SRC TR 88 - 41

STEERING POLICIES for MARKOV DECISION PROCESSES UNDER a RECURRENCE CONDITION

by

Dye-Jyun Ma
and
Armand M. Makowski

STEERING POLICIES FOR MARKOV DECISION PROCESSES UNDER A RECURRENCE CONDITION

by

Dye-Jyun Ma[1] and Armand M. Makowski[2]

Electrical Engineering Department and Systems Research Center
University of Maryland, College Park, Maryland 20742

ABSTRACT

This paper presents a class of adaptive policies in the context of Markov decision processes (MDP's) with long-run average performance measures. Under a recurrence condition, the proposed policy alternates between two stationary policies so as to adaptively track a sample average cost to a desired value. Direct sample path arguments are presented for investigating the convergence of sample average costs and the performance of the adaptive policy is discussed. The obtained results are particularly useful in discussing constrained MDP's with a single constraint. Applications include a wide class of constrained MDP's with finite state space (Beutler and Ross 1985), an optimal flow control problem (Ma and Makowski 1987) and an optimal resource allocation problem (Nain and Ross 1986).

[1] The work of this author was supported partially through NSF Grant ECS-83-51836 and partially through NSF Grant NSFD CDR-85-00108.

[2] The work of this author was supported partially through ONR Grant N00014-84-K-0614 and partially through a grant from AT&T Bell Laboratories.

Large classes of engineering problems can be cast as Markov decision processes (MDP's) with long-run average performance measures and in many situations, the analysis identifies a (possibly randomized) *stationary* policy f^* to yield the desired performance. Unfortunately, the structural properties of the policy f^* often prevent its implementability owing to computational difficulties inherent to its definition (Nain and Ross 1986) or to insufficient knowledge of the model parameters (Ma 1988). In fact, in many applications (Ma, Makowski and Shwartz 1986), solving explicitly for f^* turns out to be a difficult task which is further compounded when some of the model parameters are not exactly known.

Such difficulties naturally point to the need for an *implementation* theory within the context of MDP's. The purpose of this theory is to develop implementable strategies which yield the same performance as the policy f^*. Here, implementability is synonymous with the availability of an algorithm which produces *on-line* control values, given available feedback and model information. Such implementation issues were recently discussed in (Makowski and Shwartz 1986a), where various methods for implementation were proposed.

In this paper, the discussion is given in the context of MDP's with *countable* state space, under some recurrent structures, and the attention is focused on a class of implementable policies called *steering* policies. More concretely, let V be the (desirable) value of the long-run average cost incurred under the policy f^*, and let $\overline{g}$ and $\underline{g}$ denote two stationary policies (possibly randomized). The policy $\overline{g}$ (resp. $\underline{g}$) overshoots (resp. undershoots) the requisite performance level V in that the policy $\overline{g}$ (resp. $\underline{g}$) yields a value for the long-run average cost which is higher (resp. lower) than V. The proposed scheme assumes only the *implementability* of the two stationary policies $\overline{g}$ and $\underline{g}$, and *adaptively* alternates between $\overline{g}$ and $\underline{g}$ under the assumption that some privileged state z is visited infinitely often under both policies $\overline{g}$ and $\underline{g}$. The decision to switch policies is taken only at the times when the state of the system visits the privileged state z so as to adaptively track the *sample* average cost to the value V. At those (random) instants, the current value of the sample average cost is compared against the target value V. If the sample average is above (resp. below) the value V, the policy $\underline{g}$ (resp. $\overline{g}$) will be used until the next visit to the privileged state. Thus, between two consecutive visits to that particular state, one and *only* one of the two policies is used.

This steering policy α is analyzed under the assumption that for both policies $\overline{g}$ and $\underline{g}$, the privileged state z is recurrent for the induced Markov chain and that there is "no escape at infinity". It is shown that the policy α indeed steers the sample cost averages to the desired value V, and under additional growth conditions, that the long-run expected averages under f^* and α coincide.

Direct sample path arguments are presented. They take advantage of the very form of the steering policy α and exploits some hidden regenerative properties of the state process under the steering policy α. This discussion is inspired by the proof of the Ergodic Theorem for recurrent Markov chains based on the Strong Law of Large Numbers as given by Chung (1967).

The obtained results are of interest in the context of constrained MDP's with a single constraint, where an optimal stationary policy is often found by simple randomization between two pure stationary policies $\overline{g}$ and $\underline{g}$ with the abovementioned properties. Typically, these two pure policies are identified through Lagrangian arguments and the randomization bias is chosen so as to meet the constraint value (Beutler and Ross 1985, Ma and Makowski 1987, Nain and Ross 1986). The very form of this solution lends itself to an implementation via the steering policy α, which requires no knowledge of the randomization bias value, and as such can be viewed as an indirect adaptive policy (Ma, Makowski and Shwartz 1986). The steering policy α considered here should also be contrasted against the so-called *time-sharing* implementation of f^* proposed by Altman and Shwartz (1986), whereby the decision-maker alternates between the two policies $\overline{g}$ and $\underline{g}$ according to some *deterministic* (thus *non-adaptive*) mechanism associated with the recurrence cycles. The instrumentation of these time-sharing policies requires the explicit evaluation of certain cost functionals, so that the proposed steering policy α could be interpreted as providing an adaptive version of time sharing.

The work reported here was motivated by an idea proposed by Ross (1985) in the context of an optimal resource allocation problem with a constraint. Ross suggested a scheme whereby the decision-maker could possibly switch between two static priority assignments at any decision epoch so as to steer the long-run average cost to the value V. The analysis in that case seems more involved and as of the writing this paper, the question of its performance still remains open. However, in some specific situations, which include a class of constrained MDP's with finite state spaces, the results obtained here translate into results for Ross scheme.

The paper is organized as follows. The underlying MDP formulation is stated in Section 1. The problem of steering the cost to a specific value is precisely formulated in Section 2.1, the steering policy α is introduced in Section 2.2 and the key technical assumptions are discussed in Section 2.3. The main results of the paper are presented in Section 3.1, and are proved in Section 3.3 using some key intermediate results which are summarized in Section 3.2. While the proof of these intermediate results is delayed until Section 5, Section 4 first outlines applications to constrained MDP's. The situation of finite state spaces and compact action space is discussed in Section 4.1, while problems

in optimal flow control and resource allocation are considered in Sections 4.2 and 4.3, respectively. Section 5 closes the paper with a detailed discussion of the sample path arguments.

A word on the notation: The set of real numbers is denoted by $\mathbb{R}$, and $\mathbb{N}$ denotes the set of all non-negative integers. The indicator function of any set E is simply denoted by $1[E]$. Unless stated otherwise, $\lim_n$, $\underline{\lim}_n$ and $\overline{\lim}_n$ are taken with n going to infinity.

1. The model

Consider a MDP with *countable* state space S, *measurable* action space U, and *Borel measurable* transition kernel $(p_{xy}(u))$, i.e., the mappings $p_{xy}(\bullet) : U \to \mathbb{R}$ are *Borel* measurable and satisfy the standard properties $0 \leq p_{xy}(u) \leq 1$ and $\sum_y p_{xy}(u) = 1$ for all x and y in S and u in U. The probabilistic framework for this MDP is defined on the *canonical* sample space $\Omega := (S \times U)^\infty$. An element ω of Ω is viewed as a sequence $(x_0, \omega_0, \omega_1, \ldots)$ with x_0 in S and ω_n in $U \times S$ for all $n = 0, 1, \ldots$, where each block component ω_n is of the form (u_n, x_{n+1}) with u_n and x_{n+1} elements of U and S, respectively. The information spaces $\{\mathbb{H}_n\}_0^\infty$ are recursively generated by $\mathbb{H}_0 := S$ and $\mathbb{H}_{n+1} := \mathbb{H}_n \times U \times S$ for all $n = 0, 1, \ldots$, so that an element h_n in $\mathbb{H}_n$ is uniquely associated with the sample ω by $h_n := (x_0, \omega_0, \ldots, \omega_{n-1})$ with $h_0 := x_0$. The interpretation of these quantities is as follows: When the sample $\omega = (x_0, \omega_0, \omega_1, \ldots)$ is realized, the system is in state x_n at time n, and the *control action* u_n is generated according to some prespecified mechanism on the basis of the *information* vector h_n.

The coordinate mappings $\{U(n)\}_0^\infty$ and $\{X(n)\}_0^\infty$ are defined on the sample space Ω by setting $U(n, \omega) := u_n$ and $X(n, \omega) := x_n$ with the information mappings $\{H(n)\}_0^\infty$ given by $H(n, \omega) := (x_0, \omega_0, \omega_1, \ldots, \omega_{n-1}) = h_n$ for every ω in Ω and for all $n = 0, 1, \ldots$.

For every $n = 0, 1, \ldots$, let $\mathbb{F}_n$ be the σ-field generated by the mapping $H(n)$ on the sample space Ω and with standard notation, $\mathbb{F} := \vee_{n=0}^\infty \mathbb{F}_n$ is simply the natural σ-field on Ω generated by the mappings $\{(U(n), X(n))\}_0^\infty$. On the space $(\Omega, \mathbb{F})$, the mappings $U(n)$, $X(n)$ and $H(n)$ are random variables (RV) taking values in U, S and $\mathbb{H}_n$, respectively.

Let $\mathbb{M}$ denote the space of probability measures on U, when equipped with its natural Borel σ-field. Since randomization is allowed, an admissible policy π is defined as any collection $\{\pi_n\}_0^\infty$ of mappings $\pi_n : \mathbb{H}_n \to \mathbb{M}$ such that the mappings $\mathbb{H}_n \to [0, 1] : h_n \to \pi_n(A; h_n)$ are $\mathbb{F}_n$-measurable for every Borel subset A of U. For each h_n in $\mathbb{H}_n$, the quantity $\pi_n(\bullet; h_n)$ is interpreted as the conditional probability distribution of selecting the control value at time n, given that the information vector h_n is available to the decision-maker. In the sequel, denote the collection of all such admissible policies by $\mathcal{P}$.

Let $\mu(\bullet)$ be a fixed probability distribution on S. Given any policy π in $\mathcal{P}$, there exists a unique probability measure P^π on $\mathbb{F}$, with corresponding expectation operator E^π, satisfying the requirements (R1)-(R3), where

(R1) For all x_0 in S,

$$P^\pi[X(0) = x_0] := \mu(x_0),$$

(R2) For every Borel subsets A of U,

$$P^\pi[U(n) \in A | \mathbb{F}_n] = \pi_n(A; H(n)), \qquad n = 0, 1, \ldots$$

(R3) For all y in S,

$$P^\pi[X(n+1) = y | \mathbb{F}_n \vee \sigma(U(n))] = p_{X(n)y}(U(n)). \qquad n = 0, 1, \ldots$$

When $\mu(\bullet)$ is the point mass distribution at x in S, this notation is specialized to P_x^π and E_x^π, respectively, and it is then plain that $P^\pi[A|X(0) = x] = P_x^\pi[A]$ for every A in $\mathbb{F}$. It now follows readily from (R2)-(R3) that

$$P^\pi[X(n+1) = y | \mathbb{F}_n] = \int_U \pi_n(du; H(n)) p_{X(n)y}(u) \qquad n = 0, 1, \ldots (1.1)$$

for all y in S.

A policy π in $\mathcal{P}$ is said to be a *Markov* policy if there exists a family $\{g_n\}_0^\infty$ of mappings $g_n : S \to \mathbb{M}$ such that $\pi_n(\bullet; H(n)) = g_n(\bullet; X(n))$ P^π-a.s. for all $n = 0, 1, \ldots$. In the event $g_n = g$ for all $n = 0, 1, \ldots$, the Markov policy π is called *stationary* and can be identified with the mapping g itself. It is plain from (R1)-(R3) that for each stationary policy g, the RV's $\{X(n)\}_0^\infty$ form a *time-homogeneous* Markov chain under P^g, with corresponding one-step transition probability matrix $P(g) \equiv (p_{xy}(g))$ given by

$$p_{xy}(g) := \int_U p_{xy}(u) g(du; x) \qquad (1.2)$$

for all x and y in S.

A policy π in $\mathcal{P}$ is said to be a *pure* (or *non-randomized*) policy if there exists a family $\{f_n\}_0^\infty$ of mappings $f_n : \mathbb{H}_n \to U$ such that for every Borel subset A of U, $\pi_n(A; H(n)) = 1[f_n(H(n)) \in A]$ P^π-a.s. for all $n = 0, 1, \ldots$. A *pure Markov stationary* policy π in $\mathcal{P}$ is thus fully characterized by a single mapping $f : S \to U$.

For any mapping $c : S \to \mathbb{R}$, the long-run average cost $J^c(\pi)$ incurred by the admissible policy π in $\mathcal{P}$ is defined by

$$J^c(\pi) := \overline{\lim}_n \frac{1}{n} E^\pi \sum_{t=0}^{n-1} c(X(t)) \tag{1.3}$$

whenever meaningful, and for future reference, introduce the corresponding sample average costs $\{J^c(n)\}_1^\infty$ which are given by

$$J^c(n) := \frac{1}{n} \sum_{t=0}^{n-1} c(X(t)). \qquad n = 1, 2, \ldots (1.4)$$

2. Implementation via Steering Policies

2.1. The problem

Start with a mapping $d : S \to \mathbb{R}$, and let the constant V represent the desired performance level for the long-run average cost (1.3) associated with d. The discussion assumes the existence of two stationary policies $\overline{g}$ and $\underline{g}$ such that

$$J^d(\underline{g}) < V < J^d(\overline{g}). \tag{2.1}$$

The motivation for such an assumption can be found in the theory of constrained MDP's (Ma, Makowski and Shwartz 1986, Ross 1985). The problem (P_V) of interest in this paper is then formulated as

$$(P_V): \quad \text{Find a policy } \alpha \text{ in } \mathcal{P} \text{ such that } J^d(\alpha) = V.$$

Under the condition (2.1), several solutions to this problem are known and are briefly surveyed in (Makowski and Shwartz 1986a). However, as pointed out there, some of these solutions may not be readily implementable given available model and feedback information. It is the purpose of this paper to present and analyze yet another way to solve the problem (P_V), the key feature of the proposed solution being the minimal amount of information required for its implementation.

2.2. Steering policies

The policy α proposed here is of the form

$$\alpha_n(\bullet; H(n)) := \eta(n)\overline{g}(\bullet; X(n)) + (1 - \eta(n))\underline{g}(\bullet; X(n)) \qquad n = 0, 1, \ldots (2.2)$$

where $\{\eta(n)\}_0^\infty$ is a sequence of $\{0,1\}$-valued RV's to be specified shortly. In other words, the policy α alternates between the two policies $\overline{g}$ and $\underline{g}$, with the quantity $\eta(n)$ specifying which one of these two policies is to be used in the time slot $[n, n+1)$.

The policy α proposed and analyzed in this paper finds its origin in an idea proposed by Ross (1985, pp. 126) in the context of an optimal constrained resource allocation problem. In order to steer the long-run average cost to the requested value V, Ross suggested a scheme whereby the decision-maker alternates between the policies $\overline{g}$ and $\underline{g}$ so that the sample averages $\{J^d(n)\}_1^\infty$ track the value V. This policy, denoted hereafter by α_R, also has the form (2.2) but uses a sequence $\{\eta_R(n)\}_0^\infty$ given by

$$\eta_R(n) = 1[J^d(n) \leq V] \qquad n = 1, 2, \ldots (2.3)$$

with $\eta_R(0)$ arbitrary in $\{0, 1\}$. This idea was subsequently adapted by Makowski and Shwartz (1986a), and by Ross (1988) to a more general class of MDP's.

The analysis of the performance of the policy α_R appears quite involved, and to the authors' knowledge, few results are available on this issue. While the general case (Makowski and Shwartz 1986a, Ross 1985) is still open, various special situations have been handled successfully. Makowski (1987) treated the i.i.d. case by viewing the sample averages $\{J^d(n)\}_1^\infty$ as the output values of a stochastic approximations algorithm of the Robbins-Monro type. Ma (1988) in his Ph.D. dissertation solved the problem in the context of a simple flow control problem for discrete-time M/M/1 systems. A careful examination of the analysis carried in (Ma 1988) leads very naturally to the policy α investigated in this paper.

The definition of the policy α will require that the assumptions (A1) be enforced, namely

(A1) *The Markov chain $\{X(n)\}_0^\infty$ has a single recurrent class under each one of the policies $\overline{g}$ and $\underline{g}$. These recurrent classes have a non-empty intersection, and moreover starting from any transient state (if any), the time to absorption in the recurrent class is a.s finite under each policy.*

Let z denote any state in S which is recurrent under both $\overline{g}$ and $\underline{g}$. By virtue of assumption (A1), such a state z clearly exists and has the property that the system returns to it infinitely often under each policy. The RV's $\{\eta(n)\}_0^\infty$ entering the definition of α are recursively generated by the simple relation

$$\eta(n) = 1[X(n) = z]1[J^d(n) \leq V] + 1[X(n) \neq z]\eta(n-1), \qquad n = 1, 2, \ldots (2.4)$$

with $\eta(0)$ arbitrary in $\{0, 1\}$. This policy α operates according to one of the policies $\overline{g}$ and $\underline{g}$ during each cycle, where a *cycle* is defined as the time duration between two consecutive visits of the process $\{X(n)\}_0^\infty$ to the recurrent state z. The essential difference between the policies α and α_R is that although both track the sample cost averages $\{J^d(n)\}_1^\infty$ about the value V, the decision to

switch policies may be taken at every time instant under α_R while only at successive recurrence times (to the state z) under α.

2.3. The assumptions

Let g denote any one of the policies $\overline{g}$ and $\underline{g}$, unless otherwise specified. The first return time to the state z is the RV T defined by

$$T := \inf\{n \geq 1 : \ X(n) = z\}. \tag{2.5}$$

The assumption (A1) essentially amounts to saying that

(A1) *For all x in S,*

$$P_x^g[T < \infty] = 1.$$

For any mapping $c : S \to I\!R$, it will be convenient to define the corresponding cost per cycle Z^c by

$$Z^c := \sum_{t=0}^{T-1} c(X(t)) \tag{2.6}$$

(whenever meaningful); observe that under (A1) the RV Z^c is P^g-a.s well defined and finite. In order to study the performance of the policy α, the following additional technical assumptions (A2)-(A4) will be needed, where

(A2) *The mean recurrence time to the state z is finite under P^g, i.e.,*

$$E_z^g[T] < \infty,$$

(A3) *The expected cost over a cycle Z^d is finite under P_z^g, i.e.,*

$$E_z^g[|Z^d|] = E_z^g[|\sum_{t=0}^{T-1} d(X(t))|] < \infty,$$

(A4) *The equality*

$$J^d(g) := \overline{\lim}_n \frac{1}{n} E^g \sum_{t=0}^{n-1} d(X(t)) = I^d(g)$$

takes place, where

$$I^d(g) := \frac{E_z^g[Z^d]}{E_z^g[T]}. \tag{2.7}$$

The assumption (A2) implies the state z to be positive recurrent under P^g, whence the Markov chain $\{X(n)\}_0^\infty$ under P^g has a unique invariant measure. Under (A1)-(A3), the renewal arguments

given in Chung (1967) shows that the sequence $\{J^d(n)\}_1^\infty$ has a P^g-a.s. finite limit which is given by

$$\lim_n J^d(n) = \frac{E_z^g[Z^d]}{E_z^g[T]} = I^d(g). \qquad P^g - a.s. (2.8)$$

Moreover, if the RV's $\{d(X(n))\}_0^\infty$ are uniformly integrable under P^g, then (A4) is automatically guaranteed since then the convergence (2.8) also holds in $L^1(\Omega, I\!F, P^g)$ (Chung 1967, Thm. 4.5.4). In that case, the quantity $I^d(g)$ also coincides with the expected value of the RV $d(X)$ under the invariant measure induced by P^g, where X denotes a generic S-valued RV.

3. The results

3.1. Performance of the steering policy

The main results of this paper are stated in Theorems 3.1-3.3 below, and are proved in Section 3.3 using some key intermediate results which are summarized in Section 3.2. To state the results, let the RV $p(n)$ given by

$$p(n) := \frac{1}{n} \sum_{t=0}^{n-1} \eta(t) \qquad n = 1, 2, \ldots (3.1)$$

denote the fraction of time over $[0, n)$ during which the policy $\overline{g}$ is used. Set

$$p^* := \frac{V - J^d(\underline{g})}{J^d(\overline{g}) - J^d(\underline{g})} \qquad (3.2)$$

and observe from (2.1) that $0 < p^* < 1$.

Theorem 3.1 *Under* (A1)-(A4), *the convergences*

$$\lim_n p(n) = p^* \qquad P^\alpha - a.s. (3.3)$$

and

$$\lim_n J^d(n) = \lim_n \frac{1}{n} \sum_{t=0}^{n-1} d(X(t)) = V \qquad P^\alpha - a.s. (3.4)$$

take place.

Theorem 3.1 establishes the a.s. convergence of the sample cost averages to the desired value V, and the policy α will indeed constitute a solution to the problem (P_V), provided some additional integrability conditions hold to guarantee convergence of the mean. One possible set of conditions

is given in the next corollary which is based on standard facts on uniform integrability (Chung 1967, Thm. 4.5.4).

Corollary 3.1 *Under* (A1)-(A4), *whenever the RV's* $\{d(X(n))\}_0^\infty$ *are uniformly integrable under* P^α, *the convergence (3.4) also takes place in* $L^1(\Omega, \mathcal{F}, P^\alpha)$ *and consequently*

$$J^d(\alpha) = \lim_n \frac{1}{n} E^\alpha \sum_{t=0}^{n-1} d(X(t)) = V. \tag{3.5}$$

The convergence (3.4) can also be established for other cost mappings $c : S \to \mathbb{R}$ under the assumption (A3bis) stated below which is similar to (but weaker than) (A3), provided (3.3) holds. To state the condition, denote by c^+ and c^- the mappings $S \to \mathbb{R}$ defined by $c^+(x) := \max(c(x), 0)$ and $c^-(x) := \max(-c(x), 0)$ for all x in S.

(A3bis) *The cost over a cycle* Z^c *has a (possibly infinite) expectation under* P_z^g, *or equivalently, the quantities* $E_z^g[Z^{c^+}]$ *and* $E_z^g[Z^{c^-}]$ *are not both infinite under* P_z^g.

Set

$$I^c(g) := \frac{E_z^g[Z^c]}{E_z^g[T]}. \tag{3.6}$$

Theorem 3.2 *Assume (3.3) to hold and let the mapping* $c : S \to \mathbb{R}$ *satisfy* (A3bis). *If the quantity* $I^c(\overline{g}) + I^c(\underline{g})$ *is well defined (but possibly infinite), then the convergence*

$$\lim_n J^c(n) = p^* I^c(\overline{g}) + (1 - p^*) I^c(\underline{g}) \qquad P^\alpha - a.s. \tag{3.7}$$

takes place.

If the assumption (A3bis) is strengthened to (A3)-(A4) (with c replacing d), and if the RV's $\{c(X(n))\}_0^\infty$ are *uniformly integrable* under P^α, then the quantity on the righthand side of (3.7) is finite, and the convergence (3.7) holds also in $L^1(\Omega, \mathcal{F}, P^\alpha)$.

Corollary 3.2 *Assume (3.3) to hold and let the mapping* $c : S \to \mathbb{R}$ *satisfy assumptions* (A3)-(A4). *If the RV's* $\{c(X(n))\}_0^\infty$ *are uniformly integrable under* P^α, *the convergence (3.7) also takes place in* $L^1(\Omega, \mathcal{F}, P^\alpha)$, *and consequently*

$$J^c(\alpha) = \lim_n \frac{1}{n} E^\alpha \sum_{t=0}^{n-1} d(X(t)) = p^* J^c(\overline{g}) + (1 - p^*) J^c(\underline{g}). \tag{3.8}$$

This result will be particularly useful in discussing constrained MDP's in the next section.

Although this paper is devoted essentially to the study of the policy α, the results obtained here have implications for the policy α_R in some special yet important cases. Such a situation is discussed in the next proposition.

Theorem 3.3 *If the policies $\overline{g}$ and $\underline{g}$ coincide in all but one state, say x_0 in S, which is recurrent under each policy, then the policy α_R coincides with the policy α defined by (2.2) and (2.4) with $z = x_0$, and consequently Theorems 3.1-3.2 and their corollaries hold for the policy α_R under appropriate assumptions.*

The situation of Theorem 3.3 occurs in a wide class of constrained MDP's with finite state space and in some other problems as well, as illustrated in Section 4.

3.2. Convergence along recurrence times

The intermediate results which are useful in establishing the main Theorems 3.1-3.2 are summarized in this section. They are motivated by the very form of the steering policy, and represent the main technical ingredients of the paper. A complete discussion of their analysis is delayed until Section 5.

Recall that the very form of the steering policy α forces the decision for switching between policies to be taken only at the times the state process visits the state z. This suggests that the behavior of the control algorithm might be fully determined by the properties of the sample average cost sequence taken only along these recurrence epochs.

To that end, consider the state z in S entering the definition (2.4), and recursively define the recurrence time sequence $\{\tau(k)\}_0^\infty$ of $I\!N \cup \{\infty\}$-valued RV's by

$$\tau(k+1) = \begin{cases} \inf\{t > \tau(k) : X(t) = z\} & \text{if the set is non-empty;} \\ \infty & \text{otherwise} \end{cases} \qquad k = 0, 1, \ldots (3.9)$$

where $\tau(0) := 0$. With this notation, the interval $[\tau(k-1), \tau(k))$ is simply the k^{th} cycle.

The recurrence condition (A1) and the definition of the steering policy α lead readily to the following intuitive fact, the proof of which is omitted for sake of brevity.

Lemma 3.4 *Assume the recurrence condition* (A1) *to hold. The RV's $\tau(k)$ are P^α-a.s. finite for all $k = 1, 2, \ldots$, or equivalently, the state process $\{X(n)\}_0^\infty$ visits the state z infinitely often under P^α. Moreover, under the additional assumptions* (A2)-(A4), *the steering policy α alternates infinitely often between the two policies $\overline{g}$ and $\underline{g}$.*

The key intermediate results for proving Theorems 3.1-3.2 are summarized in the next proposition. Set

$$q^* := \frac{p^* E_z^{\underline{g}}[T]}{(1-p^*)E_z^{\overline{g}}[T] + p^* E_z^{\underline{g}}[T]}. \tag{3.10}$$

Theorem 3.5 *Assume* (A1)-(A4) *to hold. The convergence*

$$\lim_k p(\tau(k)) = p^* \qquad P^\alpha - a.s. \tag{3.11}$$

takes place, and for any mapping $c : S \to \mathbb{R}$ *satisfying* (A3bis), *the convergence*

$$\lim_k J^c(\tau(k)) = p^* I^c(\overline{g}) + (1-p^*) I^c(\underline{g}) \qquad P^\alpha - a.s. \tag{3.12}$$

takes place whenever the quantity $I^c(\overline{g}) + I^c(\underline{g})$ *is well defined. Moreover, the Law of Large Numbers holds true in the form*

$$\lim_k \frac{\tau(k)}{k} = q^* E_z^{\overline{g}}[T] + (1-q^*) E_z^{\underline{g}}[T]. \qquad P^\alpha - a.s. \tag{3.13}$$

When applying (3.12) to the cost mapping d, simple algebraic calculations using (A4), (2.7) and (3.2) readily yield

$$\lim_k J^d(\tau(k)) = V. \qquad P^\alpha - a.s. \tag{3.14}$$

In other words, (3.11)-(3.12) yield the convergences (3.3)-(3.4) and (3.7) along the resurrence times. Although the convergence (3.13) presents a similar version of the Law of Numbers, it should be noted that the recurrence times $\{\tau(k)\}_1^\infty$ do not form a renewal sequence under P^α.

3.3. A proof of Theorems 3.1-3.2

The proof of Theorems 3.1-3.2 is now easily recovered from Theorem 3.5. Let

$$k(n) := \max\{k \geq 0 : \ \tau(k) \leq n\} \qquad n = 1, 2, \ldots \tag{3.15}$$

be the number of cycles over the horizon $[0, n)$ including the one in progress at time n. It is plain from Lemma 3.4 that

$$\lim_n k(n) = \infty. \qquad P^\alpha - a.s. \tag{3.16}$$

For each $n = 1, 2, \ldots$, $\tau(k(n)) \leq n < \tau(k(n)+1)$ so that for any *non-negative* mapping $c : S \to \mathbb{R}$,

$$\frac{\tau(k(n))}{n} J^c(\tau(k(n))) \leq J^c(n) \leq \frac{\tau(k(n)+1)}{n} J^c(\tau(k(n)+1)), \tag{3.17}$$

and similarly,

$$\frac{\tau(k(n))}{n} p(\tau(k(n))) \leq p(n) \leq \frac{\tau(k(n)+1)}{n} p(\tau(k(n)+1)). \tag{3.18}$$

By the Law of Large Numbers (3.13), it is clear that

$$\lim_k \frac{\tau(k)}{\tau(k+1)} = \lim_k \frac{\frac{1}{k}\tau(k)}{\frac{1}{k+1}\tau(k+1)} \frac{k+1}{k} = 1. \qquad P^\alpha - a.s. \tag{3.19}$$

Since

$$\frac{\tau(k(n))}{\tau(k(n)+1)} \leq \frac{\tau(k(n))}{n} \leq 1, \tag{3.20}$$

it is now plain from (3.16) and (3.19)-(3.20) that

$$\lim_n \frac{\tau(k(n))}{n} = \lim_n \frac{\tau(k(n)+1)}{n} = 1. \qquad P^\alpha - a.s. \tag{3.21}$$

By virtue of (3.11)-(3.12), the inequalities (3.17)-(3.18) and the convergence (3.21) yield the convergences (3.3), (3.4) and (3.7) for non-negative mappings.

For a general cost mapping c, start with the decomposition $J^c(n) = J^{c^+}(n) - J^{c^-}(n)$ for all $n = 1, 2, \ldots$, and apply the result for non-negative mappings developed above, so that

$$\lim_n J^{c^+}(n) = p^* I^{c^+}(\overline{g}) + (1-p^*) I^{c^+}(\underline{g}) \qquad P^\alpha - a.s. \tag{3.22a}$$

and

$$\lim_n J^{c^-}(n) = p^* I^{c^-}(\overline{g}) + (1-p^*) I^{c^-}(\underline{g}). \qquad P^\alpha - a.s. \tag{3.22b}$$

It is now plain under the enforced assumptions that

$$\begin{aligned} \lim_n J^c(n) &= \lim_n J^{c^+}(n) - \lim_n J^{c^-}(n) \\ &= p^* I^c(\overline{g}) + (1-p^*) I^c(\underline{g}) \end{aligned} \qquad P^\alpha - a.s. \tag{3.23}$$

and the proof of Theorems 3.1-3.2 is therefore complete.

4. Applications to Constrained MDP's

This section is devoted to various applications of Theorems 3.1-3.3 and their corollaries to constrained MDP's. Let c and d be two mappings $S \to \mathbb{R}$ and for every V in $\mathbb{R}$, define the set $\mathcal{P}_V$ of constrained policies by

$$\mathcal{P}_V := \{\pi \text{ in } \mathcal{P} : J^d(\pi) \leq V\}. \tag{4.1}$$

The constrained MDP (CP_V) is then formulated as

$$(CP_V): \quad \text{Minimize } J^c(\pi) \text{ over } \mathcal{P}_V.$$

In the three situations discussed here, this constrained MDP is solved by Lagrangian arguments: For every $\gamma > 0$, define the mapping $b^\gamma : S \to \mathbb{R} : x \to b^\gamma(x) = c(x) + \gamma d(x)$, and consider the corresponding *unconstrained* Lagrangian problem (LP^γ), where

$$(LP^\gamma): \quad \text{Minimize } J^{b^\gamma}(\pi) \text{ over } \mathcal{P}.$$

In each example, under appropriate hypotheses, there exist two *pure* stationary policies $\overline{g}$ and $\underline{g}$ which both solve the same Lagrangian problem (LP^{γ^*}) for some $\gamma^* > 0$, i.e.,

$$J^{b^{\gamma^*}}(\overline{g}) = J^{b^{\gamma^*}}(\underline{g}) = \inf_{\pi \in \mathcal{P}} J^{b^{\gamma^*}}(\pi), \tag{4.2}$$

and which satisfy the cost inequalities

$$J^d(\underline{g}) < V < J^d(\overline{g}). \tag{4.3}$$

For $0 \leq \eta \leq 1$, let the *randomized* policy f^η be the stationary policy defined by $f^\eta := \eta\overline{g} + (1-\eta)\underline{g}$. If the mapping $\eta \to J^d(f^\eta)$ is continuous, the equation

$$J^d(f^\eta) = V, \quad 0 \leq \eta \leq 1 \tag{4.4}$$

has at least one solution, say η^*, in view of (4.3). The constrained problem (CP_V) is then solved by the stationary policy $f^* \equiv f^{\eta^*}$, provided (i) f^* solves the Lagrangian problem (LP^{γ^*}) and (ii) both functionals $J^c(f^*)$ and $J^d(f^*)$ exist as limits, so that

$$J^{b^{\gamma^*}}(f^*) = J^c(f^*) + \gamma^* J^d(f^*) = \inf_{\pi \in \mathcal{P}} J^{b^{\gamma^*}}(\pi). \tag{4.5}$$

In that case,

$$J^d(f^*) = V \quad \text{and} \quad J^c(f^*) = \inf_{\pi \in \mathcal{P}_V} J^c(\pi) \tag{4.6}$$

by standard arguments which are summarized in (Ma, Makowski and Shwartz 1986, Ross 1985).

Consider now the steering policy α defined in Section 2. Under (A1)-(A4), whenever the RV's $\{d(X(n))\}_0^\infty$ are uniformly integrable under P^α, Corollary 3.1 and (4.6) yield $J^d(\alpha) = J^d(f^*) = V$ and (3.3) holds by Theorem 3.1. Consequently, if the mapping c also satisfies (A3)-(A4) (thus so

does the mapping b^{γ^*}) and the RV's $\{c(X(n))\}_0^\infty$ are uniformly integrable under P^α (thus so are the RV's $\{b^{\gamma^*}(X(n))\}_0^\infty$), then Corollary 3.2 necessarily implies the relation $J^c(\alpha) = J^c(f^*)$. These remarks are summarized in the next proposition.

Theorem 4.1 *Suppose the problem* (CP_V) *admits a solution* f^* *as determined via (4.2)-(4.5). Under* (A1)-(A2), *if* (A3)-(A4) *hold for both mappings* c *and* d, *and if the RV's* $\{c(X(n))\}_0^\infty$ *and* $\{d(X(n))\}_0^\infty$ *are uniformly integrable under* P^α, *then the steering policy* α *also solves the problem* (CP_V), *with* $J^c(\alpha) = J^c(f^*)$ *and* $J^d(\alpha) = J^d(f^*) = V$.

4.1. MDP's with finite state spaces

Beutler and Ross (1985) considered MDP's under the following assumptions (H1)-(H3), where

(H1) *The state space* S *is finite, and the action space* U *is a compact metric space,*

(H2) *For every pure stationary policy* f, *there exists a common state* z *in* S *which is accesible from each state* x *in* S *under* P^f,

(H3) *The set* $\mathcal{P}_V$ *contains at least one pure stationary policy, but does not contain any pure stationary policy which achieves the minimum cost* $J^c(\pi)$ *over all admissible policies* π *in* $\mathcal{P}$.

Under (H1)-(H3), an optimal policy f^* was shown to be determined via (4.2)-(4.5), with the randomization to be performed in only one particular state, i.e., the *pure* policies $\overline{g}$ and $\underline{g}$ coincide in all but one state. As shown by Beutler and Ross (1985), (H2) holds for all *randomized* stationary policies as well, thus implying (A1). The state space being finite, the costs are necessarily *bounded*, so that the assumptions of Theorem 4.1 are immediately satisfied, and the optimality of the steering policies α and α_R easily follows.

Theorem 4.2 *Under* (H1)-(H3), *the steering policies* α *and* α_R *(coincide and) solve the constrained problem* (CP_V) *with* $J^c(\alpha) = J^c(\alpha_R) = J^c(f^*)$ *and* $J^d(\alpha) = J^d(\alpha_R) = J^d(f^*) = V$.

4.2. Optimal flow control

Ma and Makowski (1987) considered the following flow control model for discrete-time $M|M|1$ queues: At the beginning of each time slot, the controller decides either to admit or reject the potential arrival during that slot. A customer (if any) may fail to complete service in a slot with fixed probability $1 - \mu$, in which case it remains at the head of the line to await service in the next slot. This scenario is repeated until successful service completion occurs, at which time the customer leaves the system. The arrival pattern is modelled as a *Bernoulli* sequence with parameter λ, *independent* of the service process as well as of the initial queue size. Under these assumptions, a MDP formulation with state space $S = I\!N$ is readily obtained by taking the state process $\{X(n)\}_0^\infty$

to be the queue size process.

The optimal flow control problem was formulated as the search for a policy that maximizes the throughput subject to the constraint that the long-run average queue size does not exceed a given value V. Here, the throughput and the average queue size incurred by the admissible policy π in $\mathcal{P}$ are given by

$$T(\pi) := \underline{\lim}_n \frac{1}{n} E^\pi \sum_{t=0}^{n-1} \mu 1[X(t) \neq 0] \tag{4.7}$$

and

$$N(\pi) := \overline{\lim}_n \frac{1}{n} E^\pi \sum_{t=0}^{n-1} X(t), \tag{4.8}$$

respectively.

This constrained MDP can be cast as a problem of the form (CP_V) by taking $J^c(\pi) = -T(\pi)$ and $J^d(\pi) = N(\pi)$. The technical assumptions enforced in (Ma and Makowski 1987) are listed below as (H4)-(H5), where

(H4) $N((\infty, 1)) > V$, *where* $(\infty, 1)$ *denotes the policy that admits every single customer,*

(H5) *For every policy* π *in* P,

$$E^\pi[X(0)] = \sum_{x=0}^{\infty} x\mu(x) < \infty,$$

with $\mu(\bullet)$ *denoting the initial queue size distribution.*

It is a simple matter to check (Ma and Makowski 1987) that the RV's $\{X(n)\}_0^\infty$ are uniformly integrable under P^α. Under these assumptions, the constrained optimal control problem is solved by a threshold policy $f^* = (L^*, \eta^*)$ with $N(f^*) = V$. Here, a threshold policy (L, η), with L in $\mathbb{N}$ and $0 \leq \eta \leq 1$, is a stationary policy which at the beginning of each time slot admits (resp. rejects) an incoming customer if the queue size is $< L$ (resp. $> L$), while if the queue size is exactly L, this new customer is accepted (resp. rejected) with probability η (resp. $1 - \eta$).

It should be pointed out that here too the optimal threshold policy $f^* = (L^*, \eta^*)$ is obtained as a randomization with bias η^* between the pure policies $\overline{g} = (L^*, 1)$ and $\underline{g} = (L^*, 0)$, which are identical in all but one state, the state where there are L^* customers in the system. The assumptions of Theorem 4.1 now hold. The states $\{0, 1, \ldots, L^*\}$ are all recurrent under the policies $\overline{g}$ and $\underline{g}$ so that any element in the set $\{0, 1, \ldots, L^*\}$ can be selected as the state z. The optimality of the corresponding steering policy α now follows immediately.

Theorem 4.3 *Under* (H4)-(H5), *the steering policies* α *and* α_R *(coincide and) solve the constrained optimal control problem* (CP_V), *with* $T(\alpha) = T(\alpha_R) = T(f^*)$ *and* $N(\alpha) = N(\alpha_R) = N(f^*) = V$.

4.3. Optimal resource allocation

Consider a system of $K+1$ infinite-capacity queues that compete in discrete-time for the service attention of a single server. At the beginning of each time slot, the controller gives priority to one of the queues. If the k^{th} queue is given service attention during that slot, with probability μ_k the serviced customer (if any) completes service and leaves the system, while with probability $1-\mu_k$, the customer fails to complete service and remains in the queue. The arrival pattern $\{A(n)\}_0^\infty$ of $\mathbb{N}^{K+1}$-valued RV's, with $A_k(n)$ denoting the number of arrivals to the k^{th} queue in the slot $[n, n+1)$, is independent of the initial queue size and of the service processes, and is modelled as a *renewal* process, in that the batch sizes of customers arriving into the system in each slot are independent and identically distributed from slot to slot. Under these assumptions, the MDP of interest is modelled by the $\mathbb{N}^{K+1}$-valued process $\{X(n)\}_0^\infty$, where $X_k(n)$, $0 \le k \le K$, represents the queue sizes of the k^{th} queue at the beginning of the slot $[n, n+1)$, $n = 0, 1, \ldots$.

Nain and Ross (1986) identified the service allocation policy that minimizes the long-run average of a linear expression in the queue sizes of the K queues $\{1, \ldots, K\}$ subject to the constraint that the long-run average queue size of the 0^{th} queue does not exceed a given value V. With the notation used here, they considered the constrained problem (CP_V) with cost functionals

$$J^c(\pi) := \overline{\lim}_n \frac{1}{n} E^\pi \sum_{t=0}^{n-1} \sum_{k=1}^{K} c_k X_k(t), \tag{4.9}$$

and

$$J^d(\pi) := \overline{\lim}_n \frac{1}{n} E^\pi \sum_{t=0}^{n-1} X_0(t), \tag{4.10}$$

where $c_k, 1 \le k \le K$, are non-negative weights.

A work conserving static priority assignment policy is a non-idling service allocation policy with fixed priority. With this notation, the results are given under the following assumptions (H6)-(H8), where

(H6) *The stability condition*

$$\rho := \sum_{k=0}^{K} \frac{\lambda_k}{\mu_k} < 1$$

holds, where $\lambda_k := E^\pi[A_k(n)]$ *for all* π *in* $\mathcal{P}$ *and every* $n = 0, 1, \ldots$,

(H7) *The set $\mathcal{P}_V$ contains at least one work conserving static priority assignment policy that gives the highest priority to the 0^{th} queue, but does not contain any work conserving static priority assignment policy which gives the lowest priority to the 0^{th} queue,*

(H8) *For some $r > 2$, the finite moment conditions*

$$\sum_{k=0}^{K} E^{\pi}|X_k(0)|^r < \infty \quad \text{and} \quad \sum_{k=0}^{K} E^{\pi}|A_k(n)|^r < \infty$$

hold for all π in $\mathcal{P}$ and every $n = 0, 1, \ldots$.

Under (H6)-(H7), the problem (CP_V) admits (Nain and Ross 1986) an optimal stationary policy f^* which is obtained by simple randomization between two work conserving static priority assignment policies $\overline{g}$ and $\underline{g}$, as determined by (4.2)-(4.5). Under (H8), Makowski and Shwartz (1986b) have shown that the RV's $\{X(n)\}_0^\infty$ are *uniformly integrable* under P^π for any *non-idling* policy π in $\mathcal{P}$. Moreover, for any non-idling stationary policy g, the Markov chain $\{X(n)\}_0^\infty$ forms a single ergodic class under P^g over the state space $\mathbb{N}^{K+1}$. These facts imply readily that under (H6)-(H8), Theorem 4.1 applies to the steering policy α defined by (2.2) and (2.4), where the state z is chosen arbitrarily in $\mathbb{N}^{K+1}$.

Theorem 4.4 *Under* (H6)-(H8), *the steering policy α solves the constrained optimal resource allocation problem (CP_V) with $J^c(\alpha) = J^c(f^*)$ and $J^d(\alpha) = J^d(f^*) = V$.*

For $K = 1$, the system is composed of two queues and the policy $\overline{g}$ (resp. $\underline{g}$) specializes to the work conserving static priority assignment policy giving higher priority to the 1^{st} queue (resp. the 0^{th} queue). In that case, the steering policy α constitutes an adaptive policy in the restrictive sense understood in the literature of adaptive control of Markov chains (Kumar and Varaiya 1986) in that no knowledge of the model parameters is needed for implementing the policy α.

5. A Proof of Theorem 3.5 by Sample Path Arguments

In this section, Theorem 3.5 is established through direct sample path arguments. The discussion is carried out through a series of technical lemmas.

5.1. Regenerative properties of $\{X(n)\}_0^\infty$

To study the performance of the policy α, start with the following observation: Under the recurrence assumption (A1), the process $\{X(n)\}_0^\infty$ is *regenerative* under each one of the measures $P^{\overline{g}}$ and $P^{\underline{g}}$ (Chung 1974), while it need not be so under P^α owing to its non-stationarity. It thus seems reasonable to try a *decomposition* of this non-stationary process into two *regenerative*

processes. This is done by connecting together the cycles corresponding to the use of each one of the policies so that results from the theory of regenerative processes may be applied. This idea is made precise in the lemma below and the arguments that follow it.

Let $\overline{t}(m)$ (resp. $\underline{t}(m)$) be the left boundary of the slot during which the policy $\overline{g}$ (resp. $\underline{g}$) is used for the m^{th} time so that $\eta(\overline{t}(m)) = 1$ (resp. $\eta(\underline{t}(m)) = 0$). Note that the RV's $\overline{t}(m)$ and $\underline{t}(m)$ are $I\!F_n$-stopping times, and the RV's $\overline{X}(m)$ and $\underline{X}(m)$ given by

$$\overline{X}(m) = X(\overline{t}(m)) \quad \text{and} \quad \underline{X}(m) = X(\underline{t}(m)) \qquad m = 1, 2, \ldots (5.1)$$

are thus $I\!F_{\overline{t}(m)}$ and $I\!F_{\underline{t}(m)}$-measurable, respectively.

Lemma 5.1 *Assume the recurrence condition* (A1) *to hold. Under* P^{α}, *the RV's* $\{\overline{X}(m)\}_1^{\infty}$ *(resp.* $\{\underline{X}(m)\}_1^{\infty}$*) form a time-homogeneous Markov chain with one-step transition probability matrix* $P(\overline{g})$ *(resp.* $P(\underline{g})$*).*

Proof. The result will be established for the sequence $\{\overline{X}(m)\}_1^{\infty}$, provided the equality

$$P^{\alpha}[X(\overline{t}(m+1)) = y | I\!F_{\overline{t}(m)}] = p_{X(\overline{t}(m))y}(\overline{g}) \qquad m = 1, 2 \ldots (5.2)$$

can be shown to hold. In fact, it suffices to show the set equality

$$[X(\overline{t}(m+1)) = y] = [X(\overline{t}(m) + 1) = y], \qquad m = 1, 2 \ldots (5.3)$$

since then

$$P^{\alpha}[X(\overline{t}(m+1)) = y | I\!F_{\overline{t}(m)}] = P^{\overline{g}}[X(\overline{t}(m) + 1) = y | I\!F_{\overline{t}(m)}] \qquad m = 1, 2 \ldots (5.4)$$

by the very definition of α and of the stopping time $\overline{t}(m)$, and the strong Markov property now readily yields (5.2) from (5.4).

The proof of (5.3) is now given and considers two cases. If $y \neq z$, then necessarily $\overline{t}(m+1) = \overline{t}(m) + 1$ and (5.3) is trivially true. If on the other hand $y = z$, the set equality (5.3) (with $y = z$) is seen to hold by the following observations: On the event $[X(\overline{t}(m+1)) = z]$, it is *not* possible that $X(\overline{t}(m) + 1) \neq z$, for this would imply $\overline{t}(m+1) = \overline{t}(m) + 1$ by the very definition of α, thus leading to the contradiction $X(\overline{t}(m+1)) \neq z$! Consequently,

$$[X(\overline{t}(m+1)) = z] \subseteq [X(\overline{t}(m) + 1) = z]. \qquad m = 1, 2 \ldots (5.5)$$

Conversely, on the event $[X(\overline{t}(m)+1) = z]$, the epoch $\overline{t}(m)+1$ corresponds to the end of a cycle and the next time the policy $\overline{g}$ is used necessarily marks the beginning of a cycle, so that $X(\overline{t}(m+1)) = z$ and

$$[X(\overline{t}(m)+1) = z] \subseteq [X(\overline{t}(m+1)) = z]. \qquad m = 1, 2 \ldots (5.6)$$

The result (5.3) is now obtained by combining (5.5) and (5.6).

These arguments apply *mutatis mutandis* to the sequence $\{\underline{X}(m)\}_1^\infty$. Details are left to the interested reader. □

5.2. The key convergence results

For any mapping $c : S \to \mathbb{R}$, in order to study the convergence of $\{J^c(\tau(k))\}_1^\infty$ as k, the number of cycles, goes to ∞, let $\overline{T}(l)$ (resp. $\underline{T}(l)$) denote the length of the l^{th} cycle during which the policy $\overline{g}$ (resp. $\underline{g}$) is used, and set

$$\overline{\tau}(l) := \sum_{s=1}^{l} \overline{T}(s) \quad \text{and} \quad \underline{\tau}(l) := \sum_{s=1}^{l} \underline{T}(s). \qquad l = 1, 2, \ldots (5.7)$$

In words, $\overline{\tau}(l)$ (resp. $\underline{\tau}(l)$) represents the total number of slots in the l first cycles during which $\overline{g}$ (resp. $\underline{g}$) is used. Moreover, let the RV's $\overline{Z}^c(l)$ and $\underline{Z}^c(l)$ defined by

$$\overline{Z}^c(l) := \sum_{m=\overline{\tau}(l-1)+1}^{\overline{\tau}(l)} c(\overline{X}(m)) \quad \text{and} \quad \underline{Z}^c(l) := \sum_{m=\underline{\tau}(l-1)+1}^{\underline{\tau}(l)} c(\underline{X}(m)) \qquad l = 1, 2, \ldots (5.8)$$

represent the total costs over the l^{th} cycle during which the policies $\overline{g}$ and $\underline{g}$ are used, respectively. In the definition of (5.8), it is convenient to set $\overline{Z}^c(l) = 0$ (resp. $\underline{Z}^c(l) = 0$) if $\overline{\tau}(l-1) = \infty$ (resp. $\underline{\tau}(l-1) = \infty$). Thus, under Lemma 3.4, the quantities $\overline{Z}^c(l)$ and $\underline{Z}^c(l)$ are P^α-a.s. well defined and finite for all $l = 1, 2, \ldots$.

The next lemma is an immediate consequence of Lemma 5.1.

Lemma 5.2 *Assume the recurrence condition* (A1) *to hold. For any mapping $c : S \to \mathbb{R}$, the RV's $\{\overline{Z}^c(l)\}_1^\infty$ (resp. $\{\underline{Z}^c(l)\}_1^\infty$) form a (possibly delayed) renewal sequence under P^α. Moreover, if the mapping c satisfies the condition* (A3bis), *then*

$$E^\alpha[\overline{Z}^c(l)] = E_z^{\overline{g}}[Z^c] \quad \text{and} \quad E^\alpha[\underline{Z}^c(l)] = E_z^{\underline{g}}[Z^c]. \qquad l = 2, 3, \ldots (5.9)$$

Let the RV's $\overline{\nu}(k)$ and $\underline{\nu}(k)$ count the total number of cycles in the first k cycles that $\overline{g}$ and $\underline{g}$ are used, respectively. It is now plain that

$$\sum_{t=0}^{\tau(k)-1} c(X(t)) = \sum_{l=1}^{\overline{\nu}(k)} \overline{Z}^c(l) + \sum_{l=1}^{\underline{\nu}(k)} \underline{Z}^c(l) \qquad P^\alpha - a.s. (5.10a)$$

for each $k = 1, 2, \ldots$, and with $c(x) = 1$ for all x in S, this last relation specializes to

$$\tau(k) = \sum_{l=1}^{\overline{\nu}(k)} \overline{T}(l) + \sum_{l=1}^{\underline{\nu}(k)} \underline{T}(l). \qquad P^\alpha - a.s.(5.10b)$$

By virtue of Lemma 3.4, $\lim_k \overline{\nu}(k) = \lim_k \underline{\nu}(k) = \infty$ so that the next lemma is now immediate from Lemma 5.2 and the Strong Law of Large Numbers.

Lemma 5.3 *Assume* (A1)-(A4) *to hold. For any mapping* $c : S \to \mathbb{R}$ *satisfying* (A3bis), *the convergences*

$$\lim_k \frac{1}{\overline{\nu}(k)} \sum_{l=1}^{\overline{\nu}(k)} \overline{Z}^c(l) = E_z^{\overline{g}}[Z^c] \quad \text{and} \quad \lim_k \frac{1}{\underline{\nu}(k)} \sum_{l=1}^{\underline{\nu}(k)} \underline{Z}^c(l) = E_z^{\underline{g}}[Z^c] \qquad P^\alpha - a.s.(5.11a)$$

take place, and in particular,

$$\lim_k \frac{1}{\overline{\nu}(k)} \sum_{l=1}^{\overline{\nu}(k)} \overline{T}(l) = E_z^{\overline{g}}[T] \quad \text{and} \quad \lim_k \frac{1}{\underline{\nu}(k)} \sum_{l=1}^{\underline{\nu}(k)} \underline{T}(l) = E_z^{\underline{g}}[T]. \qquad P^\alpha - a.s.(5.11b)$$

Set

$$q(k) := \frac{\overline{\nu}(k)}{k} \qquad k = 1, 2, \ldots (5.12)$$

and note that $\frac{\underline{\nu}(k)}{k} = 1 - q(k)$. The relations (5.10) imply that

$$\frac{\tau(k)}{k} = q(k) \frac{1}{\overline{\nu}(k)} \sum_{l=1}^{\overline{\nu}(k)} \overline{T}(l) + (1 - q(k)) \frac{1}{\underline{\nu}(k)} \sum_{l=1}^{\underline{\nu}(k)} \underline{T}(l), \qquad P^\alpha - a.s.(5.13a)$$

$$p(\tau(k)) = \frac{q(k) \frac{1}{\overline{\nu}(k)} \sum_{l=1}^{\overline{\nu}(k)} \overline{T}(l)}{q(k) \frac{1}{\overline{\nu}(k)} \sum_{l=1}^{\overline{\nu}(k)} \overline{T}(l) + (1 - q(k)) \frac{1}{\underline{\nu}(k)} \sum_{l=1}^{\underline{\nu}(k)} \underline{T}(l)} \qquad P^\alpha - a.s.(5.13b)$$

and

$$J^c(\tau(k)) = \frac{q(k) \frac{1}{\overline{\nu}(k)} \sum_{l=1}^{\overline{\nu}(k)} \overline{Z}^c(l) + (1 - q(k)) \frac{1}{\underline{\nu}(k)} \sum_{l=1}^{\underline{\nu}(k)} \underline{Z}^c(l)}{q(k) \frac{1}{\overline{\nu}(k)} \sum_{l=1}^{\overline{\nu}(k)} \overline{T}(l) + (1 - q(k)) \frac{1}{\underline{\nu}(k)} \sum_{l=1}^{\underline{\nu}(k)} \underline{T}(l)} \qquad P^\alpha - a.s.(5.13c)$$

for all $k = 1, 2, \ldots$, where the convention $\frac{0}{0} = 0$ is used.

For any mapping $c : S \to \mathbb{R}$ to satisfy (A3bis) with the quantity $E_z^{\overline{g}}[Z^c] + E_z^{\underline{g}}[Z^c]$ being well defined (but possibly infinite), it is now plain from Lemma 5.3 and (5.13) that under P^α,

the sequences of RV's $\{\frac{\tau(k)}{k}\}_1^\infty$, $\{p(\tau(k))\}_1^\infty$ and $\{J^c(\tau(k))\}_1^\infty$ converge a.s. if the RV's $\{q(k)\}_1^\infty$ converge a.s.. This key convergence result of the RV's $\{q(k)\}_1^\infty$ is taken on in the next proposition whose proof is delayed until the next section.

Theorem 5.4 *Under* (A1)-(A4), *the RV's* $\{q(k)\}_1^\infty$ *converge* P^α*-a.s. to the constant* q^* *given by (3.10), i.e.,*

$$\lim_k q(k) = q^*. \qquad P^\alpha - a.s. \quad (5.14)$$

With the help of Theorem 5.4, Theorem 3.5 can now be proved easily.

A proof of Theorem 3.5. From the the remarks made earlier, it follows readily that under (5.14), the P^α-a.s. limits of the sequences of RV's $\{\frac{\tau(k)}{k}\}_1^\infty$, $\{p(\tau(k))\}_1^\infty$ and $\{J^c(\tau(k))\}_1^\infty$ are necessarily given by

$$\lim_k \frac{\tau(k)}{k} = q^* E_z^{\overline{g}}[T] + (1-q^*)E_z^{\underline{g}}[T], \qquad P^\alpha - a.s. \quad (5.15a)$$

$$\lim_k p(\tau(k)) = \frac{q^* E_z^{\overline{g}}[T]}{q^* E_z^{\overline{g}}[T] + (1-q^*)E_z^{\underline{g}}[T]} \qquad P^\alpha - a.s. \quad (5.15b)$$

and

$$\lim_k J^c(\tau(k)) = \frac{q^* E_z^{\overline{g}}[Z^c] + (1-q^*)E_z^{\underline{g}}[Z^c]}{q^* E_z^{\overline{g}}[T] + (1-q^*)E_z^{\underline{g}}[T]}, \qquad P^\alpha - a.s. \quad (5.15c)$$

respectively. While (5.15a) gives (3.13), simple algebraic calculations based on (3.6) and (3.10) easily yield (3.11)-(3.12) from (5.15b)-(5.15c). The proof of Theorem 3.5 is therefore complete. □

5.3. A proof of Theorem 5.4

Crucial to the proof of Theorem 5.4 is the following *deterministic* lemma.

Lemma 5.5 *Let* $\{a(k)\}_1^\infty$, $\{\overline{b}(k)\}_1^\infty$ *and* $\{\underline{b}(k)\}_1^\infty$ *be* $\mathbb{R}$*-valued sequences satisfying the conditions*

$$\overline{b}(k) > 0 \quad \text{and} \quad \underline{b}(k) > 0 \qquad k = 1, 2, \ldots \quad (5.16a)$$

and

$$\lim_k \overline{b}(k) = 0, \quad \lim_k \underline{b}(k) = 0 \quad \text{and} \quad \lim_k a(k) = a \quad (5.16b)$$

for some a *in* $\mathbb{R}$. *If the* $\mathbb{R}$*-valued sequence* $\{\theta(k)\}_1^\infty$ *is defined recursively by*

$$\theta(k+1) = \begin{cases} \theta(k) - \overline{b}(k) & \text{if } \theta(k) > a(k); \\ \theta(k) + \underline{b}(k) & \text{if } \theta(k) \le a(k), \end{cases} \qquad k = 1, 2, \ldots \quad (5.17)$$

with $\theta(1)$ arbitrary in $\mathbb{R}$, then either $\{\theta(k)\}_1^\infty$ converges monotonically (in the tail) to some constant $\theta(\infty) \neq a$, or $\lim_k \theta(k) = a$.

Proof. By assumption, given $\epsilon > 0$, there exists a positive integer k_ϵ such that $\overline{b}(k) < \epsilon$, $\underline{b}(k) < \epsilon$ and $|a(k) - a| < \epsilon$ for all $k \geq k_\epsilon$, and define

$$m_\epsilon = \inf\{k \geq k_\epsilon : \theta(k) \in (a - \epsilon, a + \epsilon)\}. \tag{5.18}$$

If $m_\epsilon = \infty$, then $\theta(k)$ is *not* in the interval $(a - \epsilon, a + \epsilon)$ for all $k \geq k_\epsilon$. If $\theta(k_\epsilon) \leq a - \epsilon$, then $\theta(k) \leq a - \epsilon < a(k)$ for *all* $k \geq k_\epsilon$. To see this, recall that $\underline{b}(k) < \epsilon$ for all $k \geq k_\epsilon$, and from (5.17) this implies $\theta(k_\epsilon + 1) < a$, whence $\theta(k_\epsilon + 1) \leq a - \epsilon$ by the definition of m_ϵ. An induction argument now shows that $\theta(k) \leq a - \epsilon$ for all $k \geq k_\epsilon$, so that by (5.17) the sequence $\{\theta(k)\}_1^\infty$ is monotone increasing from time k_ϵ onward, and must converge to some value $\theta(\infty) \leq a - \epsilon$. The case $\theta(k_\epsilon) \geq a + \epsilon$ is similarly discussed.

Suppose $m_\epsilon < \infty$ so that $\theta(m_\epsilon)$ now lies in $(a - \epsilon, a + \epsilon)$. From (5.17) again, it follows that

$$a - \epsilon - \overline{b}(m_\epsilon) < \theta(m_\epsilon + 1) < a + \epsilon + \underline{b}(m_\epsilon). \tag{5.19}$$

If in (5.19), $a - \epsilon < \theta(m_\epsilon + 1) < a + \epsilon$, then the inequalities

$$a - \epsilon - \overline{b}(m_\epsilon + 1) < \theta(m_\epsilon + 2) < a + \epsilon + \underline{b}(m_\epsilon + 1) \tag{5.20a}$$

hold. On the other hand, if in (5.19), $\theta(m_\epsilon + 1) \notin (a - \epsilon, a + \epsilon)$, then two cases are possible: Either (i) $a - \epsilon - \overline{b}(m_\epsilon) < \theta(m_\epsilon + 1) \leq a - \epsilon$ in which case $\theta(m_\epsilon + 1) < a(m_\epsilon + 1)$ and therefore

$$a - \epsilon - \overline{b}(m_\epsilon) + \underline{b}(m_\epsilon + 1) < \theta(m_\epsilon + 2) < a - \epsilon + \underline{b}(m_\epsilon + 1), \tag{5.20b}$$

by making use of (5.17) or (ii) $a + \epsilon \leq \theta(m_\epsilon + 1) < a + \epsilon + \underline{b}(m_\epsilon)$ in which case $\theta(m_\epsilon + 1) > a(m_\epsilon + 1)$ and therefore

$$a + \epsilon - \overline{b}(m_\epsilon + 1) < \theta(m_\epsilon + 2) < a + \epsilon + \underline{b}(m_\epsilon) - \overline{b}(m_\epsilon + 1). \tag{5.20c}$$

It follows easily from (5.20) that

$$a - \epsilon - \max\{\overline{b}(m_\epsilon), \overline{b}(m_\epsilon + 1)\} < \theta(m_\epsilon + 2) < a + \epsilon + \max\{\underline{b}(m_\epsilon), \underline{b}(m_\epsilon + 1)\}.$$

An induction argument now implies that the inequalities

$$a - \epsilon - \max_{0 \leq i < l} \overline{b}(m_\epsilon + i) \leq \theta(m_\epsilon + l) \leq a + \epsilon + \max_{0 \leq i < l} \underline{b}(m_\epsilon + i) \tag{5.21}$$

holds for all $l = 1, 2, \ldots$. Since $m_\epsilon \geq k_\epsilon$, the definition of k_ϵ yields

$$a - 2\epsilon < \theta(k) < a + 2\epsilon$$

for all $k \geq m_\epsilon$, and ϵ being arbitrary, the proof is now complete. □

A proof of Theorem 5.4 is now presented.

A proof of Theorem 5.4. Define the RV's $\{Y(k)\}_1^\infty$ by

$$Y(k) := \frac{\frac{\sum_{l=1}^{\underline{\nu}(k)} \underline{T}(l)}{\underline{\nu}(k)} V - \frac{\sum_{l=1}^{\underline{\nu}(k)} \underline{Z}^d(l)}{\underline{\nu}(k)}}{\left(\frac{\sum_{l=1}^{\overline{\nu}(k)} \overline{Z}^d(l)}{\overline{\nu}(k)} - \frac{\sum_{l=1}^{\underline{\nu}(k)} \underline{Z}^d(l)}{\underline{\nu}(k)}\right) - \left(\frac{\sum_{l=1}^{\overline{\nu}(k)} \overline{T}(l)}{\overline{\nu}(k)} - \frac{\sum_{l=1}^{\underline{\nu}(k)} \underline{T}(l)}{\underline{\nu}(k)}\right) V}, \qquad k = 1, 2, \ldots (5.22)$$

and observe from (5.13c) (with d replacing c) that $J^d(\tau(k)) > V$ if and only if $q(k) > Y(k)$. The definition of α implies that the RV's $\{q(k)\}_1^\infty$ are defined recursively by

$$q(k+1) = \begin{cases} q(k) - \frac{1}{k+1} q(k) & \text{if} \quad q(k) > Y(k); \\ q(k) + \frac{1}{k+1}(1 - q(k)) & \text{if} \quad q(k) \leq Y(k). \end{cases} \qquad k = 1, 2, \ldots (5.23)$$

Under (A1)-(A4), it follows from Lemma 5.3 that

$$\lim_k Y(k) = \frac{E_z^{\underline{g}}[T]V - E_z^{\underline{g}}[Z^d]}{(E_z^{\overline{g}}[Z^d] - E_z^{\underline{g}}[Z^d]) - (E_z^{\overline{g}}[T] - E_z^{\underline{g}}[T])V} \qquad P^\alpha - a.s. (5.24)$$

so that

$$\lim_k Y(k) = \frac{p^* E_z^{\underline{g}}[T]}{(1 - p^*) E_z^{\overline{g}}[T] + p^* E_z^{\underline{g}}[T]} = q^* \qquad P^\alpha - a.s. (5.25)$$

by simple algebraic manipulations based on (A4), (2.7), (3.2) and (3.10).

Pick a sample ω not in the P^α-null set on which (5.25) fails and set $\theta(k) = q(k, \omega)$, $a(k) = Y(k, \omega)$, $\overline{b}(k) = \frac{q(k,\omega)}{k+1}$ and $\underline{b}(k) = \frac{(1 - q(k,\omega))}{k+1}$ for all $k = 1, 2, \ldots$, and note that $a = q^*$. Since $0 \leq q(k, \omega) \leq 1$, the assumptions of Lemma 5.5 are immediately satisfied, and the a.s. convergence of the RV's $\{q(k)\}_1^\infty$ follows. It is not possible for the values $\{q(k, \omega)\}_1^\infty$ to converge monotonically (in the tail) to some value not equal to q^*, for this would imply that the policy α sticks to one policy from some cycle onward, in clear contradiction with Lemma 3.4. □

Acknowledgement

The authors would like to thank Keith W. Ross for his comments on an earlier version of this paper.

REFERENCES

Altman, E. and A. Shwartz. 1986. Optimal priority assignment with general constraints. In *Proceedings of the 24th Allerton Conference on Communication, Control and Computing.* Allerton, Illinois, 1147-1148.

Beutler, F. J. and K. W. Ross. 1985. Optimal policies for controlled Markov chains with a constraint. *J. Math. Anal. Appl.*. 112, 236-252.

Chung, K. L. 1967. *Markov Chains with Stationary Transition Probabilities.* Second Edition, Springer-Verlag, New York.

Chung, K. L. 1974. *A Course in Probability Theory.* Second Edition, Academic Press, New York.

Kumar, P. R. and P. Varaiya. 1986. *Stochastic Systems; Estimation, Identification and Adaptive Control.* Prentice-Hall, New Jersey.

Ma, D.-J. 1988. A Simple Problem of Flow Control: Optimality and Adaptive Implementations. Ph.D. Thesis, Electrical Engineering Department, University of Maryland, College Park, Maryland.

Ma, D.-J. and A. M. Makowski. 1987. Optimality results for a simple flow control problem. In *Proceedings of the 26th IEEE Conference on Decision and Control.* Los Anges, California, 1852-1857.

Ma, D.-J., A. M. Makowski and A. Shwartz. 1986. Estimation and optimal control for constrained Markov chains. In *Proceedings of the 25th IEEE Conference on Decision and Control.* Athens, Greece, 994-999.

Makowski, A. M. 1987. How to randomize between two policies: The i.i.d. case. unpublished manuscript.

Makowski, A. M. and A. Shwartz. 1986a. Implementation issues for Markov decision processes. In *Proceedings of a Workshop on Stochastic Differential Systems.* Institute of Mathematics and its Applications, University of Minnesota. Eds. W. Fleming and P.-L. Lions, Springer-Verlag Lecture Notes in Control and Information Sciences.

Makowski, A. M. and A. Shwartz. 1986b. Recurrence properties of a system of competing queues with applications. *Adv. Appl. Prob.*. submitted.

Nain, P. and K. W. Ross. 1986. Optimal priority assignment with hard constraint. *IEEE Trans. Auto. Control.* 31, 883-888.

Ross, K. W. 1985. Constrained Markov Decision Processes with Queueing Applications. Ph.D. Thesis, CICE Program, University of Michigan, Ann Arbor, Michigan.

Ross, K. W. 1988. Randomized and past-dependent policies for Markov decision processes with multiple constraints. *Opns. Res.*. to appear.