(X_n) \mathbb{1}[T > n]\} \xrightarrow{n \rightarrow \infty} 0,$$

where  $T := \min\{m > 0 | X_m = z^*\}$  is the first passage time to state  $z^*$ , and  $\mathbb{1}[A]$  denotes the indicator function for the event  $A$ .  $\square$

The LFC was introduced by Foster (1953) for noncontrolled Markov Chains, and by Hordijk (1974) for CMC, and has been extensively used in the study of denumerable CMC with an EAR criterion; see Arapostathis et al. (1993), Section 5.2. Furthermore, Cavazos-Cadena and Hernández-Lerma (1992) have shown its equivalence, under additional conditions, to several other stability/ergodicity conditions on the transition law of the system. There are two main results derived under the LFC: a) EAR optimal stationary policies are shown to exist, and such policies can be obtained as minimizers in the EAR optimality equation (EAROE); and b) an ergodic structure is induced in the stream of states/rewards. We summarize these well known results in the two lemmas below.

*Lemma 3.1:* Under Assumptions 2.1 and 3.1, there exist  $\rho^* \in \mathbb{R}$  and  $h: S \rightarrow \mathbb{R}$  such that the following holds:

- (i)  $J^*(x) = \rho^*$ ,  $\forall x \in S$ ;
- (ii)  $h(\cdot) \leq (1 + \ell(z^*)) \cdot \ell(\cdot)$ ;
- (iii) The pair  $(\rho^*, h(\cdot))$  is a (possibly unbounded) solution to the EAROE, i.e.,

$$\rho^* + h(x) = \sup_{a \in A} \left[ r(x, a) + \sum_{y \in S} p_{x,y}(a) h(y) \right], \quad \forall x \in S; \quad (3.3)$$

- (iv) For each  $x \in S$ , the term within brackets in (3.3) is a continuous function of  $a \in A$ , and thus it has a maximizer  $f^*(x) \in A$ . Moreover, the policy  $f^* \in \Pi_{SD}$  thus prescribed is EAR optimal.

*Proof:* For a proof of results (i), (iii), and (iv) see Arapostathis et al. (1993), Cavazos-Cadena and Hernández-Lerma (1992), Hordijk (1974). On the other hand, the result in (ii) is essentially contained in the proof of Theorem 5.1 in Hordijk (1974); see especially Lemmas 5.3–5.7. However, we have not found an explicit statement of this inequality, and thus we provide one next. As shown in the above references,  $h(\cdot)$  can be defined as

$$h(x) := \mathbb{E}_x^* \left[ \sum_{t=0}^{T-1} (r(X_t, A_t) - \rho^*) \right], \quad x \in S,$$

where  $T$  is as in Assumption 3.1. Observe that by iterating the equation in Assumption 3.1 (i) we get, for each  $x \in S$ ,

$$\mathbb{E}_x^{\pi^*} \left[ \sum_{t=0}^{T-1} (|r(X_t, A_t)| + 1) \right] \leq \ell(x), \quad x \in \mathbf{S}. \quad (3.4)$$

Therefore,

$$\begin{aligned} |h(x)| &\leq \mathbb{E}_x^{\pi^*} \left[ \sum_{t=0}^{T-1} (|r(X_t, A_t)| + |\rho^*|) \right] \\ &\leq (1 + |\rho^*|) \mathbb{E}_x^{\pi^*} \left[ \sum_{t=0}^{T-1} (|r(X_t, A_t)| + 1) \right] \\ &\leq (1 + |\rho^*|) \ell(x), \end{aligned} \quad (3.5)$$

where the third inequality follows from (3.4). Now observe that

$$\rho^* = \frac{\mathbb{E}_x^{\pi^*} \left[ \sum_{t=0}^{T-1} r(X_t, A_t) \right]}{\mathbb{E}_x^{\pi^*} [T]},$$

which follows from the theory of renewal reward processes (see Ross (1970)): under the action of  $f^*$ , successive visits to the state  $z^*$  determine a renewal process; see also Hordijk (1974), pp. 41–42. Then using (3.4) with  $x = z^*$ , it follows that

$$|\rho^*| \leq \frac{\ell(z^*)}{\mathbb{E}_x^{\pi^*} [T]},$$

since  $T \geq 1$ . Combining this inequality with (3.5), we obtain the desired result.  $\square$

**Lemma 3.2:** Let Assumption 3.1 hold.

(i) Let  $x \in \mathbf{S}$  and  $\pi \in \mathbf{\Pi}$  be arbitrary. Then:

$$\frac{1}{n+1} \mathbb{E}_x^{\pi} [\ell(X_n)] \xrightarrow{n \rightarrow \infty} 0.$$

(ii) Let  $x \in \mathbf{S}$  and  $\pi \in \mathbf{\Pi}$  be arbitrary. Then, for  $T$  as in Assumption 3.1,

$$1 \leq \mathbb{E}_x^{\pi} [T] \leq \ell(x).$$

In particular, for every stationary policy  $f \in \mathbf{\Pi}_{SD}$ , the Markov chain induced by  $f$  has a unique invariant distribution  $q_f \in \mathcal{P}(\mathbf{S})$ , such that:

$$q_f(z^*) = \frac{1}{\mathbb{E}_x^f [T]} \geq \frac{1}{\ell(z^*)} > 0;$$

moreover, the mapping  $f \mapsto q_f(z^*)$  is continuous on  $f \in \mathbf{\Pi}_{SD}$ .

(iii) For each  $f \in \mathbf{\Pi}_{SD}$ , the following holds:

$$\frac{1}{n+1} \mathbb{E}_x^f \left[ \sum_{t=0}^n r(X_t, f(X_t)) \right] \xrightarrow{n \rightarrow \infty} \sum_{y \in \mathbf{S}} q_f(y) r(y, f(y)),$$

and

$$\frac{1}{n+1} \mathbb{E}_x^f \left[ \sum_{t=0}^n r(X_t, f(X_t)) \right] \xrightarrow{n \rightarrow \infty} \sum_{y \in \mathbf{S}} q_f(y) r(y, f(y)), \quad \mathcal{P}_x^f - a.s.$$

(iv) Let  $\gamma \in \mathbf{\Pi}_{SR}$ . The Markov chain induced by  $\gamma$  has a unique invariant distribution  $q_\gamma \in \mathcal{P}(\mathbf{S})$ , and:

$$\frac{1}{n+1} \mathbb{E}_x^\gamma \left[ \sum_{t=0}^n r(X_t, A_t) \right] \xrightarrow{n \rightarrow \infty} \sum_{y \in \mathbf{S}} q_\gamma(y) r^\gamma(y),$$

and

$$\frac{1}{n+1} \sum_{t=0}^n r(X_t, A_t) \xrightarrow{n \rightarrow \infty} \sum_{y \in \mathbf{S}} q_\gamma(y) r^\gamma(y), \quad \mathcal{P}_x^\gamma - a.s.,$$

where

$$r^\gamma(y) := \int_{\mathbf{A}} r(y, a) \gamma(da|x).$$

(v) Let  $W \in \mathcal{C}(\mathbf{K})$  be such that  $|W| \leq |r| + L$ , for some positive constant  $L$ . Then  $(L+1) \cdot \ell(\cdot)$  is a Lyapunov function corresponding to  $W$ , i.e., it satisfies Assumption 3.1, with  $W(\cdot, \cdot)$  taken as the one-stage reward function. Furthermore, there exist  $\rho_\#^\pi \in \mathbb{R}$  and  $h_{\rho_\#^\pi}: \mathbf{S} \rightarrow \mathbb{R}$  such that:

$$\rho_{\#}^* + h_W(x) = \sup_{a \in A} \left[ W(x, a) + \sum_{y \in S} p_{x,y}(a) h_W(y) \right], \forall x \in S,$$

that is,  $(\rho_{\#}^*, h_W(\cdot))$  is a solution to the EAROE, for the CMC with reward function  $W(\cdot, \cdot)$ .

*Remark 3.3:* The result in Lemma 3.2 (i) is due to Hordijk (1974); see also Cavazos-Cadena (1992). For the results in Lemma 3.2 (ii), see Lemma 5.3 and the equivalence between conditions  $L_1$  and  $L_2$  in Cavazos-Cadena and Hernández-Lerma (1992); see also Theorem 5.8 in Arapostathis et al. (1993). On the other hand, the convergence results in (iii) and (iv) can be easily derived by using the theory of (delayed) renewal reward processes; see Theorem 3.16 and the remarks on pp. 53–54 in Ross (1970). Finally, by applying Lemma 3.1 to the reward function  $W(\cdot, \cdot)$ , the result in Lemma 3.2 (v) follows immediately.

#### 4 Sample Path Optimality

The EAR criterion of (3.1)–(3.2) is commonly used as an approximation of undiscounted optimization problems when the planning horizon is very long. However, this criterion can be grossly underselective, in that the finite horizon behavior of the stream of costs is completely neglected. Moreover, it can be the case that EAR optimal policies not only fail to induce a desirable (long) finite horizon performance, but that the performance actually degrades as the horizon increases; see examples of this pathology in Flynn (1980). Thus, *stronger* EAR criteria have been considered, see Arapostathis et al. (1993), Cavazos-Cadena and Fernández-Gaucherand (1993), Dynkin and Yushkevich (1979), Flynn (1980) and Ghosh and Marcus (1992). Also, *weighted* criteria, which introduce sensitivity to both finite and asymptotic behaviour, have been recently introduced, see Fernández-Gaucherand et al. (1994), and references therein.

If there exist a *bounded* solution  $(\rho, h(\cdot))$  to the optimality equation, i.e., with  $h(\cdot)$  a bounded function, then EAR stationary optimal policies derived as maximizers in the optimality equation have been shown to be also (strong) average optimal and *sample path* optimal, i.e., the long-run average of rewards along almost all sample paths is optimal; see Arapostathis et al. (1993), Dynkin and Yushkevich (1979), Georgin (1978), and Yushkevich (1973). Undoubtedly, sample path average reward (SPAR) optimality is a much more desirable property than just EAR optimality, since a policy has to actually be used by the decision-maker along nature's selected sample path. However, bounded solutions to the optimality equation necessarily impose very restrictive conditions on the ergodic

structure of the controlled chain; see Arapostathis et al. (1993), Cavazos-Cadena (1991), and Fernández-Gaucherand et al. (1990). Under the conditions used in this paper, the solutions to the optimality equation obtained in Lemma 3.1 are possibly unbounded. For similar results, but under a different set of conditions as those used here, see the summary in Arapostathis et al. (1993), Section 5.3. After precisely defining SPAR optimality (see also Arapostathis et al. (1993)) we show in the sequel the SPAR optimality of the EAR optimal stationary policies in Lemma 3.1 (iv).

*Sample Path Average Reward (SPAR):* The long-run sample path average reward obtained by  $\pi \in \Pi$ , when the initial state of the system is  $x \in S$ , is given by

$$J_S(x, \pi) := \liminf_{N \rightarrow \infty} \frac{1}{N+1} \sum_{i=0}^N r(X_i, A_i). \quad (4.1)$$

A policy  $\bar{\pi}^* \in \Pi$  is said to be SPAR optimal if there exists a constant  $\bar{\rho}$  such that for all  $x \in S$  we have that:

$$J_S(x, \bar{\pi}^*) = \bar{\rho}, \mathcal{P}_x^{\bar{\pi}^*} - a.s.,$$

while, for all  $\pi \in \Pi$  and  $x \in S$ ,

$$J_S(x, \pi) \leq \bar{\rho}, \mathcal{P}_x^{\pi} - a.s..$$

The constant  $\bar{\rho}$  is the optimal sample path average reward.

We present next our main result, showing the SPAR optimality of the EAR optimal stationary policy obtained in Lemma 3.1 (iv); its proof is presented in Section 6, after some technical preliminaries given in the next section.

*Theorem 4.1:* Let Assumptions 2.1 and 3.1 hold. Let  $f^*$  and  $\rho^*$  be as in Lemma 3.1. Then:

- (i)  $f^*$  is SPAR optimal, and  $\rho^*$  is the optimal sample path average reward.
- (ii) For all  $\gamma \in \Pi_{SR}$ , we have that

$$\limsup_{N \rightarrow \infty} \frac{1}{N+1} \sum_{i=0}^N r(X_i, A_i) \leq \rho^*, \mathcal{P}_x^{\gamma} - a.s..$$

*Remark 4.1:* According to the above result, regardless of the initial state  $x \in \mathbf{S}$  and policy  $\pi \in \Pi$  being used, with probability 1 the limit inferior of the sample average reward over  $N$  periods does not exceed  $\rho^*$ , the expected average reward.

Moreover:

$$\lim_{N \rightarrow \infty} \frac{1}{N+1} \sum_{t=0}^N r(X_t, A_t) = \rho^*, \text{ a.s.},$$

and the limit inferior and the limit superior of sample path average rewards lead to the same optimal policy and optimal value, when the optimization is restricted to only policies in  $\Pi_{sr}$ .

## 5 Preliminaries

This section contains some technical results that will be used in the proof of Theorem 4.1. The next dominated convergence result is similar to that in Proposition 18, p. 232 of Royden (1968), although the latter requires stronger assumptions.

*Lemma 5.1:* Let  $\{v_n\} \subset \mathbf{P}(\mathbf{K})$  be a sequence converging weakly to  $v \in \mathbf{P}(\mathbf{K})$ . Let  $R \in \mathcal{G}(\mathbf{K})$  be a nonnegative function such that:

$$\int_{\mathbf{K}} R d v_{n \rightarrow \infty} \int_{\mathbf{K}} R d v < \infty.$$

Then, for each  $C \in \mathcal{G}(\mathbf{K})$  satisfying  $|C(\cdot)| \leq R(\cdot)$ , it follows that:

$$\int_{\mathbf{K}} C d v_{n \rightarrow \infty} \int_{\mathbf{K}} C d v.$$

*Proof:* Let  $\varepsilon > 0$  be fixed, and pick a finite set  $G \subset \mathbf{S}$  such that

$$\int_{G^c \times \Lambda} R d v < \varepsilon, \quad (5.1)$$

where  $G^c$  stands for the complement of  $G$ . Since  $\{v_n\}$  converges weakly to  $v$ , it follows that

$$\int_{G \times \Lambda} R d v_{n \rightarrow \infty} \int_{G \times \Lambda} R d v.$$

Then, we have that

$$\begin{aligned} \int_{G^c \times \Lambda} R d v_n &= \int_{\mathbf{K}} R d v_n - \int_{G \times \Lambda} R d v_{n \rightarrow \infty} \int_{\mathbf{K}} R d v - \int_{G^c \times \Lambda} R d v \\ &= \int_{G^c \times \Lambda} R d v. \end{aligned}$$

Therefore, there exists  $M \in \mathbb{N}$  such that for all  $n \geq M$

$$\int_{G^c \times \Lambda} R d v_n < \varepsilon, \quad (5.2)$$

and thus

$$\begin{aligned} \left| \int_{\mathbf{K}} C d v_n - \int_{\mathbf{K}} C d v \right| &\leq \left| \int_{G \times \Lambda} C d v_n - \int_{G \times \Lambda} C d v \right| + \int_{G^c \times \Lambda} |C| d v_n \\ &\quad + \int_{G^c \times \Lambda} |C| d v \\ &\leq \left| \int_{G \times \Lambda} C d v_n - \int_{G \times \Lambda} C d v \right| + 2\varepsilon, \end{aligned} \quad (5.3)$$

where the second inequality follows from (5.1) and (5.2), together with the fact that  $|C(\cdot)| \leq R(\cdot)$ . On the other hand, since  $\{v_n\}$  converges weakly to  $v$  and  $C(\cdot)$  is continuous, it follows that

$$\int_{G \times \Lambda} C d v_{n \rightarrow \infty} \int_{G \times \Lambda} C d v, \quad (5.4)$$

since  $G$  is a finite set. Taking the limit superior, as  $n \rightarrow \infty$ , in both sides of (5.3) and using (5.4), we obtain that

$$\limsup_{n \rightarrow \infty} \left| \int_{\mathbf{K}} Cd v_n - \int_{\mathbf{K}} Cd v \right| \leq 2\varepsilon,$$

and the conclusion follows since  $\varepsilon > 0$  was arbitrary.  $\square$

*Definition 5.1:* Let  $\rho_1^*$  be the optimal EAR associated with the reward function  $|r|$ , and set

$$\delta := \sup_{r \in \Pi_{\mathbf{S}_0}} \{1 - q_r(z^*)\}, \quad (5.5)$$

where  $q_r(\cdot)$  is as in Lemma 3.2; notice that  $0 \leq \delta < 1$ , by Lemma 3.2 (ii). Next, define  $L \in \mathbb{R}$  as

$$L := 1 + \frac{\rho_1^*}{1 - \delta}, \quad (5.6)$$

and  $R \in \mathcal{C}(\mathbf{K})$  as

$$R(x, a) := \begin{cases} |r(x, a)| + L, & x \neq z^*, a \in A; \\ |r(x, a)|, & x = z^*, a \in A. \end{cases} \quad (5.7)$$

The function  $R$  defined above will play an important role in the proof of Theorem 4.1. The following result will also be very useful.

*Lemma 5.2:* Let  $\rho_R^*$  be the optimal EAR corresponding to the reward function  $R$ . Then  $\rho_R^* < L$ .

*Proof:* Let  $\bar{f} \in \Pi_{\mathbf{S}D}$  be an EAR optimal policy for the reward function  $R$ . In this case

$$\begin{aligned} \rho_R^* &= \sum_{x \in \mathbf{S}} q_{\bar{f}}(x) R(x, \bar{f}(x)) \\ &= \sum_{x \in \mathbf{S}} q_{\bar{f}}(x) |r(x, \bar{f}(x))| + L \sum_{x \neq z^*} q_{\bar{f}}(x) \\ &\leq \rho_1^* + L(1 - q_{\bar{f}}(z^*)), \end{aligned}$$

where the second equality follows from (5.7), and the inequality from the fact that  $\rho_1^*$  is the optimal EAR corresponding to  $|r(\cdot, \cdot)|$ . Then (5.5) and (5.6) yield that

$$\rho_R^* \leq \rho_1^* + L\delta = \delta + \frac{\rho_1^*}{1 - \delta} < L,$$

since  $\delta < 1$ .  $\square$

*Definition 5.2:* The sequence of state/action empirical measures  $\{v_n\} \subset \mathcal{P}(\mathbf{K})$  is computed as follows: for each pair of Borel sets  $G \in \mathcal{A}(\mathbf{S})$  and  $B \in \mathcal{A}(A)$ , let

$$v_n(G \times B) := \frac{1}{n+1} \sum_{i=0}^n \mathbf{1}[X_i \in G, A_i \in B], \quad n \in \mathbb{N}_0.$$

*Remark 5.1:* Notice that  $\{v_n\}$  is a stochastic process, adapted to the filtration  $\{\sigma(H_n, A_n)\}$ , and that  $v_n \in \mathcal{P}(\mathbf{K})$ . Also, for each  $W: \mathbf{K} \rightarrow \mathbb{R}$ , we have that

$$\frac{1}{n+1} \sum_{i=0}^n W(X_i, A_i) = \int_{\mathbf{K}} W d v_n. \quad (5.8)$$

Next, let  $\bar{\mathbf{S}} := \mathbf{S} \cup \{\infty\}$  be the one-point compactification of  $\mathbf{S}$ , and observe that  $v_n$  can be naturally considered as an element of  $\mathcal{P}(\bar{\mathbf{S}} \times A)$ . Since  $\bar{\mathbf{S}} \times A$  is compact, then  $\{v_n\}$  is a tight sequence in  $\mathcal{P}(\bar{\mathbf{S}} \times A)$ . The following result, summarized from Borkar (1991), Chapter 5, describes the asymptotic behavior of the sequence of empirical measures; see also Arapostathis et al (1993), Section 5.3.

*Lemma 5.3:* Let  $x \in \mathbf{S}$ , and  $\pi \in \Pi$  be arbitrary. Then the following holds for  $\mathcal{P}^x$ -almost all sample paths: If  $\nu \in \mathcal{P}(\bar{\mathbf{S}} \times A)$  is a limit point of  $\{v_n\}$ , then  $\nu$  can be written as

$$\nu = (1 - \alpha)\mu_1 + \alpha\mu_2, \quad (5.9)$$

where  $0 \leq \alpha \leq 1$ , and  $\mu_1, \mu_2 \in \mathcal{P}(\bar{\mathbf{S}} \times A)$  satisfy the following:

- (i)  $\mu_1(\mathbf{S} \times A) = 1 = \mu_2(\{\infty\} \times A)$ ;
- (ii)  $\mu_1$  can be decomposed as

$$\mu_1(\{y\} \times B) = \bar{\mu}(y) \cdot \gamma(B|y), \quad (5.10)$$

for each  $y \in \mathbf{S}$  and  $B \in \mathcal{A}(A)$ , where  $\bar{\mu} \in \mathcal{P}(\mathbf{S})$  and  $\gamma \in \mathcal{P}(A|\mathbf{S})$ ;

(iii) if  $\bar{\mu}$  and  $\gamma$  are as in (5.10), then  $\bar{\mu}$  is the unique invariant distribution of the Markov chain induced by  $\gamma$ , when  $\gamma$  is viewed as a policy in  $\Pi_{SR}$ . Thus, we have that  $\bar{\mu} = q_\gamma$ , using the notation in Lemma 3.2.

Now, let  $R(\cdot, \cdot)$  be the function in Definition 5.1, and recall that  $(L + 1)\gamma(\cdot)$  is a Lyapunov function for  $R(\cdot, \cdot)$ . Let  $\rho_R^\# \in \mathbb{R}$  and  $h_R: \mathbf{S} \rightarrow \mathbb{R}$  be a solution to the EAROE corresponding to the reward function  $R(\cdot, \cdot)$ , i.e.,

$$\rho_R^\# + h_R(x) = \sup_{a \in \mathbf{A}} \left[ R(x, a) + \sum_{y \in \mathbf{S}} p_{x,y}(a) h_R(y) \right], \forall x \in \mathbf{S};$$

recall that such a solution exists, by Lemma 3.1. Furthermore, by Lemma 3.2 we obtain that

$$\frac{1}{n+1} \mathbb{E}_x^\pi [h_R(X_n)] \xrightarrow{n \rightarrow \infty} 0, \forall x \in \mathbf{S}, \pi \in \Pi. \quad (5.11)$$

Next, define Mandl's discrepancy function  $\Phi: \mathbf{K} \rightarrow [0, \infty)$ , by (see Arapostathis et al. (1993), Hernández-Lerma (1989))

$$\Phi(x, a) = \rho_R^\# + h_R(x) - R(x, a) - \sum_{y \in \mathbf{S}} p_{x,y}(a) h_R(y), \forall (x, a) \in \mathbf{K}.$$

Note that  $\Phi(x, a) \geq 0$ , for all  $(x, a) \in \mathbf{K}$  and that  $\Phi(x, a) = 0$  if and only if the action  $a \in \mathbf{A}$  attains the maximum in the EAROE. With the above definitions, standard arguments show that for each  $n \in \mathbb{N}_0$ ,  $x \in \mathbf{S}$ , and  $\pi \in \Pi$ ,

$$\begin{aligned} \rho_R^\# + \frac{h_R(x)}{n+1} &= \frac{1}{n+1} \mathbb{E}_x^\pi \left[ \sum_{i=0}^n R(X_i, A_i) + \Phi(X_n, A_n) \right] + \frac{1}{n+1} \mathbb{E}_x^\pi [h_R(X_{n+1})] \\ &= \mathbb{E}_x^\pi \left[ \int_{\mathbf{K}} (R + \Phi) d\nu_n + \frac{1}{n+1} \mathbb{E}_x^\pi [h_R(X_{n+1})] \right], \end{aligned}$$

where the second equality follows from (5.8). Then, we have that (5.11) yields

$$\mathbb{E}_x^\pi \left[ \int_{\mathbf{K}} (R + \Phi) d\nu_n \right] \xrightarrow{n \rightarrow \infty} \rho_R^\#. \quad (5.12)$$

In particular, for any policy  $\gamma \in \Pi_{SR}$ ,

$$\mathbb{E}_x^\gamma \left[ \int_{\mathbf{K}} (R + \Phi) d\nu_n \right] \xrightarrow{n \rightarrow \infty} \rho_R^\# = \sum_{y \in \mathbf{S}} q_\gamma(y) (R^\gamma(y) + \Phi^\gamma(y)), \quad (5.13)$$

where

$$R^\gamma(y) = \int_{\mathbf{K}} R(x, a) \gamma(da|x), x \in \mathbf{S},$$

with a similar definition for  $\Phi^\gamma$ . The following technical result will also be used in the proof of Theorem 4.1; we present its proof in the Appendix.

*Theorem 5.1:* Let  $x \in \mathbf{S}$  and  $\pi \in \Pi$  be arbitrary. Then for  $\mathcal{P}_x^\pi$ -almost all sample paths  $\{X_i(h_{\infty}), A_i(h_{\infty})\}$ ,  $h_\infty \in \mathbf{H}_\infty$ , there exists a sequence  $\{n_k\} \subset \mathbb{N}$ , with  $n_k \rightarrow \infty$  as  $k \rightarrow \infty$ , such that the following holds:

- (i)  $\{\nu_{n_k}\}$  converges weakly to  $\nu \in \mathcal{P}(\mathbf{K})$ ;
- (ii)

$$\int_{\mathbf{K}} (R + \Phi) d\nu_{n_k} \xrightarrow{n_k \rightarrow \infty} \int_{\mathbf{K}} (R + \Phi) d\nu. \quad (5.14)$$

□

*Remark 5.2:* Note that Theorem 5.1 (i) says that  $\nu \in \mathcal{P}(\mathbf{K}) = \mathcal{P}(\mathbf{S} \times \mathbf{A})$ , a stronger assertion than  $\nu \in \mathcal{P}(\mathbf{S} \times \mathbf{A})$ .

## 6 Proof of the Main Result

Now we are ready to present the proof to Theorem 4.1, our main result.

*Proof of Theorem 4.1:*

(i) By Lemma 3.2 (iii), we have that

$$\frac{1}{n+1} \sum_{i=0}^n r(X_i, f^*(X_i)) \xrightarrow{n \rightarrow \infty} \sum_{y \in \mathbf{S}} q_{f^*}(y) r(y), f^*(y) = \rho^*, \mathcal{P}_x^{f^*} - a.s.,$$

ie., when the system is driven by policy  $f^*$ , the average of the sample path rewards converges to the optimal EAR, almost surely.

Now, let  $x \in \mathbf{S}$  and  $\pi \in \Pi$  be arbitrary but fixed, and let  $U \in \mathbf{H}_\infty$  be the set of sample paths along which the conclusions in Theorem 5.1 and Lemma 5.3 are valid; observe that  $\mathcal{P}_x^x\{U\} = 1$ . Select next a sample path  $\{X(h_\infty), A(h_\infty)\}$ ,  $h_\infty \in U$ , and pick a sequence  $\{n_k\} \in \mathbb{N}$  as in Theorem 5.1; note that  $\{n_k\}$  depends on the sample path selected. Recall that  $\Phi(x, a) \geq 0$ , for any  $(x, a) \in \mathbf{K}$ , and that  $|V| \leq R + \Phi$ , by (5.7). Then by (5.14) and Lemma 5.1, we get that

$$\int_{\mathbf{K}} r d\nu_{n_k} \xrightarrow{n_k \rightarrow \infty} \int_{\mathbf{K}} r d\nu. \quad (6.1)$$

On the other hand, by Lemma 5.3,  $\nu$  can be decomposed in such a way that for all  $y \in \mathbf{S}$  and  $B \in \mathcal{B}(\mathbf{A})$ ,

$$\nu(\{y\} \times B) = q_\nu(y) \cdot \gamma(B|y); \gamma \in \mathbb{P}(\mathbf{A}|\mathbf{S}).$$

Therefore, by Lemma 3.2 (iv),

$$\begin{aligned} \int_{\mathbf{K}} r d\nu &= \sum_{y \in \mathbf{S}} q_\nu(y) \int_{\mathbf{A}} r(y, a) \gamma(da|y) \\ &= \sum_{y \in \mathbf{S}} q_\nu(y) r^\gamma(y) \\ &= \lim_{n \rightarrow \infty} \frac{1}{n+1} \mathbb{E}_x^\pi \left[ \sum_{l=0}^n r(X_l, A_l) \right], \end{aligned}$$

and thus

$$\int_{\mathbf{K}} r d\nu \leq \rho^*, \quad (6.2)$$

since  $\rho^*$  is the optimal expected average reward associated with the reward function  $r(\cdot, \cdot)$ . Since

$$\begin{aligned} \liminf_{n \rightarrow \infty} \frac{1}{n+1} \sum_{l=0}^n r(X_l, A_l) &= \liminf_{n \rightarrow \infty} \int_{\mathbf{K}} r d\nu_n \\ &\leq \liminf_{k \rightarrow \infty} \int_{\mathbf{K}} r d\nu_{n_k}, \end{aligned}$$

then using (6.1) and (6.2) it follows that

$$\liminf_{n \rightarrow \infty} \frac{1}{n+1} \sum_{l=0}^n r(X_l, A_l) \leq \rho^*.$$

Therefore, the conclusion follows since the sample path  $\{X(h_\infty), A(h_\infty)\}$ ,  $h_\infty \in U$ , was arbitrary, and  $\mathcal{P}_x^x\{U\} = 1$ .

(ii) This follows immediately from (i) and Lemma 3.2 (iii)–(iv).  $\square$

### Appendix: Proof of Theorem 5.1

Let  $x \in \mathbf{S}$  and  $\pi \in \Pi$  be arbitrary but fixed. Notice that (5.12) and Fatou's Lemma (see Royden (1968), p. 231) imply that

$$\rho_{\mathbf{K}}^* \geq \mathbb{E}_x^\pi \left[ \liminf_{n \rightarrow \infty} \int_{\mathbf{K}} (R + \Phi) d\nu_n \right]. \quad (A.1)$$

Next, let  $V \subset \mathbf{H}_\infty$  be the set of sample paths along which the conclusions in Lemma 5.3 are valid, and observe that  $\mathcal{P}_x^x\{V\} = 1$ . Now select and fix arbitrarily a sample path  $\{X(h_\infty), A(h_\infty)\}$ ,  $h_\infty \in V$ , and pick a sequence  $\{n_k\} \in \mathbb{N}$  such that (e.g., see Royden (1968), p. 36),

$$\lim_{n_k \rightarrow \infty} \int_{\mathbf{K}} (R + \Phi) d\nu_{n_k} = \liminf_{n \rightarrow \infty} \int_{\mathbf{K}} (R + \Phi) d\nu_n. \quad (A.2)$$

Taking a subsequence if necessary, we can assume that  $\{\nu_{n_k}\}$  converges weakly to  $\nu = (1 - \alpha)\mu_1 + \alpha\mu_2 \in \mathbb{P}(\bar{\mathbf{S}} \times \mathbf{A})$ , where  $0 \leq \alpha \leq 1$  and  $\mu_1, \mu_2$  are as in Lemma 5.3. Now let  $G \subset \mathbf{S}$  be a finite set, chosen arbitrarily except that  $z^* \in G$ . Define  $R_G$  and  $\Phi_G$  in  $\mathcal{B}(\bar{\mathbf{S}} \times \mathbf{A})$  as follows:

$$R_G(x, a) := \begin{cases} R(x, a), & (x, a) \in G \times \mathbf{A}; \\ L, & (x, a) \in (\bar{\mathbf{S}} \setminus G) \times \mathbf{A}, \end{cases} \quad (A.3)$$

and

$$\Phi_G(x, a) := \begin{cases} \Phi(x, a), & (x, a) \in G \times \mathbf{A}; \\ 0, & (x, a) \in (\bar{\mathbf{S}} \setminus G) \times \mathbf{A}. \end{cases} \quad (A.4)$$

Therefore, we have that

$$0 \leq R_G(x, a) + \Phi_G(x, a) + \Phi(x, a) ,$$

as  $G$  increases to  $S$ . Thus, by the weak convergence of  $\{v_{n_k}\}$  to  $v$ , (A.3)–(A.4), and since  $R_G + \Phi_G \in \mathcal{C}(\bar{S} \times A)$  is a bounded function, we obtain that,

$$\begin{aligned} \liminf_{n \rightarrow \infty} \int_{\mathbf{K}} (R + \Phi)d\nu_n &= \lim_{n_k \rightarrow \infty} \int_{\mathbf{K}} (R + \Phi)d\nu_{n_k} \\ &\geq \lim_{n_k \rightarrow \infty} \int_{\bar{S} \times A} (R_G + \Phi_G)d\nu_{n_k} \\ &= \int_{\bar{S} \times A} (R_G + \Phi_G)d\nu \\ &= (1 - \alpha) \int_{\mathbf{K}} (R_{\mathbf{K}} + \Phi_G)d\mu_1 + \alpha \int_{(\infty) \times A} (R_G + \Phi_G)d\mu_2 \\ &= (1 - \alpha) \int_{\mathbf{K}} (R_G + \Phi_G)d\mu_1 + \alpha L . \end{aligned} \quad (\text{A.5})$$

Now, let  $G$  increase to  $S$ . In this case the Monotone Convergence Theorem and (A.5) yield that

$$\lim_{n_k \rightarrow \infty} \int_{\mathbf{K}} (R + \Phi)d\nu_{n_k} \geq (1 - \alpha) \int_{\mathbf{K}} (R + \Phi)d\mu_1 + \alpha L . \quad (\text{A.6})$$

To continue, as in Lemma 5.3 (ii), let  $\mu_1$  be decomposed as,

$$\mu_1(\{y\} \times B) = q_y(y) \cdot \gamma(B|y) \gamma \in \mathcal{P}(A|S) .$$

Then, straightforward calculations yield that

$$\int_{\mathbf{K}} (R + \Phi)d\mu_1 = \sum_{y \in S} q_y(y)(R^\gamma(y) + \Phi^\gamma(y)) = \rho_{\mathbf{K}}^* , \quad (\text{A.7})$$

where the second equality follows from (5.13). Hence, combining (A.6) and (A.7), it follows that

$$\lim_{n_k \rightarrow \infty} \int_{\mathbf{K}} (R + \Phi)d\nu_{n_k} \geq (1 - \alpha)\rho_{\mathbf{K}}^* + \alpha L . \quad (\text{A.8})$$

Since by Lemma 5.2  $\rho_{\mathbf{K}}^* < L$ , (A.8) and (A.2) yield that:

$$\liminf_{n \rightarrow \infty} \int_{\mathbf{K}} (R + \Phi)d\nu_n \geq \rho_{\mathbf{K}}^* ; \quad (\text{A.9})$$

Note that (A.9) has been established for an arbitrary sample path  $\{X_i(h_\infty), A_i(h_\infty)\}$ ,  $h_\infty \in V$ . Since  $\mathcal{P}_{\mathbf{K}}^*\{V\} = 1$ , we have that (A.9) holds  $\mathcal{P}_{\mathbf{K}}^*$ -almost surely; this together with (A.1) yields that

$$\liminf_{n \rightarrow \infty} \int_{\mathbf{K}} (R + \Phi)d\nu_n = \rho_{\mathbf{K}}^* \mathcal{P}_{\mathbf{K}}^* - a.s. . \quad (\text{A.10})$$

Hence, by (A.8) and since  $\rho_{\mathbf{K}}^* < L$ ,  $\alpha$  above must be zero, i.e.,  $\{v_{n_k}\}$  converges weakly to  $\mu_1 \in \mathcal{P}(S \times A)$ .

Let  $W$  be the set of all sample paths for which (A.10) holds; notice that  $\mathcal{P}_{\mathbf{K}}^*[W \cap V] = 1$ . Then, for each sample path  $\{X_i(h_\infty), A_i(h_\infty)\}$ ,  $h_\infty \in W \cap V$ , equality is attained in (A.8). Thus choosing a sequence  $\{v_{n_k}\}$  which converges weakly to  $v \in \mathcal{P}(\bar{S} \times A)$  and satisfying (A.2), the following holds:

- (i)  $\{v_{n_k}\}$  converges weakly to  $v = \mu_1 \in \mathcal{P}(S \times A)$ ; and
- (ii)

$$\begin{aligned} \lim_{n_k \rightarrow \infty} \int_{\mathbf{K}} (R + \Phi)d\nu_{n_k} &= \liminf_{n \rightarrow \infty} \int_{\mathbf{K}} (R + \Phi)d\nu_n = \rho_{\mathbf{K}}^* \\ &= \int_{\mathbf{K}} (R + \Phi)d\mu_1 , \end{aligned}$$

where the second equality follows from (A.7).

In conclusion, it has been shown that for any sample path  $\{X_i(h_\infty), A_i(h_\infty)\}$ ,  $h_\infty \in W \cap V$ , there exists a subsequence  $\{v_{n_k}\}$  satisfying the conclusions in Theorem 5.1. Since  $\mathcal{P}_{\mathbf{K}}^*\{W \cap V\} = 1$ , the proof is complete.  $\square$

*Acknowledgements:* We want to thank the referee for several comments and corrections that helped us improve the paper. We also want to thank Professor A. A. Yushkevich for some elucidating comments in relation to Remark 3.2.

## References

- Arapostathis A, Borkar VS, Fernández-Gaucherand E, Ghosh MK, Marcus SI (1993) Discrete-time controlled Markov processes with an average cost criterion: A survey. *SIAM J Control Optim* 31: 282–344
- Bertsekas DP (1987) *Dynamic programming: Deterministic and stochastic models*. Prentice-Hall Englewood Cliffs
- Borkar VS (1991) *Topics in controlled Markov chains*. Pitman Research Notes in Mathematics Series 240, Longman Scientific & Technical UK
- Cavazos-Cadena R (1991) Recent results on conditions for the existence of average optimal stationary policies. *Annals Operat Res* 28:3–28
- Cavazos-Cadena R (1992) Existence of optimal stationary policies in average reward Markov decision processes with a recurrent state. *Appl Math Optim* 26:171–194
- Cavazos-Cadena R, Hernández-Lerma O (1992) Equivalence of Lyapunov stability criteria in a class of Markov decision processes. *Appl Math Optim* 26:113–137
- Cavazos-Cadena R, Fernández-Gaucherand E (1993) Denumerable controlled Markov chains with strong average optimality criterion: Bounded & unbounded costs. SIE Working paper 93-15, SIE Department The University of Arizona
- Dynkin EB, Yushkevich AA (1979) *Controlled Markov processes*. Springer-Verlag New York
- Ephremides A, Verdú S (1989) Control and optimization methods in communication networks. *IEEE Trans Automat Control* 34:930–942
- Fernández-Gaucherand E, Arapostathis A, Marcus SI (1990) Remarks on the existence of solutions to the average cost optimality equation in Markov decision processes. *Syst Control Lett* 15:425–432
- Fernández-Gaucherand E, Ghosh MK, Marcus SI (1994) Controlled Markov processes on the infinite planning horizon: Weighted and overtaking cost criteria. *ZOR: Methods and Models of Operations Research* 39:131–155
- Flynn J (1980) On optimality criteria for dynamic programs with long finite horizon. *J Math Anal Appl* 76:202–208
- Foster FG (1953) On the stochastic processes associated with certain queueing processes. *Ann Math Stat* 24:355–360
- Georgin J-P (1978) Contrôle des chaînes de Markov sur des espaces arbitraires. *Ann Inst H Poincaré, Sect B*, 14:255–277
- Ghosh MK, Marcus SI (1992) On strong average optimality of Markov decision processes with unbounded costs. *Operat Res Lett* 11:99–104
- Hernández-Lerma O (1989) *Adaptive Markov control processes*. Springer-Verlag New York
- Hinderer K (1970) *Foundations of non-stationary dynamic programming with discrete time parameters*. Lect Notes Operat Res Math Syst 33, Springer-Verlag Berlin
- Hordijk A (1974) *Dynamic programming and Markov potential theory*. Math Centre Tracts 51, Mathematisch Centrum Amsterdam
- Ross SM (1970) *Applied probability models with optimization applications*. Holden-Day San Francisco
- Ross SM (1983) *Introduction to stochastic dynamic programming*. Academic Press New York
- Royden HL (1968) *Real analysis*, 2nd ed. Macmillan New York
- Stidham S, Weber R (1993) A survey of Markov decision models for control of networks of queues. *Queueing Syst* 13:291–314
- Tijms HC (1986) *Stochastic modelling and analysis: A computational approach*. John Wiley Chichester
- Yushkevich AA (1973) On a class of strategies in general Markov decision models. *Theory Prob Applications* 18:777–779

Received: November 1993

Revised version received: July 1994