MATHEMATICS OF OPERATIONS RESEARCH
Vol. 16, No. 1, February 1991
Printed in U.S.A.

MULTICHAIN MARKOV DECISION PROCESSES WITH A SAMPLE PATH CONSTRAINT: A DECOMPOSITION APPROACH*

KEITH W. ROSS† AND RAVI VARADARAJAN

We consider finite-state finite-action Markov decision processes which accumulate both a reward and a cost at each decision epoch. We study the problem of finding a policy that maximizes the expected long-run average reward subject to the constraint that the long-run average cost be no greater than a given value with probability one. We establish that if there exists a policy that meets the constraint, then there exists an ϵ-optimal stationary policy. Furthermore, an algorithm is outlined to locate the ϵ-optimal stationary policy. The proof of the result hinges on a decomposition of the state space into maximal recurrent classes and a set of transient states.

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

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

subject to the sample path constraint

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

Here, $\{X_m\}$ is the state process taking values in the finite state-space; $\{A_m\}$ is the action process taking values in the finite action space; Δ is the initial state assumed to be fixed and given; $r(\cdot,\cdot)$ 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. We refer to the constraint (2) as the "sample path constraint".

For general multichain Markov Decision Processes (MDPs) we establish that if there exists a policy that meets the sample path constraint, then there exists an ϵ-optimal stationary policy. Furthermore, an algorithm to locate the ϵ-optimal policy is outlined. The results presented here rely on our results in [12] where communicating MDPs with a sample path constraint are considered.

The sample path constraint resembles a more commonly studied constraint of requiring the time average expected cost to be $\leqslant \alpha$ (see [5], [6], [7], [8], [15], [4], [11]), which we refer to as the "expectation constraint." Indeed, if the MDP satisfies the unichain condition, then an optimal policy with the sample path constraint is optimal with the expectation constraint and vice versa [12]. But the two criteria are different

*Received March 30, 1987; revised September 21, 1989.
AMS 1980 subject classification. Primary: 90C47.
IAOR 1973 subject classification. Main: Programming/Markov decision.
OR/MS Index 1978 subject classification. Primary: 118 Dynamic programming/Markov/finite state.
Key words. Markov decision processes, long-run average reward, sample path constraint, decompositions.
†Supported partially through NSF Grant NCR-8707620.

0364-765X/91/1601/0195/$01.25

for many nontrivial problems, as a policy which is optimal with the expectation constraint may not even satisfy the sample path constraint [12]. One frustrating property of the expectation constraint is that there does not in general exist an optimal or ϵ-optimal stationary policy [7], [8]. Therefore it is difficult, if not impossible, to develop efficient algorithms to locate ϵ-optimal policies in the presence of an expectation constraint.

The main idea behind the proof of our result is to decompose the state space into "strongly communicating classes" and a set of transient states. This decomposition is similar in spirit to a classification given by Bather [2]. For each of the strongly communicating classes, an ϵ-optimal stationary policy is then found in the corresponding "restricted MDP". Finally, the ϵ-optimal stationary policy for the original problem is obtained by solving an unconstrained "intermediate" MDP. The intermediate MDP is shown to be equivalent to an "aggregated" MDP, where there is one state for each strongly communicating class and one state for each transient state.

The decomposition of the state space into strongly communicating classes, along with the properties of this decomposition and the intermediate MDP, can be applied to other multichain MDP problems. The technique appears particularly appropriate for MDPs with nonstandard criteria, such as sample path constraints as studied here or variability sensitive MDPs as studied in [3].

This paper is organized as follows. We give our notation for MDPs in §2. In §3 the decomposition is introduced and is shown to have several interesting properties. An algorithm to locate the strongly communicating classes is outlined and an example is given. In §4 the existence of an ϵ-optimal stationary policy for MDPs with a sample path constraint is established. In §5, we briefly discuss the equivalence between the intermediate MDP and the aggregated MDP.

2. Preliminaries. Let $\mathcal{S}$ and $\mathcal{A}$ denote the finite state and action space, respectively. The underlying sample space for the MDP is

$$\Omega = \{(x_1, a_1, x_2, a_2, \dots) : x_n \in \mathcal{S}, a_n \in \mathcal{A} \text{ for all } n = 1, 2, \dots\}.$$

Throughout, the sample space Ω will be equipped with the σ-algebra generated by the random variables $(X_1, A_1, X_2, A_2, \dots)$. In this paper we consider optimizing over the set of all policies, which includes randomized as well as past-dependent policies (see [5], [8]). Denote p_{xay}, $x \in \mathcal{S}$, $a \in \mathcal{A}$, $y \in \mathcal{S}$, for the law of motion for MDP, i.e., for all policies $\mathbf{u}$ and all epochs n

$$P_{\mathbf{u}}(X_{n+1} = y \mid X_1, A_1, \dots, X_{n-1}, A_{n-1}, X_n = x, A_n = a) = p_{xay}.$$

In general a policy may be difficult to implement since it depends on the entire past history of the process. We are therefore motivated to study smaller and more appealing classes of policies. A policy $\mathbf{f}$ is said to be *stationary* if the choice of action depends only on the current state of the process; denote $f(x, a)$ for the probability of choosing action a when in state x. A stationary policy $\mathbf{f}$ is said to be *pure* or *nonrandomized* if for each $x \in \mathcal{S}$ there is an $a \in \mathcal{A}$ such that $f(x, a) = 1$. A pure policy can be represented by a mapping $\mathbf{g}$ from the state space $\mathcal{S}$ to the action space $\mathcal{A}$.

Under any stationary policy $\mathbf{f}$, the state process $\{X_n\}$ is a Markov chain with transition matrix $\mathbf{P}(\mathbf{f})$ with components given by

$$P_{xy}(\mathbf{f}) = \sum_{a \in \mathcal{A}} p_{xay} f(x, a).$$

A transition matrix $\mathbf{P}(\mathbf{f})$ is said to be *unichain* if it has at most one recurrent class plus a (perhaps empty) set of transient states.

3. Decomposition of MDPs.

DEFINITION 1. A set of states $\mathscr{C} \subseteq \mathscr{S}$ is said to be a *strongly communicating class* if (i) $\mathscr{C}$ is a recurrent class for some stationary policy; (ii) $\mathscr{C}$ is not a proper subset of some set $\mathscr{C}'$ for which (i) holds true.

Let $\mathscr{T}$ be the (possibly empty) set of states that are transient under all stationary policies. Let $\{\mathscr{C}_1, \ldots, \mathscr{C}_l\}$ be the collection of all strongly communicating classes.

PROPOSITION 1. *The collection of sets* $\{\mathscr{C}_1, \ldots, \mathscr{C}_l, \mathscr{T}\}$ *forms a partition of the state space* $\mathscr{S}$.

PROOF. It is clear that the collection covers $\mathscr{S}$ and that none of the strongly communicating classes intersect $\mathscr{T}$. It remains to show that if $\mathscr{C}$ and $\mathscr{C}'$ are two strongly communicating classes that are not disjoint, then $\mathscr{C} = \mathscr{C}'$. Let $\mathbf{f}$ (respectively $\mathbf{f}'$) be a stationary policy under which $\mathscr{C}$ (respectively $\mathscr{C}'$) is a recurrent class. Consider a policy $\tilde{\mathbf{f}}$ defined as follows:

$$\tilde{f}(x,a) = \begin{cases} f(x,a), & x \in \mathscr{C} - \mathscr{C}',\ a \in \mathscr{A}, \\ f'(x,a), & x \in \mathscr{C}' - \mathscr{C},\ a \in \mathscr{A}, \\ \frac{1}{2}\left[f(x,a) + f'(x,a)\right] & \text{otherwise.} \end{cases}$$

The transition matrix associated with $\tilde{\mathbf{f}}$ is given by

$$P_{xy}(\tilde{\mathbf{f}}) = \begin{cases} P_{xy}(\mathbf{f}), & x \in \mathscr{C} - \mathscr{C}', \\ P_{xy}(\mathbf{f}'), & x \in \mathscr{C}' - \mathscr{C}, \\ \frac{1}{2}\left[P_{xy}(\mathbf{f}) + P_{xy}(\mathbf{f}')\right] & \text{otherwise.} \end{cases} \tag{3}$$

Let $\tilde{\mathscr{C}} = \mathscr{C} \cup \mathscr{C}'$. Since $\mathscr{C}$ and $\mathscr{C}'$ are not disjoint, it follows from (3) that all states in $\tilde{\mathscr{C}}$ are accessible from each other under $\tilde{\mathbf{f}}$. Moreover, it follows from (3) and the fact that $\mathscr{C}$ and $\mathscr{C}'$ are closed under $\mathbf{f}$ and $\mathbf{f}'$, respectively, that $\tilde{\mathscr{C}}$ is closed. Hence $\tilde{\mathscr{C}}$ is recurrent class under $\tilde{\mathbf{f}}$, which implies that $\mathscr{C} = \mathscr{C}'$ since both $\mathscr{C}$ and $\mathscr{C}'$ are strongly communicating classes. □

Bather [2] gives a classification of states that is in the same spirit as ours. (Bather considers unconstrained optimization over pure policies of finite-state MDPs with compact action spaces.) Although Bather's classification is very insightful, it is cumbersome to work with in terms of establishing our main existence result in the subsequent section. Platzman [10] defines the notion of a "connected class" for a finite MDP. A connected class is always a strongly communicating class but not vice versa. (In fact, the connected classes are the strongly communicating classes obtained at level 1 in the following algorithm.)

We now outline an algorithm which determines the strongly communicating classes. We refer to the original MDP as being in level 1. First, partition the state space into communicating sets, where x and y are in the same set of the partition if and only if there exists a stationary policy $\mathbf{f}$ such that x is accessible from y and y is accessible from x. (A state is always assumed accessible from itself.) This partition is easily found with a standard depth-first algorithm. One or more of these communicating sets is guaranteed to be closed. (A communicating set is closed if it is impossible to leave it.) Each of these closed communicating sets is then labeled a strongly commu-

nicating class. It then remains to classify the states in the "open" communicating classes. We consider each open communicating class, say $\mathscr{D}$, separately. We now delete all actions a associated with states $x \in \mathscr{D}$ for which $p_{xay} > 0$ where y is not in $\mathscr{D}$. Having reduced the action spaces, we then remove from $\mathscr{D}$ any state x for which the associated action space is now empty and put that state in $\mathscr{T}$. We continue to reduce $\mathscr{D}$ and the actions associated with $\mathscr{D}$ in this manner until we obtain a set $\mathscr{D}' \subseteq \mathscr{D}$ such that (i) each state in $\mathscr{D}'$ has at least one action associated with it; (ii) it is not possible to leave $\mathscr{D}'$ with any of the remaining actions. Refer to the MDP over $\mathscr{D}'$ with the reduced action space as being in level 2. We then repeat the entire procedure over $\mathscr{D}'$, i.e., we find the closed and open communicating classes in $\mathscr{D}'$, label the closed classes as strongly communicating classes, and then examine each of the open communicating classes in $\mathscr{D}'$ separately. In general, this depth-first algorithm could descend as many as $|\mathscr{S}|$ levels.

A more formal statement of the algorithm along with a proof of its correctness is given in [14]. Moreover, it is shown in [14] that the worst case complexity of the algorithm is $O(|\mathscr{S}|^3|\mathscr{A}|)$.

Figure 1 depicts an example with six states. States 1 and 2 each have one action, whereas the remaining states each have two actions. A single arrow corresponds to action 1 and a double arrow corresponds to action 2. The numbers next to the arrows are the probabilities of the corresponding transitions. For example, when in state 3 and action 1 is chosen, the subsequent state is 2 with probability $1/2$ and is 4 with probability $1/2$; and when in state 3 and action 2 is chosen, the subsequent state is 3 with probability 1. The communicating sets in this example are $\{1\}$ and $\{2, 3, 4, 5, 6\}$. There are four strongly communicating classes: $\{1\}$, $\{3\}$, $\{4\}$ and $\{5, 6\}$. The set of transient states is given by the singleton $\{2\}$. The algorithm would consider four levels in this example.

The concept of strongly communicating classes will also enable us to characterize the probabilistic behaviour of the MDP under general policies. Lemma 1 below states that the set $\mathscr{T}$ is "transient under all policies." Lemma 2 states that it is impossible (with probability 1) to be and not to be in a given strongly communicating class an infinite number of times. Let i.o. abbreviate infinitely often and a.a. abbreviate almost always. Let $1(\cdot)$ denote the indicator function.

LEMMA 1. *For all policies* $\mathbf{u}$ *and all initial states* x,

$$P_{\mathbf{u}}^{x}(X_n \in \mathscr{T} \text{ i.o.}) = 0. \tag{4}$$

PROOF. Fix an initial state x, define the expected total reward

$$v(\mathbf{u}) := E_{\mathbf{u}}^{x}\left[\sum_{n=1}^{\infty} 1(X_n \in \mathscr{T})\right],$$

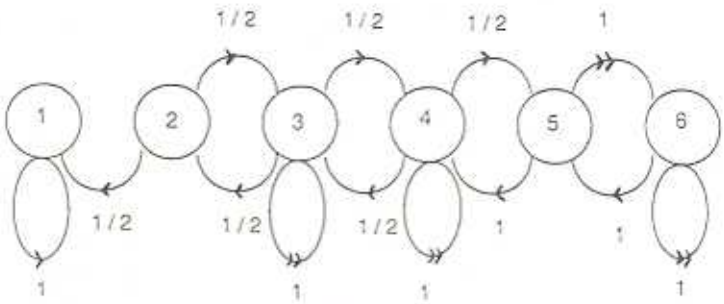

FIGURE 1

and consider the problem of maximizing $v(\mathbf{u})$ over all policies $\mathbf{u}$. This is a total reward Markov decision problem (e.g., see [13, p. 133]), which is maximized by some pure policy $\mathbf{g}$. But since $\mathscr{T}$ is transient under all stationary policies, we have for all $\mathbf{u}$

$$v(\mathbf{u}) \leqslant v(\mathbf{g}) = \sum_{n=1}^{\infty} P_{\mathbf{g}}^{x}(X_n \in \mathscr{T}) < \infty,$$

which in turn implies (4). □

LEMMA 2. *Let $\mathscr{C}$ be a strongly communicating class. Then for all policies $\mathbf{u}$ and all initial states x,*

$$P_{\mathbf{u}}^{x}(\{X_n \in \mathscr{C} \text{ i.o.}\} \cap \{X_n \notin \mathscr{C} \text{ i.o.}\}) = 0. \tag{5}$$

PROOF. Fix an initial state x. Suppose that for some policy $\mathbf{v}$,

$$P_{\mathbf{v}}^{x}(\{X_n \in \mathscr{C} \text{ i.o.}\} \cap \{X_n \notin \mathscr{C} \text{ i.o.}\}) > 0. \tag{6}$$

Since the state space is finite, (6) implies the existence of an $\bar{x} \in \mathscr{C}$ and a $\bar{y} \notin \mathscr{C}$ such that

$$P_{\mathbf{v}}^{x}((X_n, X_{n-1}) = (\bar{x}, \bar{y}) \text{ i.o.}) > 0. \tag{7}$$

Now consider the expected total reward Markov decision problem of maximizing over all policies $\mathbf{u}$

$$v(\mathbf{u}) = E_{\mathbf{u}}^{x} \sum_{n=1}^{\infty} [r(X_n, A_n)],$$

where

$$r(x, a) = 1(x = \bar{x}) p_{\bar{x}a\bar{y}}.$$

A straightforward calculation gives for all policies $\mathbf{u}$,

$$E_{\mathbf{u}}^{x}\left[\sum_{n=1}^{\infty} 1(X_n = \bar{x}, X_{n+1} = \bar{y})\right] = v(\mathbf{u}). \tag{8}$$

It follows from (7) and (8) that $v(\mathbf{v}) = \infty$. This combined with the fact that $v(\mathbf{u})$ is maximized by some pure policy implies that

$$\infty = v(\mathbf{g}) = \sum_{n=1}^{\infty} P_{\mathbf{g}}^{x}(X_n = \bar{x}) P_{\bar{x}\bar{y}}(\mathbf{g}) \tag{9}$$

for some pure policy $\mathbf{g}$. Equation (9) implies that $\bar{x}$ is recurrent under $\mathbf{g}$ and that $\bar{y}$ is accessible from $\bar{x}$ under $\mathbf{g}$. Hence, $\bar{x}$ and $\bar{y}$ belong to the same recurrent class under $\mathbf{g}$. But this gives a contradiction since $\mathscr{C}$ is a strongly communicating class such that $\bar{x} \in \mathscr{C}$ and $\bar{y} \notin \mathscr{C}$. □

For each $i = 1, \ldots, I$, denote for all $x \in \mathscr{C}_i$ the set

$$\mathscr{F}_x = \{a \in \mathscr{A}: p_{xay} = 0 \text{ for all } y \in \mathscr{C}_i\}.$$

Thus beginning in a state $x \in \mathscr{C}_i$, the set $\mathscr{F}_x$ contains the actions which guarantee that the state process will remain in the strongly communicating class $\mathscr{C}_i$. Proposition

2 states that any "recurrent class associated with an arbitrary policy $\mathbf{u}$" is contained within one of the strongly communicating classes. It also gives a key property of the action process $\{A_n\}$.

PROPOSITION 2. *For all policies* $\mathbf{u}$ *and all initial states* x,

$$\sum_{i=1}^{I} P_{\mathbf{u}}^{x}(X_n \in \mathscr{C}_i \text{ a.a.}) = 1 \tag{10}$$

and

$$P_{\mathbf{u}}^{x}(A_n \in \mathscr{S}_{X_n} \text{ a.a.}) = 1. \tag{11}$$

PROOF. Equation (10) is an immediate consequence of Lemmas 1 and 2. Suppose that (11) is false for some $\mathbf{u}$ and some x. As in the proof of Lemma 2, this implies that there is a state $\tilde{x}$ belonging to some strongly communicating class $\mathscr{C}$ and an action $\tilde{a} \notin \mathscr{S}_{\tilde{x}}$ such that

$$P_{\mathbf{u}}^{x}((X_n, A_n) = (\tilde{x}, \tilde{a}) \text{ i.o.}) > 0.$$

Continuing to parallel the proof of Lemma 2, the above implies the existence of a pure policy $\mathbf{g}$ such that $\tilde{x}$ is a recurrent state under $\mathbf{g}$ and $g(\tilde{x}) = \tilde{a}$. Since $\tilde{a} \notin \mathscr{S}_{\tilde{x}}$, this in turn implies the existence of a $\tilde{y} \notin \mathscr{C}$ that is accessible from $\tilde{x}$ under $\mathbf{g}$. But this is a contradiction since $\mathscr{C}$ is a strongly communicating class. □

DEFINITION 2. For each $i = 1, \ldots, I$ we define a new MDP, called MDP-i, as follows:

1. The state space is $\mathscr{C}_i$.
2. For each $x \in \mathscr{C}_i$, the set of available actions is given by the state-dependent action spaces $\mathscr{S}_x$.
3. The laws of motion, reward function, and cost function are the same as for the original MDP but restricted to the state-dependent action spaces $\mathscr{S}_x$, $x \in \mathscr{C}_i$.

The following proposition gives meaning to the previous definition. Its proof follows directly from Definition 1 and the definition of $\mathscr{S}_x$, $x \in \mathscr{C}_i$. Recall that an MDP is *communicating* if the state space constitutes a single communicating set (e.g., see [1]).

PROPOSITION 3. *For all* $i = 1, \ldots, I$,

1. $\mathscr{S}_x$ *is nonempty for all* $x \in \mathscr{C}_i$.
2. $\sum_{y \in \mathscr{C}_i} p_{xy(a)} = 1$ *for all* $a \in \mathscr{S}_x$, $x \in \mathscr{C}_i$.
3. *MDP-*i *is a communicating MDP.*

4. Optimization with a sample path constraint. We now address the constrained optimization problem discussed in the Introduction. Let Δ be a fixed and given initial state. We shall say that a policy $\mathbf{u}$ is feasible if (2) is satisfied. Denote ϕ^* for the supremum of $\phi(\mathbf{u})$ over the class of all feasible policies. A policy $\mathbf{u}$ is said to be *optimal* if $\mathbf{u}$ is feasible and $\phi(\mathbf{u}) = \phi^*$. For a fixed $\epsilon > 0$, a policy $\mathbf{u}$ is said to be *ϵ-optimal* if it is feasible and $\phi(\mathbf{u}) > \phi^* - \epsilon$.

We shall show that if there exists a feasible policy, then for each $\epsilon > 0$ there exists an ϵ-optimal stationary policy. Verifiable sufficient conditions are also given for the existence of an optimal stationary policy. The proof of these existence results is carried out in three steps: first, for each $i = 1, \ldots, I$, find an ϵ-optimal stationary policy for MDP-i whenever a feasible policy for MDP-i exists; second, for each

$i = 1, \ldots, I$, associate a constant reward over all states in $\mathscr{C}_i$, and then find an optimal pure policy for the corresponding intermediate MDP; third, combine the ϵ-optimal policies for the restricted MDPs with the optimal policy for the intermediate MDP to give an ϵ-optimal stationary policy.

4.1. *The restricted MDPs.* Throughout this subsection fix an $i \in \{1, \ldots, I\}$. Consider the evolution of the state and action processes for MDP-i. For all $n = 1, 2, \ldots$ we have

$$X_n \in \mathscr{C}_i \quad \text{and} \quad A_n \in \mathscr{S}_{X_n}.$$

A policy $\mathbf{u}$ and a fixed initial state $x \in \mathscr{C}_i$ determine a probability measure $P^x_{\mathbf{u},i}$ for the sample space associated with MDP-i. The corresponding expected average reward for MDP-i is given by

$$\phi_i^x(\mathbf{u}) := E^x_{\mathbf{u},i}\left[\liminf_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} r(X_m, A_m)\right].$$

A policy $\mathbf{u}$ is said to be *feasible for MDP-i* if

$$P^x_{\mathbf{u},i}\left(\limsup_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} c(X_m, A_m) \leqslant \alpha\right) = 1$$

for all $x \in \mathscr{C}_i$.

For each MDP-i we also need to introduce an associated Linear Program (LP) with decision variables $z(x, a)$, $a \in \mathscr{S}_x$, $x \in \mathscr{C}_i$. Let $\delta_{xy} = 1$ if $x = y$ and $\delta_{xy} = 0$ otherwise.

LP-i:

$$t_i = \max \sum_{x \in \mathscr{C}_i} \sum_{a \in \mathscr{S}_x} r(x, a) z(x, a)$$

s.t.

$$\sum_{x \in \mathscr{C}_i} \sum_{a \in \mathscr{S}_x} (\delta_{xy} - p_{xay}) z(x, a) = 0, \qquad y \in \mathscr{C}_i,$$

$$\sum_{x \in \mathscr{C}_i} \sum_{a \in \mathscr{S}_x} z(x, a) = 1,$$

$$\sum_{x \in \mathscr{C}_i} \sum_{a \in \mathscr{S}_x} c(x, a) z(x, a) \leqslant \alpha,$$

$$z(x, a) \geqslant 0, \qquad a \in \mathscr{S}_x, x \in \mathscr{C}_i.$$

The relationship between MDP-i and LP-i is expressed in the following,

LEMMA 3. *There exists a feasible policy MDP-i if and only if LP-i is feasible. If LP-i is feasible, then for each $\epsilon > 0$ there exists a feasible stationary policy $\mathbf{f}_i$ such that*

$$\phi_i^x(\mathbf{f}_i) = \phi_i(\mathbf{f}_i) > t_i - \epsilon \quad \text{for all } x \in \mathscr{C}_i.$$

Moreover, the transition matrix $P(\mathbf{f}_i)$ is unichain.

PROOF. From Proposition 3 we know that MDP-i is a communicating MDP. The result therefore follows from Theorem 2 of [12].

The following lemma links the original constrained optimization problem with MDP-i. In words it says: given that the state process settles down in the strongly communicating class $\mathscr{C}_i$, the average reward for a feasible policy cannot exceed t_i. Since much of the proof is similar to the proof of Proposition 2 of [12] only an outline is given.

LEMMA 4. *Suppose* $\mathbf{u}$ *is feasible and* $P_{\mathbf{u}}^{\lambda}(X_n \in \mathscr{C}_i \text{ a.a.}) > 0$. *Then* (i) *LP-i is feasible, and* (ii)

$$P_{\mathbf{u}}^{\lambda}\left(\liminf_{n\to\infty} \frac{1}{n}\sum_{m=1}^{n} r(X_m, A_m) \leqslant t_i \mid X_n \in \mathscr{C}_i \text{ a.a.}\right) = 1.$$

PROOF. Fix a feasible policy $\mathbf{u}$. The strong Law of Large Numbers for Martingale Differences (e.g., see [9]) implies that $P_{\mathbf{u}}^{\lambda}$-almost surely

(12)

$$\lim_{n\to\infty} \frac{1}{n}\sum_{m=2}^{n}\left[1(X_m = y) - \sum_{x,a} 1(X_{m-1} = x, A_{m-1} = a)p_{xay}\right] = 0 \quad \text{for all } y \in \mathscr{S}.$$

Let Γ be the set of sample paths $(x_1, a_1, x_2, a_2, \ldots) \in \Omega$ that satisfy

$$a_n \in \mathscr{S}_{x_n} \text{ for all } n \geqslant N, \text{ for some positive integer } N,$$

$$\lim \frac{1}{n}\sum_{m=2}^{n}\left[1(x_m = y) - \sum_{x,a} 1(x_{m-1}, a_{m-1} = a)p_{xay}\right] = 0, \quad \text{and}$$

$$\limsup \frac{1}{n}\sum_{m=1}^{n} c(x_m, a_m) \leqslant \alpha.$$

Due to Proposition 2, (12) and the fact that $\mathbf{u}$ is feasible, we have

$$P_{\mathbf{u}}^{\lambda}(\Gamma) = 1.$$

It therefore suffices to show that LP-i is feasible and

$$\{X_n \in \mathscr{C}_i \text{ a.a.}\} \cap \Gamma \subset \left\{\liminf_{n\to\infty} \frac{1}{n}\sum_{m=1}^{n} r(X_m, A_m) \leqslant t_i\right\}.$$

Let $(x_1, a_1, x_2, a_2, \ldots) \in \{X_n \in \mathscr{C}_i \text{ a.a.}\} \cap \Gamma$ and define

$$z_n(x, a) := \frac{1}{n}\sum_{m=1}^{n} 1(x_m = x, a_m = a).$$

The proof is then completed by showing that any limit point $\{z'(x, a)\}$ of $\{z_n(x, a)\}$, $n = 1, 2, \ldots$ is a feasible solution for LP-i and that

$$\sum_{x\in\mathscr{C}_i}\sum_{a\in\mathscr{A}} r(x, a)z'(x, a) \leqslant t_i \quad \square$$

4.2. *The intermediate MDP.* Let

$$G := \{i: \text{LP-}i \text{ is feasible}\}.$$

Over the original sample space, define for each policy $\mathbf{u}$ and $M \leqslant 0$ the following expected time-average reward

$$\beta^M(\mathbf{u}) := E_{\mathbf{u}}^{\lambda}\left[\liminf_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} \sum_{i=1}^{I} t_i^M 1(X_m \in \mathscr{C}_i)\right],$$

where

$$t_i^M := \begin{cases} t_i & \text{if } i \in G, \\ M & \text{if } i \notin G. \end{cases}$$

In this subsection we consider the unconstrained problem of maximizing $\beta^M(\mathbf{u})$ over all policies $\mathbf{u}$ with M given and fixed. Intuitively, setting $M \ll 0$ would discourage the state process from remaining in any of the strongly communicating classes for which LP-i (and hence MDP-i) is infeasible. Let β^M be the supremum of $\beta^M(\mathbf{u})$ over the class of *all* policies. Clearly, β^M is nonincreasing as $M \to -\infty$. Denote

$$\beta := \lim_{M\to-\infty} \beta^M.$$

From Proposition 2 and Lebesgue's Dominated Convergence Theorem we have for all policies $\mathbf{u}$ and $M \leqslant 0$,

$$(13)\quad \beta^M(\mathbf{u}) = \sum_{i=1}^{I} t_i^M P_{\mathbf{u}}^{\lambda}(X_n \in \mathscr{C}_i \text{ a.a.}) = \lim_{n\to\infty} \frac{1}{n} E_{\mathbf{u}}^{\lambda}\left[\sum_{m=1}^{n} \sum_{i=1}^{I} t_i^M 1(X_m \in \mathscr{C}_i)\right].$$

Thus, the expected time average reward $\beta^M(\mathbf{u})$ is equivalent to the more commonly studied time average expected reward. It is well known that there exists a pure policy that maximizes the time average expected reward for any unconstrained MDP with finite state and action space [8]. Thus for each $M \leqslant 0$ there exists a pure policy $\mathbf{g}^M$ such that $\beta^M(\mathbf{g}^M) = \beta^M$.

In words the following lemma states that ϕ^*, the maximum average reward for the constrained optimization problem, is bounded by β. Subsequently we shall show that this bound is tight. It also states that β can be determined by maximizing $\beta^N(\mathbf{u})$ for a specific finite N. Finally, it states that under the pure policy that maximizes $\beta^N(\mathbf{u})$ the strongly communicating classes $\mathscr{C}_i, i \notin G$, are transient.

LEMMA 5. *Suppose there exists a feasible policy. Then* (i) $\phi^* \leqslant \beta$; (ii) *there exists an* $N \leqslant 0$ *such that* $\beta = \beta^N$; *and* (iii)

$$P_{\mathbf{g}^N}^{\lambda}(X_n \in \mathscr{C}_i \text{ a.a.}) = 0 \quad \textit{for all } i \notin G.$$

PROOF. (i) Let $\mathbf{u}$ be a feasible policy. From Lemma 4 (i) we have $P_{\mathbf{u}}^{\lambda}(X_n \in \mathscr{C}_i$ a.a.) $= 0$ for $i \notin G$. This combined with (13) gives

$$\beta^M(\mathbf{u}) = \sum_{i\in G} t_i P_{\mathbf{u}}^{\lambda}(X_n \in \mathscr{C}_i \text{ a.a.}).$$

From Proposition 2 and Lemma 4 (ii) we have

$$\begin{aligned}\phi(\mathbf{u}) &= \sum_{i=1}^{I} E_{\mathbf{u}}^{\lambda}\left[\liminf_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} r(X_m, A_m) | X_n \in \mathscr{C}_i \text{ a.a.}\right] P_{\mathbf{u}}^{\lambda}(X_n \in \mathscr{C}_i \text{ a.a.}) \\ &\leqslant \sum_{i\in G} t_i P_{\mathbf{u}}^{\lambda}(X_n \in \mathscr{C}_i \text{ a.a.}).\end{aligned}$$

Thus $\phi(\mathbf{u}) \leqslant \beta^M(\mathbf{u})$ for all feasible policies $\mathbf{u}$ and all $M \leqslant 0$, whence (i). Since the state and action spaces are finite, there exists a sequence $\{M_k\}$ and an $N \leqslant 0$ such that $\lim M_k = -\infty$ and $\mathbf{g}^{M_k} = \mathbf{g}^N$ for all k. Thus, by (13) we have

$$\beta = \lim_{k \to \infty} \sum_{i=1}^{I} t_i^{M_k} P_{\mathbf{g}^N}^{\lambda}(X_n \in \mathscr{C}_i \text{ a.a.}).$$

This combined with the fact that β is finite establishes (ii) and (iii). □

Let

$$H = \{i \in G: \mathscr{C}_i \text{ contains a recurrent class under } \mathbf{g}^N\}.$$

Without loss of generality, we may assume that each $\mathscr{C}_i$, $i \in H$, is closed under $\mathbf{g}^N$. (Otherwise, modify $\mathbf{g}^N$ so that $\mathbf{g}^N(x) \in \mathscr{S}_x$ for all $x \in \mathscr{C}_i$, $i \in H$. Clearly the modified policy has the desired property, and it is not difficult to show that it continues to maximize $\beta^N(\mathbf{u})$.)

4.3. *Existence results.* Let $\epsilon > 0$. Construct a stationary policy $\mathbf{f}^*$ as follows.

1. Let G be the set of i such that LP-i is feasible. For each $i \in G$ let $\mathbf{f}_i$ be the stationary policy for MDP-i as given in Lemma 3.

2. Let $\mathbf{g}^N$ be as in Lemma 5. Let H be the set of $i \in G$ such that $\mathscr{C}_i$ is closed under $\mathbf{g}^N$.

3. Define a stationary policy $\mathbf{f}^*$ as follows: when in state $x \in \mathscr{C}_i$ with $i \in H$, apply $\mathbf{f}_i$; otherwise apply $\mathbf{g}^N$.

THEOREM 1. *There exists a feasible policy if and only if β is finite. If β is finite, then the stationary policy $\mathbf{f}^*$ as constructed above is ϵ-optimal.*

PROOF. From Lemma 5 (i) we know that if there exists a feasible policy then β is finite. Now suppose that β is finite. Since $\mathbf{g}^N$ does not have a recurrent class in any strongly communicating class $\mathscr{C}_i$, $i \in G - H$, it follows from Lemma 5 (iii) that

$$P_{\mathbf{g}^N}^{\lambda}(X_n \in \mathscr{C}_i \text{ a.a.}) = 0 \quad \text{for all } i \notin H. \tag{14}$$

Since $\mathbf{f}^*$ is identical to $\mathbf{g}^N$ outside of $\bigcup_{i \in H} \mathscr{C}_i$ and since each $\mathscr{C}_i$, $i \in H$, is closed under both $\mathbf{f}^*$ and $\mathbf{g}^N$, we have

$$P_{\mathbf{f}^*}^{\lambda}(X_n \in \mathscr{C}_i \text{ a.a.}) = P_{\mathbf{g}^N}^{\lambda}(X_n \in \mathscr{C}_i \text{ a.a.}) \quad \text{for all } i = 1, \ldots, I. \tag{15}$$

We now show that $\mathbf{f}^*$ is feasible. Since for each $i \in H$, $\mathbf{P}(\mathbf{f}_i)$ is unichain, there is a unique stationary distribution, $\pi_i(x)$, $x \in \mathscr{C}_i$, for the state process of MDP-i. From (14)–(15) and the fact that $\mathbf{f}^*$ is identical to $\mathbf{f}_i$ over $\mathscr{C}_i$ for each $i \in H$ we have $P_{\mathbf{f}^*}^{\lambda}$-almost surely

$$\lim_{n \to \infty} \frac{1}{n} \sum_{m=1}^{n} c(X_m, A_m) = \sum_{i \in H} 1(X_n \in \mathscr{C}_i \text{ a.a.}) \sum_{x \in \mathscr{C}_i} \sum_{a \in \mathscr{S}_x} \pi_i(x) f_i(x, a) c(x, a). \tag{16}$$

But since $\mathbf{f}_i$ is feasible for MDP-i we also have

$$\sum_{x \in \mathscr{C}_i} \sum_{a \in \mathscr{S}_x} \pi_i(x) f_i(x, a) c(x, a) \leqslant \alpha. \tag{17}$$

Combining (16) and (17) establishes the feasibility of $\mathbf{f}^*$.

We now show that $\mathbf{f}^*$ is ϵ-optimal. Recall that $\phi_i(\mathbf{f}_i)$ is the long-run average reward associated with $\mathbf{f}_i$ for MDP-i. Using the same reasoning above, we obtain

$$\phi(\mathbf{f}^*) = \sum_{i \in H} P_{\mathbf{f}}^{\Delta}(X_n \in \mathscr{C}_i \text{ a.a.}) \phi_i(f_i)$$

$$\geqslant \sum_{i \in H} P_{\mathbf{f}}^{\Delta}(X_n \in \mathscr{C}_i \text{ a.a.})(t_i - \epsilon). \tag{18}$$

where the last inequality follows from Lemma 3. Combining (15) and (18) gives

$$\phi(\mathbf{f}^*) \geqslant \sum_{i \in H} t_i P_{\mathbf{g}}^{\Delta}(X_n \in \mathscr{C}_i \text{ a.a.}) - \epsilon = \beta^N - \epsilon. \tag{19}$$

The proof is then completed by combining (19) with Lemma 5 (i) and (ii). □

We conclude this section with a condition for the existence of an *optimal* stationary policy. An MDP is said to be *unichain* if every stationary policy $\mathbf{f}$ gives rise to a unichain transition matrix $\mathbf{P}(\mathbf{f})$.

THEOREM 2. *Suppose that there exists a feasible policy and that MDP-i is unichain for all $i \in H$. Then there exists an optimal stationary policy.*

PROOF. In this case for each $i \in H$ there exists an $\mathbf{f}_i$ for MDP-i that satisfies $\phi_i(\mathbf{f}_i) = t_i$. The proof then mimics the proof of Theorem 1. □

5. Computational considerations. The arguments up to Theorem 1 also lead to a recipe for the construction of an ϵ-optimal policy.

1. Determine the strongly communicating classes $\mathscr{C}_i$, $i = 1, \ldots, I$.
2. Determine G the set of i such that LP-i is feasible. Determine t_i, the maximum long-run average reward for MDP-i, for $i \in G$.
3. For each $i \in G$ determine $\mathbf{f}_i$, the stationary ϵ-optimal policy, for MDP-i.
4. Determine a penalty factor N satisfying $\beta^N = \beta$.
5. Determine an optimal policy $\mathbf{g}^N$ which maximizes $\beta^N(\mathbf{u})$.
6. Combine $\mathbf{g}^N$ with $\mathbf{f}_i$, $i \in H$ (as discussed at the beginning of §5) to obtain an ϵ-optimal stationary policy $\mathbf{f}^*$.

Step 1 can be carried out in $O(|\mathscr{S}|^3|\mathscr{A}|)$ worst case time with the algorithm outlined in §3. Step 2 involves solving I LPs over reduced state and action spaces. Step 3 can be carried out by solving $|G|$ parametric LPs; see [12].

Step 4 can be solved by choosing some $M \ll 0$, solving the corresponding intermediate problem, and then checking whether the recurrent classes associated with $\mathbf{g}^M$ are in strongly communicating classes in G. If not, then multiply M by 2 and repeat the procedure until a pure policy $\mathbf{g}^M$ with the desired property is found.

For the remainder of this section we discuss how Step 5 can be solved more efficiently. In order to simplify the discussion we suppose that the following condition holds true.

Condition 1. *LP-i is feasible for all $i = 1, \ldots, I$.*

Note that Condition 1 does not necessarily imply that all policies are feasible since there may exist both feasible and infeasible policies for any of the restricted MDPs. Also note that Condition 1 is trivially satisfied for the unconstrained problem.

Under Condition 1, Step 5 is equivalent to finding a pure policy that maximizes

$$\beta(\mathbf{g}) = \lim_{n \to \infty} \frac{1}{n} E_{\mathbf{g}}^{\Delta}\left[\sum_{m=1}^{n} \sum_{i=1}^{I} t_i 1(X_m \in \mathscr{C}_i)\right].$$

over the class of pure policies. This is a standard unconstrained time-average MDP problem that can be solved by either policy improvement, successive approximations or linear programming. However, these classical techniques do not take into account the special structure of the problem. Since the reward is constant over each of the strongly communicating classes, a more computationally efficient approach would be to aggregate all the states within such a class and then solve a corresponding "aggregated MDP".

DEFINITION 3. The *aggregated MDP* is defined as follows.

(i) The state space is $\bar{\mathcal{S}} = \{1, 2, \ldots, I + J\}$, where $J = |\mathcal{T}|$.

(ii) The state-dependent action spaces $\bar{\mathcal{A}}_i$, $i \in \bar{\mathcal{S}}$, are

$$\bar{\mathcal{A}}_i = \begin{cases} \{\theta\} \cup \{(x, a) x \in \mathcal{C}_i, a \notin \mathcal{S}_x\} & 1 \leqslant i \leqslant I, \\ \mathcal{A}, & I + 1 \leqslant i \leqslant I + J. \end{cases}$$

(iii) For $i = 1, \ldots, I$, the law of motion is given by $\bar{p}_{i\theta i} = 1$ and

$$\bar{p}_{i(x,a)j} = \begin{cases} \sum_{y \in \mathcal{C}_j} p_{xay}, & 1 \leqslant j \leqslant I, (x, a) \in \bar{\mathcal{A}}_i, \\ p_{xaj}, & I + 1 \leqslant j \leqslant I + J, (x, a) \in \bar{\mathcal{A}}_i. \end{cases}$$

(iv) For $i = I + 1, \ldots, I + J$, the law of motion is given by

$$\bar{p}_{iaj} = \begin{cases} \sum_{y \in \mathcal{C}_j} p_{iay}, & 1 \leqslant j \leqslant I, a \in \bar{\mathcal{A}}_i, \\ p_{iaj}, & I + 1 \leqslant j \leqslant I + J, a \in \bar{\mathcal{A}}_i. \end{cases}$$

(v) For $i = 1, \ldots, I$, the reward in state i is r_i; the rewards in states $j = I + 1, \ldots, I + J$ are arbitrary.

Thus, in the aggregated MDP, there is one state corresponding to each strongly communicating class plus one state corresponding to each transient state in $\mathcal{T}$. For each state $i = 1, \ldots, I$, action θ is available, which keeps the state process in the same state with probability 1. The actions of the form (x, a) are also available for the aggregated MDP, which, in the intermediate MDP, correspond to a movement to state x and then a selection of action a.

Let $\bar{\mathbf{g}}$ be a pure policy that maximizes the long-run reward for the aggregated MDP. (Again, one of many classical algorithms can be employed to obtain $\bar{\mathbf{g}}$.) Denote

$$\bar{H} = \{i \in \bar{\mathcal{S}} : \bar{g}(i) = \theta\}.$$

Note that all states in $\bar{H}$ are absorbing under $\bar{\mathbf{g}}$.

Now consider a pure policy $\mathbf{g}$ for the original problem with state space $\mathcal{S}$ and action space $\mathcal{A}$ defined as follows. (i) For $x \in \mathcal{C}_i$ with $\bar{g}(i) = \theta$, let $g(x)$ be equal to an arbitrary element in $\mathcal{S}_x$. (ii) For $x \in \mathcal{C}_i$ with $\bar{g}(i) = (w, a)$ let $\mathbf{g}$ be such that it drives the state process from x to w while remaining in $\mathcal{C}_i$; once in state w, it chooses action $a \in \mathcal{A}$. (iii) For $x \in \mathcal{T}$, $g(x) = \bar{g}(x)$.

By "driving" the process from state x to w we mean that $\mathbf{g}$ chooses actions that keep the state process in $\mathcal{C}_i$ and which bring the state process to w in finite expected time. (This is possible since MDP-i is a communicating MDP.) We remark that these actions can be determined directly from the transition graph of the law of motion in $O(|\mathcal{C}_i|)$ time by starting at w and moving outwards.

It is a straightforward exercise to show

THEOREM 3. *The pure policy* $\mathbf{g}$ *constructed from an optimal pure policy for the aggregated MDP according to the above procedure is optimal for the intermediate MDP. Consequently, the stationary policy* $\mathbf{f}^*$ *which applies* $\mathbf{f}_i$ *over* $\mathcal{C}_i$, $i \in H$, *and applies* $\mathbf{g}$ *otherwise is* ϵ*-optimal for the original constrained optimization problem.*

We conclude by making the observation that the theory developed in this paper can also be employed to obtain an optimal pure policy for the classical unconstrained problem of maximizing the long-run average reward. In this case, an optimal pure policy $\mathbf{g}_i$ can be found for each MDP-i, which, when combined with the optimal pure policy $\mathbf{g}$ of the aggregated MDP, gives an optimal pure policy $\mathbf{g}^*$ for the original (unconstrained) problem.

Acknowledgement. We would like to thank Ward Whitt and Martin L. Puterman for helping us improve the presentation of the paper.

References

[1] Bather, J. (1973). Optimal Decision Procedures in Finite Markov Chains. Part II: Communicating Systems. *Adv. Appl. Probab.* **5**, 521–552.

[2] _____ (1973). Optimal Decision Procedures in Finite Markov Chains. Part III: General Convex Systems. *Adv. Appl. Probab.* **5**, 541–553.

[3] Baykal-Gursoy, M. and Ross, K. W. Variability Sensitive Markov Decision Processes, to appear in *Math. Oper. Res.*

[4] Beutler, F. J. and Ross, K. W. (1986). Time-average Optimal Constrained Semi-Markov Decision Processes. *Adv. Appl. Probab.* **18**, 341–359.

[5] Derman, C. (1970). *Finite State Markovian Decision Processes*. Academic Press, New York.

[6] _____ and Veinott, A. F. Jr. (1972). Constrained Markov Decisions Chains. *Management Sci.* **19**, 389–390.

[7] Hordijk, A. and Kallenberg, L. C. M. (1984). Constrained Undiscounted Dynamic Programming. *Math. Oper. Res.* **9**, 276–289.

[8] Kallenberg, L. C. M. (1983). *Linear Programming and Finite Markovian Control Problems*. Vol. 148, Math. Centre Tracts, Amsterdam.

[9] Loeve, M. (1978). *Probability Theory*. Vol. 2, Springer-Verlag, New York.

[10] Platzman, L. (1977). Improved Conditions for Convergence in Undiscounted Markov Renewal Programming. *Oper. Res.* **25**, 529–533.

[11] Ross, K. W. (1989). Randomized and Past-Dependent Policies for Markov Decision Processes with Multiple Constraints. *Oper. Res.* **37**, 474–477.

[12] _____ and Varadarajan, R. (1989). Markov Decision Processes with Sample Path Constraints: The Communicating Case. *Oper. Res.* **37**, 780–790.

[13] Ross, S. (1971). Applied Probability Models with Optimization Applications. Holden-Day, San Francisco, CA.

[14] Varadarajan, R. (1987). Reliability and Performance Models for Reconfigurable Computer Systems. PhD thesis, University of Pennsylvania, Philadelphia, PA.

[15] White, D. J. Dynamic Programming and Probabilistic Constraints. *Oper. Res.* **22**, 654–664.

ROSS: DEPARTMENT OF SYSTEMS, UNIVERSITY OF PENNSYLVANIA, PHILADELPHIA, PENNSYLVANIA 19104

VARADARAJAN: DEPARTMENT OF COMPUTER AND INFORMATION SCIENCE, UNIVERSITY OF FLORIDA, GAINESVILLE, FLORIDA 32611