

10. R. Cavazos-Cadena and **E. Fernández-Gaucherand**, “Denumerable Controlled Markov Chains with Strong Average Optimality Criterion: Bounded & Unbounded Costs,” *ZOR: Mathematical Methods of Operations Research* **43** (1996) 281-300.

Denumerable Controlled Markov Chains with Strong Average Optimality Criterion: Bounded & Unbounded Costs¹

ROLANDO CAVAZOS-CADENA

Departamento de Estadística y Cálculo, Universidad Autónoma Agraria Antonio Narro,
Buenvista 25315, Saltillo, COAH, Mexico

EMMANUEL FERNÁNDEZ-GAUCHERAND²

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

Abstract: This paper studies discrete-time nonlinear controlled stochastic systems, modeled by controlled Markov chains (CMC) with denumerable state space and compact action space, and with an infinite planning horizon. Recently, there has been a renewed interest in CMC with a long-run, expected average cost (AC) optimality criterion. A classical approach to study average optimality consists in formulating the AC case as a limit of the discounted cost (DC) case, as the discount factor increases to 1, i.e., as the discounting effect *vanishes*. This approach has been rekindled in recent years, with the introduction by Sennott and others of conditions under which AC optimal stationary policies are shown to exist. However, AC optimality is a rather underselective criterion, which completely neglects the finite-time evolution of the controlled process. Our main interest in this paper is to study the relation between the notions of AC optimality and *strong* average cost (SAC) optimality. The latter criterion is introduced to assess the performance of a policy over long but finite horizons, as well as in the long-run average sense. We show that for bounded one-stage cost functions, Sennott's conditions are sufficient to guarantee that *every* AC optimal policy is also SAC optimal. On the other hand, a detailed counterexample is given that shows that the latter result does not extend to the case of unbounded cost functions. In this counterexample, Sennott's conditions are verified and a policy is exhibited that is both average and Blackwell optimal and satisfies the average cost inequality.

Key Words: Denumerable Controlled Markov Chains, Strong Average Cost Criterion, Bounded & Unbounded Costs, Sufficient Conditions.

1 Introduction

This paper studies discrete-time nonlinear controlled stochastic systems, modeled by controlled Markov chains (CMC) with denumerable state space and

¹ 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.

² Research partially supported by the Engineering Foundation under grant RI-A-93-10, and by a grant from the AT&T Foundation.

compact action space, and with an infinite planning horizon. These models are a very important class of stochastic decision and control problems, with numerous applications in many diverse disciplines; see Bertsekas (1987), Ross (1983), Puterman (1994). One optimality criterion often used is the long-run expected average cost (AC); see Arapostathis et al. (1993). A classical approach to study AC optimality consists in formulating the AC case as a limit of the discounted cost (DC) case, as the discount factor increases to 1, i.e., as the discounting effect vanishes. This approach has been rekindled in recent years, with the introduction by Sennott and others of conditions under which AC optimal stationary policies are shown to exist; see Arapostathis et al. (1993), Borkar (1991), Cavazos-Cadena (1991a), Cavazos-Cadena and Sennott (1992), Hernández-Lerma and Lasserre (1990), Puterman (1994), Schäl (1993), Sennott (1986), Sennott (1989), Ritt and Sennott (1992). However, the AC criterion is a rather unselective criterion, since it completely neglects the finite-time evolution of the state/cost process; see Arapostathis et al. (1993), Fernández-Gaucheraud et al. (1994). Therefore, it is of much theoretical and practical interest to obtain conditions under which stronger results than AC optimality can be obtained.

Our main interest in this paper is to study the relation between the notions of AC optimality and strong average cost (SAC) optimality. The latter criterion is introduced to assess the performance of a policy over long but finite horizons, as well as in the long-run average sense. We show that for bounded one-stage cost functions, Sennott's conditions are sufficient to guarantee that every AC optimal policy is also SAC optimal. On the other hand, a detailed counterexample is given that shows that the latter result does not extend to the case of unbounded cost functions. In this counterexample, Sennott's conditions are verified and a policy is exhibited that is both average and Blackwell optimal and satisfies the average cost optimality inequality.

The paper is organized as follows. In section 2 the model is presented. Section 3 defines the stochastic control problem, with an expected average cost criterion. Section 4 introduces the strong average optimality criterion. In section 5, the case of bounded cost functions is studied, while unbounded costs are treated in section 6.

2 The Model

This paper studies controlled Markov chains (CMC) described by the four-tuples $\langle S, U, \mathcal{Q}, p \rangle$, where the state space S is a denumerable set, endowed with the discrete topology; the metric space U denotes the control or action set. Furthermore, for each $x \in S$, a nonempty compact set $\mathcal{Q}(x) \subset U$ specifies the admissible actions when the system is in state x . Let $K := \{(x, u) | x \in S, u \in \mathcal{Q}(x)\}$ denote the space of admissible state-action pairs, which is considered as a

topological subspace of $S \times U$. The state transition law is given by $p: (x, y, u) \mapsto P_{x,y}(u)$, an stochastic kernel on S given K ; see Arapostathis et al. (1993), Hernández-Lerma (1989), Puterman (1994).

In addition, to assess the performance of the system, a measurable³ (and possibly unbounded) one-stage cost function $c: K \rightarrow \mathbb{R}$ is chosen by the decision-maker. 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 control/decision $U_t = u \in \mathcal{Q}(x)$ is taken. Then a cost $c(x, u)$ is incurred, and by the next control/decision epoch $t + 1$, the state of the system will have evolved to $X_{t+1} = y$ with probability $P_{x,y}(u)$. The following is an standard assumption.

Assumption 2.1:

- (i) For each $x, y \in S$, the mappings $u \mapsto c(x, u)$ and $u \mapsto P_{x,y}(u)$ are continuous in $u \in \mathcal{Q}(x)$.
- (ii) The cost function $c: K \rightarrow \mathbb{R}$ is bounded below. □

The available information for control at time $t \in \mathbb{N}_0$ is given by the history of the process up to that time $H_t := (X_0, U_0, X_1, U_1, \dots, U_{t-1}, X_t)$, which is a random process taking values in the history spaces H_t , given by

$$H_0 := S, \quad H_t := H_{t-1} \times K,$$

and the canonical sample space is given by $\Omega_\infty := (S \times U)^\infty$; see Arapostathis et al. (1993).

An *admissible control policy* is a (possibly randomized) rule for choosing actions, which may depend on the entire history of the process H_t up to the present time; see Arapostathis et al. (1993) for a more extensive discussion. Thus, a policy is specified by a sequence $\pi = \{\pi_t\}_{t \in \mathbb{N}_0}$ of stochastic kernels π_t on U given H_t , such that for each $h_t \in H_t$, $\pi_t(\cdot | h_t)$ is a probability measure on $\mathcal{Q}(U)$, concentrated on $\mathcal{Q}(x)$. The set of all admissible policies will be denoted by Π . In the sequel, the class of *stationary deterministic* policies will be of particular interest. A policy $\pi \in \Pi$ is said to be stationary deterministic if there exists a control/decision function (or selector) $f: S \rightarrow U$, such that $U_t = f(x) \in \mathcal{Q}(x)$ is the action prescribed by π at time t , if $X_t = x$; we may identify such policy π with the function f . The set of all stationary deterministic policies is denoted as Π_{sd} . Given the initial state $X_0 = x \in S$, and a policy $\pi \in \Pi$, the corresponding state/action and history processes, $\{X_t, U_t\}$ and $\{H_t\}$ respectively, are random

³ Given a topological space W , its Borel σ -algebra will be denoted by $\mathcal{B}(W)$; measurability will be always understood as Borel measurability henceforth.

processes defined on the canonical probability space $(\Omega_\infty, \mathcal{F}(\Omega_\infty), \mathcal{P}_x^x)$ via the projections $X_i(h_\infty) := x_i$, $U_i(h_\infty) := u_i$ and $H_i(h_\infty) := h_i$, for each $h_\infty = (x, u_0, \dots, x_1, u_1, \dots) \in \Omega_\infty$; where \mathcal{P}_x^x is uniquely determined; see Arapostathis et al. (1993), Bertsekas and Shreve (1978), Hernández-Lerma (1989). The corresponding expectation operator is denoted by E_x^x .

3 The Stochastic Control Problem

Our main interest in this paper is to study the relation between the notions of AC optimality and *strong average cost* (SAC) optimality. The latter criterion is introduced to assess the performance of a policy over long but finite horizons, as well as in the long-run average sense. The standard approach to study the AC stochastic control problem as a limit of the discounted cost (DC) problem will be followed; see Arapostathis et al. (1993). The criteria that will be used in subsequent developments are given below.

Discounted Cost (DC): For a discount factor $0 < \alpha < 1$, the DC incurred by $\pi \in \Pi$, when the initial state of the system is $x \in \mathbf{S}$, is given by

$$V_\alpha(x, \pi) := \lim_{n \rightarrow \infty} E_x^x \left[\sum_{i=0}^n \alpha^i c(X_i, U_i) \right],$$

and the optimal α -discounted *value function* is defined as

$$V_\alpha^*(x) := \inf_{\pi \in \Pi} \{V_\alpha(x, \pi)\}. \quad (3.1)$$

A policy $\pi \in \Pi$ is said to be DC optimal, for the discount factor α , if $V_\alpha(x, \pi) = V_\alpha^*(x)$, for all $x \in \mathbf{S}$.

Under Assumption 3.1 below, $V_\alpha^*(\cdot)$ satisfies *Bellman's Optimality Equation* (also called the discounted cost optimality equation (DCOE)), i.e.,

$$V_\alpha^*(x) = \inf_{u \in \mathcal{U}(x)} \left\{ c(x, u) + \alpha \sum_{y \in \mathbf{S}} P_{x,y}(u) V_\alpha^*(y) \right\}, \quad \forall x \in \mathbf{S}; \quad (3.2)$$

see Arapostathis et al. (1993), Bertsekas (1987), Bertsekas and Shreve (1978), Hernández-Lerma (1989), Puterman (1994).

Assumption 3.1: For each $x \in \mathbf{S}$, $V_\alpha^*(x) < \infty$. □

Total Expected Cost Over Finite Horizons (FHC): The total expected cost incurred by the policy $\pi \in \Pi$ over a planning horizon of $n \in \mathbb{N}$ stages, when the initial state of the system is $x \in \mathbf{S}$, is given by

$$V_n(x, \pi) := E_x^x \left[\sum_{i=0}^n c(X_i, U_i) \right], \quad (3.3)$$

and the optimal n -horizon *value function* is defined as

$$V_n^*(x) := \inf_{\pi \in \Pi} \{V_n(x, \pi)\}. \quad (3.4)$$

From Assumptions 2.1 and 3.1 it follows that $V_n^*(x)$ is finite for each $x \in \mathbf{S}$. Moreover, there exists an n -horizon optimal policy $\pi_n^* \in \Pi$, i.e.,

$$V_n(x, \pi_n^*) = V_n^*(x), \quad (3.5)$$

see Arapostathis et al. (1993), Bertsekas (1987), Bertsekas and Shreve (1978), Hernández-Lerma (1989), Puterman (1994).

Average Cost: The long-run expected average cost obtained by $\pi \in \Pi$, when the initial state of the system is $x \in \mathbf{S}$, is given by

$$J(x, \pi) := \limsup_{n \rightarrow \infty} \frac{1}{n} E_x^x \left[\sum_{i=0}^{n-1} c(X_i, U_i) \right], \quad (3.6)$$

and the optimal expected average cost is defined as

$$J^*(x) := \inf_{\pi \in \Pi} \{J(x, \pi)\}. \quad (3.7)$$

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

Remark 3.1: In (3.6) above, the limit inferior could also have been used, instead of the limit superior. Although both alternatives are equivalent under some additional conditions (see Theorem 7.2 in Cavazos-Cadena (1991 a)), in general

the behavior of the inferior and superior limits is distinct; see Dynkin and Yushkevich (1979), pp. 181–183. Moreover, in choosing the limit superior, a conservative philosophy is being embraced, in that what is chosen to be minimize is the *worst expectations* for the costs. In addition, the limit superior has the economic interpretation of *guaranteed expected minimal costs*, even if the (long) horizon is not known, or if the termination point is not determined by the controller or decision-maker.

A classical approach to study average optimality consists in formulating the AC case as a limit of the DC case, as $\alpha \uparrow 1$, i.e., as the discounting effect vanishes; see Arapostathis et al. (1993) and references therein for a comprehensive discussion on the subject. This approach has been rekindled in recent years, with the introduction by Sennott and others of conditions under which AC optimal stationary policies are shown to exist; see Arapostathis et al. (1993), Cavazos-Cadena and Sennott (1992), Hernández-Lerma and Lasserre (1990), Gaterek and Stettner (1990), Puterman (1994), Sennott (1986), Sennott (1989), Schäl (1993). Our analysis will be carried out under the set of assumptions introduced by Sennott (1986), Sennott (1989), which are minimal with respect to several other competing assumptions; see Cavazos-Cadena and Sennott (1992). In particular, the assumptions in Ghosh and Marcus (1992) imply those in Sennott (1986), Sennott (1989). Hence, in addition to Assumptions 2.1 and 3.1, the following assumptions will be used in the sequel.

Assumption 3.2:

(i) There exists $z^* \in S$, $\beta \in (0, 1)$, and $N \in [0, \infty)$ such that for all $\alpha \in (\beta, 1)$,

$$h_\alpha(x) := V_\alpha^*(x) - V_\alpha^*(z^*) \geq -N, \quad \forall x \in S. \quad (3.8)$$

(ii) There exists a function $b: S \rightarrow [0, \infty)$ such that $h_\alpha(\cdot) \leq b(\cdot)$, for all $\alpha \in (\beta, 1)$, and furthermore $\sum_{j \in S} p_{x,j}(u)b(y) < \infty$, for some $(x, u) \in K$. \square

The main results deriving from the above assumption are summarized below; for a proof see Arapostathis et al. (1993), Cavazos-Cadena (1991.a) – Cavazos-Cadena (1991.b), Puterman (1994), Sennott (1986), Sennott (1989).

Lemma 3.1: Let

$$\rho_\alpha := (1 - \alpha)V_\alpha^*(z^*), \quad (3.9)$$

for $\alpha \in (0, 1)$. Then under Assumptions 2.1, 3.1–3.2, the following holds.

(i) There exists $\rho^* \in \mathbb{R}$ such that

$$\rho^* = J^*(x), \quad \forall x \in S.$$

(ii) Moreover,

$$\lim_{\alpha \uparrow 1} \rho_\alpha = \rho^*.$$

(iii) There exists $h: S \rightarrow \mathbb{R}$, with $-N \leq h(\cdot) \leq b(\cdot)$ for all $x \in S$, such that the AC Optimality Inequality (ACOI) holds:

$$\rho^* + h(x) \geq \inf_{u \in \mathcal{U}(x)} \left[c(x, u) + \sum_{j \in S} p_{x,j}(u)h(y) \right]. \quad (3.10)$$

(iv) For each $x \in S$, the term within brackets in (3.10) is a lower semi-continuous function of $u \in \mathcal{U}(x)$, and thus it has a minimizer $f^*(x) \in \mathcal{U}(x)$. Moreover, any policy $f^* \in \Pi_{SD}$ attaining the minimum in the ACOI is AC optimal.

Remark 3.2: Assumptions 3.1–3.2 impose restrictions on the cost and probabilistic structure of the model in an *indirect* way, since these are given in terms of the derived quantity $V_\alpha^*(\cdot)$. Therefore, every attempt at verifying these assumptions must study in detail the cost and probabilistic structure, and perhaps impose further assumptions. In the literature, assumptions 3.1–3.2 have been verified when the cost function has a monotone or quasi-monotone structure which penalizes instability, i.e., large deviations from the special state z^* ; see Arapostathis et al. (1993), Borkar (1991), Cavazos-Cadena and Sennott (1992), Puterman (1994), Sennott (1989), Weber and Stidham (1987). Furthermore, the inequality in (3.10) is in general strict; see Cavazos-Cadena (1991.b).

Remark 3.3: As mentioned before, the limit inferior could be used instead of the limit superior in (3.6). However, under assumptions 2.1(i), 3.1–3.2, and if in addition the cost function $c(\cdot, \cdot)$ is bounded, then both average criteria are equivalent, as shown in Cavazos-Cadena (1991.a). Furthermore, in the latter situation average optimal policies are also asymptotically optimal, in the sense of Dynkin and Yushkevich (1979); the same result is also true when $c(\cdot, \cdot)$ is a quasi-monotone function, and other mild conditions are satisfied; see Ghosh and Marcus (1992).

4 Strong Average Optimality Criterion

Flynn (1980) introduced the following criterion for problems with long finite horizons, which was also used by Ghosh and Marcus (1992).

Definition 4.1: A policy $\pi^* \in \Pi$ is said to be *strong average cost* (SAC) optimal if

$$\limsup_{n \rightarrow \infty} \frac{1}{n+1} [V_n(x, \pi^*) - V_n^*(x)] = 0, \quad \forall x \in \mathbf{S}. \quad (4.1)$$

Notice that since $[V_n(x, \pi^*) - V_n^*(x)] \geq 0$, then (4.1) can be equivalently formulated as:

$$\frac{1}{n+1} [V_n(x, \pi^*) - V_n^*(x)] \xrightarrow{n \rightarrow \infty} 0, \quad \forall x \in \mathbf{S}. \quad (4.2)$$

Thus, a policy $\pi^* \in \Pi$ is SAC optimal if the difference between the average cost for horizon n incurred under π^* and the optimal average cost for horizon n vanishes as $n \rightarrow \infty$. This property thus ensures that π^* is a policy inducing good performance for long but finite horizons, which is indeed a very desirable property to look for in infinite horizon average optimal policies. Notice that every policy that is SAC optimal is also AC optimal:

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{1}{n+1} [V_n^*(x) - V_n(x, \pi^*)] &= 0 \\ \Rightarrow \limsup_{n \rightarrow \infty} \frac{1}{n+1} [V_n(x, \pi) - V_n(x, \pi^*)] &\geq 0 \quad \forall x \in \mathbf{S}, \pi \in \Pi. \end{aligned} \quad (4.3)$$

However the opposite is not necessarily true; see Flynn (1980), Ghosh and Marcus (1992). Therefore, in the sequel we study the following.

Question: Is an arbitrary AC optimal policy also SAC optimal under Assumptions 2.1 and 3.1–3.2?

We will show that the answer to the above question is affirmative, when the one-stage cost function is bounded. However, for unbounded cost functions, a counterexample is provided which shows that the answer is negative, in general.

and in particular for a policy as in Lemma 3.1(iv). A related study to the question above is Ghosh and Marcus (1992), where SAC optimality was shown for every AC optimal policy, under conditions stronger than those used in the sequel.

Remark 4.1: Related optimality criteria are those of the *overtaking cost* (OC) type; see Fernández-Gaucherand et al. (1994), Puterman (1994). A policy π^* may be said to be *average OC* optimal if:

$$\liminf_{n \rightarrow \infty} \frac{1}{n+1} [V_n(x, \pi) - V_n(x, \pi^*)] \geq 0, \quad \forall x \in \mathbf{S}, \pi \in \Pi. \quad (4.4)$$

Clearly, (4.2) implies (4.4), and thus SAC optimality is in general a more selective criterion than average OC optimality.

5 Bounded Costs Case

The following result is similar to results in Ghosh and Marcus (1992), but here it is proved under a different set of assumptions; this result will be fundamental to prove the main result in this section.

Lemma 5.1: Suppose that the one-stage cost function $c(\cdot, \cdot)$ is (uniformly) bounded in \mathbf{K} . Then, under Assumptions 2.1, 3.1–3.2, the following holds:

$$\frac{V_n^*(x)}{n+1} \xrightarrow{n \rightarrow \infty} \rho^*, \quad \forall x \in \mathbf{S}. \quad (5.1)$$

Proof: Let $\beta \in (0, 1)$ be as in Assumption 3.2, and pick $\alpha \in (\beta, 1)$. The DCOE (3.2) can equivalently be written as (see Arapostathis et al. (1993), Hernández-Lerma (1989)),

$$\rho_\alpha + h_\alpha(x) = \inf_{u \in \mathfrak{K}(x)} \left\{ c(x, u) + \alpha \sum_{y \in \mathfrak{S}} P_{x,y}(u) h_\alpha(y) \right\}, \quad \forall x \in \mathbf{S},$$

where $h_\alpha(\cdot)$ and ρ_α are as in (3.8) and (3.9), respectively. Then, for every $(x, u) \in \mathbf{K}$

it follows that

$$\rho_\alpha + h_\alpha(x) \leq c(x, u) + \alpha \sum_{y \in \mathcal{S}} p_{x,y}(u) h_\alpha(y). \quad (5.2)$$

Then, by (3.8),

$$-N \leq h_\alpha(x) \leq \frac{2\|c\|}{1-\alpha} < \infty,$$

where $\|c\| := \max\{c(x, u) \mid (x, u) \in \mathbf{K}\}$, which is finite by the boundedness assumption. Therefore, defining $\tilde{h}_\alpha(\cdot) := h_\alpha(\cdot) + N$, it follows that

$$0 \leq \tilde{h}_\alpha(x) \leq \frac{2\|c\|}{1-\alpha} + N < \infty. \quad (5.3)$$

Moreover, it is not difficult to see that (5.2) is equivalent to the following

$$\rho_\alpha - N(1-\alpha) + \tilde{h}_\alpha(x) \leq c(x, u) + \alpha \sum_{y \in \mathcal{S}} p_{x,y}(u) \tilde{h}_\alpha(y),$$

and since $\tilde{h}(\cdot) \geq 0$, this last inequality implies that

$$\rho_\alpha - N(1-\alpha) + \tilde{h}_\alpha(x) \leq c(x, u) + \sum_{y \in \mathcal{S}} p_{x,y}(u) \tilde{h}_\alpha(y), \quad \forall (x, u) \in \mathbf{K}.$$

Then, using standard arguments (see Arapostathis et al. (1993), Hernández-Lerma (1989), Ross (1983), Sennott (1986), Sennott (1989)), it follows that

$$\rho_\alpha - N(1-\alpha) + \frac{\tilde{h}_\alpha(x)}{n+1} \leq \frac{\mathbb{E}_x^*[\sum_{i=0}^n c(X_i, U_i)]}{n+1} + \frac{\mathbb{E}_x^*[\tilde{h}_\alpha(X_{n+1})]}{n+1},$$

for all $x \in \mathcal{S}$, $\pi \in \Pi$, and $n \in \mathbb{N}$. Using the policy π_n^* in (3.5), (5.3) and the above inequality yield

$$\rho_\alpha - N(1-\alpha) + \frac{\tilde{h}_\alpha(x)}{n+1} \leq \frac{V_n^*(x)}{n+1} + \frac{2\|c\|}{(1-\alpha)(n+1)} + \frac{N}{n+1},$$

and taking the limit inferior in both sides as $n \rightarrow \infty$, it follows that

$$\rho_\alpha - N(1-\alpha) \leq \liminf_{n \rightarrow \infty} \frac{V_n^*(x)}{n+1},$$

and then letting $\alpha \uparrow 1$, Lemma 3.1(ii) yields that

$$\rho^* \leq \liminf_{n \rightarrow \infty} \frac{V_n^*(x)}{n+1}. \quad (5.4)$$

Now let $f^* \in \Pi_{SD}$ be the policy in Lemma 3.1(iv), then

$$\rho^* + h(x) \geq c(x, f^*(x)) + \sum_{y \in \mathcal{S}} p_{x,y}(f^*(x)) h(y), \quad \forall x \in \mathcal{S}.$$

A simple induction argument then gives that

$$\rho^* + \frac{h(x)}{n+1} \geq \frac{\mathbb{E}_x^*[\sum_{i=0}^n c(X_i, U_i)]}{n+1} + \frac{\mathbb{E}_x^*[h(X_{n+1})]}{n+1}.$$

From the above, together with (3.3), (3.4) and (3.8), it follows that

$$\rho^* + \frac{h(x)}{n+1} \geq \frac{V_n^*(x)}{n+1} - \frac{N}{n+1}, \quad \forall x \in \mathcal{S}, n \in \mathbb{N}.$$

Therefore, one obtains that

$$\rho^* \geq \limsup_{n \rightarrow \infty} \frac{V_n^*(x)}{n+1}. \quad (5.5)$$

Thus, the result follows by combining (5.4) and (5.5). \square

The following is the main result of this section, which gives a positive partial answer to the question posed in Section 4.

Theorem 5.1. Suppose that the one-stage cost function $c(\cdot, \cdot)$ is (uniformly) bounded in \mathbf{K} . Then, under Assumptions 2.1, 3.1–3.2, every AC optimal policy is SAC optimal.

Proof: Let $\pi^* \in \Pi$ be any average optimal policy. Hence, for each $x \in \mathbf{S}$,

$$\limsup_{n \rightarrow \infty} \frac{V_n(x, \pi^*)}{n+1} = J^*(x) = \rho^*, \quad (5.6)$$

by Lemma 3.1(i). On the other hand $V_n(x, \pi^*) \geq V_n^*(x)$ and thus

$$\liminf_{n \rightarrow \infty} \frac{V_n(x, \pi^*)}{n+1} \geq \liminf_{n \rightarrow \infty} \frac{V_n^*(x)}{n+1} = \rho^*,$$

by Lemma 5.1. The last inequality and (5.6) combined give that

$$\frac{V_n(x, \pi^*)}{n+1} \xrightarrow{n \rightarrow \infty} \rho^*, \quad \forall x \in \mathbf{S}. \quad (5.7)$$

Finally, from (5.1) and (5.7) it follows that

$$\frac{V_n(x, \pi^*) - V_n^*(x)}{n+1} \xrightarrow{n \rightarrow \infty} (\rho^* - \rho^*) = 0.$$

Thus π^* is strong average optimal. \square

6 A Counterexample for the Unbounded Costs Case

In this section an example is given showing that the result in Theorem 5.1 does not extend to the case of unbounded one-stage cost functions. In particular, a policy as in Lemma 3.1(iv) is exhibited, which is not SAC optimal. This example exploits the fact that Assumptions 3.1–3.2 involve the one-stage cost and the probabilistic structure of the controlled Markov chain only indirectly, i.e., through the derived quantities $V_n^*(\cdot, \cdot)$, but no explicit conditions are given on the primary model parameters $c(\cdot, \cdot)$ and $[P_{x,y}(u)]$. In the literature, e.g., Borkar (1991), Sennott (1986), Sennott (1989), Weber and Stidham (1987), explicit conditions have been imposed on $c(\cdot, \cdot)$ and $[P_{x,y}(u)]$ in order to verify Assumptions 3.1–3.2; see also Cavazos-Cadena and Sennott (1992). In Ghosh and Marcus (1992) explicit conditions, which imply Assumptions 3.1–3.2, were imposed on $c(\cdot, \cdot)$ and $[P_{x,y}(u)]$ to show SAC optimality of AC optimal policies.

Example 6.1: Consider a CMC with state space $\mathbf{S} = \mathbb{N}_0$. To specify the other components of the model, let $\beta \in (0, 1)$ be fixed, and select a sequence $\{t_k\} \subset \mathbb{N}_0$ such that

- (a) $0 = t_0 < t_1 < t_2 < \dots$, and
 (b)

$$t_k > k \left[\sum_{s=0}^{k-1} \frac{(1+t_s)}{\beta^{t_s}} \right], \quad k \in \mathbb{N}.$$

Next, set $\mathbf{U} = \{0, 1\}$ (endowed with the discrete topology) and define the action set, the cost function and the transition law as follows:

- (i) for $x \neq t_k$, $k \in \mathbb{N}_0$, let $\mathcal{Q}(x) := \{1\}$, $c(x, 1) := 0$, and $P_{x,x+1}(1) := 1$; whereas
 (ii) for $x = t_k$, $k \in \mathbb{N}_0$, let $\mathcal{Q}(t_k) := \{0, 1\}$, $c(t_k, 0) = c(t_k, 1) := \frac{(1+t_k)}{\beta^{t_k}}$, and $P_{t_k,0}(0) := 1$, $P_{t_k,1+t_k}(1) = 1$.

Note that $c(0, u) = 1$, for $u = 0, 1$. Thus, in state $X_t = x \neq t_k$, $k \in \mathbb{N}_0$, the only available action is $U_t = 1$, which produces a state transition to $X_{t+1} = x + 1$, at no cost. On the other hand, when $X_t = t_k$, for some $k \in \mathbb{N}_0$, both actions are available: $U_t = 0$ produces a state transition to $X_{t+1} = 0$, and $U_t = 1$ to $X_{t+1} = 1 + t_k$; in either case the cost incurred is $c(t_k, U_t) = \frac{(1+t_k)}{\beta^{t_k}}$.

In the example above, Assumption 2.1 clearly holds. Furthermore, we will show the Assumptions 3.1–3.2 are also satisfied. To accomplish this, let f_0 and f_1 be stationary policies given by

$$f_0(t_k) = 0; \quad f_1(t_k) = 1, \quad k \in \mathbb{N}_0. \quad (6.1)$$

Thus, under action of policy f_0 , the state $z^* = 0$ is absorbing, and it is clear that

$$V_n(0, f_0) = \frac{1}{1-\alpha}, \quad (6.2)$$

and, for $k \in \mathbb{N}$,

$$\begin{aligned} V_n(t_k, f_0) &= c(t_k, 0) + \alpha V_n(0, f_0) \\ &= \frac{1+t_k}{\beta^{t_k}} + \frac{\alpha}{1-\alpha}. \end{aligned} \quad (6.3)$$

More generally, if the initial state $x \in \mathbb{N}$ is such that $t_k < x < t_{k+1}$, then under the action of policy f_0 no cost will be incurred until state t_{k+1} is reached, which occurs after $(t_{k+1} - x)$ time periods. Therefore, one obtains that for $t_k < x < t_{k+1}$, $k \in \mathbb{N}_0$,

$$V_k(x, f_0) = \alpha^{t_{k+1}-x} V_k(t_{k+1}, f_0). \tag{6.4}$$

The proof of the following result is rather technical, and is given in the Appendix.

Proposition 6.1: For each $\alpha \in (\beta, 1)$, the policy f_0 in (6.1) is DC optimal for the CMC in Example 6.1.

Proposition 6.2: For the CMC in Example 6.1, Assumptions 3.1–3.2 are satisfied. In addition, the policy f_0 in (6.1) is both AC and Blackwell optimal, and attains the minimum in the ACOI.

Proof: Assume that $\alpha \in (\beta, 1)$. By (6.2)–(6.4), and Proposition 6.1, we have that, for all $x \in \mathbb{N}_0$,

$$V_\alpha^*(x) = V_\alpha(x, f_0) < \infty,$$

and hence Assumption 3.1 holds. To verify assumption 3.2, set $z^* = 0$. Since $0 < t_k < t_{k+1}$, then from (6.3) it follows that

$$V_\alpha(t_k, f_0) < V_\alpha(t_{k+1}, f_0),$$

which together with (6.4) shows that $V_\alpha(\cdot, f_0)$ is increasing in $x \in \mathbb{N}_0$. Hence

$$V_\alpha^*(x) - V_\alpha^*(0) \geq 0, \quad \alpha \in (\beta, 1), \quad x \in \mathbb{N}_0.$$

Therefore, Assumption 3.2(i) holds with $z^* = 0$ and $N = 0$, for the CMC in Example 6.1.

Now, from Proposition 6.1 and (6.2)–(6.4), it follows that, for $\alpha \in (\beta, 1)$,

$$0 \leq V_\alpha^*(x) - V_\alpha^*(0) < b(x),$$

where the function $b: \mathbb{N}_0 \rightarrow (0, \infty)$ is defined as

$$b(x) := \frac{1 + t_{k+1}}{\beta^{t_{k+1}}}, \quad t_k \leq x < t_{k+1}, \quad k \in \mathbb{N}_0.$$

Since one always has that

$$\sum_{j \in \mathbb{N}_0} p_{x,j}(u)b(y) \leq b(0) + b(x + 1) < \infty,$$

for all $(x, u) \in \mathbb{K}$, then $b(\cdot)$ satisfies Assumption 3.2(ii) (actually, it satisfies a stronger version of this assumption, see Puterman (1994), pp. 415–420). Thus both Assumptions 3.1 and 3.2 are satisfied for the CMC in Example 6.1.

Finally, since f_0 is DC optimal for all $\alpha \in (\beta, 1)$, i.e., it is Blackwell optimal (see Arapostathis et al. (1993), Bertsekas (1987)) then it is also AC optimal, and moreover it satisfies the ACOI (actually, it satisfies (3.10) with equality); see Arapostathis et al. (1993), Cavazos-Cadena (1991.a), Puterman (1994), Sennott (1986), Sennott (1989). \square

Proposition 6.3: For the CMC in Example 6.1, and the policy f_0 in (6.1), the following holds:

$$\limsup_{n \rightarrow \infty} \frac{V_n(0, f_0) - V_n^*(0)}{n + 1} = 1. \tag{6.5}$$

Therefore, f_0 is not SAC optimal.

Proof: Since state $z^* = 0$ is absorbing under f_0 and $c(0, 0) = 1$, then clearly

$$\frac{V_n(0, f_0)}{n + 1} = 1, \quad n \in \mathbb{N},$$

and thus (6.5) is equivalent to

$$\liminf_{n \rightarrow \infty} \frac{V_n^*(0)}{n + 1} = 0. \tag{6.6}$$

To verify (6.6), set $n_k := t_k - 1$, $k \in \mathbb{N}$, and consider the policy f_1 in (6.1). When the system is under the control of policy f_1 , then the state increases by one unit

in every step. Then, it follows that

$$\begin{aligned} V_{n_k}(0, f_1) &= \sum_{t=0}^{n_k} c(t, f_1(t)) \\ &= \sum_{s=0}^{k-1} c(t_s, 1) \\ &= \sum_{s=0}^{k-1} \frac{1+t_s}{\beta^t} \\ &< \frac{t_k}{k} = \frac{n_k + 1}{k}, \end{aligned}$$

where the second equality follows from the definition of the cost function, and the inequality is due to part (b) in Example 6.1. Therefore, we have that

$$0 \leq \frac{V_{n_k}(0, f_1)}{n_k + 1} < \frac{1}{k} \xrightarrow{k \rightarrow \infty} 0,$$

and thus

$$\begin{aligned} \liminf_{n \rightarrow \infty} \frac{1}{n+1} \liminf_{m_k \rightarrow \infty} \frac{V_{m_k}^*(0)}{m_k + 1} \\ \leq \liminf_{n_k \rightarrow \infty} \frac{V_{n_k}(0, f_1)}{n_k + 1} = 0, \end{aligned}$$

which yields (6.6), since $V_n^*(0) \geq 0$ given that $c(\cdot, \cdot) \geq 0$. \square

In summary, it has been shown that the CMC in Example 6.1 satisfies Assumptions 2.1, 3.1 and 3.2, but the AC optimal policy f_0^* is not SAC optimal. Furthermore, it is not difficult to verify that, for all $x \in \mathbb{N}_0$,

$$\frac{V_n(x, f_0^*)}{n+1} \xrightarrow{n \rightarrow \infty} 1,$$

and that

$$\liminf_{n \rightarrow \infty} \frac{V_n(x, f_1)}{n+1} = 0, \quad (6.7)$$

and thus

$$\limsup_{n \rightarrow \infty} \frac{V_n(x, f_0^*) - V_n^*(x)}{n+1} = 1.$$

Remark: In addition, (6.7) shows that using the limit inferior or superior in the definition of the AC criterion does *not* lead to equivalent criteria, answering in the negative a question posed in Cavazos-Cadena (1991.a).

Appendix: Proof of Proposition 6.1

Let $\alpha \in (\beta, 1)$ be fixed, and let $f_\alpha^* \in \Pi_{SD}$ be an arbitrary stationary DC optimal policy for the discount factor α ; see Bertsekas (1987), Bertsekas and Shreve (1978), Puterman (1994). For each $k \in \mathbb{N}_0$, define

$$m_k := \min \{s \geq k \mid f_\alpha^*(t_s) = 0\}. \quad (A.1)$$

Note that the statement $f_0 = f_\alpha^*$ is equivalent to $f_\alpha^*(t_k) = 0$, for all $k \in \mathbb{N}_0$, which in turn is equivalent to

$$m_k = k, \quad \forall k \in \mathbb{N}_0, \quad (A.2)$$

which we verify in the sequel. The proof of (A.2) is divided into three steps.

Step 1: For each $k \in \mathbb{N}_0$, $m_k < \infty$.

This assertion is established by contradiction, as follows: suppose that $m_k = \infty$, for some $k \in \mathbb{N}_0$. In this case $f_\alpha^*(t_s) = 1$, for all $s \geq k$, and starting at state $X_1 = t_k$, the state of the system increases by one unit in every step, incurring in a nonzero cost $c(t_s, 1)$ when state $X_t = t_s$ is visited after $t = t_s - t_k$ time periods, $s \geq k$. Therefore

$$\begin{aligned} V_\alpha^*(t_k) &= V_\alpha(t_k, f_\alpha^*) = \sum_{s=k}^{\infty} \alpha^{(t_s - t_k)} c(t_s, 1), \\ &= \sum_{s=k}^{\infty} \alpha^{(t_s - t_k)} \frac{1+t_s}{\beta^t} \\ &\geq \alpha^{-t_k} \sum_{s=k}^{\infty} (1+t_s) = \infty, \end{aligned}$$

where the inequality follows since $\alpha \in (\beta, 1)$. Thus, comparing the above with (6.3), a contradiction is obtained.

Step 2: $m_0 = 0$.

This assertion is established also by contradiction, as follows: suppose that $0 < m_0 < \infty$. Starting at $X_0 = t_0 = 0$, under the control of policy f_α^* the state of the system will increase by one unit every time step, reaching the state $X_{t_{m_0}} = t_{m_0}$ at time t_{m_0} and resetting to state $X_{t_{m_0}+1} = 0$, at time $t_{m_0} + 1$. Then by Bellman's Principle of Optimality (see Bertsekas (1987), Bertsekas and Shreve (1978)) it follows that

$$\begin{aligned} V_\alpha^*(0) &= V_\alpha(t_0, f_\alpha^*) \\ &= \sum_{t=0}^{t_{m_0}} \alpha^t c(t, f_\alpha^*(t)) + \alpha^{t_{m_0}+1} V_\alpha(0, f_\alpha^*) \\ &= \sum_{k=0}^{m_0} \alpha^{t_k} \frac{1+t_k}{\beta^{t_k}} + \alpha^{t_{m_0}+1} V_\alpha^*(0), \end{aligned}$$

since $c(t, u) = 0$, when $t \neq t_s, s \in \mathbb{N}_0$. Therefore

$$(1 - \alpha^{t_{m_0}+1}) V_\alpha^*(0) = \sum_{k=0}^{m_0} \left(\frac{\alpha}{\beta} \right)^{t_k} (1 + t_k),$$

and since $\alpha \in (\beta, 1)$, it follows that

$$(1 - \alpha^{t_{m_0}+1}) V_\alpha^*(0) > 1 + t_{m_0}. \tag{A.3}$$

To conclude, note that

$$\begin{aligned} (1 - \alpha^{t_{m_0}+1}) V_\alpha^*(0) &\leq (1 - \alpha^{t_{m_0}+1}) V_\alpha(0, f_0) \\ &= \frac{1 - \alpha^{t_{m_0}+1}}{1 - \alpha} \\ &= \alpha^{t_{m_0}} (1 + t_{m_0}), \end{aligned}$$

for some value $\hat{\alpha} \in (\alpha, 1)$, by the Mean Value Theorem; for the first equality see

(6.2). Then

$$(1 - \alpha^{t_{m_0}+1}) V_\alpha^*(0) < 1 + t_{m_0},$$

which contradicts (A.3). Therefore $m_0 = 0$.

Note that since $m_0 = 0$, then $f_\alpha^*(0) = 0$. Thus, under control of policy f_α^* , if the initial state is $X_0 = 0$, then the state of the system will remain $X_t = 0$, and a unit cost per stage will be incurred; hence

$$V_\alpha^*(0) = \frac{1}{1 - \alpha}. \tag{A.4}$$

Step 3: for each $k > 0, m_k = k$.

As before, the proof of this statement is established by contradiction, as follows: suppose that $m_k > k$, for some integer $k > 0$. Under the control of policy f_α^* , when the initial state of the system is $X_0 = t_k$, the state of the system will increase until state $X_{t_{m_k}-t_k} = t_{m_k}$ is reached, at time $t_{m_k} - t_k$; then the state will be reset to $X_{t_{m_k}-t_k+1} = 0$, at time $t_{m_k} - t_k + 1$. However, between these events state $X_{t_s-t_k} = t_s$ will be reached at time $t_s - t_k, k \leq s \leq m_k$, at which time a cost $(1 + t_s)/\beta^{t_s}$ will be incurred. Therefore, by Bellman's Principle of Optimality, it follows that

$$\begin{aligned} V_\alpha^*(t_k) &= \sum_{t=0}^{t_{m_k}-t_k} \alpha^t c(t + t_k, f_\alpha^*(t + t_k)) + \alpha^{t_{m_k}-t_k+1} V_\alpha^*(0) \\ &= \sum_{s=k}^{m_k} \alpha^{(t_s-t_k)} \frac{1+t_s}{\beta^{t_s}} + \frac{\alpha^{t_{m_k}-t_k+1}}{1 - \alpha}, \end{aligned}$$

where the second equality follows from (A.4). Therefore, since $\alpha \in (\beta, 1)$, it follows that

$$\begin{aligned} V_\alpha^*(t_k) &> \frac{1+t_k}{\beta^{t_k}} + (1+t_{m_k}) + \frac{\alpha^{t_{m_k}-t_k+1}}{1 - \alpha} \\ &> \frac{1+t_k}{\beta^{t_k}} + (t_{m_k} - t_k) + \frac{\alpha^{t_{m_k}-t_k+1}}{1 - \alpha} \\ &> \frac{1+t_k}{\beta^{t_k}} + \sum_{s=1}^{t_{m_k}-t_k} \alpha^s + \frac{\alpha^{t_{m_k}-t_k+1}}{1 - \alpha} \\ &= \frac{1+t_k}{\beta^{t_k}} + \frac{\alpha}{1 - \alpha} = V_\alpha(t_k, f_0), \end{aligned}$$

where the last equality follows from (6.3). The above contradicts the optimality of $V_{\alpha}^*(\cdot)$, concluding our proof. \square

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
- Bertsekas DP, Shreve SE (1978) *Stochastic optimal control: The discrete time case*. Academic Press, New York
- Borkar VS (1991) *Topics in controlled Markov chains*. Pitman Research Notes in Mathematics Series #240, Longman Scientific & Technical, UK
- Cavazos-Cadena R (1991a) Recent results on conditions for the existence of average optimal stationary policies. *Annals Operat Res* 28:3–28
- Cavazos-Cadena R (1991b) A counterexample on the optimality equation in Markov decision chains with the average cost criterion. *Syst Control Lett* 16:387–392
- Cavazos-Cadena R, Sennott LI (1992) Comparing recent assumptions for the existence of average optimal stationary policies. *Operat Res Lett* 26:33–37
- Dynkin EB, Yushkevich AA (1979) *Controlled Markov processes*. Springer-Verlag, New York
- 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
- Ghosh MK, Marcus SI (1992) On strong average optimality of Markov decision processes with unbounded costs. *Operat Res Lett* 11:99–104
- Gatarek D, Stettner L (1990) On the compactness method in general ergodic impulsive control of Markov processes. *Stoch and Stoch Reports* 31:15–25
- Hernández-Lerma O (1989) *Adaptive Markov control processes*. Springer-Verlag, New York
- Hernández-Lerma O, Lasserre JB (1990) Average cost optimal policies for Markov control processes with Borel state space and unbounded costs. *Syst Control Lett* 15:349–356
- Puterman ML (1994) *Markov decision processes*. John Wiley, New York
- Ritt RK, Sennott LI (1992) Optimal stationary policies in general state space Markov decision chains with finite action sets. *Math Operat Res* 17:901–909
- Schäl M (1993) Average optimality in dynamic programming with general state space. *Math Operat Res* 18:163–172
- Sennott LI (1986) A new condition for the existence of optimal stationary policies in average cost Markov decision processes. *Oper Res Lett* 5:17–23
- Sennott LI (1989) Average cost optimal stationary policies in infinite state Markov decision processes with unbounded costs. *Oper Res* 37:626–633
- Ross SM (1983) *Introduction to stochastic dynamic programming*. Academic Press, New York
- Weber RR, Stidham S Jr (1987) Optimal control of service rates in networks of queues. *Advances in Appl Probab* 19:202–218

Received: January 1995

Revised version received: May 1995