Proceedings of the 28th Conference
on Decision and Control
Tampa, Florida • December 1989



# Percentile Objective Criteria in Limiting Average Markov Control Problems

Jerzy A. Filar
Department of Mathematics and Statistics
University of Maryland Baltimore County
Baltimore, MD 21228

Dmitry Krass
Faculty of Management
University of Toronto
Toronto, Canada

Keith W. Ross
Department of Systems
University of Pennsylvania
Philadelphia, PA 19104

## 1. Introduction and Definitions

Infinite horizon Markov Control Problems, or 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 controller aims to maximize the expected value of the limit-average ("long-run average") of an infinite stream of single-stage rewards or outputs. There are now a number of good algorithms for computing optimal deterministic policies in the limiting average MDP's (e.g., see Blackwell [2], Derman [3], and Kallenberg [6]).

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 controller 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.

It will be seen 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 5 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: 0 or 1 (see Theorem 3 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. Our analysis is made possible by the recently developed decomposition theory due to Ross and Varadarajan [7], and the logical development of the results is along the lines of Filar [4]. The latter paper, to the best 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 theory of [7] was not known at that time. In the remainder of this section we shall introduce the notation of the limiting average Markov Decision Process.

A finite *Markov Decision Process*, $\Gamma$, is observed at discrete time points $n = 1, 2, \ldots$. The *state space* is denoted by $S = \{1, 2, \ldots, |S|\}$. With each state $s \in S$ we associate a finite *action set* $A(s)$. At any time point $n$, the system is in one of the states

[1]This work was supported in part by the AFOSR and the NSF under the grant ECS-8704954. We are grateful to Marc Teboulle for discussing this subject with us.

[2]The work of this author was supported in part by NSF grant NCR-8707620

and an action has to be chosen by the controller. If the system is in state $s$ and the action $a \in A(s)$ is chosen, then an immediate reward $r(s,a)$ is earned and the process moves to a state $t \in S$ with transition probability $p_{sat}$, where $p_{sat} \geq 0$ and $\sum_{t \in S} p_{sat} = 1$.

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: $u = (u^1, u^2, \ldots, u^n, \ldots)$. A policy is *stationary* if each $u^n$ depends only on the current state at time $n$, and $u^1 = u^2 = \ldots = u^n = \ldots$ . A *pure* (or *deterministic*) policy is a stationary policy with nonrandomized decision rules. 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 actual *limiting average reward* as the random variable

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

It should now be clear that once a policy $u$ and an initial state $X_1 = s_1$ are fixed, the expectation $\phi(u, s_1) := E_u(R|X_1 = s_1)$ 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 $u$

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

It is well-known (e.g., see [2]) that there always exists a pure optimal policy $u^*$.

## 2. Problems Relating To Percentile Objective Criteria

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

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

$$P_u\{R \geq k | X_1 = s_1\} \geq \alpha. \tag{2.1}$$

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

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

$$k_\alpha := \sup\{k | k \text{ is } \alpha\text{-achievable}\}. \tag{2.2}$$

**Problem 3.** Given $k \in \mathbb{R}$ find

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

**Remark 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 $k_\alpha$ is a non-increasing function of $\alpha$, while $\alpha_k$ is a non-increasing function of $k$.

## 3. Preliminaries

We shall develop out results within the framework of the decomposition theory due to Ross and Varadarajan [7]. (For a related decomposition, see Bather [1]). In this section we collect some results from [7] that will be needed for the proofs in the subsequent sections.

In [7], it is shown that the state space $S$ has a unique partition $C_1, C_2, \ldots, C_p, T$, whose properties are summarized below.

**Theorem 1** *(Thm 2.7 of [7]). For any policy $u$, we have*

$$\sum_{i=1}^{p} P_u(\Phi_i | X_1 = s_1) = 1,$$

*where*

$$\Phi_i := \{X_n \in C_i \text{ almost always }\}.$$

The sets $C_1, \ldots, C_p$ are referred to as strongly communicating classes. For a given strongly communicating class $C_1$, denote $\Gamma(i)$ for the MDP restricted to $C_i$. Thus, the state space of $\Gamma(i)$ is $C_i$ and the action space $A_i(s), s \in C_i$, is given by

$$A_i(s) = \{a \in A(s) : p_{sat} = 0 \ \forall\, y \notin C_i\}.$$

From [7] we know that $A_i(s)$ is nonempty for all $s \in C_i$ and that $i$ is a communicating MDP. (Recall that a communicating MDP is such that for any pair of states $s, t \in s$, there is a pure policy under which $t$ is accessible from $s$.) Now consider the following linear program $LP(i)$:

$$\begin{aligned}
\max \quad & \sum_{s \in C_i} \sum_{a \in A_i(s)} r(s,a) z_{s,a} \\
\text{s.t.} \quad & \sum_{s \in C_i} \sum_{a \in A_i(s)} (\delta_{st} - p_{sat}) z_{sa} = 0 \qquad t \in C_i \\
& \sum_{s \in C_i} \sum_{a \in A_i(s)} z_{sa} = 1 \\
& z_{sa} \geq 0, \quad s \in C_i, \ a \in A_i(s).
\end{aligned}$$

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

**Theorem 2** *(Thm 3.5 [7]) For all policies $u$, all initial states $s_1 \in S$, and all $i = 1, \ldots, p$, we have*

$$P_u(R \leq v_i | \Phi_i, X_1 = s_1) = 1,$$

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

## 4. Basic Results

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 $T$ empty; thus, $S = C_1$.

Consider then $LP(1)$ and denote $v := v_1$ in order to simplify notation. Also let $\{z^*_{sa}\}$ be an optimal solution of $LP(1)$ and $g^*$ be a stationary optimal policy constructed from $\{z^*_{sa}\}$ (e.g., see [6], or [5]). Clearly, $g^*$ satisfies

$$\phi(g^*, s) = v, \quad s \in S. \tag{4.1}$$

Moreover, the Markov chain associated with the policy $g^*$ has at most one recurrent class plus (a perhaps empty) set of transient states.

**Theorem 3** *In a communicating MDP $\Gamma$ there exists a policy that achieves the target level $k$ with percentile $\alpha$ if and only if $k \leq v$. If $k \leq v$, then the pure policy $g^*$ achieves the target level $k$ with percentile $\alpha$, for any $\alpha \in [0,1]$.*

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

$$P_{g^*}(R = \phi(g^*, s) | X_1 = S_1) = 1 \tag{4.2}$$

(e.g., see [8], Proposition 1 (iii)). Combining this with (4.1) gives

$$P_u(R \leq v | X_1 = S_1) = 1. \tag{4.3}$$

The result then follows from (4.3) combined with (2.1).

□

As a direct consequence of Theorem 3 we have

**Corollary 1** *In a communicating MDP $\Gamma$, $k_\alpha = v$ for all $\alpha \in [0,1]$, and*

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

Problems 1-3 have now been solved for communicating MDPs. We return to the general case, where we have strongly communication classes $C_1, \ldots, C_p$ and the set $T$ of transient states. Denote by $g_i^*$ the pure policy of Thm 4.1 associated with $\Gamma(i)$, the MDP restricted to $C_i$.

**Corollary 2** *For a fixed $i \in \{1, \ldots, p\}$ let $g$ be a pure policy that coincides with $g_i^*$ on $C_i$ and is defined arbitrarily elsewhere. Then*

$$P_g(R = v_i | \Phi_i, X_1 = s_1) = 1$$

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

*Proof*: This follows easily from the proof of Theorem 3.

□

Next we shall consider a fixed target level $k$, and associate with it an index set $I_k = \{i : 1 \leq i \leq p, v_i \geq k\}$, and an auxillary 0-1 MDP, $\Gamma_k$, whose states, actions and transition law are the same as $\Gamma$, but with rewards defined by

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

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

$$\begin{aligned} \phi^k(u, s_1) &= E_u[\lim_{N\to\infty} \inf \frac{1}{N} \sum_{n=1}^{N} \sum_{i \in I_k} 1(X_n \in C_i) | X_1 = s] \\ &= \sum_{i \in I_k} P_u(\Phi_i | X_1 = s_1), \end{aligned} \tag{4.4}$$

where the last equality follows from Theorem 1.

**Theorem 4** *Let $g^*$ be an optimal stationary policy in $\Gamma_k$ which coincides with $g_i^*$ on $C_i$ for $i \in I_k$.*[1] *There exists a policy $u$ satisfying*

$$P_u(R \geq k | X_1 = s_1) \geq \alpha, \tag{4.5}$$

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

*Proof*: From Theorem 1 we have that for any policy $u$

$$P_u(R \geq k | X_1 = s_1) = \sum_{i=1}^{p} P_u(R \geq k | \Phi_i, X_1 = s_1) P_u(\Phi_i | X_1 = s_1) \tag{4.6}$$

[1] Note that there is no loss of generality here, because $g_i^*$ yields the maximal reward (1 or 0) for every state $s \in C_i$, $i \in I_k$.

From Theorem 2 we have

$$P_u(R \geq k | \Phi_i, X_1 = s_1) = 0 \quad i \notin I_k. \tag{4.7}$$

From Corollary 2

$$1 = P_{g^*}(R \geq k | \Phi_i, X_1 = s_1) \geq P_u(R \geq k | \Phi_i, X_1 = s_1) \quad i \in I_k, \tag{4.8}$$

where the inequality follows from the optimality of $g_i^*$ for $\Gamma(i)$. Combining (4.6)-(4.8) gives

$$\begin{aligned} P_u(R \geq k | X_1 = s_1) &\leq P_{g^*}(R \geq k | X_1 = s_1) \\ &= \sum_{i \in I_k} P_{g^*}(\Phi_i | X_1 = s_1) \\ &= \phi^k(g^*, s_1), \end{aligned}$$

from which the result follows.

□

It is important to note that Theorem 4 provides a constructive answer to Problem 1 of Section 2 concerning $\alpha$-achievability of the target level $k$. We shall now address the problems of determining $k_\alpha$ the maximal achievable percentile for the fixed level $k$. Towards this goal we assume without loss of generality that the strong communicating classes $C_1, \ldots, C_p$ are ordered so that

$$v_1 \geq v_2 \geq \ldots \geq v_p. \tag{4.9}$$

Now define $I_j = \{1, \ldots, j\}$ for each $j = 1, 2, \ldots, p$. In analogy with (4.4) define

$$\phi^j(u, s_1) = \sum_{i=1}^{j} P_u(\Phi_i | X_1 = s_1), \tag{4.10}$$

with corresponding MDP $\Gamma_j$ defined over the state space $S$.

**Theorem 5** *Let $g_j^*$ be an optimal pure policy for MDP $\Gamma_j$ which coincide with $g_i^*$ (optimal in $\Gamma(i)$) on each $C_i, i = 1, \ldots, j$. We have for $\alpha \in (0,1]$ that*

$$k_\alpha = k^* := \max\{v_j | \phi^j(g_j^*, s_1) \geq \alpha, \ j = 1, \ldots, p\} \tag{4.11}$$

*Proof*: Let $\ell$ be the largest index that achieves the maximum in (4.11), which is well-defined since $\phi^p(g^*, s_1) = 1$. Since $k^* = v_\ell$, we have $I_{k^*} = \{1, 2, \ldots, \ell\}$. Thus, from Theorem 4, we know that

$$P_{g_\ell^*}(R \geq k^* | X_1 = s_1) \geq \alpha. \tag{4.12}$$

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

$$P_u(R \geq k' | X_1 = s_1) \geq \alpha. \tag{4.13}$$

Now let

$$m := \max\{i : v_i \geq k'\},$$

noting that if $v_i < k'$ for all $i = 1, \ldots, p$, then the left side of (4.13) equals 0 contradicting the hypothesis $\alpha > 0$. By the definition of $m$ we have

$$v_m \geq k' > k^*. \tag{4.14}$$

Applying Theorem 1 to (4.14) yields

$$\begin{aligned}\alpha &\leq \sum_{i=1}^{m} P_u(R \geq k' | \Phi_i, X_1 = s_1) P_u(\Phi_i | X_1 = s_1) \\ &\leq \sum_{i=1}^{m} P_u(R \geq k' | \Phi_i, X_1 = s_1) P_u(\Phi_i | X_1 = s_1) \\ &\leq \sum_{i=1}^{m} P_{g_m^*}(R \geq k' | \Phi_i, X_1 = s_1) P_{g_m^*}(\Phi_i | X_1 = s_1) \\ &= \phi(g_m^*, s_1). \qquad (4.15)\end{aligned}$$

But, by the definition of $k^*$, (4.15) implies $k^* \geq v_m$, which contradicts (4.14).

□

**Corollary 3** *The maximal $\alpha$-achievable target level, $k_\alpha$, is a monotone nonincreasing step-function of $\alpha$, defined on the interval (0,1].*

*Proof*: Let $\alpha_j := \phi^j(g_j^*, s_1)$ for $j = 1, \ldots, p$, so that $0 \leq \alpha_1 \leq \alpha_2 \leq \ldots \leq \alpha_p = 1$. If we define $k_0 := \max\{v_i | i = 1, \ldots, p\}$, then by Theorem 5 $k_\alpha = k_0$ for all $\alpha \in (0, \alpha_1]$. Similarly, $k_\alpha = k_j$, a constant for all $\alpha \in (\alpha_j, \alpha_j + 1]$, where $k_j \geq k_{j+1}$ for each $j = 1, \ldots, p-1$.

□

**Corollary 4** *The maximum percentile for a given target level , $\alpha_k$, is a monotone nonincreasing step function of k defined in the interval $[v_p, v_1]$. In particular for $k \in [v_{j+1}, v_j]$ we have*

$$\alpha_k = \phi^j(g_j^*, s_1)$$

*for each $j = 1, \ldots, p-1$.*

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

□

**Remark 2** Corollaries 3 and 4 demonstrate the strength of the percentile objective criteria. Namely, the decomposition of states in $C_1, \ldots, C_p$ and $T$, and the subsequent computation of policies $g_j^*$ together with "break-points" $k_i$ and $v_i$ for $k_\alpha$ and $\alpha_k$ respectively, allows for a flexible and practical evaluation of gain-risk trade-offs in an average reward MDP.

## References


[1] J. Bather, *Optimal Decision Procedures in Finite Markov Chains. Part III: General Convex Systems*, Advances in Applied Probability, **5** (1973), pp. 541–553.

[2] D. Blackwell, *Discrete Dynamic Programming*, Annals of Math. Stat., **33** (1962), pp. 719–726.

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

[4] J. A. Filar, *Percentiles and markovian decision processes*, Operations Research Letters, **2** (1983), pp. 13–15.

[5] J. A. Filar and T. A. Schultz, *Communicating mdp's: Equivalence and lp Properties*, Operations Research Letters, **7** (1988), pp. 303–307.

[6] L. C. M. Kallenberg, *Linear Programming and Finite Markovian Control Problems*, Mathematical Center Tracts 148, Amsterdam, 1983.

[7] K. W. Ross and R. Varadarajan, *The Decomposition of Time-Average Markov Decision Processes: Theory, Algorithms and Applications*, Tech. Rep., Department of Systems, University of Pennyslvannia, 1986.

[8] K. W. Ross and R. Varadarajan, *Markov Decision Processes with Sample-Path Constraints: The Communicating Case*, Tech. Rep., Department of Systems, University of Pennsylvannia, (to appear in Operations Research), 1986.