TR-86-63

IMPLEMENTATION ISSUES FOR MARKOV DECISION PROCESSES

by

Armand M. Makowski and Adam Shwartz

This paper is based on a presentation made in the
Workshop on Stochastic Differential Systems with Applications
in Electrical/Computer Engineering, Control Theory and Operations Research,
Institute for Mathematics and its Applications
University of Minnesota,
Minneapois, Minnesota 55455
June 1986

IMPLEMENTATION ISSUES FOR MARKOV DECISION PROCESSES

by

Armand M. Makowski [1] and Adam Shwartz [2]

University of Maryland and Technion

ABSTRACT

In this paper, the problem of steering a lon-run average cost functional to a prespecified value is discussed in the context of Markov decision processes with countable state-space; this problem naturally arises in the study of constrained Markov decision processes by Lagrangian arguments. Under reasonable assumptions, a Markov stationary steering control is shown to exist, and to be obtained by fixed memoryless randomization between two Markov stationary policies. The implementability of this randomized policy is investigated in view of the fact that the randomization bias is solution to a (highly) nonlinear equation, which may not even be available in the absence of full knowledge of the model parameter values. Several proposals for implementation are made and their relative properties discussed. The paper closes with an outline of a methodology that was found useful in investigating properties of Certainty Equivalence implementations

[1] Electrical Engineering Department and Systems Research Center, University of Maryland, College Park, Maryland 20742, U.S.A. The work of this author was supported partially through NSF Grant ECS-83-51836, partially through a Grant from AT&T Bell Laboratories and partially through ONR Grant N00014-84-K-0614.

[2] Electrical Engineering Department, Technion-Israel Institute of Technology, Haifa 32000, ISRAEL. The work of this author was supported partially through a Grant from AT&T Bell Laboratories and partially through United States-Israel Binational Science Foundation Grant BSF 85-00306.

1. INTRODUCTION:

Consider a Markov decision process (MDP) with countable state space S, action space U and one-step transition mechanism $P = (p_{xy}(u))$, as commonly understood in the literature [5, 14, 15] ; a precise statement of the model and of the underlying assumptions is given in Section 2.

The discussion assumes the system performance to be quantified through a long-run average cost criterion associated with some instantaneous cost $c : S \times U \to \mathbb{R}$. If the sequence of states $\{X_n\}_1^\infty$ describes the evolution of the system while the decisions are encoded in the sequence $\{U_n\}_1^\infty$, then for every admissible control policy π, pose

$$J(\pi) := \overline{\lim_{n \uparrow \infty}} \frac{1}{n} E^\pi \sum_{i=1}^{n} c(X_i, U_i) \tag{1.1}$$

with the usual meaning for the notation E^π. The quantity $J(\pi)$ represents a measure of system performance when the policy π is in use.

Over the years, a great deal of effort has been devoted to characterizing and evaluating policies which exhibit desirable performance properties with respect to the cost functional (1.1). These questions are typically (but not exclusively) formulated as optimization problems; they have been abundantly studied and a variety of results are available in the literature, ranging from conditions for existence to algorithmic solutions [5, 14, 15] . There, as in many other situations, analysis often identifies the policy of interest to be a *Markov stationary* policy g . Unfortunately, this Markov stationary policy is usually *not* readily *implementable* (sometimes in spite of strong structural properties), with the encountered difficulties falling essentially into one of the two following categories:

(D1): The form of the policy g is a function of the various parameters determining the statistical description of the model. The actual *values* of these parameters are often *not* available to the decision-maker and need to be estimated as part of the system operation, possibly given some prior distribution on the parameters. Non-Bayesian versions of this problem have been worked out for specific control models, and the reader is referred to the survey paper by Kumar [6] and the monograph by Kumar and Varaiya [7] for additional information on this subject.

(D2): Even in the event the actual parameter values are available, the Markov stationary policy g may still not be implementable due to *computational* difficulties inherent to its definition. The situation treated by Nain and Ross [12] is a good case in point, for there non-trivial off-line computations are required in order to actually compute the value of a bias that enters the definition of a seemingly simple randomized policy.

This state of affairs very naturally suggests the formulation of the following *implementation* problem whose discussion constitutes the main subject of this paper: If g is a *given* Markov stationary policy, design an *implementable* policy α such that $J(\alpha) = J(g)$; such a policy α will be called an *implementation* of g . Here, implementability of a policy is synonymous with the availability of an *algorithm* which produces *on-line* control values, given *available* feedback and model information.

This question constitutes a broadening of the usual formulation of the adaptive control problem for Markov chains, in both its *direct* and *indirect* versions [4, 7] . Since this problem is somewhat amorphous in its stated generality, attention will be focused in this paper on the specific problem of *steering* the cost (1.1) to a particular value. Versions of this problem arise naturally in the solution of *constrained* MDP's via Lagrangian arguments, as demonstrated by Ross [12, 13] and by others [8, 9] in several specific instances. In all the *constrained* MDP's considered in the literature, at least to the authors' knowledge, the discussion proceeds according to the following arguments: Two Markov stationary policies $\underline{g}$ and $\overline{g}$ are identified which are optimal for some Lagrangian problem with the property

$$J(\underline{g}) < V < J(\overline{g}) \tag{1.2}$$

where V denotes the value of the constraint. The constrained optimal policy g then turns out to be a Markov stationary policy which solves the same Lagrangian problem as $\underline{g}$ and $\overline{g}$ but with $J(g) = V$. In all known references, such a Markov stationary policy g with $J(g) = V$ is constructed by suitably *randomizing* between the policies $\underline{g}$ and $\overline{g}$. Simple *memoryless* randomization at each step between the policies $\underline{g}$ and $\overline{g}$ produces a one-parameter family of Markov stationary policies $\{f^{\eta}, 0 \leq \eta \leq 1\}$. When the mapping $\eta \rightarrow J(f^{\eta})$ is *continuous* on the interval $[0,1]$, there exists at least one randomized strategy f^{η^*} that meets the value V and its corresponding bias value η^* is a solution of the equation $J(f^{\eta}) = V$, η in $[0,1]$, whence the policy $g = f^{\eta^*}$ solves the constrained MDP.

The determination of the optimal bias value η^* requires the evaluation of the expression $J(f^{\eta})$ for all values of η in the unit interval $[0,1]$, and this is a non-trivial task even in the simplest of situations, when all the parameter values are available [12] .

The general problem of steering the cost to some specified value is formulated in Section 3 under assumptions which are motivated by the situation encountered in constrained MDP's. Here too, the policy g that steers the cost to the prespecified value V suffers from similar implementation difficulties. In order to remedy to them, several implementations are proposed which exploit the structure of the obtained randomized steering policy, under the assumption that the policies $\underline{g}$ and $\overline{g}$ themselves *are* implementable. The properties of each proposed implementation are briefly reviewed. The paper closes with Section 4 where a useful methodology is outlined for establishing that the performance measure under both policies g and α coincide. These ideas were found useful in studying Certainty Equivalence policies.

A few words on the notation used throughout the paper: The notation $\mathbb{R}$ stands for the set of all real numbers, and the indicator function of a set A is denoted by $I(A)$. For any mapping $h : S \rightarrow \mathbb{R}$, pose $| h | := \sup_{x} | h(x) |$.

2. MODEL AND ASSUMPTIONS:

Assume the state-space S to be *countable* set and the control space U to be a *complete separable metric* space. The one-step transition mechanism P is defined through the one-step transition probability functions $p_{xy}(\cdot) : U \rightarrow \mathbb{R}$ which are assumed to be *Borel* measurable and to satisfy the standard properties

$$0 \leq p_{xy}(u) \leq 1, \quad \sum_{y} p_{xy}(u) = 1 \tag{2.1}$$

for all x and y in S, and all u in U. The space of probability measures on U (when equipped with its natural Borel σ-field) is denoted by $I\!M$.

The canonical sample space

In this paper, all probabilistic elements are defined on a single sample space $\Omega := S \times (U \times S)^\infty$ which acts as the *canonical* space for the MDP (S, U, P) under consideration. The *information* spaces $\{\mathbb{H}_n\}_1^\infty$ are recursively defined by $\mathbb{H}_1 := S$ and $\mathbb{H}_{n+1} := \mathbb{H}_n \times U \times S$ for all $n = 1,2,\ldots$. With a slight abuse of notation, Ω is clearly identified with $\mathbb{H}_\infty$.

A generic element ω of the sample space Ω is viewed as a sequence $(x_1, \omega_1, \omega_2, \cdots)$ with x_1 in S and blocks $\{\omega_n\}_1^\infty$ in $U \times S$. Each one of the blocks ω_n is decomposed into a ordered pair (u_n, x_{n+1}), where u_n and x_{n+1} are elements of U and S, respectively. Moreover, for each $n = 1,2,\ldots$, an element h_n in $\mathbb{H}_n$ is uniquely associated with the sample ω by posing $h_n := (x_1, \omega_1, \omega_2, \cdots, \omega_{n-1})$, with $h_1 := x_1$.

These quantities can be readily interpreted in the context of the situation described in the introduction: As the sample $\omega = (x_1, \omega_1, \omega_2, \cdots)$ is realized, the state of the system at time n is represented by x_n, and the decision-maker keeps track of the past system states x_i, $1 \leq i \leq n$, and past decisions u_i, $1 \leq i < n$. Thus, the controller has knowledge of the information vector h_n which is used to generate the control action u_n implemented at time n. The selection of this control is done according to a prespecified mechanism, which may be either deterministic or random, i. e., the controller thus uses h_n for selecting a probability measure in $\mathbb{M}$.

The basic random variables

The *coordinate* mappings $\{U_n\}_1^\infty$ and $\{X_n\}_1^\infty$ are defined on the sample space Ω by setting

$$U_n(\omega) := u_n, \quad X_n(\omega) := x_n \qquad \text{n=1,2,...(2.2)}$$

for all ω in Ω, with the corresponding *information* mappings $\{H_n\}_1^\infty$ given by

$$H_n(\omega) := (x_1, \omega_1, \omega_2, \cdots, \omega_{n-1}) = h_n \qquad \text{n=1,2,...(2.3)}$$

for all ω in Ω.

For each $n = 1,2,\ldots$, the mapping H_n generates a σ-field $\mathbb{F}_n$ on the sample space Ω, with $\mathbb{F}_n \subseteq \mathbb{F}_{n+1}$. With standard notation, $\mathbb{F} := \bigvee_{n=1}^{\infty} \mathbb{F}_n$ is simply the natural σ-field on the infinite cartesian product $\mathbb{H}_\infty$ generated by the mappings $\{U_n, X_n\}_1^\infty$. The sample space Ω is always equipped with this σ-field $\mathbb{F}$ and in that event, the mappings $\{U_n\}_1^\infty$ and $\{X_n\}_1^\infty$ are all *random variables* (RV) taking values in U and S, respectively.

The probabilistic structure

Since randomized strategies are allowed, an *admissible* control policy π is defined as any collection $\{\pi_n\}_1^\infty$ 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. Here, $\pi_n(\cdot, h_n)$ is interpreted as the conditional probability distribution for selecting the control value at time n, given that the information vector h_n is available to the decision-maker. Denote the collection of all such admissible policies by Π.

Let $\mu(\cdot)$ be a fixed probability distribution on S and let P be the one-step transition kernel $(p_{xy}(u))$ specified earlier. Under the assumptions made, the Kolmogorov Extension Theorem guarantees the existence (and uniqueness) of a family $\{P^\pi, \pi \in \Pi\}$ of probability measures on the σ-field $\mathbb{F}$ which satisfies the following requirements (R1)-(R3), i. e., for every policy π in Π, (R1): For all x_1 in S,

$$P^{\pi}[\, X_1 = x_1\,] = \mu(x_1),$$

(R2): For all Borel subsets A of U,

$$P^{\pi}[\, U(n) \in A \mid \mathbb{F}_n \,] = \pi_n(A\,;\, H_n), \qquad n=1,2,...$$

and

(R3): For all x and y in S,

$$P^{\pi}[\, X_{n+1} = y \mid \mathbb{F}_n \vee \sigma(U_n)\,] = p_{X_n\, y}(U_n) \qquad n=1,2,...$$

With these requirements (R1)-(R3), it is plain that $\pi_n(\bullet, H_n)$ is a *regular* conditional distribution of U_n given $\mathbb{F}_n$, and that

$$P^{\pi}[\, X_{n+1} = y \mid \mathbb{F}_n\,] = \int_U \pi_n(du\,;\, H_n) p_{X_n\, y}(u) \qquad n=1,2,...(2.4)$$

for all x and y in S. Motivated by (2.4), it is convenient to extend the notation $(p_{xy}(u))$ to elements of $\mathbb{M}$, i. e., for every probability measure ν in $\mathbb{M}$, the notation

$$p_{xy}(\nu) := \int_U \nu(du) p_{xy}(u) \qquad n=1,2,...(2.5)$$

is adopted for all x and y in S.

Subclasses of policies

A policy π in Π is said to be a *Markov* or *memoryless* policy if there exists a family $\{g_n\}_1^{\infty}$ of mappings $g_n : S \to \mathbb{M}$ such that

$$\pi_n(\bullet, H_n) = g_n(\bullet, X_n) \qquad P^{\pi}\text{-}a.s. \qquad n=1,2,...(2.6)$$

In the event the mappings $\{g_n\}_1^{\infty}$ are all identical to a given mapping $g : S \to \mathbb{M}$, the Markov policy π is termed *stationary* and can be identified with the mapping g itself, as will be done repeatedly in the sequel.

A policy π in Π will be said to be a *pure* strategy if there exists a family $\{f_n\}_1^{\infty}$ of mappings $f_n : \mathbb{H}_n \to U$ such that for every Borel subset A of U,

$$\pi_n(A\,;\, H_n) = \begin{cases} 1 & \text{if } f_n(H_n) \in A \\ 0 & \text{otherwise} \end{cases} \qquad P^{\pi}\text{-}a.s. \qquad n=1,2,...(2.7)$$

A pure policy π can thus be identified with the sequence of deterministic mappings $\{f_n\}_1^{\infty}$. A *pure Markov stationary* policy π in Π is thus fully characterized by a single mapping $f : S \to U$ to which it is substituted in the notation.

3. IMPLEMENTATIONS:

In this section, several proposals are made for obtaining an implementable control policy which solves the problem of steering the cost to a given value.

Steering the cost to a prespecified value

The problem of interest here is to find a Markov stationary policy g such that $J(g) = V$, with V some real constant determined through various design considerations. The discussion assumes the existence of two *implementable* Markov (possibly randomized) stationary policies $\underline{g}$ and $\overline{g}$ such that

$$J(\underline{g}) < V < J(\overline{g}), \tag{3.1}$$

i. e., the Markov stationary policy $\underline{g}$ (resp. $\overline{g}$) undershoots (resp. overshoots) the requisite performance level V. For every η in the unit interval [0,1], the policy f^{η} obtained by simply randomizing between the two policies $\underline{g}$ and $\overline{g}$ with *bias* η is the Markov stationary policy determined through the mapping $f^{\eta} : S \rightarrow \mathbb{M}$ where

$$f^{\eta}(\cdot, x) := \eta\overline{g}(\cdot, x)+(1-\eta)\underline{g}(\cdot, x) \tag{3.2}$$

for all x in S. Note that for $\eta=1$ (resp. $\eta=0$) the randomized policy f^{η} coincides with $\overline{g}$ (resp. $\underline{g}$). Owing to the condition (3.1), if the mapping $\eta \rightarrow J(f^{\eta})$ is *continuous* on the interval [0,1], then at least one randomized strategy f^{η^*} meets the value V and its corresponding bias value η^* is a solution of the equation

$$J(f^{\eta}) = V, \quad \eta \text{ in } [0,1], \tag{3.3}$$

whence $g = f^{\eta^*}$ steers (1.1) to the value V.

As pointed out in the introduction, solving the (highly) nonlinear equation (3.3) for the bias value η^* is usually a non-trivial task, even in the simplest of situations [12] . The implementations of the policy g which are defined below circumvent this difficulty by bypassing a direct solution of the equation (3.3). All the proposed implementations $\alpha=\{\alpha_n\}_1^{\infty}$ have the form

$$\alpha_n(\cdot, H_n) := \eta_n \overline{g}(\cdot, X_n)+(1-\eta_n)\underline{g}(\cdot, X_n) \qquad \text{n=1,2,...(3.4)}$$

where $\{\eta_n\}_1^{\infty}$ is some sequence of [0,1]-valued RV's which play the role of "estimates" for the bias value η^*. It will become apparent in what follows that the quality of such a sequence of "estimates" $\{\eta_n\}_1^{\infty}$ is not necessarily measured in terms of its possible convergence to the bias value η^*, but rather in terms of the proximity of $J(\alpha)$ to V.

Implementation by steering sample costs

In order to define this implementation of the policy f^{η^*}, define the sample costs $\{J_n\}_1^{\infty}$ by

$$J_n := \frac{1}{n}\sum_{i=1}^{n} c(X_i, U_i). \qquad \text{n=1,2,...(3.5)}$$

From the very definition of the policy f^{η^*}, it is expected that when using f^{η^*},

$$\lim_{n\uparrow\infty} J_n = J(f^{\eta^*}) = V \qquad P^{\eta^*}\text{-}a.s. \tag{3.6}$$

under reasonable recurrence conditions on the Markov chain induced by the Markov stationary policy f^{η^*}. This seems to suggest that the value V could be achieved by keeping the sample cost values $\{J_n\}_1^{\infty}$ as closely as possible to the desired value V. The corresponding implementation α_{SSC} exploits this idea by alternating between the policies $\underline{g}$ and $\overline{g}$ on the basis of the sign of $\{J_n - V\}_1^{\infty}$. Formally, the implementation α_{SSC} is defined as the admissible policy $\{\alpha_{SSC,n}\}_1^{\infty}$ in Π given by (3.4) where

$$\eta_{\mathrm{SSC},n} = I[\, J_n < V \,] \qquad \mathbf{n=1,2,...(3.7)}$$

This certainly defines an implementable policy since the values $\{J_n\}_1^\infty$ can be recursively computed upon observing the system.

The policy α_{SSC} was proposed by Ross [13] , without any analysis of its performance, in the context of a simple constrained problem for two competing queues. Makowski [10] shows that $J(\alpha_{\mathrm{SSC}}) = V$, not only for the situation considered by Ross but for more general MDP's as well. In a simple flow control problem [8] , Ma and Makowski obtain a similar result for threshold policies. Finally, it should be clear that the sequence of $\{0,1\}$-valued "estimates" $\{\eta_{\mathrm{SSC},n}\}_1^\infty$ cannot converge pathwise to the bias value η^*, and that such convergence can only take place in some weaker sense, say in the sense of Cesaro convergence.

Implementation by mixing policies

In the event the cost $J(\underline{g})$ and $J(\overline{g})$ are *readily* computable, an alternate implementation can be found in the one-parameter family $\{\pi(p),\ p \in [0,1]\}$ of *mixing* policies: For every p in $[0,1]$, consider a two-sided coin biased so that the events Head and Tail occur with probability p and $1-p$, respectively. To define the corresponding mixing policy $\pi(p)$, throw the coin exactly *once* at the beginning of times, and this *independently* of the other randomness defining the operation of the system. The policy $\pi(p)$ is defined as the policy that uses $\underline{g}$ (resp. $\overline{g}$) if the outcome is Tail (resp. Head). It is plain that

$$J(\pi(p)) = (1-p)J(\underline{g}) + pJ(\overline{g}) \tag{3.8}$$

provided that $J(\underline{g})$ and $J(\overline{g})$ are defined as *limits* in (1.1). An easy computation now shows that the mixing policy with mixing parameter p^* given by

$$p^* := \frac{V - J(\underline{g})}{J(\overline{g}) - J(\underline{g})} \tag{3.9}$$

steers the cost to V. This certainly defines an implementable policy α_{MIX} under the computability assumption made earlier; the calculations involved in the determination of the mixing parameter are trivial, in contradistinction with what is typically needed for evaluating the bias η^*. It is noteworthy that this policy α_{MIX} is also of the form (3.4), the sequence $\{\eta_{\mathrm{MIX},n}\}_1^\infty$ being given by

$$\eta_{\mathrm{MIX},n} = U. \qquad n=1,2,...(3.10)$$

Here U denotes a $\{0,1\}$-valued RV defined on the (augmented) sample space $\Omega \times [0,1]$, with $P^\pi[U = 1] = p^*$ under any policy π.

The idea of mixing policies was originally proposed by Altman and Shwartz [1] for a problem of competing queues with a single constraint; the general formulation is available in [9] . Interestingly enough, in some constrained MDP's with *multiple* constraints, the idea of mixing policies can be used to determine a constrained optimal solution by solving a linear programming problem; this approach is taken by Altman and Shwartz [1] for the same problem of competing queues.

Implementation by time-sharing policies

The reader will observe that mixing policies are not strictly speaking admissible policies in Π as defined in Section 2, owing to the initial randomization. The second drawback suffered by the policy α_{MIX}, possibly the most severe one, at least from an operational standpoint, lies in the fact that the sample costs $\{J_n\}_1^\infty$ cannot be expected to converge pathwise to V.

To improve on these structural deficiencies of the mixing policy α_{MIX}, consider the situation where the Markov chains induced by both $\underline{g}$ and $\overline{g}$ are *positive recurrent.* If x_0 denotes some privileged state in S, define a cycle as the time interval T between consecutive visits to the state x_0. The expectation of a cycle under policies $\underline{g}$ and $\overline{g}$ are denoted by $\underline{T} := E^{\underline{g}}\, T$ and $\overline{T} := E^{\overline{g}}\, T$, respectively.

For every r in [0,1], consider two sequences of non-negative integers $\{\underline{n}_j\}_1^\infty$ and $\{\overline{n}_j\}_1^\infty$ with the property that

$$\lim_{J \uparrow \infty} n(J) = \infty \quad \text{and} \quad \lim_{J \uparrow \infty} \frac{\overline{n}(J)}{n(J)} = r \tag{3.11}$$

where the notation

$$\underline{n}(J) = \sum_{j=1}^{J} \underline{n}_j, \quad \overline{n}(J) = \sum_{j=1}^{J} \overline{n}_j, \quad n(J) = \underline{n}(J) + \overline{n}(J) \tag{3.12}$$

is used for all J in $\mathbb{N}$. The discrete-time axis is divided into contiguous *control frames,* the $(J+1)^{rst}$ such control frame starts upon completion of the $n(J)^{th}$ cycle and is made up of $\underline{n}_{J+1} + \overline{n}_{J+1}$ cycles. The policy $\alpha_{\mathrm{TS}}(r)$ is defined as the policy in Π that during the J^{th} frame first operates $\underline{g}$ for $\underline{n}_J$ cycles and then operates $\overline{g}$ for $\overline{n}_J$ cycles. It follows from well-known properties of return times for Markov chains, that

$$\lim_{n \to \infty} J_n = J(\alpha_{\mathrm{TS}}(r)) = \frac{(1-r)\underline{T}J(\underline{g}) + r\overline{T}J(\overline{g})}{(1-r)\underline{T} + r\overline{T}} \tag{3.13}$$

where the first convergence in (3.13) takes place (under weak conditions) in the $P^{\alpha_{\mathrm{TS}}}$-a.s. sense.

Now, for every p in [0,1], let $r(p)$ be the element of [0,1] uniquely determined through the relation

$$p = \frac{r(p)\overline{T}}{(1-r(p))\underline{T} + r(p)\overline{T}}. \tag{3.14}$$

It is plain from (3.13)-(3.14) that

$$J(\alpha_{\mathrm{TS}}(r(p))) = \frac{(1-r(p))\underline{T}J(\underline{g}) + r(p)\overline{T}J(\overline{g})}{(1-r(p))\underline{T} + r(p)\overline{T}} = J(\pi(p)) \tag{3.15}$$

where the second equality follows from (3.8). In short, the policies $\pi(p)$ and $\alpha_{\mathrm{TS}}(r(p))$ both achieve the same cost, and in particular, the policy $\alpha_{\mathrm{TS}}(r(p^*))$ with p^* given by (3.9) steers the cost (1.1) to V. An easy computation that combines (3.9) and (3.14) readily shows that the value $r^* = r(p^*)$ is given by

$$r^* = \frac{\underline{T}[V - J(\underline{g})]}{\overline{T}[J(\overline{g}) - V] + \underline{T}[V - J(\underline{g}]} \tag{3.16}$$

The implementability of this time-sharing policy depends only on the availability of the costs values and mean return times under the given policies $\underline{g}$ and $\overline{g}$.

The reader will note that the policy $\alpha_{\mathrm{TS}}(r^*)$ can also be put in the form (3.4) with the sequence $\{\eta_{\mathrm{TS},n}\}_1^\infty$ being given this time by

$$\eta_{\mathrm{TS},n} = \begin{cases} 0 & \text{if } \tau_J \leq n < \sigma_J \text{ for some} J = 1,2,\ldots \\ 1 & \text{if } \sigma_J \leq n < \tau_{J+1} \text{ for some} J = 1,2,\ldots \end{cases} \qquad \mathbf{n=1,2,...(3.17)}$$

Here the $\mathbb{F}_n$-stopping times $\{\tau_J\}_1^\infty$ and $\{\sigma_J\}_1^\infty$ are defined as the time epochs at which the J^{th} control frame starts and the control policy switches from $\underline{g}$ to $\overline{g}$ (in the J^{th} control frame), respectively. Although the sequence $\{\eta_{\mathrm{TS},n}\}_1^\infty$ will certainly not converge pathwise, it is expected to converge in some distributional sense to a Bernoulli RV with parameter p^* given by (3.9).

The time-sharing implementation was introduced by Altman and Shwartz [1] for solving a constrained MDP with multiple constraints associated with a system of competing queues.

Implementation by Certainty Equivalence policies

In many situations of practical interest, it is possible to design (simple recursive) schemes for estimating the value η^* which solves (3.3). In that case, the *Certainty Equivalence Principle* is naturally invoked for defining the so-called "naive feedback" policy α_{CE}. Such a policy is of the form (3.4) where the sequence of estimates $\{\eta_n\}_1^\infty$ is the one generated through the given estimation scheme. It is hoped that the effects of controlling and learning about the system will combine to produce a (weakly) *consistent* estimation scheme. In such a case, the sequence of estimates $\{\eta_n\}_1^\infty$ converges to the value η^* in some sense, thus providing increasingly better approximations to it. Certainty Equivalence implementations (as defined here) are intrinsically different from the other implementations introduced earlier in this section, despite their common form (3.4). Indeed, the estimation schemes that appear in Certainty Equivalence implementations, exist *independently* of the policies used, and are selected with the hope that the estimates $\{\eta_n\}_1^\infty$ converge to the bias η^*.

At this point, the reader may wonder as to how such an estimation scheme can be selected. The remainder of this section is devoted to some of the approaches taken in the literature:

(i): Sometimes, it is feasible to compute the bias value η^* as an explicit function $\eta^*(\theta)$ of some external parameter θ whose value is not available to the decision-maker. In that case, in the spirit of *indirect* adaptive control [4, 7] , the designer may want to consider using the Certainty Equivalence Principle in conjunction with a parameter estimation scheme, say based on the *Maximum Likelihood Principle* or the *Method of Least-Squares,* i. e., if the parameter estimates are given by $\{\theta_n\}_1^\infty$, then the estimates $\{\eta_n\}_1^\infty$ are simply determined by $\eta_n = \eta^*(\theta_n)$ for all $n = 1,2,\ldots$ This program has been carried out by the authors for a variety of situations, including problems of server allocation in models of competing queues [16, 17] and problems of flow control in discrete-time $M \mid M \mid 1$ systems [8] .

(ii): In many applications, the function $\eta \rightarrow J(f^\eta)$ turns out to be *continuous* and *strictly monotone,* say increasing for sake of definiteness. The search for η^* can then be interpreted as finding the zero of the continuous, strictly monotone function $\eta \rightarrow J(f^\eta) - V$ and this brings to mind ideas from the theory of *Stochastic Approximations*. Here, this circle of ideas suggests generating a sequence of bias values $\{\eta_n\}_1^\infty$ through the recursion

$$\eta_{n+1} = \left[\eta_n + a_{n+1} \Big(V - c\,(X_{n+1}, U_{n+1}) \Big) \right]_0^1 \qquad \text{n=1,2,...(3.18)}$$

with η_1 given in [0,1]. In (3.18), the notation $(\,x\,)_0^1 = 0 \vee (x \wedge 1)$ is used for every x in $\mathbb{R}$, and the sequence of step sizes $\{a_n\}_1^\infty$ satisfies

$$0 < a_n \downarrow 0, \quad \sum_{n=1}^{\infty} a_n = \infty, \quad \sum_{n=1}^{\infty} | a_{n+1} - a_n | < \infty \tag{3.19}$$

Certainty Equivalence controllers based on Stochastic Approximation schemes such as (3.18)-(3.19) can be viewed as *indirect* adaptive control schemes; they have been studied by the authors for the models mentionned earlier [8, 17] .

4. A CONVERGENCE RESULT:

The design of Certainty Equivalence implementations raises the following natural questions:

Which notion of convergence in the space of policies is appropriate?

Which systems lend themselves to Certainty Equivalence type controllers?

To this date, no general answers are available on these questions. However, in a variety of situations studied in the references [8, 16, 17] , a useful methodology was developed for establishing the performance of Certainty Equivalence policies. This section summarizes the approach taken in these special cases, with the results being given in as generic a form as possible to emphasize the broad applicability of the proposed methodology. The underlying idea is due to Mandl [11] who introduced it in his seminal paper on the (optimal) adaptive control of finite-state Markov chains. The discussion given here provides an extension of this idea to the case of unbounded costs over countable state-spaces and randomized policies. A detailed exposition of the method is given by the authors in the context of a discrete-time model for competing queues [18] to which the reader is referred for addititonal information.

A convergence condition

The effect of two different admissible policies, say α and β, on the chain transitions is captured simply by the differences

$$p_{xy}(\alpha(H_n)) - p_{xy}(\beta(H_n)) = \int_U p_{xy}(u)[\alpha(du\,;H_n) - \beta(du\,;H_n)] \qquad \text{n=1,2,...(4.1)}$$

for all x and y in S. It is expected, maybe somewhat naively, that if this effect (4.1) becomes negligible in time, then the sample cost sequence $\{J_n\}_1^\infty$ will have same limiting behavior under both probability measures P^α and P^β. Roughly speaking, it is anticipated, of course under appropriate technical conditions, that "convergence of the controls" should imply "convergence of the cost" in some sense.

The methodology outlined below indicates that such a result is indeed possible under reasonable convergence assumptions. This approach is well suited (but not limited) to studying the cost performance of Certainty Equivalence policies when convergence of the corresponding estimates $\{\eta_n\}_1^\infty$ to η^* can be directly established by independent problem-specific arguments.

Throughout this section, let g denote a fixed Markov stationary policy in Π, and let α be a second admissible policy. The comments made earlier motivate the following notion of convergence in the space of policies: The policy α is said to satisfy the *convergence condition* (C) with respect to the Markov stationary policy g if the sequence of (signed) random measures $\{\alpha_n(\cdot;H_n) - g(\cdot;X_n)\}_1^\infty$ converges *weakly* [2] *in probability* to zero, i. e., for every bounded continuous function $f : U \to R$,

$$\lim_{n\to\infty} P^{\alpha}\left[\ \left|\int_U f(u)\,[\alpha_n(du;H_n) - g(du;X_n)]\right| > \epsilon\ \right] = 0 \qquad (4.2)$$

for every $\epsilon>0$. If g and α are given by (3.2) and (3.4), respectively, then (4.2) reduces to

$$\lim_{n\to\infty} P^{\alpha}\left[\ |\eta - \eta_n| \ \ \left|\int_U f(u)\,[\overline{g}(du;X_n) - \underline{g}(du;X_n)]\right| > \epsilon\ \right] = 0. \qquad (4.3)$$

and roughly speaking, condition (C) is seen to hold provided the sequence of estimates $\{\eta_n\}_1^{\infty}$ converges to η^* at least in probability under P^{α}.

The discussion starts with the following version of a standard result in the theory of MDP's with lon-run average criterion ([14] , Thm. 6.17, pp. 144-145)

Lemma 4.1. *If the mapping* $h : S \to \mathbb{R}$ *and the constant* J *solve the equations*

$$h(x) + J = \sum_y \int_U p_{xy}(u)[h(y) + c(x,u)]\, g(du;x) \qquad (4.4)$$

for all x in S *under the conditions*

$$E^g\left[\ |h(X_n)|\ \right] < \infty \qquad n=1,2,...(4.5a)$$

and

$$\lim_{n\to\infty} \frac{1}{n} E^g[h(X_n)] = 0, \qquad (4.5b)$$

then necessarily

$$J = J(g) = \lim_{n\to\infty} \frac{1}{n} E^g \sum_{i=1}^{n} c(X_i, U_i). \qquad (4.6)$$

Proof: The convention (2.5) for the transition probabilities allows a rewriting of (4.4) in the form

$$h(X_i) + J = E^g[h(X_{i+1}) \mid \mathbb{F}_i] + c(X_i, U_i), \qquad i=1,2,...(4.7)$$

and a direct iteration then gives

$$E^g[h(X_1)] + nJ = E^g[h(X_{n+1})] + E^g[\sum_{i=1}^{n} c(X_i, U_i)]. \qquad n=1,2,...(4.8)$$

The result now follows readily upon dividing by n in (4.8) and letting n go to infinity. □

Bounded costs

When the cost function c is *bounded,* arguments similar to the ones given in Section 6.7 of the monograph by Ross [14] can be used to prove the existence of a solution pair (h, J) which satisfies the conditions of Lemma 4.1. Such argument assumes the existence of a privileged state in S, say x_0, which is *positive recurrent* under g. The return time to the state x_0 is defined as the $\mathbb{F}_n$-stopping time τ given by

$$\tau := \inf \{ n \geq 1 : X(n) = x_0 \}. \qquad \textbf{(4.9)}$$

The following technical conditions (H1)-(H3) can now be defined, where

(H1): The *finiteness* condition

$$Z(x) := E^g [\tau \mid X_1 = x] < \infty$$

holds for all x in S;

(H2): The *boundedness* condition

$$\sup_n E^g \left[Z(X_n) \right] < \infty$$

holds, and

(H3): The *integrability* condition

$$\sum_y p_{xy}(g) Z(y) < \infty$$

holds for all y in S.

The hypotheses (H1)-(H3) are clearly satisfied when the state space S is *finite*. The authors have also verified these hypotheses in a variety of problems where the state-space is *not* finite, including the competing queue problem [18] and a flow control problem for discrete-time $M \mid M \mid 1$ systems [8]. In the proposed set-up, the arguments establishing Lemma 7.2 of [18] lead to the following result

Lemma 4.2. *Under the assumptions (H1)-(H3), there exist a mapping* $h : S \rightarrow \mathbb{R}$ *and a constant* J *which satisfy (4.4)-(4.5) and the bound*

$$| h(x) | \leq C Z(x) \qquad (4.10)$$

for all x *in* S, *with some* $C > 0$.

For the remainder of this section, the conditions of Lemma 4.1 are assumed to hold, i. e., there exists a pair (h, J) satisfying (4.4)-(4.5), together with the additional condition (H4), where

(H4): The sequence of RV's $\{h(X_n)\}_1^\infty$ is *integrable* under both P^g and P^α.

The sequences $\{\Phi_n\}_1^\infty$ and $\{Y_n\}_1^\infty$ of $\mathbb{R}$-valued RV's are now defined to be

$$\Phi_n = E^\alpha [h(X_{n+1}) \mid \mathbb{F}_n] - E^g [h(X_{n+1}) \mid \mathbb{F}_n] \qquad n=1,2,...(4.11)$$

and

$$Y_n = h(X_{n+1}) - E^\alpha [h(X_{n+1}) \mid \mathbb{F}_n] \qquad n=1,2,...(4.12)$$

with $Y_1 = h(X_1) - E^\alpha [h(X_1)]$; these definitions are well posed under the integrability assumptions (H4). With these definitions, the argument proposed by Mandl [11] takes the form

$$h(X_i) + J(g) = -\Phi_i - Y_i + h(X_{i+1}) + c(X_i, U_i) \qquad i=1,2...(4.13)$$

upon adding and substracting both RV's $E^\alpha [h(X_{i+1}) \mid \mathbb{F}_i]$ and $h(X_{i+1})$ on the right handside of (4.7). Iteration of (4.13) implies the relation

$$J_n = J(g) + \frac{1}{n}\sum_{i=1}^{n}\Phi_i + \frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{1}{n}\Big(h(X_{n+1}) - h(X_1) \Big), \quad \mathbf{n=1,2,...(4.14)}$$

which is key for establishing $J(\alpha) = J(g)$.

Theorem 4.1. *Assume there exists a pair (h, J) which satisfies the conditions of Lemma 4.1 together with the condition (H4). If the convergences*

$$\frac{1}{n}\sum_{i=1}^{n} Y_i \to 0 \qquad \textit{in expectation under } P^{\alpha} \tag{4.15a}$$

$$\frac{1}{n}\Big(h(X_{n+1}) - h(X_1) \Big) \to 0 \qquad \textit{in expectation under } P^{\alpha} \tag{4.15b}$$

$$\frac{1}{n}\sum_{i=1}^{n} \Phi_i \to 0 \qquad \textit{in expectation under } P^{\alpha} \tag{4.15c}$$

take place, then

$$J(\alpha) = \lim_{n\to\infty} E^{\alpha}\frac{1}{n}\sum_{i=1}^{n} c(X_i, U_i) = J(g) \tag{4.16}$$

Note that if the convergences (4.15) are in $L^1(\Omega, \mathbb{F}, P^{\alpha})$, instead of being only in expectation under P^{α}, then the convergence (4.16) of the sample cost sequence $\{J_n\}_1^{\infty}$ is also in $L^1(\Omega, \mathbb{F}, P^{\alpha})$.

The verification of the convergences (4.15) often turns out to be a fairly technical task which depends heavily on the problem at hand. As will become from the discussion given below, many of the difficulties are related to the fact that the state space S may *not* be finite, and the RV's of interest therefore *not* necessarily bounded. The convergence proof typically proceeds along the following lines.

(4.15a): The very definition of the RV's $\{Y_n\}_1^{\infty}$ implies $E^{\alpha}Y_n = 0$ for all $n = 1,2,...$, whence the convergence (4.15a) is automatic.

(4.15b): This convergence is obtained, for example, when S is finite. More generally, it can also be established under the conditions (H1)-(H3) provided the bound

$$\sup_n E^{\alpha}\Big[Z(X_n) \Big] < \infty \tag{4.17}$$

holds, by virtue of (4.10).

(4.15c): Suppose first that S is *finite*. In that case, (4.11) takes the form

$$\Phi_n = \sum_y h(y) \int_U p_{X_n y}(u) \Big[\alpha_n(du, H_n) - g(du, X_n) \Big] \qquad \text{n=1,2,...(4.18)}$$

with the sum being *finite*. If the one-step transition probability functions $p_{xy}(\cdot) : U \to \mathbb{R}$ are *continuous* for all x and y in S, a property assumed hereafter, then condition (C) and the *finiteness* of S imply $\lim_n \Phi_n = 0$ in *probability* under P^{α}. Moreover, the RV's $\{\Phi_n\}_1^{\infty}$ form a sequence of bounded RV's (since S is finite), thus are uniformly integrable under P^{α} [3], and the convergence (4.15c) also takes place in $L^1(\Omega, \mathbb{F}, P^{\alpha})$, as well as in expectation under P^{α}.

When S is *not* finite, the arguments are more involved and require *bounds* on the sequence of RV's $\{h(X_n)\}_1^\infty$, as well as *tightness* conditions on the state sequence $\{X_n\}_1^\infty$. One possible set of such conditions is formulated as condition (H5), where

(H5a): The RV's $\{X_n\}_1^\infty$ form a *tight* sequence under the probability measure P^α, i. e., for every $\epsilon>0$, there exists a *finite* subset K_ϵ of S such that

$$\sup_n \; P^\alpha[\, X_n \notin K_\epsilon \,] < \epsilon,$$

(H5b): The RV's $\{h(X_n)\}_1^\infty$ are *uniformly integrable* under the probability measure P^α, *and*

(H5c): The RV's $\{(h(X_n), P^{(\alpha \mid n \mid g)})\}_1^\infty$ are *uniformly integrable*, where the policy $(\alpha \mid n \mid g)$ is defined by the sequence $\{\alpha_1, \alpha_2, \ldots, \alpha_{n-1}, g, g, \cdots\}$ for all $n=1,2,\ldots$, i. e.,

$$\lim_{B\uparrow\infty} \sup_n \; E^\alpha\left[E^g\left[\; |h(X_{n+1})| \, I[\; |h(X_{n+1})| > B \;] \mid \mathbb{F}_n \right]\right] = 0$$

or equivalently,

$$\lim_{B\uparrow\infty} \sup_n \; E^\alpha\left[\sum_{y:|h(y)|>B} |h(y)| \int_U p_{X_n y}(u)\, g(du, X_n)\right] = 0.$$

Such conditions are typically verified by establishing bounds on the moments of the sequences of RV's $\{X_n\}_1^\infty$ and $\{h(X_n)\}_1^\infty$, respectively, over some larger, but easier to handle, class of policies [18] , or by using the specific structure of the policies g and α [8] .

To proceed in the discussion, observe that

$$\Phi_n = \sum_y h(y)\, \Delta_n(X_n, y). \qquad \text{n=1,2,...(4.19)}$$

where the RV's $\{\Delta_n(x,y)\}_1^\infty$ are defined by

$$\Delta_n(x,y) := \int_U p_{xy}(u) \left[\alpha_n(du, H_n) - g(du, X_n)\right]. \qquad \text{n=1,2,...(4.20)}$$

for all x and y in S. Two situations then naturally arise.

If h is *bounded*, the representation (4.20) implies

$$|\Phi_n| \le |h| \sum_y |\Delta_n(X_n,y)| \le 2|h| \qquad \text{n=1,2,...(4.21)}$$

and therefore for every $\epsilon>0$, with K_ϵ a finite subset of S entering the tightness condition (H5a), it follows that

$$E^\alpha[\,|\Phi_n|\,] \le |h| E^\alpha[\, I[X_n \in K_\epsilon] \max_{x\in K_\epsilon} \sum_y |\Delta_n(x,y)|\,]+2|h| P^\alpha[\, X_n \notin K_\epsilon\,]$$

$$\le |h| \sum_{x\in K_\epsilon} E^\alpha[\, \sum_y |\Delta_n(x,y)|\,]+2|h|\epsilon. \qquad \text{n=1,2,...(4.22)}$$

With the enforced continuity assumption on the one-step transition probability functions, condition (C) implies $\lim_n |\Delta_n(x,y)| = 0$ in probability under P^α for all x and y in S, and the Bounded Convergence Theorem thus gives

$$\lim_{n \to \infty} E^{\alpha}[\sum_{y} | \Delta_n(x,y) |] = 0 \tag{4.23}$$

owing to the bound (4.21). It is now clear from (4.22)-(4.23), since ϵ is arbitrary, that $\lim_{n} E^{\alpha}[| \Phi_n |] = 0$.

If h is *not* bounded, define the sequence $\{\Phi_n^B\}_1^{\infty}$ of $\mathbb{R}$-valued RV's for every $B > 0$, where

$$\Phi_n^B := E^{\alpha}\left[h(X_{n+1})I[\; | h(X_{n+1}) | \leq B \;] \mid \mathbb{F}_n \right] - E^{g}\left[h(X_{n+1})I[\; | h(X_{n+1}) | \leq B \;] \mid \mathbb{F}_n \right].$$

n=1,2,...(4.24)

For every $B > 0$, it is clear that

$$E^{\alpha}[| \Phi_n |] \leq E^{\alpha}[| \Phi_n^B |] + E^{\alpha}[| h(X_{n+1}) | I[\; | h(X_{n+1}) | > B \;]]$$

$$+ E^{\alpha}\left[E^{g}[| h(X_{n+1}) | I[\; | h(X_{n+1}) | > B \;] \mid \mathbb{F}_n]\right], \quad \text{n=1,2,...(4.25)}$$

and use of the uniform integrability assumptions (H5b)-(H5c) readily implies

$$E^{\alpha}[| \Phi_n |] \leq E^{\alpha}[| \Phi_n^B |] + \epsilon_B \quad \text{n=1,2,...(4.26)}$$

with $\lim_{B} \epsilon_B = 0$ monotonically as $B \uparrow \infty$. By the first part of the argument (for h bounded), the convergence $\lim_{n} E^{\alpha}[| \Phi_n^B |] = 0$ takes place for every $B > 0$, and therefore $\overline{\lim_{n}} E^{\alpha}[| \Phi_n |] \leq \epsilon_B$, the conclusion $\lim_{n} E^{\alpha}[| \Phi_n |] = 0$ being now immediate.

Note that under (H5), the RV's $\{\Phi_n\}_1^{\infty}$ are uniformly integrable under P^{α}, and the convergence (4.15c) therefore takes place in the stronger $L^1(\Omega,\mathbb{F},P^{\alpha})$ sense, as does the convergence (4.15b). To obtain $L^1(\Omega,\mathbb{F},P^{\alpha})$ convergence in (4.16), such convergence in (4.15a) needs to be established. It is plain from the definition (4.12) that the sum in (4.15a) forms a $(P^{\alpha},\mathbb{F}_n)$ *-martingale* and the Law of Large Numbers for martingales ([11] , Theorem 3), the so-called Stability Theorem, thus yields the desired convergence under the condition

$$E^{\alpha}\left(\sum_{n=1}^{\infty} \frac{| Y_n |^2}{n^2} \right) < \infty. \tag{4.27}$$

These remarks can be summarized in the following proposition.

Theorem 4.1 *Assume the conditions (H1)-(H5) to hold and the one-step transition probabilities to be continous. Whenever the policy α satisfies the convergence condition (C) with respect to the policy g, the convergences (4.15) take place in expectation under P^{α}. Moreover, if (4.28) holds as well, then the convergences (4.15) take place in $L^1(\Omega,\mathbb{F},P^{\alpha})$.*

Unbounded costs

When the cost c is not bounded, it may not be possible to establish the existence of a pair (h,J) that satisfies the conditions (4.4)-(4.5) of Lemma 4.1. As shown by the authors [18] , an indirect route can be taken in such a case by imposing a uniform integrability condition on the sequence $\{c(X_n, U_n)\}_1^{\infty}$ under P^{α}. This condition provides a means to carry out a standard truncation argument: The one-step cost is first approximated by a sequence of bounded costs obtained by truncating the original (unbounded) cost. Assumptions are imposed so that the previous arguments, which culminated in a version of Theorem 4.1, apply on each one of the truncated costs. A double limiting argument is then validated through the uniform integrability condition on the sequence $\{c(X_n, U_n)\}_1^\infty$. This approach can be summarized in the following

Theorem 4.2. *Assume conditions (H1)-(H5) to hold, continuity of the one-step transition probabilities and uniform integrability of the RV's* $\{c(X_n, U_n)\}_1^\infty$ *under* P^α. *Whenever the policy* α *satisfies the convergence condition (C) with respect to the policy* g, *the convergences (4.15) take place in expectation under* P^α. *Moreover, if (4.27) holds as well, then the convergences (4.15) take place in* $L^1(\Omega, \mathbb{F}, P^\alpha)$, *and so does the convergence in (4.16).*

REFERENCES:

[1] E. Altman and A. Shwartz, "Optimal priority assignment with general constraints," *Proceedings of the 24th Allerton Conference on Communication, Control and Computing,* (October 1986.).

[2] P. Billingsley, *Convergence of Probability Measures,* John Wiley, New York (1968).

[3] K. L. Chung, *A course in probability theory,* Second Edition, Academic Press, New York (1974).

[4] G. C. Goodwin and K. W. Sin, *Adaptive Filtering, Prediction and Control,* Prentice-Hall, Englewood Cliffs, New Jersey (1984).

[5] D. P. Heyman and M. J. Sobel, *Stochastic Models in Operations Research, Volume II: Stochastic Optimization,* MacGraw-Hill, New York (1984).

[6] P. R. Kumar, "A survey of some results in stochastic adaptive control," *SIAM J. Control Opt.* vol. 23, no. 3, pp. 329-380 (May 1985).

[7] P. R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification and Adaptive Control,* Prentice-Hall, Englewood Cliffs, New Jersey (1986).

[8] D.-J. Ma and A. M. Makowski, *A simple problem of flow control II: Implementation of threshold policies,* Systems Research Report, In preparation, Systems Research Center, University of Maryland at College Park (1986).

[9] D. -J. Ma, A. M. Makowski, and A. Shwartz, "Estimation and optimal control for constrained Markov chains," *25st IEEE Conference on Decision and Control,* (December 1986). Invited paper

[10] A. M. Makowski, *How to randomize between two policies so as to meet a constraint,* Systems Research Report, In preparation, Systems Research Center, University of Maryland at College Park (1986).

[11] P. Mandl, "Estimation and control in Markov chains," *Adv. Appl. Prob.* vol. 6, pp. 40-60 (1974).

[12] P. Nain and K. W. Ross, "Optimal priority assignment with hard constraint," *IEEE Trans. Auto. Control* vol. AC-31, no. 10, pp. 883-888 (October 1986).

[13] K. W. Ross, *Constrained Markov Decision Processes with Queueing Applications,* Ph. D. thesis,

Computer, Information and Control Engineering, University of Michigan, Ann Arbor, Michigan (1985).

[14] S. M. Ross, *Applied Probability Models with Optimization Applications,* Holden-Day, San Francisco (1970).

[15] S. M. Ross, *Introduction to Stochastic Dynamic Programming,* Academic Press (1984).

[16] A. Shwartz and A. M. Makowski, "An optimal adaptive scheme for two competing queues with constraints," pp. 515-532 in *Proceedings of the 7th International Conference on Analysis and Optimization of Systems,* ed. A. Bensoussan and J. -L. Lions, Springer Verlag Lecture Notes in Control and Information Sciences, Antibes, France (June 1986).

[17] A. Shwartz and A. M. Makowski, *Adaptive policies for a system of competing queues II: Implementable schemes for optimal server allocation.,* Systems Research Report, In preparation, Systems Research Center, University of Maryland at College Park (1986).

[18] A. Shwartz and A. M. Makowski, "Adaptive policies for a system of competing queues I: Convergence results for the long-run average cost," *J. Appl. Prob.*, (November 1986). Submitted