

DENUMERABLE CONTROLLED MARKOV CHAINS WITH AVERAGE REWARD CRITERION: SAMPLE PATH OPTIMALITY

Rolando Cavazos-Cadena
Departamento de Estadística y Cálculo
UAAAN
Buenvista 25315
Saltillo, COAH
MEXICO

Emmanuel Fernández-Gaucherand
Systems & Industrial Eng. Dept.
The University of Arizona
Tucson, AZ 85721
USA
Email: emmanuel@sie.arizona.edu.

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 and actions. Under a Lyapunov function condition, we show that 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. Technical results used in the proof of our main result are given in section 5. A complete version of this paper will appear in **ZOR: Methods and Models in Operations Research**, Vol. 41, issue 2, 1995.

2. The Model

We study discrete-time controlled stochastic dynamical systems, modeled by CMC described by the triplet (S, A, P) , 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† (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 $\mathcal{C}(B)$ denote the set of all real-valued and continuous functions on B . The following continuity assumptions are standard.

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

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

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 π_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 $B(A)$; and b) for each $B \in B(A)$, the map $h_t \mapsto \pi_t(B | h_t)$ is measurable.

† Given a topological space W , its Borel σ -algebra will be denoted by $B(W)$; measurability will always be understood as Borel measurability henceforth.

The set of all admissible policies will be denoted by Π . 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 $f : \mathbf{S} \rightarrow \mathbf{A}$ such that $A_t = f(x)$ is the action prescribed by π at time t , if $X_t = x$. The set of all stationary deterministic policies is denoted as Π_{SD} . On the other hand, a policy $\pi \in \Pi$ is said to be a stationary randomized policy if there exists a stochastic kernel γ on \mathbf{A} given \mathbf{S} , such that for each $B \in \mathcal{B}(\mathbf{A})$, $\gamma(B | X_t)$ 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 Π_{SR} ; $\pi \in \Pi_{SD}$ or $\pi \in \Pi_{SR}$ will be equivalently identified by the appropriate decision function f or stochastic kernel γ , respectively.

Given the 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 $(\mathbf{H}_\infty, \mathcal{B}(\mathbf{H}_\infty), \mathcal{P}_x^\pi)$ via the projections $X_t(h_\infty) := x_t$, $A_t(h_\infty) := a_t$ and $H_t(h_\infty) := h_t$, for each $h_\infty = (x, a_0, \dots, x_t, a_t, \dots) \in \mathbf{H}_\infty$, where \mathcal{P}_x^π is uniquely determined; see Arapostathis et al. (1993), Bertsekas/Shreve (1978), Hinderer (1970), Hernández Lerma (1989). The corresponding expectation operator is denoted by \mathbf{E}_x^π . The following notation will also be used in the sequel: given Borel spaces B and D (see Arapostathis et al. (1993)), then a) $\mathbb{P}(B)$ denotes the set of all probability measures on B ; b) $\mathbb{P}(B | D)$ 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 $x \in \mathbf{S}$, is given by

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

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 $x \in \mathbf{S}$.

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 $\ell : \mathbf{S} \rightarrow [0, \infty)$, and a fixed state z^* such that:

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

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

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

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

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

where $T := \min\{m > 0 | X_m = z^*\}$ is the first passage time to state z^* , and $\mathbf{1}[A]$ 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 2.1-3.1, there exist $\rho^* \in \mathbf{R}$ and $h : \mathbf{S} \rightarrow \mathbf{R}$ such that the following holds:

(i) $J^*(x) = \rho^*$, $\forall x \in \mathbf{S}$;

(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.,

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

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

Lemma 3.2: Let Assumption 3.1 hold.

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

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

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

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

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

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

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

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

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

and, $\mathcal{P}_x^f - a.s.$,

$$\frac{1}{n+1} \sum_{t=0}^n r(X_t, f(X_t)) \xrightarrow{n \rightarrow \infty} \sum_{y \in \mathbf{S}} q_f(y) r(y, f(y)).$$

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

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

and

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

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 $\rho_W^* \in \mathbf{R}$ and $h_W : \mathbf{S} \rightarrow \mathbf{R}$ such that:

$$\rho_W^* + h_W(x) = \sup_{a \in \mathbf{A}} [W(x, a) + \sum_{y \in \mathbf{S}} p_{x,y}(a) h_W(y)],$$

$\forall x \in \mathbf{S}$, that is, $(\rho_W^*, h_W(\cdot))$ is a solution to the EAROE, for the CMC with reward function $W(\cdot, \cdot)$.

Remark 3.1: 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_5 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 $x \in \mathbf{S}$, is given by

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

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

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

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

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

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); some technical preliminaries are given in the next section.

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

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

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

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 ρ^* , the expected average reward. Moreover:

$$\lim_{N \rightarrow \infty} \frac{1}{N+1} \sum_{t=0}^N r(X_t, A_t) = \rho^*, \quad \mathcal{P}_x^{f^*} - a.s.,$$

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 Π_{SR} .

5. Preliminaries

This section contains some technical results that are used to prove 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 $\{\nu_n\} \subset \mathbb{P}(\mathbf{K})$ be a sequence converging weakly to $\nu \in \mathbb{P}(\mathbf{K})$. Let $R \in \mathcal{C}(\mathbf{K})$ be a nonnegative function such that:

$$\int_{\mathbf{K}} R d\nu_n \xrightarrow{n \rightarrow \infty} \int_{\mathbf{K}} R d\nu < \infty.$$

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

$$\int_{\mathbf{K}} C d\nu_n \xrightarrow{n \rightarrow \infty} \int_{\mathbf{K}} C d\nu.$$

Definition 5.1: Let ρ_1^* be the optimal EAR associated with the reward function $|r|$, and set

$$\delta := \sup_{f \in \Pi_{SD}} \{1 - q_f(z^*)\}, \quad (5.1)$$

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

$$L := 1 + \frac{\rho_1^*}{1 - \delta}, \quad (5.2)$$

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

$$R(x, a) := \begin{cases} |r(x, a)| + L, & x \neq z^*, a \in \mathbf{A}; \\ |r(x, a)|, & x = z^*, a \in \mathbf{A}. \end{cases} \quad (5.3)$$

The function R defined above plays an important role in the proof of Theorem 4.1. The following result is also very useful.

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

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

$$\nu_n(G \times B) := \frac{1}{n+1} \sum_{t=0}^n \mathbf{1}[X_t \in G, A_t \in B], \quad n \in \mathbb{N}_0.$$

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

$$\frac{1}{n+1} \sum_{t=0}^n W(X_t, A_t) = \int_{\mathbf{K}} W d\nu_n. \quad (5.4)$$

Next, let $\bar{\mathbf{S}} := \mathbf{S} \cup \{\infty\}$ be the one-point compactification of \mathbf{S} , and observe that ν_n can be naturally considered as an element of $\mathbb{P}(\bar{\mathbf{S}} \times \mathbf{A})$. Since $\bar{\mathbf{S}} \times \mathbf{A}$ is compact, then $\{\nu_n\}$ is a tight sequence in $\mathbb{P}(\bar{\mathbf{S}} \times \mathbf{A})$.

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^π -almost all sample paths: If $\nu \in \mathbb{P}(\bar{\mathbf{S}} \times \mathbf{A})$ is a limit point of $\{\nu_n\}$, then ν can be written as

$$\nu = (1 - \alpha)\mu_1 + \alpha\mu_2, \quad (5.5)$$

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

- (i) $\mu_1(\mathbf{S} \times \mathbf{A}) = 1 = \mu_2(\{\infty\} \times \mathbf{A})$;
- (ii) μ_1 can be decomposed as

$$\mu_1(\{y\} \times B) = \bar{\mu}(y) \cdot \gamma(B | y), \quad (5.6)$$

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

(iii) if $\bar{\mu}$ and γ are as in (5.6), then $\bar{\mu}$ is the unique invariant distribution of the Markov chain induced by γ , when γ is viewed as a policy in Π_{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 $(L + 1)\ell(\cdot)$ 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.,

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

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

$$\frac{1}{n+1} \mathbb{E}_x^\pi [h_R(X_n)] \xrightarrow{n \rightarrow \infty} 0, \quad \forall x \in \mathbf{S}, \pi \in \Pi. \quad (5.7)$$

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 $\pi \in \Pi$,

$$\begin{aligned} \rho_R^* + \frac{h_R(x)}{n+1} &= \frac{1}{n+1} \mathbb{E}_x^\pi \left[\sum_{t=0}^n R(X_t, A_t) + \Phi(X_t, A_t) \right] \\ &+ \frac{1}{n+1} \mathbb{E}_x^\pi [h_R(X_{n+1})] \\ &= \mathbb{E}_x^\pi \left[\int_{\mathbf{K}} (R + \Phi) d\nu_n \right] \\ &+ \frac{1}{n+1} \mathbb{E}_x^\pi [h_R(X_{n+1})], \end{aligned}$$

where the second equality follows from (5.4). Then, we have that (5.7) yields

$$\mathbb{E}_x^\pi \left[\int_{\mathbf{K}} (R + \Phi) d\nu_n \right] \xrightarrow{n \rightarrow \infty} \rho_R^*. \quad (5.8)$$

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

$$\begin{aligned} \mathbb{E}_x^\gamma \left[\int_{\mathbf{K}} (R + \Phi) d\nu_n \right] &\xrightarrow{n \rightarrow \infty} \rho_R^* \\ &= \sum_{y \in \mathbf{S}} q_\gamma(y) (R^\gamma(y) + \Phi^\gamma(y)), \end{aligned} \quad (5.9)$$

where

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

with a similar definition for Φ^γ . The following technical result is also used in the proof of Theorem 4.1.

Theorem 5.1: Let $x \in \mathbf{S}$ and $\pi \in \Pi$ be arbitrary. Then for \mathcal{P}_x^π -almost all sample paths $\{X_t(h_\infty), A_t(h_\infty)\}$, $h_\infty \in \mathbf{H}_\infty$, 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 $\nu \in \mathbb{P}(\mathbf{K})$;

(ii)

$$\int_{\mathbf{K}} (R + \Phi) d\nu_{n_k} \xrightarrow{n_k \rightarrow \infty} \int_{\mathbf{K}} (R + \Phi) d\nu. \quad (5.10)$$

□

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

Acknowledgements

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. The first author was also partially supported by the MAXTOR Foundation for applied Probability and Statistics, under grant No. 01-01-56/04-93. The second author was partially supported by the Engineering Foundation under grant RI-A-93-10, and by a grant from the AT&T Foundation.

References

- A. Arapostathis, V.S. Borkar, E. Fernández Gaucherand, M.K. Ghosh and S.I. Marcus (1993) Discrete-Time controlled Markov processes with an average cost criterion: a survey. *SIAM J. Control Optim.* 31:282-344
- D.P. Bertsekas (1987) *Dynamic programming: deterministic and stochastic models.* Prentice-Hall, Englewood Cliffs.
- V.S. Borkar (1991) *Topics in controlled Markov chains.* Pitman Research Notes in Mathematics Series #240, Longman Scientific & Technical, UK.
- R. Cavazos-Cadena (1991) Recent results on conditions for the existence of average optimal stationary policies. *Annals Operat. Res.* 28:3-28.
- R. Cavazos-Cadena (1992) Existence of optimal stationary policies in average reward Markov decision processes with a recurrent state. *Appl. Math. Optim.* 26:171-194.
- R. Cavazos-Cadena and O. Hernández-Lerma (1992) Equivalence of Lyapunov stability criteria in a class of Markov decision processes. *Appl. Math. Optim.* 26:113-137.
- R. Cavazos-Cadena and E. Fernández Gaucherand (1993) Denumerable controlled Markov chains with strong average optimality criterion: bounded & unbounded costs. SIE Working paper #93-15, SIE Department, The University of Arizona.
- E.B. Dynkin and A.A. Yushkevich (1979) *Controlled Markov processes.* Springer-Verlag, New York.
- A. Ephremides and S. Verdú (1989) Control and optimization methods in communication networks. *IEEE Trans. Automat. Control* 34:930-942.
- E. Fernández-Gaucherand, A. Arapostathis and S.I. Marcus (1990) Remarks on the existence of solutions to the average cost optimality equation in Markov decision processes. *Syst. Control Lett.* 15:425-432.
- E. Fernández-Gaucherand, M.K. Ghosh and S.I. Marcus (1994) Controlled Markov processes on the infinite planning horizon: weighted and overtaking cost criteria. *ZOR: Methods and Models of Operations Research* 39:131-155.
- J. Flynn (1980) On optimality criteria for dynamic programs with long finite horizon. *J. Math. Anal. Appl.* 76:202-208.
- F.G. Foster (1953) On the stochastic processes associated with certain queueing processes. *Ann. Math. Stat.* 24:355-360.
- J.-P. Georgan (1978) Contrôle des chaînes de Markov sur des espaces arbitraires. *Ann. Inst. H. Poincaré, Sect. B*, 14:255-277.
- M.K. Ghosh and S.I. Marcus (1992) On strong average optimality of Markov decision processes with unbounded costs. *Operat. Res. Lett.* 11:99-104.
- O. Hernández-Lerma (1989) *Adaptive Markov control processes.* Springer-Verlag, New York.
- K. Hinderer (1970) Foundations of non-stationary dynamic programming with discrete time parameters. *Lect. Notes Operat. Res. Math. Syst.* #33, Springer-Verlag, Berlin.
- A. Hordijk (1974) *Dynamic programming and Markov potential theory.* Math. Centre Tracts, No. 51, Mathematisch Centrum, Amsterdam.
- S.M. Ross (1970) *Applied probability models with optimization applications.* Holden-Day, San Francisco.
- S.M. Ross (1983) *Introduction to stochastic dynamic programming.* Academic Press, New York.
- H.L. Royden (1968) *Real analysis*, 2nd. ed. Macmillan, New York.
- S. Stidham and R. Weber (1993) A survey of Markov decision models for control of networks of queues. *Queueing Syst.* 13:291-314.
- H.C. Tijms (1986) *Stochastic modelling and analysis: a computational approach.* John Wiley, Chichester.
- A.A. Yushkevich (1973) On a class of strategies in general Markov decision models. *Theory Prob. Applications* 18:777-779.