To see the relation between Theorem 1 and Proposition 1, let $P_\alpha = \alpha^{-1}P$, where α is a positive number that satisfies (3) and P is the solution to the Lyapunov equation (4). Then, it follows from (2) that P_α is a solution to (5) for $\eta < \eta_{k_0}$. In other words, the Riccati inequality (5) is solvable if conditions (2)–(4) are satisfied.

Let $\hat{\eta}$ be the maximum value for which (5) is solvable, or equivalently, the linear matrix inequalities [1]

$$\begin{pmatrix} I & \eta P \\ \eta P & -I - A^T P - PA \end{pmatrix} > O \quad \text{and} \quad P > O \tag{8}$$

are solvable. As is well known [2], the maximum value $\hat{\eta}$ is obtained by checking whether or not the corresponding Hamiltonian matrix

$$H(\eta) = \begin{pmatrix} A & I \\ -\eta^2 I & -A^T \end{pmatrix} \tag{9}$$

has an eigenvalue on the imaginary axis.

Corollary 1: The system

$$\dot{x}(t) = Ax(t) + \sum_{i=1}^{k} E_i(t)x(t - h_i) \tag{10}$$

is asymptotically stable if $\|(E_1, \cdots, E_k)\| < \hat{\eta}/\sqrt{k}$ for all t.

Proof: Let

$$V(x(t)) = x^T(t)Px(t) + \sum_{i=1}^{k} \int_{t-h_i}^{t} x^T(\theta)x(\theta)\, d\theta. \tag{11}$$

Then

$$\begin{aligned} \dot{V}(x(t)) = {} & x^T(t)(kI + A^T P + PA + \eta^2 P^2)x(t) \\ & - x^T(t)P(\eta^2 I - \tilde{E}\tilde{E}^T)Px(t) \\ & - [\tilde{E}^T Px(t) - \tilde{x}]^T [\tilde{E}^T Px(t) - \tilde{x}] \end{aligned} \tag{12}$$

where $\tilde{E} = (E_1, \cdots, E_k)$ and $\tilde{x} = (x^T(t - h_1), \cdots, x^T(t - h_k))^T$. From this, we know that (10) with $\|\tilde{E}\| < \eta$ is asymptotically stable if

$$kI + A^T P + PA + \eta^2 P^2 < O \tag{13}$$

is solvable. Then, the result follows by noticing that the solvability of (13) is equivalent to that of

$$I + A^T P + PA + k\,\eta^2 P^2 < O. \tag{14}$$

□

References

[1] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*. Philadelphia, PA: SIAM, 1994.

[2] J. C. Doyle, K. Glover, P. P. Khargonekar, and B. A. Francis, "State-space solutions to standard H_2 and H_∞ control problems," *IEEE Trans. Automat. Contr.*, vol. 34, pp. 831–847, 1989.

Author's Reply

We thank Ooba and Funahashi for their interest in our article. Indeed our result needs the selection of a matrix Q and a scalar α which maximize the bound $\eta_{k0}(Q, \alpha)$. Ooba and Funahashi propose a result that does not need this selection, and their result is summarized in *Proposition 1*: the system (1) is stable if the Riccati inequality (5) is solvable and it is equivalent to the statement that the matrix $H(\eta)$ in (9) has no eigenvalue on the imaginary axis. Also, η is not expressed explicitly. However, the solvability of (5) can be more easily checked by the well-known Bounded Real Lemma [1], i.e., the solvability of (5) is equivalent to

$$\eta\|(sI_n - A)^{-1}\|_\infty < 1 \tag{A1}$$

where $\|G(s)\|_\infty = \sup_\omega \sigma_{\max}[G(j\omega)]$. Therefore, we may conclude that (1) is stable if $\eta > 0$ satisfies condition (A1), and we can obtain the explicit bound of η when we use (A1).

References

[1] M. Green and D. J. N. Limebeer, *Linear Robust Control*. Englewood Cliffs, NJ: Prentice-Hall, 1995, p. 109.

Publisher Item Identifier S 0018-9286(99)05446-X.

Steering Policies for Controlled Markov Chains Under a Recurrence Condition

Dye-Jyun Ma and Armand M. Makowski

***Abstract*— The authors consider the class of steering policies for controlled Markov chains under a recurrence condition. A steering policy is defined as one adaptively alternating between two stationary policies in order to track a sample average cost to a desired value. Convergence of the sample average costs is derived via direct sample path arguments, and the performance of the steering policy is discussed. Steering policies are motivated by, and particularly useful in, the discussion of constrained Markov chains with a single constraint.**

***Index Terms*—Adaptive control, implementation, Markov decision processes.**

I. Introduction

We introduce the class of steering policies as a means to adaptively control a class of Markov chains. This steering policy adaptively alternates between two stationary (possibly randomized) policies, say g_* and g^*, and the decision to switch policies is taken only when the system visits some distinguished state z. At those (random) instants,

Manuscript received April 25, 1997. Recommended by Associate Editor, J. B. Lasserre. The work of D.-J. Ma was supported in part under R.O.C. National Science Council Grants NSC83-0404-E005-004 and NSC84-2213-E-005-009. The work of A. M. Makowski was supported in part by the NSF under Grant NSFD CDR-88-03012.

D.-J. Ma is with the Department of Electrical Engineering, National Chung-Hsing University, Taichung, Taiwan, R.O.C.

A. M. Makowski is with the Electrical Engineering Department and the Institute for Systems Research, University of Maryland, College Park, MD 20742 USA.

Publisher Item Identifier S 0018-9286(99)05444-6.

the current value of a sample average cost is compared against a target value, say V. If the sample average is above (respectively, below) the value V, the policy g_* (respectively, g^*) will be used until the next visit to the state z.

Steering policies find their motivation in the control of Markov chains with a single constraint. In that context, Lagrangian arguments can often be invoked to construct an optimal stationary policy by simple randomization between two (pure) stationary policies g^* and g_* with g^* (respectively, g_*) overshooting (respectively, undershooting) the constraint value; the randomization bias is selected so that the constraint value is met exactly [2], [5], [6], [9]–[11]. The steering policy represents an alternative *implementation* to this optimal policy in situations where the randomization bias is neither available nor easily computable. In that context, the optimality of the steering policy thus reduces to whether it steers the sample average cost to the target value V.

When the distinguished state z is recurrent under both policies g^* and g_*, we show that the sample average cost is steered (in an a.s. sense) to the desired value V. The analysis relies on sample path arguments and takes advantage of hidden regenerative properties induced by the steering policy. The discussion is in the spirit of the proof of the ergodic theorem for recurrent Markov chains based on the strong law of large numbers [3] and is given under minimal conditions, namely the integrability of the cost over a return cycle under both policies g_* and g^*.

A preliminary version of the work was reported in [7], where the one-step cost function was assumed to depend only on the state. Here, we extend the result to the general case in which the cost function depends on *both* state and control variables. In [8], we gave another proof for the convergence of the sample average cost to the desired value V in the framework of [7]. This was done by noting that the sample average cost, when evaluated at the return times to z, can be related to the output of a two-dimensional stochastic approximation scheme of the Robbins–Monro type. This stochastic approximation was shown to converge a.s. under some L^2-type conditions, namely square-integrability of the return cycle to z and of the total accumulated cost over a return cycle under both policies g_* and g^*. The convergence results of [8] are obtained under assumptions stronger than the ones used here.

The steering policy should be contrasted against the so-called *time-sharing* implementation of [1], whereby the decision-maker alternates between the two policies g^* and g_* according to some deterministic (thus nonadaptive) mechanism associated with the recurrence cycles. As the instrumentation of time-sharing policies requires the explicit evaluation of certain cost functionals, we can interpret the steering policy as an adaptive version of time sharing.

The steering policy is adapted from an idea originally proposed by Nain and Ross [10] in the context of an optimal resource allocation problem with a constraint. In a subsequent paper [11], Ross conjectured the optimality of a version of the steering policy in the case of finite-state Markov chains with a single constraint. The steering policy proposed by Ross in [11] differs from the one here in that the decision to switch policies is taken at every time epoch. However, in the context of finite-state Markov chains discussed in [11], the two policies g^* and g_* coincide in all but one state [2], and Ross' version simply reduces to the steering policy considered here with the distinguished state z chosen as the single state where g^* and g_* differ. Thus, the results given here imply Ross's conjecture while providing an alternative adaptive solution to the original resource allocation problem [10].

A word on the notation: The indicator function of any set E is simply denoted by $\mathbf{1}[E]$, and $\lim_n$ is understood with n going to infinity.

II. The Model

Consider an MDP (S, U, P) as defined in the literature [12] with countable state space S, measurable action space $(U, \mathcal{B}(U))$, and Borel measurable transition kernel $P \equiv (p_{xy}(u))$, i.e., for all x and y in S, the mappings $u \to p_{xy}(u)$, are Borel measurable on U, and satisfy the standard properties $0 \leq p_{xy}(u) \leq 1$ and $\Sigma_y\, p_{xy}(u) = 1$ for all u in U. The state process $\{X(n), n = 0, 1, \cdots\}$ and the control process $\{U(n), n = 0, 1, \cdots\}$ are defined on some measurable space $(\Omega, \mathcal{F})$. The feedback information available to the decision-maker is encoded through the rvs $\{H(n), n = 0, 1, \cdots\}$ which are defined recursively by $H(0) \equiv X(0)$ and $H(n+1) \equiv (H(n), U(n), X(n+1))$. For each $n = 0, 1, \cdots$, the rvs $X(n)$, $U(n)$, and $H(n)$ take values in S, U, and $S \times (U \times S)^n$, respectively; we also introduce the information σ-fields $\mathcal{F}_n \equiv \sigma\{H(n)\}$ and $\mathcal{G}_n \equiv \sigma\{H(n), U(n)\} = \mathcal{F}_n \vee \sigma\{U(n)\}$.

The space of probability measures on $(U, \mathcal{B}(U))$ is denoted by $\mathcal{M}(U)$. An *admissible* control policy π is any collection $\{\pi_n, n = 0, 1, \cdots\}$ of mappings π_n: $S \times (U \times S)^n \to \mathcal{M}(U)$ such that for all $n = 0, 1, \cdots$ and every Borel subset B of U, the mapping $S \times (U \times S)^n \to [0, 1]$: $h_n \to \pi_n(h_n; B)$ is Borel measurable. The collection of all such admissible policies is denoted by $\mathcal{P}$.

Let μ be a probability measure on S. The definition of the MDP (S, U, P) then postulates the existence of a collection of probability measures $\{\boldsymbol{P}^\pi, \pi \in \mathcal{P}\}$ on $(\Omega, \mathcal{F})$ such that (1), (2) below are satisfied. For every admissible policy π in $\mathcal{P}$, the probability measure $\boldsymbol{P}^\pi$ is constructed so that under $\boldsymbol{P}^\pi$, the rv X_0 has probability distribution μ, the control actions are selected according to

$$\boldsymbol{P}^\pi[U(n) \in B | \mathcal{F}_n] = \pi_n(H(n); B), \qquad B \in \mathcal{B}(U) \tag{1}$$

for all $n = 0, 1, \cdots$, and the state transitions are realized according to

$$\boldsymbol{P}^\pi[X(n+1) = y | \mathcal{G}_n] = p_{X(n)y}(U(n)), \quad y \in S \tag{2}$$

for all $n = 0, 1, \cdots$. The expectation operator associated with $\boldsymbol{P}^\pi$ is denoted by $\boldsymbol{E}^\pi$. When μ is the point mass distribution at x in S, this notation is specialized to $\boldsymbol{P}^\pi_x$ and $\boldsymbol{E}^\pi_x$, respectively.

Following standard usage, a policy π in $\mathcal{P}$ is said to be a Markov stationary policy if there exists a mapping g: $S \to \mathcal{M}(U)$ such that for each B in $\mathcal{B}(U)$, we have $\pi_n(H(n); B) = g(X(n); B)$ $\boldsymbol{P}^\pi$ a.s. for all $n = 0, 1, \cdots$, in which case the policy is identified with the mapping g itself. For each Markov stationary policy g, the rvs $\{X(n), n = 0, 1, \cdots\}$ form a time-homogeneous Markov chain under $\boldsymbol{P}^g$.

Any Borel mapping c: $S \times U \to \mathbb{R}$ is interpreted as a one-step cost function, and the corresponding sample cost averages $\{J_c(n), n = 1, 2, \cdots\}$ are defined by

$$J_c(n) \equiv \frac{1}{n} \sum_{t=0}^{n-1} c(X(t), U(t)), \qquad n = 1, 2, \cdots. \tag{3}$$

The following assumptions, A1)–A4), are enforced throughout the discussion.

A1) There exist two (possibly randomized) stationary policies g^* and g_* such that the Markov chain $\{X(n), n = 0, 1, \cdots\}$ has a single recurrent class under each one of the policies g^* and g_*. These recurrent classes have a nonempty intersection, and 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 g^* and g_*. By A1), such a state z exists and has the property that the system returns to it infinitely often under each policy. The first return time to the state z is the rv T defined by $T \equiv \inf\{n \geq 1\colon X(n) = z\}$. From now on, in statements, relations, and definitions which hold for *both* policies g^* and g_*, we use the compact notation g for either policy.

A2) The mean recurrence time to the state z is finite under $\boldsymbol{P}^g$, i.e., $\boldsymbol{E}_z^g[T] < \infty$.

By A2), the state z is positive recurrent under $\boldsymbol{P}^g$. Define a *cycle* as the time period that elapses between two consecutive visits of the process $\{X(n), n = 0, 1, \cdots\}$ to the recurrent state z. With some *prespecified* Borel cost function v: $S \times U \to \mathbb{R}$, we associate the cost per cycle $Z_v \equiv \Sigma_{t=0}^{T-1} v(X(t), U(t))$.

A3) The expected cost per cycle is finite under $\boldsymbol{P}^g$, i.e., $\boldsymbol{E}_z^g[|Z_v|] < \infty$.

Under A1)–A3), standard renewal arguments [3] already imply

$$\lim_n J_v(n) = \frac{\boldsymbol{E}_z^g[Z_v]}{\boldsymbol{E}_z^g[T]} \equiv I_v(g) \qquad \boldsymbol{P}^g\text{-a.s.} \tag{4}$$

A4) There exists a scalar V such that $I_v(g_*) < V < I_v(g^*)$.

Hence, under A4) the policy g^* (respectively, g_*) overshoots (respectively, undershoots) the value V which represents a desired performance level. We are interested in designing an admissible policy which steers the sample average cost $J_v(n)$ to V and which requires no additional statistical knowledge about the system other than that needed for implementing the policies g_* and g^*. One candidate solution is provided by the steering policy introduced in the next section.

III. The Steering Policy

The steering policy $\alpha = \{\alpha_n, n = 0, 1, \cdots\}$ is of the form

$$\alpha_n(H(n); \cdot) \equiv \eta(n) g^*(X(n); \cdot) + (1 - \eta(n)) g_*(X(n); \cdot)$$

for $n = 0, 1, \cdots$, where the $\{0, 1\}$-valued rv $\eta(n)$ specifies which of the two policies g^* and g_* is used in the time slot $[n, n+1)$. These rvs $\{\eta(n), n = 0, 1, \cdots\}$ are generated through the recursion

$$\eta(n) = \mathbf{1}[X(n) = z] \mathbf{1}[J_v(n) \le V] + \mathbf{1}[X(n) \ne z] \eta(n-1)$$

for all $n = 1, 2, \cdots$, with $\eta(0)$ an arbitrary $\{0, 1\}$-valued rv. During each cycle the steering policy α operates according to one of the policies g^* and g_*.

Set

$$p^* \equiv \frac{V - I_v(g_*)}{I_v(g^*) - I_v(g_*)} \tag{5}$$

and observe from A4) that $0 < p^* < 1$. For each $n = 1, 2, \cdots$, the rv

$$p(n) \equiv \frac{1}{n} \sum_{t=0}^{n-1} \eta(t) \tag{6}$$

represents the fraction of time over $[0, n)$ during which the policy g^* is used. The main properties of the steering policy α are now stated. All a.s. convergence statements are taken under $\boldsymbol{P}^\alpha$ unless otherwise specified.

Theorem 1: Under A1)–A4), we have

$$\lim_n p(n) = p^* \quad \text{a.s.} \tag{7}$$

and

$$\lim_n J_v(n) = p^* I_v(g^*) + (1 - p^*) I_v(g_*) = V \quad \text{a.s.} \tag{8}$$

Moreover, for any Borel mapping c: $S \times U \to \mathbb{R}$ such that $\boldsymbol{E}_z^g[|Z_c|] < \infty$, we get

$$\lim_n J_c(n) = p^* I_c(g^*) + (1 - p^*) I_c(g_*) \quad \text{a.s.} \tag{9}$$

where $Z_c \equiv \Sigma_{t=0}^{T-1} c(X(t), U(t))$ and $I_c(g) \equiv (\boldsymbol{E}_z^g[Z_c] / \boldsymbol{E}_z^g[T])$.

By (8) the steering policy indeed steers the sample average cost $J_v(n)$ to the desired value V. As will be apparent from the derivation of Theorem 1, the convergence (9) for general cost mappings is a direct result of (7). This proof is relegated to subsequent sections and proceeds by first establishing the convergence results (7)–(9) along the sequence of times at which state z is visited.

IV. Convergence Along Recurrence Times

Under the steering policy α, the decisions for switching between policies are taken only at the times the state process visits state z. This suggests that the behavior of this control algorithm might be fully determined by properties of the sample average cost sequence taken only at these recurrence epochs.

Consider the distinguished state z entering the definition of the policy α, and recursively define the sequence of $\mathbb{N} \cup \{\infty\}$-valued recurrence times $\{\tau(k), k = 0, 1, \cdots\}$ by $\tau(0) \equiv 0$ and

$$\tau(k+1) \equiv \inf\{t > \tau(k) \colon X(t) = z\} \tag{10}$$

for $k = 0, 1, \cdots$, with the usual convention that $\tau(k+1) = \infty$ if the set $\{t > \tau(k) \colon X(t) = z\}$ is empty. The interval $[\tau(k), \tau(k+1))$ is simply the $(k+1)^{rst}$ 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 1: Assume the recurrence condition A1) to hold. For all $k = 1, 2, \cdots$, the rvs $\tau(k)$ is $\boldsymbol{P}^\alpha$ a.s. finite, so that the state process $\{X(n), n = 0, 1, \cdots\}$ visits the state z infinitely often under $\boldsymbol{P}^\alpha$. Under the additional assumptions A2)–A4), the steering policy α alternates infinitely often between the two policies g^* and g_*.

Under the recurrence assumption A1), the process $\{(X(n), U(n)), n = 0, 1, \cdots\}$ is a *regenerative* process with regeneration epochs $\{\tau(k), k = 1, 2, \cdots\}$ under each one of the measures $\boldsymbol{P}^{g^*}$ and $\boldsymbol{P}^{g_*}$ [4, p. 298], while this may not be the case under $\boldsymbol{P}^\alpha$ owing to the nonstationarity of α. It thus seems reasonable to try decomposing this nonstationary process into two regenerative ones by connecting together the cycles associated with the use of each one of the policies. This idea is made precise in Lemma 2 below; its proof is omitted in the interest of brevity.

Fix $m = 1, 2, \cdots$. Let $t^*(m)$ [respectively, $t_*(m)$] be the left boundary of the slot during which the policy g^* (respectively, g_*) is used for the mth time so that $\eta(t^*(m)) = 1$ (respectively, $\eta(t_*(m)) = 0$). The rv $t^*(m)$ [respectively, $t_*(m)$] being a $\mathcal{F}_n$-stopping time, the rv $X_*(m) \equiv X(t^*(m))$ [respectively, $X_*(m) \equiv X(t_*(m))$] is $\mathcal{F}_{t^*(m)}$-measurable (respectively, $\mathcal{F}_{t_*(m)}$-measurable). Similarly, the rv $t^*(m)$ [respectively, $t_*(m)$] being also a $\mathcal{G}_n$-stopping time, the rv $U^*(m) \equiv U(t^*(m))$ [respectively, $U_*(m) \equiv U(t_*(m))$] is $\mathcal{G}_{t^*(m)}$-measurable (respectively, $\mathcal{G}_{t_*(m)}$-measurable). If $X^*(m) = X(t^*(m)) = z$, then $t^*(m)$ marks the beginning of a cycle, and the policy g^* is used throughout that cycle by definition of α. For each $\ell = 1, 2, \cdots$, let $T^*(\ell)$ [respectively, $T_*(\ell)$] denote the length of the ℓth cycle during which the policy g^* (respectively, g_*) is used, and set $\tau^*(\ell) \equiv \Sigma_{s=1}^{\ell} T^*(s)$ [respectively, $\tau_*(\ell) \equiv \Sigma_{s=1}^{\ell} T_*(s)$]. The rv $\tau^*(\ell)$ [respectively, $\tau_*(\ell)$] represents the total number of slots in the ℓ first cycles during which g^* (respectively, g_*) is used. With this notation we have the following.

Lemma 2: Assume A1) holds. Under $\boldsymbol{P}^\alpha$, the rvs $\{(X^*(m), U^*(m)), m = 1, 2, \cdots\}$ [respectively, $\{(X_*(m), U_*(m)), m = 1, 2, \cdots\}$] form a regenerative process with regeneration epochs at successive visits to the set $\{z\} \times U$.

For an arbitrary one-step cost function c: $S \times U \to \mathbb{R}$, we now study the convergence of $J_c(\tau(k))$ as k, the number of cycles, goes to ∞. For each $\ell = 1, 2, \cdots$, the rvs $Z_c^*(\ell)$ and $Z_{c*}(\ell)$ defined by

$$Z_c^*(\ell) \equiv \sum_{m=\tau^*(\ell-1)+1}^{\tau^*(\ell)} c(X^*(m), U^*(m))$$

and

$$Z_{c*}(\ell) \equiv \sum_{m=\tau_*(\ell-1)+1}^{\tau_*(\ell)} c(X_*(m), U_*(m))$$

represent the total costs over the ℓth cycle during which the policies g^* and g_* are used, respectively.

For each $k = 1, 2, \cdots$, if the rv $\nu^*(k)$ [respectively, $\nu_*(k)$] counts the total number of cycles in the first k cycles during which g^* (respectively, g_*) is used, then we find

$$\sum_{t=0}^{\tau(k)-1} c(X(t), U(t)) = \sum_{\ell=1}^{\nu^*(k)} Z_c^*(\ell) + \sum_{\ell=1}^{\nu_*(k)} Z_{c*}(\ell). \tag{11}$$

With $c \equiv 1$, this last relation specializes to

$$\tau(k) = \sum_{\ell=1}^{\nu^*(k)} T^*(\ell) + \sum_{\ell=1}^{\nu_*(k)} T_*(\ell). \tag{12}$$

For each $k = 1, 2, \cdots$, we also find it convenient to introduce the averaged quantities

$$\mathcal{A}Z_c^*(k) \equiv \frac{1}{\nu^*(k)} \sum_{\ell=1}^{\nu^*(k)} Z_c^*(\ell) \tag{13}$$

and

$$\mathcal{A}Z_{c*}(k) \equiv \frac{1}{\nu_*(k)} \sum_{\ell=1}^{\nu_*(k)} Z_{c*}(\ell) \tag{14}$$

as well as

$$\mathcal{A}T^*(k) \equiv \frac{1}{\nu^*(k)} \sum_{\ell=1}^{\nu^*(k)} T^*(\ell) \tag{15}$$

and

$$\mathcal{A}T_*(k) \equiv \frac{1}{\nu_*(k)} \sum_{\ell=1}^{\nu_*(k)} T_*(\ell). \tag{16}$$

It is plain from Lemma 2 that the rvs $\{Z_c^*(\ell), \ell = 1, 2, \cdots\}$ [respectively, $\{Z_{c*}(\ell), \ell = 1, 2, \cdots\}$] form a (possibly delayed) renewal sequence under $\boldsymbol{P}^\alpha$, with $\boldsymbol{E}^\alpha[Z_c^*(\ell)] = \boldsymbol{E}_z^{g^*}[Z_c]$ and $\boldsymbol{E}^\alpha[Z_{c*}(\ell)] = \boldsymbol{E}_z^{g_*}[Z_c]$ for each $\ell = 2, 3, \cdots$. Keeping in mind that a.s. convergence statements are taken under $\boldsymbol{P}^\alpha$, we have by Lemma 1 that $\lim_k \nu^*(k) = \lim_k \nu_*(k) = \infty$ while by Lemma 2 the strong law of large numbers under $\boldsymbol{P}^\alpha$ yields

$$\lim\nolimits_k \mathcal{A}Z_c^*(k) = \boldsymbol{E}_z^{g^*}[Z_c] \quad \text{a.s.} \tag{17}$$

and

$$\lim\nolimits_k \mathcal{A}Z_{c*}(k) = \boldsymbol{E}_z^{g_*}[Z_c] \quad \text{a.s.} \tag{18}$$

as well as

$$\lim\nolimits_k \mathcal{A}T^*(k) = \boldsymbol{E}_z^{g^*}[T] \quad \text{a.s.} \tag{19}$$

and

$$\lim\nolimits_k \mathcal{A}T_*(k) = \boldsymbol{E}_z^{g_*}[T] \quad \text{a.s.} \tag{20}$$

Fix $k = 1, 2, \cdots$. Set $q(k) \equiv (\nu^*(k)/k)$ and note $(\nu_*(k)/k) = 1 - q(k)$. From (11), (12), we have

$$\frac{\tau(k)}{k} = q(k)\mathcal{A}T^*(k) + (1 - q(k))\mathcal{A}T_*(k)$$
$$p(\tau(k)) = \frac{q(k)\mathcal{A}T^*(k)}{q(k)\mathcal{A}T^*(k) + (1 - q(k))\mathcal{A}T_*(k)} \tag{21}$$

and

$$J_c(\tau(k)) = \frac{q(k)\mathcal{A}Z_c^*(k) + (1 - q(k))\mathcal{A}Z_{c*}(k)}{q(k)\mathcal{A}T^*(k) + (1 - q(k))\mathcal{A}T_*(k)} \tag{22}$$

with the convention $(0/0) = 0$. Letting k go to infinity in these expressions, we see from (17)–(20) that under $\boldsymbol{P}^\alpha$ the sequences $\{(\tau(k)/k), k = 1, 2, \cdots\}$, $\{p(\tau(k)), k = 1, 2, \cdots\}$, and $\{J_c(\tau(k)), k = 1, 2, \cdots\}$ converge a.s. as soon as $\{q(k), k = 1, 2, \cdots\}$ does.

Theorem 2: Under A1)–A4), we have $\lim_k q(k) = q^*$ a.s. where

$$q^* \equiv \frac{p^* \boldsymbol{E}_z^{g_*}[T]}{(1 - p^*)\boldsymbol{E}_z^{g^*}[T] + p^* \boldsymbol{E}_z^{g_*}[T]}. \tag{23}$$

This key convergence result is proved in the next section and leads to the following convergence results along recurrence epochs by the calculations outlined earlier.

Theorem 3: Under A1)–A4), we have

$$\lim\nolimits_k p(\tau(k)) = p^* \quad \text{a.s.} \tag{24}$$

and for any mapping c: $S \times U \to \mathbb{R}$ such that $\boldsymbol{E}_z^g[|Z_c|] < \infty$, it holds that

$$\lim\nolimits_k J_c(\tau(k)) = p^* I_c(g^*) + (1 - p^*) I_c(g_*) \quad \text{a.s.} \tag{25}$$

Moreover, the law of large numbers holds in the form

$$\lim\nolimits_k \frac{\tau(k)}{k} = q^* \boldsymbol{E}_z^{g^*}[T] + (1 - q^*)\boldsymbol{E}_z^{g_*}[T] \quad \text{a.s.} \tag{26}$$

Applying (25) to the prespecified cost mapping v, we get $\lim_k J_v(\tau(k)) = V$ from (4) and (5).

V. A Proof of Theorem 2

Crucial to the proof of Theorem 2 is the following deterministic lemma.

Lemma 3: Let $\{a(k), k = 1, 2, \cdots\}$, $\{b^*(k), k = 1, 2, \cdots\}$ and $\{b_*(k), k = 1, 2, \cdots\}$ be $\mathbb{R}$-valued sequences such that $b^*(k) > 0$ and $b_*(k) > 0$ for $k = 1, 2, \cdots$, and $\lim_k b^*(k) = \lim_k b_*(k) = 0$ and $\lim_k a(k) = a$ for some a in $\mathbb{R}$. If the $\mathbb{R}$-valued sequence $\{\theta(k), k = 1, 2, \cdots\}$ is defined recursively by

$$\theta(k+1) = \begin{cases} \theta(k) - b^*(k), & \text{if } \theta(k) > a(k) \\ \theta(k) + b_*(k), & \text{if } \theta(k) \le a(k) \end{cases} \tag{27}$$

for $k = 1, 2, \cdots$, with $\theta(1)$ arbitrary in $\mathbb{R}$, then either $\{\theta(k), k = 1, 2, \cdots\}$ converges monotonically (in the tail) to some constant $\theta(\infty) \ne a$, or $\lim_k \theta(k) = a$.

Proof: By assumption, given $\varepsilon > 0$, there exists a positive integer k_ε such that $b^*(k) < \varepsilon$, $b_*(k) < \varepsilon$, and $|a(k) - a| < \varepsilon$ for all $k \ge k_\varepsilon$, and define $m_\varepsilon = \inf\{k \ge k_\varepsilon\text{: } \theta(k) \in (a - \varepsilon, a + \varepsilon)\}$. If $m_\varepsilon = \infty$, then $\theta(k)$ is not in the interval $(a - \varepsilon, a + \varepsilon)$ for *all* $k \ge k_\varepsilon$. If $\theta(k_\varepsilon) \le a - \varepsilon$, an easy induction argument based on (27) shows that for all $k \ge k_\varepsilon$, we have $\theta(k) \le a$, thus $\theta(k) \le a - \varepsilon$. Hence, the sequence $\{\theta(k), k = 1, 2, \cdots\}$ is monotone increasing from time k_ε onward and thus must converge to some value $\theta(\infty) \leq a - \varepsilon$. The case $\theta(k_\varepsilon) \geq a + \varepsilon$ is handled similarly.

Suppose $m_\varepsilon < \infty$ so that $\theta(m_\varepsilon)$ now lies in $(a - \varepsilon, a + \varepsilon)$. A worst case argument based on (27) then gives $a - \varepsilon - b^*(m_\varepsilon) < \theta(m_\varepsilon + 1) < a + \varepsilon + b_*(m_\varepsilon)$. If $a - \varepsilon < \theta(m_\varepsilon + 1) < a + \varepsilon$, then the same worst case argument also yields

$$a - \varepsilon - b^*(m_\varepsilon + 1) < \theta(m_\varepsilon + 2) < a + \varepsilon + b_*(m_\varepsilon + 1).$$

If $\theta(m_\varepsilon + 1)$ does not belong to $(a - \varepsilon, a + \varepsilon)$, then two cases are possible: either 1) $a - \varepsilon - b^*(m_\varepsilon) < \theta(m_\varepsilon + 1) \leq a - \varepsilon$, in which case $\theta(m_\varepsilon + 1) < a(m_\varepsilon + 1)$, thus $\theta(m_\varepsilon + 2) = \theta(m_\varepsilon + 1) + b_*(m_\varepsilon + 1)$ by (27), and we get

$$a - \varepsilon - b^*(m_\varepsilon) + b_*(m_\varepsilon + 1) < \theta(m_\varepsilon + 2) < a - \varepsilon + b_*(m_\varepsilon + 1)$$

or 2) $a + \varepsilon \leq \theta(m_\varepsilon + 1) < a + \varepsilon + b_*(m_\varepsilon)$, in which case $\theta(m_\varepsilon + 1) > a(m_\varepsilon + 1)$, thus $\theta(m_\varepsilon + 2) = \theta(m_\varepsilon + 1) - b^*(m_\varepsilon + 1)$ by (27), and we get

$$a + \varepsilon - b^*(m_\varepsilon + 1) < \theta(m_\varepsilon + 2) < a + \varepsilon + b_*(m_\varepsilon) - b^*(m_\varepsilon + 1).$$

Collecting these inequalities we conclude $a - \varepsilon - \max\{b^*(m_\varepsilon), b^*(m_\varepsilon + 1)\} < \theta(m_\varepsilon + 2) < a + \varepsilon + \max\{b_*(m_\varepsilon), b_*(m_\varepsilon + 1)\}$. An induction argument now implies

$$a - \varepsilon - \max_{0 \leq i < \ell} b^*(m_\varepsilon + i) \leq \theta(m_\varepsilon + \ell) \leq a + \varepsilon + \max_{0 \leq i < \ell} b_*(m_\varepsilon + i)$$

for all $\ell = 1, 2, \cdots$. Because $m_\varepsilon \geq k_\varepsilon$, the definition of k_ε yields $a - 2\varepsilon < \theta(k) < a + 2\varepsilon$ for all $k \geq m_\varepsilon$, and ε being arbitrary, the proof is now complete. ■

A proof of Theorem 2 is now feasible: as we note from the definition of α that

$$\nu^*(k+1) = \nu^*(k) + \mathbf{1}[J_v(\tau(k)) \leq V], \qquad k = 0, 1, \cdots \tag{28}$$

it is plain that the rvs $\{q(k), k = 1, 2 \cdots\}$ can be defined recursively by

$$q(k+1) = \begin{cases} q(k) - \dfrac{1}{k+1} q(k), & \text{if } J_v(\tau(k)) > V \\ q(k) + \dfrac{1}{k+1}(1 - q(k)), & \text{if } V \geq J_v(\tau(k)) \end{cases}$$

for $k = 1, 2, \cdots$. For each $k = 1, 2, \cdots$, we set

$$Y(k) \equiv \frac{\mathcal{A}T_*(k)V - \mathcal{A}Z_{c*}(k)}{(\mathcal{A}Z_c^*(k) - \mathcal{A}Z_{c*}(k)) - (\mathcal{A}T^*(k) - \mathcal{A}T_*(k))V}.$$

Invoking the convergence (17)–(20), we get

$$\lim_k Y(k) = \frac{\boldsymbol{E}_z^{g_*}[T]V - \boldsymbol{E}_z^{g_*}[Z_v]}{(\boldsymbol{E}_z^{g^*}[Z_v] - \boldsymbol{E}_z^{g_*}[Z_v]) - (\boldsymbol{E}_z^{g^*}[T] - \boldsymbol{E}_z^{g_*}[T])V}$$

so that $\lim_k Y(k) = q^*$ a.s. by simple algebra based on (4), (5), and (23).

Pick a sample ω in the set of $\boldsymbol{P}^\alpha$-measure ones where (17)–(20) *simultaneously* hold, thus $\lim_k Y(k, \omega) = q^*$ as well. Under condition A4), the denominator of $Y(k, \omega)$ is positive for large k. It is now plain from (22) (with $c = v$) that for large k, say $k \geq k_*(\omega)$, the condition $J_v(\tau(k, \omega), \omega) > V$ holds if and only if $q(k, \omega) > Y(k, \omega)$, and the sequence $\{q(k + k_*(\omega), \omega), k = 1, 2 \cdots\}$ indeed satisfies the recursion (27) with $\theta(k) = q(k + k_*(\omega), \omega)$, $a(k) = Y(k + k_*(\omega), \omega)$, $b^*(k) = (q(k + k_*(\omega), \omega)/k + k_*(\omega) + 1)$ and $b_*(k) = ((1 - q(k + k_*(\omega), \omega))/k + k_*(\omega) + 1)$ for all $k = 1, 2, \cdots$. Because $0 \leq q(k + k_*(\omega), \omega) \leq 1$, the assumptions of Lemma 3 are immediately satisfied with $a = q^*$, and the sequence $\{q(k + k_*(\omega), \omega), k = 1, 2, \cdots\}$ does converge. It is not possible for the values $\{q(k + k_*(\omega), \omega), k = 1, 2, \cdots\}$ 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 1. Hence, $\lim_k q(k, \omega) = q^*$.

VI. Performance of the Steering Policy

Theorem 1 is easily recovered from Theorem 3. For each $n = 1, 2, \cdots$, let $k(n) \equiv \max\{k \geq 0\colon\ \tau(k) \leq n\}$ be the number of cycles over the horizon $[0, n)$ including the one in progress at time n, and note that $\tau(k(n)) \leq n < \tau(k(n) + 1)$ so that

$$\frac{\tau(k(n))}{n} J_c(\tau(k(n))) \leq J_c(n) \leq \frac{\tau(k(n)+1)}{n} J_c(\tau(k(n)+1))$$

for any nonnegative mapping $c\colon\ S \times U \to \mathbb{R}_+$, and in particular

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

By Lemma 1, $\lim_n k(n) = \infty$, while by the law of large numbers (26), it is clear that $\lim_k (\tau(k)/\tau(k+1)) = 1$ a.s. and because $(\tau(k(n))/\tau(k(n)+1)) \leq (\tau(k(n))/n) \leq 1$, we have $\lim_n (\tau(k(n))/n) = \lim_n (\tau(k(n)+1)/n) = 1$ a.s. By Theorem 3, the inequalities earlier in the discussion, together with the last convergence, yield (7) and (9) for nonnegative mappings. For a general cost mapping c (and in particular $c = v$), start with the decomposition $J_c(n) = J_{c^+}(n) - J_{c^-}(n)$ for all $n = 1, 2, \cdots$, where $c^\pm \equiv \max(\pm c, 0)$. Applying the result for nonnegative mappings developed above, we conclude $\lim_n J_c(n) = \lim_n J_{c^+}(n) - \lim_n J_{c^-}(n) = p^* I_c(g^*) + (1 - p^*) I_c(g_*)$ a.s. and Theorem 1 is established with (8) an immediate consequence of (9).

References

[1] E. Altman and A. Shwartz, "Optimal priority assignment: A time sharing approach," *IEEE Trans. Automat. Contr.*, vol. 34, pp. 1098–1102, 1989.

[2] F. J. Beutler and K. W. Ross, "Optimal policies for controlled Markov chains with a constraint," *J. Math. Anal. Appl.*, vol. 112, pp. 236–252, 1985.

[3] K. L. Chung, *Markov Chains with Stationary Transition Probabilities*, 2nd ed. New York: Springer-Verlag, 1967.

[4] E. Çinlar, *Introduction to Stochastic Processes*. Englewood Cliffs, NJ: Prentice-Hall, 1975.

[5] A. Hordijk and F. Spieksma, "Constrained admission control to a queueing system," *Advances Appl. Probability*, vol. 21, pp. 409–431, 1989.

[6] D.-J. Ma and A. M. Makowski, "Optimality results for a simple flow control problem," in *Proc. 26th IEEE Conf. Decision and Control*, Los Angeles, CA, Dec. 1987, pp. 1852–1857.

[7] ——, "A class of steering policies under a recurrence condition," in *Proc. 27th IEEE Conf. Decision and Control*, Austin, TX, Dec. 1988, pp. 1192–1197.

[8] ——, "A class of two–dimensional stochastic approximations and steering policies for Markov decision processes," in *Proc. 31st IEEE Conf. Decision and Control*, Tucson, AZ, Dec. 1992, pp. 3344–3349.

[9] A. M. Makowski and A. Shwartz, "Implementation issues for Markov decision processes," in *Proc. Workshop on Stochastic Differential Systems*, Institute of Mathematics and its Applications, Univ. Minnesota, Lecture Notes in Control and Information Sciences, W. Fleming and P.-L. Lions, Eds. Springer-Verlag, 1986, pp. 323–338.

[10] P. Nain and K. W. Ross, "Optimal priority assignment with hard constraint," *IEEE Trans. Automat. Contr.*, vol. AC-31, pp. 883–888, 1986.

[11] K. W. Ross, "Randomized and past-dependent policies for Markov decision processes with multiple constraints," *Ops. Res.*, vol. 37, pp. 474–477, 1989.

[12] S. M. Ross, *Introduction to Stochastic Programming*. New York: Academic, 1984.