MARKOV DECISION PROCESSES WITH SAMPLE PATH CONSTRAINTS: THE COMMUNICATING CASE

KEITH W. ROSS

University of Pennsylvania, Philadelphia, Pennsylvania

RAVI VARADARAJAN

University of Florida, Gainesville, Florida

(Received August 1986; revisions received August 1987, April 1988; accepted June 1988)

We consider time-average Markov Decision Processes (MDPs), which accumulate a reward and cost at each decision epoch. A policy meets the sample-path constraint if the time-average cost is below a specified value with probability one. The optimization problem is to maximize the expected average reward over all policies that meet the sample-path constraint. The sample-path constraint is compared with the more commonly studied constraint of requiring the average expected cost to be less than a specified value. Although the two criteria are equivalent for certain classes of MDPs, their feasible and optimal policies differ for many nontrivial problems. In general, there does not exist optimal or nearly optimal stationary policies when the expected average-cost constraint is employed. Assuming that a policy exists that meets the sample-path constraint, we establish that there exist nearly optimal stationary policies for communicating MDPs. A parametric linear programming algorithm is given to construct nearly optimal stationary policies. The discussion relies on well known results from the theory of stochastic processes and linear programming. The techniques lead to simple proofs of the existence of optimal and nearly optimal stationary policies for unichain and deterministic MDPs, respectively.

We consider the Markov decision problem of locating a policy that maximizes over all policies $\mathbf{u}$ the expected average reward

$$\phi_r(\mathbf{u}) := E_{\mathbf{u}}\left[\liminf_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} r(X_m, A_m)\right] \tag{1}$$

subject to the sample-path constraint

$$P_{\mathbf{u}}\left(\limsup_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} c(X_m, A_m) \leq \alpha\right) = 1. \tag{2}$$

Here, $\{X_m\}$ is the state process taking values in the finite state space S; $\{A_m\}$ is the action process taking values in the finite action space B; $r(\cdot, \cdot)$ and $c(\cdot, \cdot)$ are the rewards and costs, respectively, as functions of the current state and action chosen; α is the constraint value, below which the long-run average cost must fall with probability 1.

The above problem is to be contrasted with the more commonly studied constrained Markov decision problem of maximizing the average expected reward

$$\phi_e(\mathbf{u}) := \liminf_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} E_{\mathbf{u}}[r(X_m, A_m)] \tag{3}$$

subject to a constraint on average expected cost

$$K(\mathbf{u}) := \limsup_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} E_{\mathbf{u}}[c(X_m, A_m)] \leq \alpha. \tag{4}$$

We will refer to this problem as the expectation problem and to (1)–(2) as the sample-path problem.

Many applications of constrained Markov decision processes currently exist in the literature: see Golabi, Kulkarni and Way (1982) for highway management applications; Kolesar (1970) for hospital scheduling applications; Maglaris and Schwartz (1982); and Nain and Ross (1986) for telecommunication applications.

Derman (1970) first studied the expectation problem in the context of the theory of expected state-action frequencies. Assuming that the Markov Decision Process (MDP) is unichain, that is, every pure policy gives rise to a Markov chain with at most one recurrent class, Derman proved that there exists an optimal (randomized) stationary policy for the expectation criterion. This result leads to a simple Linear Program (LP) that pinpoints such a stationary policy. More recently, by way of the dynamic programming equation, Beutler and Ross (1985, 1986) showed that under similar recurrence conditions there exists an optimal policy that requires randomization in at most

Subject classifications: Dynamic programming, Markov finite state: time-average case. Probability, Markov processes: sample-path constraints.

0030-364X/89/3705-0780 $01.25

one state. In Ross (1989), the limited randomization property was extended to multiple expectation constraints and the simple LP was shown to produce such an optimal policy. In the context of expected state-action frequencies, Hordijk and Kallenberg (1984) and Kallenberg (1983) studied the expectation problem for MDPs with general recurrent structures. In this case, there does not always exist an optimal stationary policy, or even an ϵ-optimal stationary policy.

It appears that little work has been done on the sample-path criterion. But the two criteria are indeed different, as we shall establish by an example that a policy which is feasible for the expectation criterion is not necessarily feasible for the sample-path criterion. And it is possible that there does not exist nearly optimal stationary policies for the expectation criterion, but there exists an optimal pure policy for the sample-path criterion.

In this paper, we focus on the sample-path criterion for communicating MDPs, where for every pair of states there is a policy that will bring the process from one state to the other with positive probability. For communicating MDPs, there does not, in general, exist an optimal stationary policy for the sample-path criterion; however, there always exists an ϵ-optimal stationary policy for all $\epsilon > 0$. The ϵ-optimal policy can be determined by solving a parametric LP.

The techniques developed herein also lead to a simple proof for the existence of an optimal stationary policy for the unichain case. Moreover, as a corollary to the basic existence result for communicating MDPs, we prove the existence of ϵ-optimal policies for deterministic MDPs (where the choice of action completely determines the subsequent state). Parametric LP is combined with a graph-theoretic algorithm to construct the ϵ-optimal policies.

The proofs of these results are based on simple probabilistic arguments; they rely neither on the machinery of expected state-action frequencies nor on the indirect manipulations of the dynamic programming equation. To demonstrate the simplicity of the proofs, we give a self-contained presentation that depends only on well known results from stochastic processes and linear programming. The aforementioned results also form the foundation of a companion paper (Ross and Varadarajan 1986), in which it is proved that for MDPs with general recurrent structure there always exists nearly optimal stationary policies for the sample-path criterion.

This paper is organized as follows. In Section 1, we precisely define the problems addressed and quote some well known results from the stochastic process literature. In Section 2, we briefly compare the two criteria; an example is given that illustrates their differences. In Section 3, we introduce a parametric LP and relate it to the sample-path problem; this leads to a simple proof of the existence of an optimal stationary policy for unichain MDPs. The parametric LP is again used in Section 4 to prove the existence of an ϵ-optimal stationary policy for communicating MDPs. Results on deterministic MDPs are given as a corollary in Section 5. We conclude in Section 6 with a brief discussion on the sample-path problem with multiple constraints and a final comparison of the two criteria.

1. THE SAMPLE-PATH PROBLEM

The basic notation and precise problem definitions are given in this section.

1.1. Sample Space, Policies and Probability Measures

The underlying sample space is given by

$$\Omega := \{S \times B\}^{\infty}$$

so that a typical realization ω can be represented as

$$\omega = (x_1, a_1, x_2, a_2, x_3, a_3, \ldots).$$

The state and action random variables X_m, A_m for $m = 1, 2, \ldots$ are then defined as the coordinate mappings

$$X_m(\omega) := x_m \quad \text{and} \quad A_m(\omega) := a_m.$$

Throughout, the sample space Ω will be equipped with the σ-algebra β generated by the random variables $(X_m, A_m, m = 1, 2, \ldots)$.

In order to give a formal definition of a policy, first let Ψ be the set of all probability measures on the action space B, that is

$$\Psi := \{(p_1, p_2, \ldots, p_b):$$
$$p_1 + p_2 + \ldots + p_b = 1,\ p_a \geq 0,\ 1 \leq a \leq b\}$$

where b is the cardinality of B. Then a *policy* $\mathbf{u}$ is defined to be a sequence $\mathbf{u} = (u^1, u^2, u^3, \ldots)$ where u^k is a mapping from $\{S \times B\}^{k-1} \times S$ to Ψ. We write u_a^m, $1 \leq a \leq b$, for the ath component of u^m.

Throughout we suppose that the distribution $\mathbf{q} = (q_x : x \in S)$ for the initial state is fixed and *given* to the decision maker. For a fixed policy $\mathbf{u}$, we proceed to construct the probability measure $P_{\mathbf{u}}$ for the measurable space (Ω, β). The finite-dimensional distributions of the probability measure $P_{\mathbf{u}}$ are defined

recursively as

$$P_u(X_1 = x) = q_x, \quad x \in S \tag{5}$$

$$P_u(A_m = a \mid X_1 = x_1, A_1 = a_1, \ldots, X_{m-1} = x_{m-1}, A_{m-1} = a_{m-1}, X_m = x_m) = u_a^m(x_1, a_1, \ldots, x_{m-1}, a_{m-1}, x_m) \tag{6}$$

$$P_u(X_{m+1} = y \mid X_1 = x_1, A_1 = a_1, \ldots, X_{m-1} = x_{m-1}, A_{m-1} = a_{m-1}, X_m = x, A_m = a) = P_{xay}. \tag{7}$$

P_{xay} is the *law of motion*, which is given and determined from the physical meaning of the problem. From a standard application of the Kolmogorov consistency theorem, we know there exists a unique probability measure P_u on (Ω, β) such that (5)–(7) hold for all possible histories and all $m \geq 1$. Thus, for each policy u, we have constructed a probability space (Ω, β, P_u).

1.2. The Sample-Path Optimization Problem

Let U_s be the class of all sample-path feasible policies, that is, the set of all policies u that satisfy the sample-path constraint (2). Denote

$$R := \liminf_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} r(X_m, A_m)$$

$$C := \limsup_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} c(X_m, A_m).$$

The random variables R and C are the (worst case) long-run average reward and cost, respectively; note that $\phi_s(\mathbf{u}) = E_u[R]$ and $\mathbf{u} \in U_s$ if and only if $P_u(C \leq \alpha) = 1$.

Definition 1. A policy $\mathbf{v} \in U_s$ is said to be *optimal* [*ϵ-optimal*] *for the sample-path criterion* if there exists a real number t such that

(i) $P_u(R \leq t) = 1$ for all policies $\mathbf{u} \in U_s$;
(ii) $P_v(R = t) = 1$ [$P_v(R \geq t - \epsilon) = 1$].

It follows that if v is optimal [ϵ-optimal] for the sample-path criterion, then v satisfies $\phi_s(\mathbf{v}) = \phi_s^*$ [$\phi_s(\mathbf{v}) \geq \phi_s^* - \epsilon$], where

$$\phi_s^* := \sup_{\mathbf{u}\in U_s} \phi_s(\mathbf{u}).$$

Thus, the definition of optimality and ϵ-optimality for the sample-path criterion is stronger than that alluded to in the Introduction.

In a completely analogous way we can define U_e and ϕ_e^* for the expectation criterion.

Definition 2. A policy $\mathbf{v} \in U_e$ is *optimal* [*ϵ-optimal*] *for the expectation criterion* if $\phi_e(\mathbf{v}) = \phi_e^*$ [$\phi_e(\mathbf{v}) \geq \phi_e^* - \epsilon$].

1.3. Subclasses of Policies

In general, a policy may be difficult to implement because it depends not only on the entire past of the process, but also on the epoch at which actions are applied. We are, therefore, motivated to search for optimal and ϵ-optimal policies within smaller and more appealing classes of policies. One particularly interesting class is the stationary policies. A policy u is said to be *stationary* if there is a vector $\mathbf{f} = \{f_{xa}, x \in S, a \in A\}$ such that

$$u_a^m(x_1, a_1, \ldots, x_{m-1}, a_{m-1}, x) = f_{xa}$$

for all histories and all $m \geq 1$. An even more narrow class of policies is that of the pure or nonrandomized stationary policies. A stationary policy f is said to be *pure* if, for each state $x \in S$, there is an action $a \in B$ such that $f_{xa} = 1$. Clearly, a pure policy can be represented as a mapping g from the state-space S to the action-space B.

We will repeatedly make use of the following property. Under any stationary policy f, the state process $\{X_m\}$ is a Markov chain with transition matrix $P(\mathbf{f})$, whose xy component is given by

$$P_{xy}(f) := \sum_a P_{xay} f_{xa}.$$

A transition matrix $P(\mathbf{f})$ is said to be *unichain* if it has at most one recurrent class; we will write $\pi(\mathbf{f}) = [\pi_x(\mathbf{f}), x \in S]$ for the unique equilibrium vector associated with $P(\mathbf{f})$.

1.4. Recurrent Structures and Facts

In this paper, we will study the sample-path criterion for three classes of MDP problems: unichain, communicating and deterministic. An MDP is said to be *unichain* if the transition matrix $P(\mathbf{g})$ is unichain for each pure policy g (this implies that $P(\mathbf{f})$ is unichain for all stationary policies f). An MDP is said to be *communicating* if for each ordered pair of states (x, y) there exists a pure policy g such that y is accessible from x under $P(\mathbf{g})$. Finally, an MDP is said to be *deterministic* if for every state $x \in S$ and action $a \in B$ there is a state $y \in S$ such that $P_{xay} = 1$.

We conclude this section by stating two facts that will be needed subsequently. The first fact is usually referred to as the Stability Theorem or the Strong Law of Large Numbers for Martingale Differences (e.g., see Loeve 1978, p. 53). The second fact is derived by

observing that under any stationary policy the process $\{(X_n, A_n), n = 1, 2, \ldots\}$ is a Markov chain, to which a well known ergodic theorem can be applied (e.g., see Cinlar 1975, p. 159). The notation $1(\cdot)$ represents the indicator function.

Fact 1. For every policy $\mathbf{u}$, we have P_u-almost surely

$$\lim_n \frac{1}{n} \sum_{m=2}^{n} \left[1(X_m = y) - \sum_{x,a} 1(X_{m-1} = x, A_{m-1} = a) P_{xay} \right] = 0$$

for all $y \in S$.

Fact 2. Define the random variables denoting the state-action frequencies

$$Z_n(x, a) := \frac{1}{n} \sum_{m=1}^{n} 1(X_m = x, A_m = a).$$

Under any stationary policy $\mathbf{f}$, the sequence $\{Z_n(x, a), n = 1, 2, \ldots\}$ converges P_f-almost surely for all $x \in S$, $a \in B$ to a random variable $Z(x, a)$. If $P(\mathbf{f})$ is unichain, then P_f-almost surely $Z(x, a) = \pi_x(\mathbf{f}) f_{xa}$ for all $x \in S$, $a \in B$.

2. THE TWO CRITERIA COMPARED

The expected average reward $\phi_s(\mathbf{u})$ is, in general, less than or equal to the average expected reward $\phi_e(\mathbf{u})$, but for a large class of policies the two rewards are equal.

Proposition 1. (i) *$U_s \subseteq U_e$, and for all policies u we have $\phi_s(\mathbf{u}) \leqslant \phi_e(\mathbf{u})$; consequently, $\phi_s^* \leqslant \phi_e^*$.*

(ii) *If $\{Z_n(x, a), n = 1, 2, \ldots\}$ converges P_u-almost surely for all $x \in S$, $a \in B$ to a random variable $Z(x, a)$, then P_u-almost surely*

$$R = \sum_{x,a} r(x, a) Z(x, a) \tag{8}$$

and $\phi_s(\mathbf{u}) = \phi_e(\mathbf{u})$.

(iii) *For any stationary policy f, Equation 8 holds P_f-almost surely and $\phi_s(\mathbf{f}) = \phi_e(\mathbf{f})$. If $P(\mathbf{f})$ is unichain, then P_f-almost surely*

$$R = \phi_s(\mathbf{f}) = \phi_e(\mathbf{f}) = \sum_{x,a} r(x, a) \pi_x(\mathbf{f}) f_{xa}.$$

Proof. For i note that $\mathbf{u} \in U_s$ implies $E_u[C] \leqslant \alpha$, which, when combined with Fatou's Lemma (see Billingsly 1979, p. 180) gives $\mathbf{u} \in U_e$; $\phi_s(\mathbf{u}) \leqslant \phi_e(\mathbf{u})$ follows immediately from Fatou's Lemma. For ii observe that

$$R = \liminf_n \frac{1}{n} \sum_{m=1}^{n} r(X_m, A_m) = \liminf_n \sum_{x,a} r(x, a) Z_n(x, a)$$

$$= \sum_{x,a} r(x, a) Z(x, a)$$

whence the first statement. The second statement then follows from Lebesgue's Dominated Convergence Theorem. Result iii is obtained by combining ii with Fact 2.

Consider the constrained optimization problems. The following example demonstrates that there may not exist an ϵ-optimal stationary policy for some $\epsilon > 0$ for the expectation problem, whereas, for the same MDP, there exists an optimal pure policy for the sample-path problem.

Example 1. Consider the following deterministic MDP with $S = \{1, 2\}$, $B = \{1, 2\}$, initial distribution $q_1 = 1$, $q_2 = 0$ and constraint value $\alpha = \frac{1}{2}$ (see Figure 1). The single and double arrowheads correspond to actions 1 and 2, respectively. From the figure, we see that action 1 brings the process back to state 1, whereas action 2 brings the process to state 2. In state 2, both actions bring the process back to state 2; thus, state 2 is absorbing under all policies. Observe that this MDP is not communicating, nor is it unichain, since the pure policy

$$\mathbf{g} = [g(1), g(2)] = [1, 1]$$

gives rise to a transition matrix $P(\mathbf{g})$ that has two recurrent classes. For the expectation and sample-path criteria, all feasible stationary policies $\mathbf{f}$ must have $f_{11} = 1$. Thus, the average expected reward and cost are given by $\phi_e(\mathbf{f}) = 0$ and $K(\mathbf{f}) = 0$, respectively, for all $\mathbf{f} \in U_e$. Consider the policy $\mathbf{u}$ given by $u_1(1) = \frac{1}{2}$, $u_1^m(x_1, a_1, \ldots, x_m) = 1$ for all histories and all $m \geqslant 2$. This policy belongs to U_e since $K(\mathbf{u}) = \frac{1}{2}$, and gives $\phi_e(\mathbf{u}) = \frac{1}{2}$. Hence, there does not exist an ϵ-optimal stationary policy (if ϵ satisfies $0 \leqslant \epsilon < \frac{1}{2}$) for

Figure 1. For this example, the sample-path criterion has an optimal pure policy, whereas the expectation criterion does not have nearly optimal stationary policies.

the expectation criterion. But, for the sample-path criterion, the above policy g is an optimal pure policy (with $t = 0$).

3. PRELIMINARY RESULTS

We begin by defining for each fixed $\eta \geq 0$ the following LP with decision variables $\{z(x, a), x \in S, a \in B\}$.

Program Q_η

$$t_\eta := \max \sum_{x,a} r(x, a)z(x, a) \tag{9a}$$

subject to

$$\sum_{x,a} P_{xay} z(x, a) = \sum_{a} z(y, a) \quad \text{for all } y \in S \tag{9b}$$

$$\sum_{x,a} z(x, a) = 1 \tag{9c}$$

$$\sum_{x,a} c(x, a)z(x, a) \leq \alpha \tag{9d}$$

$$z(x, a) \geq \eta \quad \text{for all } x \in S, \quad a \in B. \tag{9e}$$

With a slight abuse of notation we also refer to Q_η as the set of feasible solutions to the above LP. Loosely speaking, the limits of the state-action frequencies $\{Z_n(x, a), n = 1, 2, \ldots\}$ are P_u-almost surely feasible solutions to Q_0 for any $u \in U_s$. The decision variable $z(x, a)$, therefore, has the interpretation of being the fraction of time that the process is in state x and action a is chosen along a *feasible* sample-path $\omega \in \Omega$. For a given solution $\{z(x, a)\}$, we will write

$$z(x) = \sum_{a} z(x, a)$$

and

$$I_z := \{x \in S: z(x) > 0\}.$$

The following two key propositions relate the LPs Q_η to the sample-path criterion. Note that the first theorem holds for MDPs with general recurrent structures.

Proposition 2. *If U_s is nonempty, then Q_0 is nonempty and $P_u(R \leq t_0) = 1$ for all $u \in U_s$.*

Proof. Since $0 \leq Z_n(x, a) \leq 1$, it follows from standard compactness arguments that for each $\omega \in \Omega$, there is a subsequence $\{N_k(\omega)\}$ along which $\{Z_n(x, a; \omega), n = 1, 2, \ldots\}$ converges to some $Z(x, a; \omega)$ for all $x \in S$, $a \in B$, that is

$$\lim_{k\to\infty} Z_{N_k}(x, a) = Z(x, a) \quad \text{for all } x \in S, a \in B. \tag{10}$$

Now let $u \in U_s$. We first show that, P_u-almost surely, $\{Z(x, a)\}$ is a feasible solution for Q_0. It follows from (10) that

$$\sum_{x,a} Z(x, a) = 1, \quad Z(x, a) \geq 0 \quad x \in S, \quad a \in B. \tag{11}$$

Combining (10) with Fact 1, we have P_u-almost surely

$$\sum_{x,a} P_{xay} Z(x, a) = \sum_{a} Z(y, a) \quad \text{for all } y \in S. \tag{12}$$

Since $u \in U_s$, we also have P_u-almost surely

$$\alpha \geq \limsup_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} c(X_m, A_m) \geq \lim_{k\to\infty} \frac{1}{N_k} \sum_{m=1}^{N_k} c(X_m, A_m)$$

$$= \lim_{k\to\infty} \sum_{x,a} c(x, a)Z_{N_k}(x, a) = \sum_{x,a} c(x, a)Z(x, a). \tag{13}$$

From (11)–(13) we see that $\{Z(x, a)\}$ is, P_u-almost surely, feasible for Q_0. Consequently, Q_0 is nonempty and P_u-almost surely

$$\sum_{x,a} r(x, a)Z(x, a) \leq t_0. \tag{14}$$

The proof is then completed by combining (14) with

$$R \leq \lim_{k\to\infty} \frac{1}{N} \sum_{m=1}^{N_k} r(X_m, A_m) = \sum_{x,a} r(x, a)Z(x, a).$$

It follows from Proposition 2 that a policy ν is optimal [ϵ-optimal] for the sample-path criterion if it is feasible and $P_\nu(R = t_0) = 1$ [$P_\nu(R \geq t_0 - \epsilon) = 1$].

Proposition 3. *Fix $\eta \geq 0$ and let $\{z^\eta(x, a)\}$ be an optimal extreme point for Q_η. Define a policy f^η by the transformation*

$$f^\eta_{xa} = \begin{cases} \dfrac{z^\eta(x, a)}{z^\eta(x)} & z^\eta(x) > 0 \\ \dfrac{1}{|B|} & \text{otherwise.} \end{cases} \tag{15}$$

Then

$$\sum_{x} z^\eta(x)P_{xy}(f^\eta) = z^\eta(y) \tag{16}$$

$$\sum_{x} z^\eta(x) = 1. \tag{17}$$

Moreover, if $P(f^\eta)$ is unichain, then $f^\eta \in U_s$ and $P_{f^\eta}(R = t_\eta) = 1$. In particular, if $P(f^0)$ is unichain, then f^0 is an optimal stationary policy.

Proof. We have

$$\sum_{x} z^\eta(x)P_{xy}(f^\eta) = \sum_{x,a} z^\eta(x)P_{xay}f^\eta_{xa}$$

$$= \sum_{x,a} P_{xay} z^\eta(x, a) = z^\eta(y)$$

where the last equality follows from (9b). By (9c), we also have

$$\sum_x z^\eta(x) = 1.$$

If $P(\mathbf{f}^\eta)$ is unichain, there is a unique probability vector $\pi(\mathbf{f}^\eta)$ associated with $P(\mathbf{f}^\eta)$; therefore, (16)–(17) imply that $\pi_x(\mathbf{f}^\eta) = z^\eta(x)$ for all $x \in S$. Combining this with statement iii of Proposition 1 we have P_{f^η}-almost surely

$$C = \sum_{x,a} c(x, a)\pi_x(\mathbf{f}^\eta)f^\eta_{xa} = \sum_{x,a} c(x, a)z^\eta(x, a) \leq \alpha$$

where the last equality follows from (9d). In a similar manner, we have P_{f^η}-almost surely

$$R = \sum_{x,a} r(x, a)z^\eta(x, a) = t_\eta$$

where the last inequality follows from the optimality of $\{z^\eta(x, a)\}$. The last statement then follows from Proposition 2.

We shall see that, in most circumstances, for η sufficiently small and positive, $\mathbf{f}^\eta$ is an ϵ-optimal policy when the MDP is communicating. Nevertheless, it follows from Proposition 3 that there exists an optimal stationary policy for a class of MDPs.

Theorem 1. *Suppose that the MDP is unichain. Then U_s is nonempty if and only if Q_0 is nonempty. If Q_0 is nonempty then $\mathbf{f}^\eta$ as defined through* (15) *is optimal for the sample-path criterion.*

4. THE COMMUNICATING CASE

In contrast with the unichain case, there does not, in general, exist an optimal stationary policy for the communicating case.

Example 2. Consider the following communicating MDP with $S = \{1, 2\}$, $B = \{1, 2\}$, initial distribution $q_1 = 1$, $q_2 = 0$ and constraint value $\alpha = -\frac{1}{2}$ (see Figure 2). Consider the (weaker) sample-path criterion of maximizing $\phi_s(u)$ over all policies $u \in U_s$. This Markov decision problem can be given the following interpretation: find a policy to maximize the fraction of time being in state 1 and choosing action 1, subject to the constraint that the fraction of time being in state 2 and choosing action 2 be at least 50%.

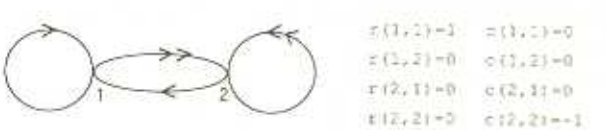

Figure 2. In general, an optimal stationary policy for communicating MDPs does not exist.

For any stationary policy $\mathbf{f}$, the corresponding transition matrix is given by

$$P(\mathbf{f}) = \begin{bmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \end{bmatrix}.$$

If $f_{12} = 0$, then $\mathbf{f}$ does not belong to U_s; if $f_{12} > 0$, then $P(\mathbf{f})$ is guaranteed to have at most one recurrent class with a corresponding equilibrium vector

$$\pi(\mathbf{f}) = \left[\frac{f_{21}}{f_{12} + f_{21}}, \frac{f_{12}}{f_{12} + f_{21}}\right]. \tag{18}$$

Combining (18) along with statement iii of Proposition 2, it is a simple exercise to show that $\phi_s(\mathbf{f}) < \frac{1}{2}$ if $\mathbf{f} \in U_s$, and that the supremum of $\phi_s(\mathbf{f})$ over all $\mathbf{f} \in U_s$ is $\frac{1}{2}$. Hence, there does not exist, in general, an optimal stationary policy for the sample-path criterion.

For the remainder of this section we suppose that the MDP is communicating. At this point, it is convenient to introduce two lemmas whose proofs are left to the reader.

Lemma 1. *An MDP is communicating if and only if $P(\mathbf{f})$ is irreducible for every stationary policy $\mathbf{f}$ that satisfies $f_{xa} > 0$, $x \in S$, $a \in B$.*

Lemma 2. *Suppose that $\mathbf{f}$ is a stationary policy and Λ is a set of states such that: (i) Λ is closed under $P(\mathbf{f})$; and (ii) $f_{xa} > 0$ for all $x \notin \Lambda$, $a \in B$. Then $S - \Lambda$ is transient under $P(\mathbf{f})$.*

Let us consider, with the communicating assumption in force, some properties of the policy $\mathbf{f}^0$, which is derived from the feasible solution $\{z^0(x, a)\}$ via transformation (15). It follows from (16) that, under $P(\mathbf{f}^0)$, I_{z^0} is a closed set and each $x \in I_{z^0}$ is a recurrent state. Furthermore, from Lemma 2 we have that all states in $S - I_{z^0}$ are transient under $P(\mathbf{f}^0)$. Thus

$$\bigcup_{k=1}^{m} \Gamma_k = I_{z^0}$$

where $\Gamma_1, \Gamma_2, \ldots, \Gamma_m$ are the recurrent classes associated with $P(\mathbf{f}^0)$. We will need to make use of the following: for all $y \in \Gamma_k$ we have

$$\sum_{x \in \Gamma_k} z^0(x)P_{xy}(\mathbf{f}^0) = \sum_x z^0(x)P_{xy}(\mathbf{f}^0) = z^0(y). \tag{19}$$

The first equality in (19) follows from $z^0(x) = 0$ for $x \notin I_{z^0}$ and from $\bigcup_{k \leq s} \Gamma_k$ is closed under $P(\mathbf{f}^0)$; the last equality follows from (16). We also need to introduce

for $k = 1, \ldots, m$

$$\lambda_k := \sum_{x \in \Gamma_k} z^0(x).$$

Note that

$$\sum_{k=1}^{m} \lambda_k = 1 \quad \text{and} \quad \lambda_k > 0 \quad k = 1, \ldots, m. \tag{20}$$

Also define

$$d_k := 1/\lambda_k \sum_{x \in \Gamma_k} \sum_{a \in B} c(x, a) z^0(x, a)$$

and let l be such that $d_l = \min_{1 \le k \le m} d_k$. The value d_k has the interpretation of being the average cost of f^0 *given* that the process has entered Γ_k. Since $\{z^0(x, a)\}$ is feasible for Q_0 we have

$$\alpha \ge \sum_{x,a} c(x, a) z^0(x, a) = \sum_{k=1}^{m} \lambda_k d_k \ge d_l \tag{21}$$

where the last inequality follows from (20).

It is first of interest to construct a sample-path feasible policy. The policy f^0 does not necessarily belong to U_s because we may have $d_k > \alpha$ for some k. We are, therefore, motivated to define the stationary policy $\tilde{f}$ as

$$\tilde{f}_{xa} = \begin{cases} f^0_{xa} & \text{if } x \in \Gamma_l \\ \dfrac{1}{|B|} & \text{if } x \notin \Gamma_l. \end{cases} \tag{22}$$

Since $P_{xy}(\tilde{f}) = P_{xy}(f^0)$ for $x \in \Gamma_l$, it follows that Γ_l is a recurrent class for $P(\tilde{f})$. Moreover, since $\tilde{f}_{xa} > 0$ for all $x \notin R_l$, $a \in B$, it follows from Lemma 2 that Γ_l is the only recurrent class for $P(\tilde{f})$. Thus, $P(\tilde{f})$ is unichain.

The existence of a sample-path feasible policy is the subject of the following lemma.

Lemma 3. *If Q_0 is nonempty, then U_s is nonempty. In particular, we have P_f-almost surely*

$$C = \sum_{x,a} c(x, a) \pi_x(\tilde{f}) \tilde{f}_{xa} \le \alpha \tag{23}$$

so that $\tilde{f} \in U_s$.

Proof. Observe that for $y \in \Gamma_l$ we have from (19) and the definition of $\tilde{f}$

$$\sum_{x \in \Gamma_l} z^0(x) P_{xy}(\tilde{f}) = z^0(y).$$

Thus, for all $x \in \Gamma_l$

$$\pi_x(\tilde{f}) = \frac{z^0(x)}{\lambda_l}$$

whereas $\pi_x(\tilde{f}) = 0$ for $x \notin \Gamma_l$. Therefore

$$\sum_{x,a} c(x, a) \pi_x(\tilde{f}) \tilde{f}_{xa} = d_l \le \alpha.$$

In light of statement iii of Proposition 1, the proof is therefore complete.

For the remainder of this section, we assume that there exists a feasible policy so that Q_0 is nonempty. The following proposition gives a simple condition to guarantee the existence of ϵ-optimal stationary policies; subsequently, we will show that the condition holds whenever there exists a *strictly feasible* policy. We will also show that if the condition does not hold, then there exists an optimal stationary policy.

In order to state the theorem, let V be the set of feasible solutions $\{z(x, a)\}$ in Q_0 that satisfy $z(x, a) > 0$ for all $x \in S$, $a \in B$. We note that if V is nonempty, then there is an $\zeta > 0$ such that for each η that satisfies $0 < \eta < \zeta$ there is a feasible solution for the LP Q_η.

Proposition 4. *Suppose that V is nonempty. Then for each $\epsilon > 0$, there exists an ϵ-optimal stationary policy for the sample-path criterion.*

Proof. From Lemma 1 it follows that $P(f^\eta)$ is irreducible for all $0 < \eta < \zeta$. From Proposition 3, we obtain $f^\eta \in U_s$ for all $0 < \eta < \zeta$ and $P_{f^\eta}(R = t_\eta) = 1$. Therefore, in view of Proposition 2, it remains to show that

$$\lim_{\eta \downarrow 0} t_\eta = t_0.$$

But the LPs Q_η with $0 \le \eta < \zeta$ can be regarded as a parametric right-hand side LP (e.g., see Murty 1983) and the desired continuity holds.

The following theorem implies that there *usually* exists an ϵ-optimal stationary policy for all $\epsilon > 0$, i.e., whenever there exists a policy that strictly meets the sample-path constraint.

Proposition 5. *Suppose that there exists a policy* u *and a $\nu > 0$ such that*

$$P_u(C \le \alpha - \nu) = 1. \tag{24}$$

Then V is nonempty.

Proof. From Proposition 2, with α replaced by $\alpha - \nu$, (24) implies that there is a $\{\hat{z}(x, a)\}$ that belongs to Q_0 such that

$$\sum_{x,a} c(x, a) \hat{z}(x, a) \le \alpha - \nu. \tag{25}$$

The basic idea of the proof is to first construct a unichain policy $\hat{f}$ from $\{\hat{z}(x, a)\}$ that satisfies (24); then perturb $\hat{f}$ in order to construct a feasible policy

that chooses every action with positive probability and gives rise to an irreducible Markov chain. The corresponding state-action frequencies will belong to V.

Recall the construction of $\tilde{\mathbf{f}}$ from $z^0(x, a)$. In an entirely analogous manner, construct the unichain stationary policy $\hat{\mathbf{f}}$ from $\hat{z}(x, a)$. From (25) we arrive at the analog of (23)

$$\sum_{x,a} c(x, a)\pi_x(\hat{\mathbf{f}})\hat{f}_{xa} \leqslant \alpha - \nu.$$

Let Γ be the recurrent class associated with $P(\hat{\mathbf{f}})$. For each $\epsilon > 0$, define the stationary policy $\mathbf{h}^\epsilon$ as

$$h^\epsilon_{xa} := \begin{cases} \hat{f}_{xa} & \text{for } x \notin \Gamma \text{ and } a \in B \\ (1 - \epsilon b_x)\hat{f}_{xa} & \text{for } x \in \Gamma \text{ and } a \notin B_x \\ \epsilon & \text{for } x \in \Gamma \text{ and } a \in B_x \end{cases}$$

where b_x is the cardinality of the set $B_x := \{a \in B: \hat{f}_{xa} = 0\}$. It follows from Lemma 1 and the fact

$$h^\epsilon_{xa} > 0 \quad x \in S, \quad a \in B$$

that $P(\mathbf{h}^\epsilon)$ is irreducible for all $\epsilon > 0$.

For $x \notin \Gamma$, we have

$$P_{xy}(\mathbf{h}^\epsilon) = P_{xy}(\hat{\mathbf{f}})$$

and for $x \in \Gamma$, we have

$$P_{xy}(\mathbf{h}^\epsilon) = (1 - \epsilon b_x) \sum_{a \notin B_x} P_{xay}\hat{f}_{xy} + \epsilon \sum_{a \in B_x} P_{xay}$$

$$= (1 - \epsilon b_x)P_{xy}(\hat{\mathbf{f}}) + \epsilon \sum_{a \in B_x} P_{xay}.$$

Thus

$$\lim_{\epsilon \downarrow 0} P_{xy}(\mathbf{h}^\epsilon) = P_{xy}(\hat{\mathbf{f}}) \tag{26}$$

for all $x, y \in S$. Since $P(\mathbf{h}^\epsilon)$, $\epsilon > 0$, and $P(\hat{\mathbf{f}})$ are all unichain transition matrices, it follows from (26) that (e.g., see Lemma 2.4 of Beutler and Ross 1985)

$$\lim_{\epsilon \downarrow 0} \pi(\mathbf{h}^\epsilon) = \pi(\hat{\mathbf{f}}).$$

This, in turn, implies

$$\lim_{\epsilon \downarrow 0} \sum_{x,a} c(x, a)\pi_x(h^\epsilon)h^\epsilon_{xa}$$

$$= \sum_{x,a} c(x, a)\pi_x(\hat{\mathbf{f}})\hat{f}_{xa} \leqslant \alpha - \nu.$$

Therefore, there exists a $\gamma > 0$ such that

$$\sum_{x,a} c(x, a)\pi_x(\mathbf{h}^\gamma)h^\gamma_{xa} \leqslant \alpha. \tag{27}$$

Thus, we have constructed the desired feasible policy $\mathbf{h}^\gamma$ that chooses each action with positive probability and is irreducible. To complete the proof, define for all $x \in S$, $a \in B$

$$w(x, a) := h^\gamma_{xa}\pi_x(\mathbf{h}^\gamma) > 0. \tag{28}$$

Utilizing (27) and (28), it is then readily verified that $\{w(x, a)\} \in V$.

From the previous theorem we see that for all practical purposes we may take V to be nonempty. For mathematical completeness, however, we need to address the degenerate case of V being empty in the following proposition.

Proposition 6. *Suppose that V is empty. Then the stationary policy $\mathbf{f}^0$ given by the transformation* (15) *is an optimal policy for the sample-path criterion.*

Proof. Recall the definitions of Γ_k, λ_k, $k = 1, \ldots, m$. Denote

$$w^k(x, a) := \begin{cases} \dfrac{z^0(x, a)}{\lambda_k} & \text{if } x \in \Gamma_k, \ a \in B \\ 0 & \text{otherwise.} \end{cases}$$

Note that

$$z^0(x, a) = \sum_{k=1}^{m} \lambda_k w^k(x, a). \tag{29}$$

Under policy $\mathbf{f}^0$, $w^k(x, a)$ is the fraction of time that the process is in state x and action a is chosen *given* that the process has entered Γ_k. We clearly have for all $k = 1, \ldots, m$

$$\sum_{x,a} w^k(x, a) = 1, \quad w^k(x, a) \geqslant 0.$$

From (19) we obtain for all $y \in S$

$$\sum_{x,a} w^k(x, a)P_{xy}(\mathbf{f}^0) = \sum_{a} w^k(y, a).$$

Thus, $\{w^k(x, a)\}$ satisfies (9b), (9c) and (9e) with $\eta = 0$, for all $k = 1, \ldots, m$. Moreover, since for all $k = 1, \ldots, m$

$$d_k = \sum_{x,a} c(x, a)w^k(x, a)$$

it follows from (21) that $\{w^l(x, a)\}$ satisfies (9d). Thus $\{w^l(x, a)\}$ belongs to Q_0.

We make use of Proposition 5 in contraposition form: If V is empty, then for every $\mathbf{u} \in U_x$ and for every $\nu > 0$, $P_{\mathbf{u}}(C \leqslant \alpha - \nu) < 1$. In particular, $P_{\tilde{\mathbf{f}}}(C \leqslant \alpha - \nu) < 1$ for all $\nu > 0$, where $\tilde{\mathbf{f}}$ is defined by (22). This combined with Lemma 3 gives $d_l = \alpha$, which, in turn, implies that $d_k \geqslant \alpha$ for all $k = 1, \ldots, m$. From (21) we then obtain $d_k = \alpha$ for all $k = 1, \ldots, m$. Thus, $\{w^k(x, a)\}$ is a solution to Q_0 for each $k = 1, \ldots, m$. But since $\{z^0(x, a)\}$ is an extreme point for Q_0, it then follows from (29) that we must have $m = 1$, i.e., $P(\mathbf{f}^0)$

is unichain. Hence, by Proposition 3, f^0 is an optimal stationary policy.

Combining Propositions 4, 5 and 6 we obtain the following theorem.

Theorem 2. *Suppose that the MDP is communicating. Then U_ϵ is nonempty if and only if Q_0 is nonempty. If Q_0 is nonempty, then for each $\epsilon > 0$ there exists an ϵ-optimal stationary policy for the sample-path criterion.*

We summarize our study of communicating MDPs with the following procedure to locate an optimal or nearly optimal stationary policy.

Step 1. Solve the LP Q_0 by the simplex method. If Q_0 is infeasible, then there does not exist a policy that meets the sample-path constraint.

Step 2. Let $\{z^0(x, a)\}$ be an optimal extreme point for the LP Q_0, and let f^0 be the corresponding stationary policy obtained via the transformation (15). If $P(f^0)$ is unichain, then f^0 is an optimal stationary policy, stop; otherwise, go to Step 3.

Step 3. Solve the parametric LP $(Q^\eta, \eta \geq 0)$ over some interval $[0, \xi]$ beginning with $\eta = 0$. Then employ the transformation (15) to obtain an ϵ-optimal stationary policy for ϵ as small as desired.

Remark. The results of this section hold if there are state-dependent action spaces B_x, $x \in S$, instead of a single action-space B.

5. DETERMINISTIC MDPs

For multichain MDPs, it is convenient to require that the given initial distribution be concentrated on a single state, say $x_1 \in S$. In this section, we make this assumption. We also assume that the MDP is deterministic.

In this case, it is useful to construct a directed graph from the MDP as follows: the set of vertices is equal to the state-space S; there is a directed arc from vertex x to vertex y if and only if there is an $a \in B$ such that $P_{xay} = 1$. Note that there is at least one arc emanating from every vertex $x \in S$ and the graph may have self-loops. Since only states that are reachable from the initial state x_1 under some policy are of interest, we assume without loss of generality that for each $y \in S$ there is a directed path from x_1 to y.

In a strongly connected component of the above graph, for any two vertices x and y such that $x \neq y$, there exists a directed path from x to y. Let us consider those strongly connected components which either have more than one state or a single state with a self-loop. For any two states x and y in a component of this kind there exists a pure policy g such that y is accessible from x under $P(g)$ through states only in the same component. Thus, the MDP *restricted* to this component is communicating. This motivates us to partition the set of states into disjoint classes C_1, $C_2, \ldots, C_m$ such that the subgraph restricted to each of these classes is a strongly connected component. The partition can be constructed via efficient depth-first algorithms (e.g., see Gondran and Minoux 1984, pp. 16–21). The idea of partitioning the state-space into communicating MDPs is due to Bather (1973).

For each $k = 1, 2, \ldots, m$ and for each $x \in C_k$, define the set

$$B_x := \{a \in B: P_{xay} = 1 \text{ for some } y \in C_k\}.$$

Thus, starting in a state $x \in C_k$, B_x is the set of actions that will keep the state process in the class C_k. Observe that if B_x is empty for some $x \in C_k$, then C_k is a singleton that consists of the state x, which is transient under all policies (in fact, x will be visited at most once under any policy). Let

$$L := \{k: 1 \leq k \leq m, B_x \text{ is nonempty for all } x \in C_k\}.$$

For each $k \in L$, we define a new MDP, denoted MDP-k, with state-space C_k and state-dependent action space B_x for each $x \in C_k$; the law of motion for the MDP-k is the law of motion for the original MDP restricted to the state-space C_k and action-spaces B_x, $x \in C_k$. It follows from the definition of C_k that the MDP-k is communicating for each $k \in \Gamma$. Moreover, once the state process leaves a given class C_k, it can never return to that class. This, combined with the assumption that all states are reachable from x_1, implies that an optimal policy will drive the state process to the class C_k which offers the highest constrained reward.

We are, therefore, motivated to define for each $k \in L$ and $\eta > 0$ the following LP.

Program Q_η^k

$$t_\eta^k := \max \sum_{x \in C_k} \sum_{a \in B_x} r(x, a) z(x, a)$$

subject to

$$\sum_{x \in R_k} \sum_{a \in B_x} P_{xay} z(x, a) = \sum_{a \in B_y} z(y, a) \quad \text{for all } y \in C_k$$

$$\sum_{x \in R_k} \sum_{a \in B_x} z(x, a) = 1$$

$$\sum_{x \in R_k} \sum_{a \in B_x} c(x, a) z(x, a) \leq \alpha$$

$$z(x, a) \geq \eta \quad \text{for all } x \in C_k, a \in B_x.$$

We also need to let l be the index such that $t_0^l = \max\{t_0^k : k \in L\}$. We construct an ϵ-optimal stationary policy as follows. The standard depth-first search algorithm easily can be modified to give as a by-product a directed path from x_1 to a state in C_l. Let $S' = \{s_1, s_2, \ldots, s_n\}$ be the set of states in this path so that $s_1 = x_1$ and $s_n \in C_l$. Let $a_1, a_2, \ldots, a_{n-1}$ be the actions such that $P_{s_m a_m s_{m+1}} = 1$ for $m = 1, 2, \ldots, n-1$. Finally, let $\mathbf{h}$ be a stationary policy for the original MDP defined as: when in state $x_m \in S'$, $\mathbf{h}$ chooses a_m; when in state $x \in C_l$, $\mathbf{h}$ follows the ϵ-optimal stationary policy for the communicating MDP-l; when in state $S - S' - C_l$, any action is chosen.

Corollary 1. *The stationary policy* $\mathbf{h}$ *is an* ϵ*-optimal policy for the sample-path criterion.*

Remark. If the LP Q_0^l gives rise to an optimal policy with at most one recurrent class, then there exists an optimal stationary policy for the deterministic MDP with the sample-path criterion. Also, since the optimal (respectively, ϵ-optimal) depends on the initial state, these policies are not uniformly optimal (respectively, ϵ-optimal) for all initial distributions, as in the unichain and communicating cases.

6. DISCUSSION AND CONCLUSION

The theory developed up to this point is incomplete because it has neither addressed MDPs with multiple sample-path constraints nor constrained MDPs with general recurrent structures. The latter problem is investigated in a companion paper (Ross and Varadarajan), where it is shown that there always exists an ϵ-optimal stationary policy for the sample-path criterion for all $\epsilon > 0$. The approach taken there is similar to that for deterministic MDPs, where the state-space is decomposed in order to create classes of communicating MDPs. However, the situation is significantly more complicated because, with positive probability, more than one state can be directly accessed for a given choice of action.

6.1. Multiple Constraints

Multiple sample-path constraints can be handled, to a large extent, by the theory presented herein; they were omitted in order not to obscure the discussion with a complicated notation. To see how this more general formulation can be treated, we introduce the multiple sample-path constraints

$$P_{\mathbf{u}}\left(\limsup_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} c^i(X_m, A_m) \leq \alpha_i\right) = 1$$

for all $i = 1, \ldots, I$. With the addition of the multiple constraints, one can then redefine the LPs Q_η, $\eta \geq 0$, in the obvious manner. In doing this, it is easily seen that all of the results through Section 3 continue to hold. In particular, if Q_0 is nonempty then there exists an optimal stationary policy for unichain MDPs.

For communicating MDPs, it is not, in general, true that $Q_0 \neq \emptyset$ implies the existence of a stationary policy in U_r if there is more than one constraint. This can be verified with an example along the lines of Example 2. However, a slight modification of the arguments in Section 4 give the following.

Theorem 3. *Suppose that the MDP is communicating. If there exists a policy* $\mathbf{u}$ *and a* $\nu > 0$ *such that*

$$P_{\mathbf{u}}\left(\liminf_{n} \frac{1}{n} \sum_{m=0}^{n} c^i(X_m, A_m) \leq \alpha_i - \nu\right) = 1$$

for all $i = 1, 2, \ldots, I$, *then for any* $\epsilon > 0$ *there exists an* ϵ*-optimal stationary policy for the sample-path criterion.*

6.2. The Expectation Criterion Revisited

Proposition 2 implies that $\phi_r(\mathbf{u}) \leq t_0$ for all $u \in U_r$. This can be strengthened as follows: apply Lebesgue's Dominated Convergence Theorem to the Stability Theorem (Fact 1) to obtain

$$\lim_{n} \frac{1}{n} \sum_{m=2}^{n} \left[P_{\mathbf{u}}(X_m = y) - \sum_{x,a} P_{\mathbf{u}}(X_{m-1} = x, A_{m-1} = a) P_{xay}\right] = 0$$

for all policies $\mathbf{u}$. Using this in place of the Stability Theorem in the proof of Proposition 2 gives $\phi_r(\mathbf{u}) \leq t_0$ for all $\mathbf{u} \in U_r$ (this result also follows from Theorem 4.7.3 of Kallenberg, where expected state-action frequencies are employed). Employing this result and the arguments of Sections 3 and 4, Theorems 1 and 2 can be established for the expectation criterion. However, as Example 1 demonstrates, Corollary 1 does not hold true for the expectation criterion.

ACKNOWLEDGMENT

This work was partially supported by NSF grant NCR-8707620. We also thank the referees for their careful reading and comments on this paper.

REFERENCES

BATHER, J. 1973. Optimal Decision Procedures for Finite Markov Chains. Part III: General Convex Systems. *Adv. Appl. Prob.* **5**, 541–553.

BEUTLER, F. J., AND K. W. ROSS. 1985. Optimal Policies for Controlled Markov Chains With a Constraint. *J. Math. Anal. Appl.* **112**, 236–252.

BEUTLER, F. J., AND K. W. ROSS. 1986. Time-Average Optimal Constrained Semi-Markov Decision Processes. *Adv. Appl. Prob.* **18**, 341–359.

BILLINGSLY, P. 1979. *Probability and Measure.* John Wiley & Sons, New York.

CINLAR, E. 1975. *Introduction to Stochastic Processes.* Prentice-Hall, Englewood Cliffs, N.J.

DERMAN, C. 1970. *Finite State Markovian Decision Processes.* Academic Press, New York.

GOLABI, K., R. B. KULKARNI AND C. B. WAY. 1982. A Statewide Pavement Management System. *Interfaces* **12**, 6, 5–21.

GONDRAN, M., AND M. MINOUX. 1984. *Graphs and Algorithms.* John Wiley & Sons, New York.

HORDIJK, A., AND L. C. M. KALLENBERG. 1984. Constrained Undiscounted Dynamic Programming. *MOR* **9**, 276–289.

KALLENBERG, L. C. M. 1983. *Linear Programming and Finite Markovian Control Problems.* Mathematical Centre Tracts 148, Amsterdam.

KOLESAR, P. 1970. A Markovian Model for Hospital Admission Scheduling. *Mgmt. Sci.* **16**, 384–396.

LOEVE, M. 1978. *Probability Theory, Vol. 2* (4th ed.). Springer-Verlag, New York.

MAGLARIS, B., AND M. SCHWARTZ. 1982. Optimal Fixed Frame Multiplexing in Integrated Line- and Packet-Switched Communication Networks. *IEEE Trans. Inform. Theory* **IT-28**, 263–273.

MURTY, K. G. 1983. *Linear Programming.* John Wiley & Sons, New York.

NAIN, P., AND K. W. ROSS. 1986. Optimal Multiplexing of Heterogeneous Traffic With Hard Constraint. *Perf. Eval. Rev.* **14**.

ROSS, K. W. 1989. Randomized and Past-Dependent Policies for Markov Decision Processes With Multiple Constraints. *Opns. Res.* **37**, 474–477.

ROSS, K. W., AND R. VARADARAJAN. 1986. The Decomposition of Time-Average MDPs: Theory, Algorithms and Applications. Technical Report, Department of Systems, University of Pennsylvania, Philadelphia.