# Percentile Performance Criteria For Limiting Average Markov Decision Processes

Jerzy A. Filar, Dmitry Krass, and Keith W. Ross, *Senior Member, IEEE*

***Abstract*— In this paper we address the following basic feasibility problem for infinite-horizon Markov decision processes (MDP's): can a policy be found that achieves a specified value (target) of the long-run limiting average reward at a specified probability level (percentile)? Related optimization problems of maximizing the target for a specified percentile and vice versa are also considered. We present a complete (and discrete) classification of both the maximal achievable target levels and of their corresponding percentiles. We also provide an algorithm for computing a deterministic policy corresponding to any feasible target-percentile pair.**

**Next we consider similar problems for an MDP with multiple rewards and/or constraints. This case presents some difficulties and leads to several open problems. An LP-based formulation provides constructive solutions for most cases.**

## I. Introduction and Definitions

INFINITE horizon Markov decision processes (MDP's, for short) have been extensively studied since the 1950's. One of the most commonly considered versions is the so-called "limiting average reward" model. In this model the decision-maker aims to maximize the expected value of the limit-average ("long-run average") of an infinite stream of single-stage rewards. There are now a number of good algorithms for computing optimal deterministic policies in the limiting average MDP's (e.g., see [4], [6], [11]).

It should be noted, however, that an optimal policy in the above "classical" sense is insensitive to the probability distribution function of the long-run average reward. That is, it is possible that an optimal policy, while yielding an acceptably high expected long-run average reward, carries with it unacceptably high probability of low values of that same random variable. This "risk insensitivity" is inherent in the formulation of the classical objective criterion as that of maximizing the expected value of a random variable, and it is not necessarily undesirable. Nonetheless, in this paper we adopt the point of view that there are many natural situations where the decision-maker is interested in finding a policy that will achieve a sufficiently high long-run average reward, that is, a target level with a sufficiently high probability, that is, a percentile. The key conceptual difference between this paper and the classical problem is that our controller is not searching for an optimal policy but rather for a policy that is "good enough," knowing that such a policy will typically fail to exist if the target level and the percentile are set too high. Conceptually, our approach is somewhat analogous to that often adopted by statisticians in testing of hypotheses where it is desirable (but usually not possible!) to simultaneously minimize both the "type 1" and the "type 2" errors. See Bouakiz [5] for a review of similar approaches in economics and operations research literature and White [16] for a review of various approaches to risk-sensitivity in MDP's.

We start out by considering a problem with a single objective. It will be seen (Section IV below) that for our target level-percentile problem it is possible to present a complete (and discrete) classification of both the maximal achievable target levels and of their corresponding percentiles (see Theorem 4.3 and its corollaries). The case of a communicating MDP is particularly interesting as here every target level can be achieved with only two possible values: zero or one (see Theorem 4.1 and its corollary). In all cases our approach is constructive in the sense that we can supply an algorithm for computing a deterministic policy for any feasible target level and percentile pair.

In Section V we turn our attention to problems with multiple objectives and/or constraints. The connection of this to the problem with sample path constraint of [13] and [14] is discussed. In Section VI, we show how the techniques developed in Section IV extend directly to these problems, provided that the problems can be solved for a communicating MDP (see Theorem 6.3). The multiobjective/constrained problems are considered in detail for communicating MDP's, with constructive solutions obtained for most cases (see Theorems 6.1 and 6.2). Conclusions and some open problems are presented in Section VII.

Our analysis is made possible by the recently developed decomposition and sample path theory due to Ross and Varadarajan [14]. The logical development of the results is along the lines of Filar [8]. The latter paper, to the best of our knowledge, introduced the percentile objective criterion in the context of a limiting average Markov control problem, but substituted the long-run expected frequencies in place of actual percentile probabilities since the decomposition and sample path theory of [14] was not known at that time (the unusual way of evaluating risk in [8] was also pointed out in [16]). Some earlier related work appeared in Mitten [12], Sobel [15], and Henig [10]. In the remainder of this section we shall introduce the notation of the limiting average Markov decision process.

Manuscript received January 22, 1992; revised April 12, 1993. This work was supported in part by the AFOSR and the NSF Grants ECS-8704954 and NCR-8707620.

J. A. Filar is with the Department of Mathematics and Statistics, University of Maryland Baltimore County, Baltimore, MD 21228 USA.

D. Krass is with the Faculty of Management, University of Toronto, Toronto, Ontario, M5S1V4.

K. W. Ross is with the Department of Systems, University of Pennsylvania, Philadelphia, PA 19104 USA.

IEEE Log Number 9406094.

A finite MDP, $\Gamma$, is observed at discrete-time points $n = 1, 2, \cdots$. The state space is denoted by $S = \{1, 2, \cdots, |S|\}$. With each state $i \in S$ we associate a finite action set $A(i)$. At any time point $n$, the system is in one of the states, and an action has to be chosen by the decision-maker. If the system is in state $i$ and the action $a \in A(i)$ is chosen, then an immediate reward $r(i, a)$ is earned and the process moves to a state $j \in S$ with transition probability $p_{iaj}$, where $p_{iaj} \geq 0$ and $\sum_{j \in S} p_{iaj} = 1$.

A decision rule $u^n$ at time $n$ is a function which assigns a probability to the event that action $a$ is taken at time $n$. In general $u^n$ may depend on all realized states up to and including time $n$. A policy (or a control) $u$ is a sequence of decision rules: $u = (u^1, u^2, \cdots, u^n, \cdots)$. A policy is stationary if each $u^n$ depends only on the current state at time $n$, and $u^1 = u^2 = \cdots = u^n = \cdots$. A pure (or deterministic) policy is a stationary policy with nonrandomized decision rules. A stationary policy $f$ induces a Markov chain $P(f)$ with the transitions $P(f)_{ij} = \sum_{a \in A(i)} p_{iaj} f_{ia}$ for $i, j \in S$, where $f_{ia}$ is the probability (under $f$) that action $a$ is chosen whenever state $i$ is visited. If $P(f)$ consists of a single recurrent class of states, then $f$ is called an irreducible policy. If $P(f)$ contains some transient states in addition to a single recurrent class, then $f$ is called a unichain policy. MDP $\Gamma$ is called unichain if every stationary policy in $\Gamma$ is unichain.

Let $X_n$ and $A_n$ be the random variables that denote the state at time $n$ and the action chosen at time $n$, and define the limiting average reward as the random variable

$$R := \lim_{N \to \infty} \inf \frac{1}{N} \sum_{n=1}^{N} r(X_n, A_n).$$

It should now be clear that once a policy $u$ and an initial state $X_1 = s_1$ are fixed, the expectation $\phi(u, s_1) := E_u[R \mid X_1 = s_1]$ of $R$ is well defined and will, from now on, be referred to as the expected average reward due to a policy $u$. The classical limiting average reward problem is to find an optimal policy $u^*$ such that for all policies $u$

$$\phi(u^*, s_1) \geq \phi(u, s_1) \qquad \text{for all} \qquad s_1 \in S. \tag{1.1}$$

It is well known (e.g., see [4]) that there always exists a pure optimal policy $u^*$.

## II. Problems Relating to Percentile Objective Criteria

We shall say that any pair $(\tau, \alpha)$ such that $\tau \in \mathbb{R}$ and $\alpha \in [0, 1]$ constitutes a target level-percentile pair. We shall address the following problems.

*Problem 1:* Fix $s_1 \in S$. Given $(\tau, \alpha) \in \mathbb{R} \times [0, 1]$ does there exist a policy $u$ such that

$$P_u(R \geq \tau \mid X_1 = s_1) \geq \alpha? \tag{2.1}$$

If (2.1) holds for some policy $u$, then we shall say that $u$ achieves the target level $\tau$ at percentile $\alpha$, and $\tau$ will be called $\alpha$-achievable.

*Problem 2:* Given $\alpha \in [0, 1]$ find

$$\tau_\alpha := \sup \{\tau \mid \tau \text{ is } \alpha - \text{achievable}\}. \tag{2.2}$$

*Problem 3:* Given $\tau \in \mathbb{R}$ find

$$\alpha_\tau := \sup \{\alpha \in [0, 1] \mid \exists \text{ a policy } u \text{ s.t. (2.1) holds}\}. \tag{2.3}$$

*Remark 2.1:* It should be clear that in many situations the natural goal of maximizing the target level will be in direct conflict with the goal of maximizing the percentile value. This is because $\tau_\alpha$ is a nonincreasing function of $\alpha$, while $\alpha_\tau$ is a nonincreasing function of $\tau$.

## III. Preliminaries

We shall develop our results within the framework of the decomposition and sample path theory due to Ross and Varadarajan [14] (for a related decomposition, see [2]). This decomposition approach has also proved instrumental in solving other limiting average MDP problems with nonstandard criteria (see [14] and [3]). In this section we collect some results from [14] that will be needed for the proofs in the subsequent sections.

In [14] it is shown that the state space $S$ has a unique partition $C_1, C_2, \cdots, C_K, T$, whose properties are summarized below.

*Theorem 3.1 (Proposition 2 of [14]):* For any policy $u$, we have

$$\sum_{k=1}^{K} P_u(\Phi_k \mid X_1 = s_1) = 1$$

where

$$\Phi_k := \{X_n \in C_k \text{ almost always}\}$$

(where "almost always" means that $X_n \notin C_k$ finitely often).

The sets $C_1, \cdots, C_K$ and $T$ are referred to as strongly communicating classes and the set of transient states, respectively. For a given strongly communicating class $C_k$, denote by $\Gamma(k)$ the MDP restricted to $C_k$. Thus, the state space of $\Gamma(k)$ is $C_k$ and the action space $A_k(i)$, $i \in C_k$, is given by

$$A_k(i) = \{a \in A(i): \quad p_{iaj} = 0 \quad \forall j \notin C_k\}.$$

From [14] we know that $A_k(i)$ is nonempty for all $i \in C_k$ and that $\Gamma(k)$ is a communicating MDP. Recall that a communicating MDP is such that for any pair of states $i, j \in S$, there is a pure policy under which $j$ is accessible from $i$. Now consider the following linear program $LP(k)$

$$\begin{aligned}
\max \quad & \sum_{i \in C_k} \sum_{a \in A_k(i)} r(i, a) x_{ia} \\
\text{s.t.} \quad & \sum_{i \in C_k} \sum_{a \in A_k(i)} (\delta_{ij} - p_{iaj}) x_{ia} = 0 \qquad j \in C_k \\
& \sum_{i \in C_k} \sum_{a \in A_k(i)} x_{ia} = 1 \\
& x_{ia} \geq 0, \qquad i \in C_k, \quad a \in A_k(i).
\end{aligned}$$

Let $v_k$ denote the optimal objective function value of $LP(k)$. We can now state the following result.

*Theorem 3.2 (Lemma 4 [14]):* For all policies $u$, all initial states $s_1 \in S$, and all $k = 1, \cdots, K$, we have

$$P_u(R \leq v_k \mid \Phi_k, X_1 = s_1) = 1$$

whenever $P_u(\Phi_k, X_1 = s_1) > 0$.

## IV. Basic Results for the Single Reward Case

We shall first solve Problems 1–3 for the case of $\Gamma$ being a communicating MDP. In this case, there is one strongly communicating class $C_1$, and $T$ is empty; thus, $S = C_1$.

Consider then $LP(1)$ and to simplify notation denote $v := v_1$. Also let $\{x_{ia}^*\}$ be an optimal basic feasible solution of $LP(1)$ and $g^*$ be a stationary optimal policy constructed from $\{x_{ia}^*\}$ (e.g., see [11], [9], or [13]). Then $g^*$ satisfies

$$\phi(g^*, s_1) = v, \qquad s_1 \in S \tag{4.1}$$

where $v$ is the maximal objective function value in $LP(1)$. Moreover, the Markov chain $P(g^*)$ associated with the policy $g^*$ has at most one recurrent class plus (a perhaps empty) set of transient states.

*Theorem 4.1:* In a communicating MDP $\Gamma$ there exists a policy that achieves the target level $\tau$ with percentile $\alpha$ if and only if $\tau \leq v$. If $\tau \leq v$, then the pure policy $g^*$ achieves the target level $\tau$ with percentile $\alpha$, for any $\alpha \in [0, 1]$

*Note:* This result (at least for the unichain case) seems to be well known in the "folklore" of MDP's. A proof for the communicating case can be found in Asriev and Rotar' [1] (who establish this result for a much more general stochastic dynamic control model). Since our proof is quite simple, we present it here for completeness.

*Proof:* Since $g^*$ gives rise to a Markov chain with one recurrent class, we have

$$P_{g^*}(R = \phi(g^*, s_1) \mid X_1 = s_1) = 1$$

(e.g., see [13, Proposition 1 (iii)]). Combining this with (4.1) gives

$$P_{g^*}(R = v \mid X_1 = s_1) = 1. \tag{4.2}$$

From Theorem 3.2, we have

$$P_u(R \leq v \mid X_1 = s_1) = 1. \tag{4.3}$$

The result then follows from (4.2) and (4.3).

As a direct consequence of Theorem 4.1 we have the following corollary.

*Corollary 4.1:* In a communicating MDP $\Gamma$, $\tau_\alpha = v$ for all $\alpha \in (0, 1]$, moreover

$$\alpha_\tau = \begin{cases} 1 & \text{if } \tau \leq v \\ 0 & \text{if } \tau > v. \end{cases}$$

Problems 1–3 have now been solved for the communicating MDP's. We return to the general case, where we have strongly communicating classes $C_1, \cdots, C_K$ and the set $T$ of transient states. Denote by $g_k^*$ the pure policy of Theorem 4.1 associated with $\Gamma(k)$, the MDP restricted to $C_k$.

*Corollary 4.2:* For a fixed $k \in \{1, \cdots, K\}$, let $g$ be a pure policy that coincides with $g_k^*$ on $C_k$ and is defined arbitrarily elsewhere. Then

$$P_g(R = v_k \mid \Phi_k, X_1 = s_1) = 1$$

if $P_g(\Phi_k, X_1 = s_1) > 0$.

*Proof:* Note that the definition of $\Phi_k$ implies that $C_k$ must be reached in finite time ($P_g$ – a.s.). Thus, the result follows easily from the proof of Theorem 4.1.

Next, we shall consider a fixed target level $\tau$ and associate with it an index set $I_\tau = \{k\colon 1 \leq k \leq K, v_k \geq \tau\}$, and an auxiliary "0-1 MDP" $\Gamma_\tau$, whose states, actions, and transition probabilities are the same as $\Gamma$, but with rewards defined by

$$r^\tau(i, a) := \begin{cases} 1 & \text{if } i \in C_k \text{ and } k \in I_\tau \\ 0 & \text{otherwise.} \end{cases}$$

It is easy to see that for an arbitrary policy $u$, the expected average reward in $\Gamma_\tau$ is given by

$$\begin{aligned} &\phi^\tau(u, s_1) \\ &\quad = E_u\left[\lim_{N\to\infty} \inf \frac{1}{N} \sum_{n=1}^{N} \sum_{k \in I_\tau} 1(X_n \in C_k) \mid X_1 = s_1\right] \\ &\quad = \sum_{k \in I_\tau} P_u(\Phi_k \mid X_1 = s_1) \end{aligned} \tag{4.4}$$

where $1(\cdot)$ is the indicator function. Note that the last equality above follows from Theorem 3.1 and the definition of $\Phi_k$, since $P_u$-a.s. after some finite time $t$ we must have $X_n \in C_k$ for all $n \geq t$ and some $k \in \{1, \cdots, K\}$.

*Theorem 4.2:* Let $g^*$ be an optimal pure policy in $\Gamma_\tau$ which coincides with $g_k^*$ on $C_k$ for $k \in I_\tau$[1] There exists a policy $u$ satisfying

$$P_u(R \geq \tau \mid X_1 = s_1) \geq \alpha \tag{4.5}$$

where $\alpha$ is the percentile, if and only if $\phi^\tau(g^*, s_1) \geq \alpha$. Further, if the target $\tau$ can be achieved at percentile $\alpha$, then it can be achieved by the pure policy $g^*$.

*Proof:* From Theorem 3.1 we have that for any policy $u$

$$P_u(R \geq \tau \mid X_1 = s_1) = \sum_{k=1}^{K} P_u(R \geq \tau \mid \Phi_k, X_1 = s_1) P_u(\Phi_k \mid X_1 = s_1). \tag{4.6}$$

From Theorem 3.2 we have

$$P_u(R \geq \tau \mid \Phi_k, X_1 = s_1) = 0 \quad k \notin I_\tau. \tag{4.7}$$

From Corollary 4.2

$$\begin{aligned} 1 &= P_{g^*}(R \geq \tau \mid \Phi_k, X_1 = s_1) \\ &\geq P_u(R \geq \tau \mid \Phi_k, X_1 = s_1) \quad k \notin I_\tau \end{aligned} \tag{4.8}$$

[1] Note that there is no loss of generality here, because $g_k^*$ ensures that once the process enters $C_k$ it remains there forever. Thus it yields the maximal reward of one for every state $i \in C_k$, $k \in I_\tau$.

where the inequality follows from the optimality of $g_k^*$ for $\Gamma(k)$. Combining (4.6)–(4.8) gives

$$\begin{aligned} P_u(R \geq \tau \mid X_1 = s_1) &\leq P_{g^*}(R \geq \tau \mid X_1 = s_1) \\ &= \sum_{k \in I_\tau} P_{g^*}(\Phi_k \mid X_1 = s_1) \\ &= \phi^\tau(g^*, s_1) \end{aligned}$$

from which the result follows.

It is important to note that Theorem 4.2 provides a constructive answer to Problem 1 of Section II concerning $\alpha$-achievability of the target level $\tau$. We shall now address the problem of determining $\tau_\alpha$—the maximal achievable percentile for the fixed level $\tau$. Towards this goal we assume without loss of generality that the strongly communicating classes $C_1, \cdots, C_K$ are ordered so that

$$v_1 \geq v_2 \geq \cdots \geq v_K. \tag{4.9}$$

Recall the definition of the MDT $\Gamma_{v_k}$ (here $v_k$ is the target level). To simplify the notation, we will refer to $\Gamma_{v_k}$ as $\Gamma_k$ and to the corresponding expected average reward as $\phi^k$ instead of $\phi^{v_k}$.

*Theorem 4.3:* Let $g_k^*$ be an optimal pure policy for MDP $\Gamma_k$ chosen as in Theorem 4.2. We have for $\alpha \in (0, 1]$ that

$$\tau_\alpha = \tau^* := \max \{v_k \mid \phi^k(g_k^*, s_1) \geq \alpha, \quad k = 1, \cdots, K\}. \tag{4.10}$$

*Proof:* Let $l$ be the largest index that achieves the maximum in (4.10), $l$ is well defined since $\phi^K(g_K^*, s_1) = 1$. Since $\tau^* = v_l$, we have $I_{\tau^*} = \{1, 2, \cdots, l\}$. Thus, from Theorem 4.2, we know that

$$P_{g_l^*}(R \geq \tau^* \mid X_1 = s_1) \geq \alpha. \tag{4.11}$$

Hence, $\tau^*$ is $\alpha$-achievable, implying that $\tau_\alpha \geq \tau^*$. If strict inequality were possible in the preceding statement, then there would exist a $\tau' > \tau^*$ and a policy $u$ such that

$$P_u(R \geq \tau' \mid X_1 = s_1) \geq \alpha. \tag{4.12}$$

Now let

$$m := \max \{k : v_k \geq \tau'\}$$

noting that if $v_k < \tau'$ for all $k = 1, \cdots, K$, then the left side of (4.12) equals zero contradicting the hypothesis $\alpha > 0$. By the definition of $m$ we have

$$v_m \geq \tau' > \tau^*. \tag{4.13}$$

Applying Theorem 3.1, Theorem 4.1 and optimality $g_m^*$ to (4.12) yields

$$\begin{aligned} \alpha &\leq \sum_{k=1}^{K} P_u(R \geq \tau' \mid \Phi_k, X_1 = s_1) P_u(\Phi_k \mid X_1 = s_1) \\ &= \sum_{k=1}^{m} P_u(R \geq \tau' \mid \Phi_k, X_1 = s_1) P_u(\Phi_k \mid X_1 = s_1) \\ &\leq \sum_{k=1}^{m} P_u(\Phi_k \mid X_1 = s_1) \\ &= \phi^m(u, s_1) \leq \phi^m(g_m^*, s_1). \end{aligned} \tag{4.14}$$

But, by the definition of $\tau^*$, (4.14) implies $\tau^* \geq v_m$, which contradicts (4.13).

*Corollary 4.3:* The maximal $\alpha$-achievable target level, $\tau_\alpha$, is a monotone nonincreasing step-function of $\alpha$, defined on the interval (0, 1].

*Proof:* Choose $g_k^*$ as in Theorem 4.3. Let $\alpha_k := \phi^k(g_k^*, s_1)$ for $k = 1, \cdots, K$, so that $0 \leq \alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_K = 1$. If we define $\tau_0 := v_1$, then by Theorem 4.3, $\tau_\alpha = \tau_0$ for all $\alpha \in (0, \alpha_1]$. Similarly, $\tau_\alpha = \tau_k$, a constant for all $\alpha \in (\alpha_k, \alpha_{k+1}]$, where $\tau_k \geq \tau_{k+1}$ for each $k = 1, \cdots, K-1$.

*Corollary 4.4:* Choose $g_k^*$ as in Theorem 4.3. The maximum percentile $\alpha_\tau$ for a given target level $\tau$, is a monotone nonincreasing step function of $\tau$ defined in the interval $[v_K, v_1]$. In particular for $\tau \in (v_{k+1}, v_k]$ we have

$$\alpha_\tau = \phi^k(g_k^*, s_1)$$

for each $k = 1, \cdots, K-1$.

*Proof:* This follows easily from the monotonicity of $\phi^k(g_k^*, s_1)$ in the index $k$.

*Remark 4.1:* Corollaries 4.3 and 4.4 demonstrate the strength of the percentile objective criteria. Namely, the decomposition of states into $C_1, \cdots, C_K$ and $T$, and the subsequent computation of policies $g_k^*$ together with "breakpoints" $\tau_k$, and $v_k$ for $\tau_\alpha$ and $\alpha_\tau$, respectively, allows for a flexible and practical evaluation of gain-risk trade-offs in an average reward MDP.

In view of Corollaries 4.3 and 4.4 the only "reasonable" choices of $\alpha$ and $\tau$ are of the special form $(\tau, \alpha)$ with $\tau = \tau_\alpha$ and $\alpha = \alpha_\tau$; these correspond to Pareto-optimal solutions.

The preceding results are summarized in the following algorithm, which (for a fixed initial state $s_1$) finds all target-level-percentile pairs of the indicated "special form"

*Step 1:* Apply the algorithm of [14] to find the decomposition $C_1, \cdots, C_K, T$.

*Step 2:* For each $k \in \{1, \cdots, K\}$, find the value $v_k$ of $\Gamma(k)$. Order the strongly communicating classes so that $v_1 \geq \cdots \geq v_K$.

*Step 3:* For each $k \in \{1, \cdots, K\}$, form the MDP $\Gamma_k$ and find the optimal policy $g_k^*$. Let $\alpha_k = \phi^k(g_k^*, s_1)$.

*Step 4:* Let $J = \{(v_k, \alpha_k) \mid k = 1, \cdots, K\}$. If $(\tau_1, \alpha), (\tau_2, \alpha) \in J$ and $\tau_1 > \tau_2$, then eliminate $(\tau_2, \alpha)$ from $J$. Continue until no further eliminations are possible.

*Step 5:* We have constructed the set $J = \{(\tau, \alpha) = (\tau_\alpha, \alpha_\tau) \mid \tau \in \mathbb{R}, \alpha \in (0, 1]\}$.

Note that Step 2 is not hard computationally, since very efficient algorithms are known for communicating MDP's. In Step 3 one should use the aggregated MDP method of [14], where each strongly communicating class is replaced by one state. In addition to computational efficiency, this method will automatically yield a deterministic optimal policy satisfying the conditions of Theorem 4.2. Also, since when $k$ is incremented by one, only one immediate reward changes in the aggregated problem (the rest of the data stay the same), perhaps a parametric solution method can be used (e.g. by using LP algorithms for solving the aggregated problem). Further computational efficiency can be gained by quitting

Step 3 as soon as $\alpha_k = 1$ for some $k$ (since all subsequent $\alpha_k$'s are automatically equal to one). Note also that steps 1 and 2 need not be repeated for different starting states. Finally, we note that before starting the algorithm, one should check whether MDP $\Gamma$ is communicating (easily verifiable conditions can be found in [9]). If so, the complete characterization is immediately available from Corollary 4.1.

## V. Problems with Multiple Rewards and Constraints—Formulations

A natural extension of the percentile objective criteria is to the case where each action carries multiple immediate rewards, i.e., each immediate reward is actually a vector of some fixed length $L$

$$\boldsymbol{r}(i, a) \in \mathbb{R}^L \text{ for } i \in S,\ a \in A(i).$$

The definition of the limiting average reward in Section I leads to a vector $\boldsymbol{R}$ of random variables.

At this point two approaches can be taken. Under the first approach, a separate target level-percentile pair would be specified for each of the $L$ components of $\boldsymbol{R} = (R_1, \cdots, R_L)$. This leads to the following multi-objective version of Problem 1 (of Section II).

*Problem 4:* Fix $s_1 \in S$. Given a pair of vectors $(\boldsymbol{\tau}, \boldsymbol{\alpha}) \in \mathbb{R}^L \times [0, 1]^L$, does there exist a policy $u$ such that

$$P_u(R_l \geq \tau_l \mid X_1 = s_1) \geq \alpha_l \text{ for } l = 1, \cdots, L?$$

This problem appears to present serious difficulties and is presently unsolved. Below we present an example showing that stationary policies do not suffice for this case, thus indicating that it is unlikely that a simple extension of the results and methods developed in Section IV will work in this case.

*Example 5.1:* Consider the following MDP with $L = 2$

| $i = 1$ | | $i = 2$ | |
|---|---|---|---|
| $\boldsymbol{r}(1, 1) = (1, 0)$, | $p_{111} = 1$ | $\boldsymbol{r}(1, 2) = (0, 0)$, | $p_{122} = 1$ |
| $\boldsymbol{r}(2, 1) = (0, 1)$, | $p_{212} = 1$. | | |

Thus action 1 is absorbing in both states, and action 2 in state 1 results in immediate rewards of $(0, 0)$ and transition to state 2 with probability one.

Suppose the initial state is one and that the target level-percentile pairs are $\left(\frac{1}{2}, \frac{1}{2}\right)$ for each component of the reward vector. Note that for any stationary policy $f$

$$P_f(R_1 = 1 \mid X_1 = 1) = 1,$$
$$P_f(R_2 = 0 \mid X_1 = 1) = 1, \text{ if } f_{11} = 1$$

and

$$P_f(R_1 = 0 \mid X_1 = 1) = 1,$$
$$P_f(R_2 = 1 \mid X_1 = 1) = 1, \text{ if } f_{11} < 1.$$

Therefore, no stationary policy suffices.

Let $g$ be the deterministic policy that takes action 1 in each state, and let $u^1$ be the decision rule that takes actions 1 and 2 with probabilities 1/2 each in state 1. Define the nonstationary policy $u$ to be $u = (u^1, g)$ (i.e., $u$ uses the decision rule $u^1$ at time 1 and follows $g$ thereafter). It is easy to see that

$$P_u\left(R_1 \geq \frac{1}{2} \middle| X_1 = 1\right) = \frac{1}{2}, \quad P_u\left(R_2 \geq \frac{1}{2} \middle| X_1 = 1\right) = \frac{1}{2}$$

and thus $u$ achieves both the specified target-levels at the corresponding percentiles. In fact the same conclusions apply for any $\tau_1, \tau_2 \in (0, 1)$.

Under the second approach to a problem with multiple rewards, a separate target level is specified for each component of $\boldsymbol{R}$, and all target levels are required to be achieved simultaneously at a single percentile level $\alpha$.

*Problem 5:* Fix $s_1 \in S$. Given a vector $\boldsymbol{\tau} \in \mathbb{R}^L$ and $\alpha \in [0, 1]$, does there exist a policy $u$ such that

$$P_u(R_l \geq \tau_l, l = 1, \cdots, L \mid X_1 = s_1) \geq \alpha? \tag{5.1}$$

Intuitively, the difference between the two approaches is that in Problem 4, the requirements are placed on the marginal distributions of the components of $\boldsymbol{R}$, while in Problem 5 the requirement is placed on the joint distribution of $\boldsymbol{R}$. Henceforth, we will refer to these two approaches as the "marginal-probability" and the "joint-probability" formulations, respectively. While the marginal-probability formulation provides for more modeling flexibility, the advantage of the joint-probability formulation is that it leads to more tractable problems: most of the results obtained for Problem 1 will be extended to Problem 5.

We note that an extension to multiple rewards is especially important, since some of the components of $\boldsymbol{r}(i, a)$ can be regarded as negatives of the costs. In this case, the corresponding components of $\boldsymbol{R}$ and $\boldsymbol{\tau}$ can be regarded as constraints which must be satisfied at the specified percentile level (either singly in Problem 4 or jointly in Problem 5). This can be seen as a generalization of the "sample path constraint" of Ross and Varadarajan [13] and [14] (which in our notation corresponds to $\alpha = 1$ and $L = 1$).

## VI. Basic Results for the Multiple Rewards Case

In this section we consider Problem 5 for the case of communicating MDP's. We employ the linear-programming-based techniques developed in [6], [11], and [13] and introduced in Section III above. It should be noted that the problem considered in this section is related to the problem of finding optimal policies in average-reward communicating MDP's with state-action frequency constraints. It is well known (see e.g., [7, example 9.3]) that stationary policies may not be optimal in this case. For a problem with one inequality constraint, [13] shows that an $\epsilon$-optimal stationary policy can always be found (the policy is generally not deterministic). This result is related to Theorem 6.2.

In the sequel, all vector inequalities and equalities are defined componentwise, that is $\boldsymbol{R} \geq \boldsymbol{\tau}$ means $R_l \geq \tau_l$, $l = 1, \cdots, L$.

We first state some preliminary definitions and results. Consider the following polyhedral set

$$\Delta = \begin{cases} \sum_{a\in A(i)} x_{ia} = \sum_{j\in S}\sum_{a\in A(j)} p_{jai}x_{ja}, & i \in S \\ \sum_{i\in S}\sum_{a\in A(i)} x_{ia} = 1 \\ x_{ia} \geq 0, \quad i \in S,\ a \in A(i) \end{cases} \tag{6.1}$$

(this is simply the feasible region of $LP(k)$ of Section III). For $\boldsymbol{x} = (x_{ia}) \in \Delta$, define a stationary policy

$$f(\boldsymbol{x})_{ia} = \begin{cases} \dfrac{x_{ia}}{\sum_{a\in A(i)} x_{ia}} & \text{if } i \in S,\ \sum a \in A(i)\, x_{ia} > 0, \\ & a \in A(i) \\ \dfrac{1}{|A(i)|} & \text{if } i \in S,\ \sum_{a\in A(i)} x_{ia} = 0. \end{cases} \tag{6.2}$$

Define the following random variable (whenever it exists), representing long-term average state-action frequencies

$$Z_{ia} = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} 1[X_t = i, \quad A_t = a]. \tag{6.3}$$

We will need the following well-known results, the proofs of which can be found in [11] and [13].

*Lemma 6.1:* Let $\Gamma$ be an arbitrary MDP.

i) If $f$ is a stationary policy then $P_f$-almost surely, $\boldsymbol{Z}$ is well-defined, $\boldsymbol{Z} \in \Delta$, and

$$\boldsymbol{R} = \sum_{i\in S}\sum_{a\in A(i)} \boldsymbol{r}(i, a)Z_{ia}.$$

ii) If $f$ is a stationary policy and $f$ is unichain, then

$$Z_{ia} = \pi(f)_i f_{ia} \quad P_f - \text{a.s.}$$

where $\boldsymbol{\pi}(f)$ is the unique stationary probability vector of $P(f)$.

iii) If $\boldsymbol{x} \in \Delta$ and $f(\boldsymbol{x})$ is unichain, then

$$\boldsymbol{Z} = \boldsymbol{x}P_{f(\boldsymbol{x})} - \text{a.s.}$$

iv) If $\Gamma$ is a communicating MDP and $\boldsymbol{x} \in \Delta$, $\boldsymbol{x} > 0$, then $f(\boldsymbol{x})$ is an irreducible stationary policy.

Let $\boldsymbol{\tau}$ be a given target vector and $\alpha$ a given percentile level. We now define the following linear program ***LP***

max $b$
Subject to

$$\begin{aligned} &\textstyle\sum_{a\in A(i)} x_{ia} = \sum_{j\in S}\sum_{a\in A(j)} p_{jai}x_{ja} && i \in S \\ &\textstyle\sum_{i\in S}\sum_{a\in A(i)} x_{ia} = 1 \\ &\textstyle\sum_{i\in S}\sum_{a\in A(i)} r(i, a)_l x_{ia} \geq \tau_l && l = 1, \cdots, L \\ &x_{ia} \geq b, \quad i \in S, \quad a \in A(i) \\ &b \geq 0. \end{aligned}$$

Note that the feasible region of ***LP*** is contained in $\Delta \times \mathbb{R}$.

*Theorem 6.1 :* Let $\Gamma$ be a communicating MDP.

i) If ***LP*** is infeasible, then for any policy $u$

$$P_u(\boldsymbol{R} \geq \boldsymbol{\tau} \mid X_1 = s_1) = 0.$$

ii) Suppose ***LP*** is feasible. Let $(\boldsymbol{x}^*, b^*)$ be an optimal solution and $f^* = f(\boldsymbol{x}^*)$. If $b^* > 0$, or $f^*$ is unichain, then

$$P_{f^*}(\boldsymbol{R} \geq \boldsymbol{\tau} \mid X_1 = s_1) = 1.$$

*Proof:*

i) The proof is analogous to the proof of Proposition 2 in [14].

ii) If $b^* > 0$ then $f^*$ is irreducible by Lemma 6.1-iv) and thus, by Lemma 6.1-iii), $\boldsymbol{x}^* = \boldsymbol{Z}\ P_{f^*}$-a.s. It now follows by Lemma 6.1-i) and the feasibility of $\boldsymbol{x}^*$ that

$$\boldsymbol{R} \geq \boldsymbol{\tau} P_{f^*} - \text{a.s.}$$

This result is a counterpart of Theorem 4.1 for the single-objective case. When conditions in part ii) of Theorem 6.1 hold, it provides a solution to Problem 5 for any $\alpha \in [0, 1]$, and if the condition in part i) holds then Problem 5 has no solutions for any $\alpha$. Unlike Theorem 4.1, however, the current result does not provide a complete characterization of solutions. If the optimal value $b^*$ of ***LP*** is equal to zero and $f(\boldsymbol{x}^*)$ is not unichain, then it is not currently known whether Problem 5 has any solutions or any solutions in stationary policies. The following example shows that it is possible to have a situation where no feasible stationary policies exist when $b^* = 0$.

*Example 6.1:* Consider the following MDP $\Gamma$ with $L = 2$

| $i = 1$ | | $i = 2$ | |
|---|---|---|---|
| $\boldsymbol{r}(1, 1) = (1, 0)$, | $p_{111} = 1$ | $\boldsymbol{r}(2, 1) = (0, 1)$, | $p_{212} = 1$ |
| $\boldsymbol{r}(1, 2) = (0, 0)$, | $p_{122} = 1$ | $\boldsymbol{r}(2, 2) = (0, 0)$, | $p_{221} = 1$ |

where $s_1 = 1$, $\boldsymbol{\tau} = (1/2, 1/2)$ and $\alpha > 0$. Then ***LP*** is feasible with $b^* = 0$ and

$$\boldsymbol{x}^* = (x_{11}, x_{12}, x_{21}, x_{22})^T = (1/2, 0, 1/2, 0)^T.$$

Thus $f(\boldsymbol{x}^*)$ is not unichain, and in fact it is easy to see that for any stationary policy $f$, $P_f(\boldsymbol{R} \geq \boldsymbol{\tau}) = 0$. It is not known whether there exists some nonstationary policy $u$ such that $P_u(\boldsymbol{R} \geq \boldsymbol{\tau}) > 0$.

It might be reasonable to suppose that when $b^* = 0$, the presence of feasible stationary policies might be detected by checking whether $f(\boldsymbol{x}^*)$ is unichain for any optimal basic feasible solution $\boldsymbol{x}^*$ of ***LP*** (this would cover Example 6.1 where there is only one optimal basic feasible solution). The following example, however, shows that it is possible that $b^* = 0$, $f(\boldsymbol{x}^*)$ is not unichain for any optimal basic feasible solution, and yet, feasible stationary policies exist for any $\alpha > 0$.

*Example 6.2:* Consider the following MDP $\Gamma$ with $L = 3$

| $(i = 1)$ | | $i = 2$ | |
|---|---|---|---|
| $\boldsymbol{r}(1,1)=(1,0,0)$, | $p_{111} = 1$ | $\boldsymbol{r}(2,1)=(0,1,0)$, | $p_{211} = 1$ |
| $\boldsymbol{r}(1,2)=(0,0,0)$, | $p_{122} = 1$ | $\boldsymbol{r}(2,2)=(0,1,0)$, | $p_{223} = 1$ |
| $\boldsymbol{r}(1,3)=(-1,0,0)$, | $p_{131} = 1$ | | |
| $i = 3$ | | | |
| $\boldsymbol{r}(3,1)=(0,0,1)$, | $p_{313} = 1$ | | |
| $\boldsymbol{r}(3,2)=(0,0,0)$, | $p_{322} = 1$ | | |

and let $s_1 = 1$. Take some $\alpha > 0$ and $\boldsymbol{\tau} = (1/4, 1/4, 1/4)$. It is not hard to verify that the resulting LP has the optimal value $b^* = 0$ with only two optimal basic feasible solutions

$$\begin{aligned} \boldsymbol{x}^1 &= (x_{11}, x_{12}, x_{13}, x_{21}, x_{22}, x_{31}, x_{32})^T \\ &= (1/4, 1/4, 0, 1/4, 0, 1/4, 0)^T \end{aligned}$$

and

$$\boldsymbol{x}^2 = (1/4,\ 0,\ 0,\ 0,\ 1/4,\ 1/4,\ 1/4)^T.$$

Neither $f(\boldsymbol{x}^1)$ nor $f(\boldsymbol{x}^2)$ is unichain and, in fact, the probability of achieving $\boldsymbol{\tau}$ is zero for both of these policies. If we take the feasible point

$$\boldsymbol{x}^* = \frac{1}{2}\boldsymbol{x}^1 + \frac{1}{2}\boldsymbol{x}^2 = (1/4,\ 1/8,\ 0,\ 1/8,\ 1/8,\ 1/4,\ 1/8)^T$$

then $f(\boldsymbol{x}^*)$ is unichain and $P_f(\boldsymbol{x}^*)(\boldsymbol{R} \geq \boldsymbol{\tau} \mid X_1 = s_1) = 1$

We will call a target vector $\boldsymbol{\tau}$ for which ***LP*** is feasible and has optimal value $b^* = 0$ an indeterminate target vector.

We note that this case does not arise in the case of unichain MDP's (since one of the two conditions of Theorem 6.1 must be met), and we have the following corollary.

*Corollary 6.1:* Suppose $\Gamma$ is a unichain MDP. Then either ***LP*** is infeasible, in which case Problem 5 has no solutions for any $\alpha > 0$, or ***LP*** is feasible with an optimal solution $(b^*, \boldsymbol{x}^*)$, in which case for $f = f(\boldsymbol{x}^*)$, $P_f(\boldsymbol{R} \geq \boldsymbol{\tau} \mid X_1 = s_1) = 1$.

In the remainder of this section we show that the difficulties associated with indeterminate target vectors can always be avoided by a slight relaxation of the target vector and that indeterminate target vectors are sufficiently "rare" in the set of all possible target vectors. We will need the following result.

*Lemma 6.2:* Suppose $\Gamma$ is a communicating MDP. Take $\boldsymbol{x} \in \Delta$. Then for any $\epsilon > 0$ and $s_1 \in S$, there exists an irreducible stationary policy $f$ such that

$$P_f(\|\boldsymbol{Z} - \boldsymbol{x}\| \leq \epsilon \mid X_1 = s_1) = 1$$

(where $\|\cdot\|$ is an arbitrary vector norm).

*Proof:* If $\boldsymbol{x} > 0$, then $f = f(\boldsymbol{x})$ is irreducible by Lemma 6.1-iv), and it follows by Lemma 6.1-iii) that $\boldsymbol{Z} = \boldsymbol{x}$ $P_f$-a.s.

Now suppose that $x_{ia} = 0$ for some $i \in S$, $a \in A(i)$. Let $g$ be a completely randomized (stationary) policy (i.e., $g_{ia} = \frac{1}{|A(i)|}$ for any $i \in S$, $a \in A(i)$). Since $\Gamma$ is a communicating MDP, $g$ must be irreducible. By Lemma 6.1-ii), $\boldsymbol{Z} = \boldsymbol{x}(g)$ $P_g$-almost surely, where $\boldsymbol{x}(g)_{ia} = \pi(g)_i g_{ia}$ for $i \in S$, $a \in A(i)$. By Lemma 6.1-i), $\boldsymbol{x}(g) > 0$ and $\boldsymbol{x}(g) \in \Delta$. Let

$$\boldsymbol{x}(\lambda) = \lambda \boldsymbol{x} + (1 - \lambda)\boldsymbol{x}(g) \text{ for } \lambda \in (0,\ 1).$$

Clearly, $\boldsymbol{x}(\lambda) \in \Delta$ and $\boldsymbol{x}(\lambda) > 0$ for any $\lambda < 1$. It follows by continuity of $\boldsymbol{x}(\lambda)$ with respect to $\lambda$ that we can choose $\lambda^* \in (0,\ 1)$ so that

$$\|\boldsymbol{x}(\lambda^*) - \boldsymbol{x}\| \leq \epsilon. \tag{6.4}$$

Let $f = f(\boldsymbol{x}(\lambda^*))$. By Lemma 6.1-iv), $f$ must be irreducible. It follows by Lemma 6.1-iii) that $Z = \boldsymbol{x}(\lambda^*)$ $P_f$-a.s. The result now follows immediately from (6.4).

We are now ready to prove the following result.

*Theorem 6.2:* Suppose $\Gamma$ is a communicating MDP and $\boldsymbol{\tau}$ is an indeterminate target vector. Let $\boldsymbol{\delta}$ be an arbitrary vector with positive components and choose $\epsilon > 0$. Then there exists an irreducible stationary policy $f$ such that

$$P_f(\boldsymbol{R} \geq \boldsymbol{\tau} - \epsilon\boldsymbol{\delta} \mid X_1 = s_1) = 1.$$

*Proof:* Let $(b^*, \boldsymbol{x}^*)$ be an optimal solution of ***LP*** with target vector $\boldsymbol{\tau}$. By assumption, $b^* = 0$. By Lemma 6.2, for any $\gamma > 0$, there exists an irreducible stationary policy $f$ such that $\|\boldsymbol{x}' - \boldsymbol{x}^*\| \leq \gamma$, where $\boldsymbol{x}' = \boldsymbol{Z}$ is a constant $P_f$-almost surely. Choose $\gamma$ small enough so that

$$\sum_{i \in S} \sum_{a \in A(i)} x'_{ia} \boldsymbol{r}(i,\ a) \geq \boldsymbol{\tau} - \epsilon\boldsymbol{\delta} \quad P_f - \text{a.s.} \tag{6.5}$$

(this can always be done since the left-hand side of (6.5) is a continuous function of $\boldsymbol{x}'$ and $\sum_{i \in S} \sum_{a \in A(i)} x^*_{ia} \boldsymbol{r}(i,\ a) \geq \boldsymbol{\delta}$.

Since $f$ is irreducible, by Lemma 6.1-ii), $P_f$-almost surely $\boldsymbol{x}' > 0$, i.e., $\boldsymbol{x}' \geq b' > 0$ for some $b' \in \mathbb{R}$. It follows from (6.5) that $(b',\ \boldsymbol{x}')$ is feasible in ***LP*** with the target vector $\boldsymbol{\tau} - \epsilon\boldsymbol{\delta}$ and consequently, the optimal value for this LP is positive. The result now follows from Theorem 6.1-ii).

Thus, any strict relaxation of an indeterminate target vector produces a target vector for which Problem 5 can be solved by an irreducible policy for any percentile level $\alpha$. Geometrically, the situation is as follows: let $T = \{\boldsymbol{\tau} \mid \boldsymbol{LP} \text{ is feasible}\}$. $T$ must be a closed set. Let

$$\begin{aligned} TS = \{&\boldsymbol{\tau} \mid \text{Problem 5 has a solution in unichain policies} \\ &\text{for any percentile } \alpha \in [0,\ 1]\}, \\ &\cdot TF = \{\boldsymbol{\tau} \mid \boldsymbol{LP} \text{ has optimal value } b^* > 0\}, \\ &\cdot TI = \{\boldsymbol{\tau} \mid \boldsymbol{\tau} \text{ is an indeterminate target vector}\}. \end{aligned}$$

Then $TS \cup TI = T$, $TF \subset TS$, and $TI \subset \text{boundary}\{T\}$. Note also that the intersection of $TS$ and $TI$ need not be empty, as shown by Example 6.2. We illustrate the relationships between these sets in the context of Example 6.1, where

$$\begin{aligned} T = &\{\tau_1 \leq 1,\ \tau_2 \leq 0\} \cup \{\tau_1 \leq 0,\ \tau_2 \leq 1\} \\ &\cup \{\tau_1 + \tau_2 \leq 1,\ \tau_1 > 0,\ \tau_2 > 0\} \end{aligned}$$

(represented by the shaded region on Fig. 1)

$$TS = T \backslash \{\tau_1 + \tau_2 = 1, \quad \tau_1 > 0,\ \tau_2 > 0\},$$

$$\begin{aligned} TI = &\{\tau_1 = 1,\ \tau_2 \leq 0\} \cup \{\tau_1 \leq 0,\ \tau_2 = 1\} \\ &\cup \{\tau_1 + \tau_2 = 1, \quad \tau_1 > 0,\ \tau_2 > 0\} \end{aligned}$$

(the boundary of $T$), and

$$TF = T \backslash TI.$$

*Remark 6.1:* In summary, our results for the communicating case with multiple rewards closely parallel the corresponding results for the single reward: for "most" target vectors, either Problem 5 has no solution or else it is solvable by irreducible policies for any percentile level. The differences from the single reward case lie in the fact that the feasible stationary policies might have to be randomized (for single reward deterministic policies sufficed), and in our inability to

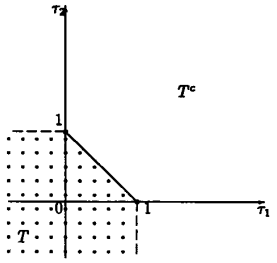



Fig. 1. The set of possible target vectors for Example 6.1.

handle (completely) the indeterminate target vector case. If the modeler has some flexibility in setting the target vector, the indeterminate case can always be avoided. In some cases, however, such flexibility might not exist. Therefore, we consider the further study of the indeterminate case to be important. Specifically, the following questions should be addressed:

1) Must nonstationary feasible policies exist when $\tau$ is indeterminate? In particular, do such policies exist in Example 6.1?
2) Can the indeterminate target vectors for which Problem 5 has a solution in unichain policies (i.e., $\tau \in TI \cap TS$) be characterized? It would be particularly interesting to find a computationally simple characterization or show that one does not exist.

We now turn our attention to the general (multichain) MDPs. Surprisingly, the extension of our communicating MDP results to this case can be done quite easily by employing the decomposition and sample path theory.

As in Section IV above, let $C_1, \cdots, C_K, T$ be the strongly communicating classes and the set of transient states of $\Gamma$, and let $\Gamma(k)$ be the MDP restricted to $C_k$. Define the index set

$$J_\tau = \{k \mid 1 \le k \le K, \text{ The optimal value of } \boldsymbol{LP} \text{ for } \Gamma(k) \text{ is positive}\}.$$

For $k \in J_\tau$, let $f_k^*$ be the stationary policy of Theorem 6.1-ii) associated with the MDP $\Gamma(k)$. Following Section IV, define MDP $\Gamma_\tau$ whose states, actions, and the transition law are the same as $\Gamma$, but with the rewards defined by

$$r^\tau(i, a) := \begin{cases} 1 & \text{if } i \in C_k,\ a \in A(i) \text{ and } k \in J_\tau \\ 0 & \text{otherwise.} \end{cases} \tag{6.6}$$

*Theorem 6.3:* Assume the target vector $\tau$ is not an indeterminate target vector for any $\Gamma(k)$, $k = \{1, \cdots, K\}$. Let $f^*$ be an optimal stationary policy in $\Gamma_\tau$ which coincides with $f_k^*$ on $C_k$ for $k \in J_\tau$.[2] There exists a policy $u$ satisfying

$$P_u(R \ge \tau \mid X_1 = s_1) \ge \alpha \tag{6.7}$$

if and only if $\phi^\tau(f^*, s_1) \ge \alpha$. Further, if the target vector $\tau$ can be achieved at percentile $\alpha$, then it can be achieved by the stationary policy $f^*$.

*Proof:* Exactly the same as in Theorem 4.2.

This result provides a constructive answer to Problem 5, provided $\tau$ satisfies the assumption of Theorem 6.3.

[2] See the footnote in Theorem 4.2.

*Remark 6.2:* Similarly to Problem 3 of Section II, define

$$\alpha_\tau = \sup\{\alpha \in [0, 1] \mid \exists \text{ a policy } u \text{ s.t. (5.1) holds}\}.$$

It is clear from the proof of Theorem 4.2 that $\alpha_\tau = \phi^\tau(f^*, s_1)$ and the policy $f^*$ constructed above achieves $\tau$ at percentile $\alpha_\tau$.

## VII. Problems for Further Research

In the preceding sections we have outlined several problems for further research. For the multiple reward case, a satisfactory treatment of the marginal-probability formulation described in Section V would be very useful. Open problems connected with the indeterminate target vector case were discussed in Remark 6.1.

Another important open problem associated with the percentile objective criterion is the satisfactory treatment of the discounted case. Analysis of a similar problem for this case can be found in Bouakiz [5]; however, no algorithmic results were obtained. Problems 1–3 of Section II appear to be much harder in this case, mainly because no equivalent of the decomposition theory is known (or perhaps, possible) for discounted MDP's.

## Acknowledgment

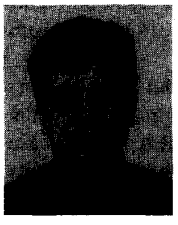
The authors thank M. Teboulle for his helpful discussions.

## References

[1] A. V. Asriev and V. I. Rotar', "On asymptotic optimality in probability and almost surely in dynamic control," *Stochastics and Stochastics Rep.*, vol. 33, pp. 1–16, 1990.
[2] J. Bather, "Optimal decision procedures in finite Markov chains. Part III: General convex systems," *Advances in Applied Prob.*, vol. 5, pp. 541–553, 1973.
[3] M. Bayal-Gursoy and K. W. Ross, "Variability sensitive Markov decision processes," *Mathematics Oper. Res.*, vol. 17, no. 3, pp. 558–572, 1992.
[4] D. Blackwell, "Discrete dynamic programming," *Annals Math. Stat.*, vol. 33, pp. 719–726, 1962.
[5] M. Bouakiz, "Risk sensitivity in stochastic optimization with applications," Ph.D. dissertation, Georgia Inst. Tech., Atlanta, 1985.
[6] C. Derman, *Finite State Markovian Decision Processes.* New York: Academic, 1970.
[7] E. A. Feinberg, "Constrained semi-Markov decision processes with average rewards," working paper, State Univ. of New York, Stony Brook, 1992.
[8] J. A. Filar, "Percentiles and Markovian decision processes," *Oper. Res. Lett.*, vol. 2, pp. 13–15, 1983.
[9] J. A. Filar and T. A. Schulz, "Communicating MDPs: equivalence and LP properties, *Op. Res. Lett.*, vol. 7, pp. 303–307, 1988.
[10] M. I. Henig, "The principle of optimality in dynamic programming with returns in partially ordered sets," *Math. Op. Res.*, vol. 10, pp. 462–470, 1985.
[11] L. C. M. Kallenberg, *Linear Programming and Finite Markovian Control Problems.* Amsterdam: Mathematical Center Tracts 148, 1983.
[12] L. G. Mitten, "Preference order dynamic programming," *Management Science*, vol. 21, pp. 43–46, 1975.
[13] K. W. Ross and R. Varadarajan, "Markov decision processes with sample path constraints: The communicating case," *Op. Res.*, vol. 37, pp. 780–790, 1989.
[14] ——, "Multichain Markov decision processes with a sample path constraint: A decomposition approach," *Math. Op. Res.*, vol. 16, no. 1, pp. 195–207, 1991.
[15] M. J. Sobel, "Ordinal dynamic programming," *Management Science*, vol. 21, pp. 967–975, 1975.
[16] D. J. White, "Mean, variance, and probabilistic criteria in finite Markov decision processes: A review," *JOTA*, vol. 56, no. 1, pp. 1–29, 1988.

**Jerzy A. Filar** was born in Warsaw in 1949. He studied Mathematics and Statistics at the University of Melbourne, Australia and at Monash University. He received the Ph.D. degree at the University of Illinois, Chicago, in 1980.

Dr. Filar is currently Professor of Mathematics and Statistics at the University of South Australia and is Director of its Centre for Mathematical Applications. He has also held academic positions at the University of Minnesota, the Johns Hopkins University, and the University of Maryland, Baltimore County. His current research interests include operations research, control theory, and environmental modeling.



**Dmitry Krass** received the B.Sc. in mathematics from the University of Chicago in 1984 and the M.S.E. and the Ph.D. in operations research from Johns Hopkins University in 1986 and 1989, respectively.

Since 1989, Dr. Krass has been as Associate Professor of Operations Management and Statistics at the Faculty of Management, University of Toronto. His current research interests include optimization and stochastic dynamic programming, and in developing advanced tools for managerial decision support in stochastic environments.

**Keith W. Ross** (S'82–M'85–SM'90) received the B.S. degree from Tufts University, Medford, MA, in 1979, the M.S. degree from Columbia University, New York, in 1981, and the Ph.D. degree from the University of Michigan, Ann Arbor, in 1985.

He is an Associate Professor in the Department of Systems Engineering, University of Pennsylvania, Philadelphia. He also holds secondary appointments in the Computer Information Science and in the Operations and Informations Management (Wharton) Departments. In 1980 he designed satellite radar systems as an employee of AVCO. He has been a visiting scholar at several research and academic institutions in France. His current research interests include protocols and traffic management in high-speed telecommunication networks, including local area networks, wide area data networks, voice networks, and broadband integrated services digital networks.

Dr. Ross is the Program Chairman of the 1995 ORSA Telecommunications Conference and is an Associate Editor for *Probability in the Engineering and the Information Sciences* and for *Telecommunication Systems*. He has published over 30 papers in leading journals and is completing a book on multiservice loss models for broadband telecommunication networks. He is the recipient of numerous grants from AT&T and NSF.