Open access • Journal Article • DOI:10.1287/OPRE.40.6.1180

A weighted Markov decision process — Source link

Dmitry Krass, Jerzy A. Filar, Sagnik Sinha

Institutions: University of Toronto, University of Maryland, Baltimore, Indian Statistical Institute

Published on: 01 Nov 1992 - Operations Research (INFORMS)

Topics: Reward-based selection, Markov decision process, Markov chain, Decision analysis and Decision theory

A WEIGHTED MARKOV DECISION PROCESS

DMITRY KRASS
University of Toronto, Toronto, Ontario, Canada

JERZY A. FILAR
University of Maryland at Baltimore County, Baltimore, Maryland

SAGNIK S. SINHA
Indian Statistical Institute, New Delhi, India

(Received April 1988; revisions received October 1989, December 1990, June 1991; accepted July 1991)

The two most commonly considered reward criteria for Markov decision processes are the discounted reward and the long-term average reward. The first tends to "neglect" the future, concentrating on the short-term rewards, while the second one tends to do the opposite. We consider a new reward criterion consisting of the weighted combination of these two criteria, thereby allowing the decision maker to place more or less emphasis on the short-term versus the long-term rewards by varying their weights. The mathematical implications of the new criterion include: the deterministic stationary policies can be outperformed by the randomized stationary policies, which in turn can be outperformed by the nonstationary policies; an optimal policy might not exist. We present an iterative algorithm for computing an ϵ-optimal nonstationary policy with a very simple structure.

Many dynamic planning problems have successfully been analyzed as Markov decision processes (MDPs); see, e.g., Bertsekas (1976), White (1985), and Tijms (1986). In these models the two most commonly adopted criteria are the expected total discounted return and the long-run average return. In the former, the sequence of single-stage expected rewards is aggregated by first discounting each successive reward using a constant factor $\alpha \in [0, 1)$ and then summing the resulting sequence (see (2)), while in the latter the long-run average of these expected rewards is considered (see (3)). It should be clear that unless the discount factor α is near 1, the main effect of discounting is to all but annihilate the importance of the rewards accumulated in the later stages of the process. For lack of a better phrase we shall refer to this, not necessarily undesirable, phenomenon as "neglect of the future." On the other hand, the long-run average model annihilates the effect of poor control of the process during any finite segment of its history. To emphasize the differences between the two models we shall call this phenomenon "neglect of the present."

To overcome this problem, Veinott (1966, 1969) and Miller and Veinott (1969) introduced a series of sensitive discounted or overtaking optimality criteria, where the terms of the Laurent series expansion of the total discounted reward (in powers of $(1 - \alpha)$) are lexicographically maximized. In this lexicographic approach, priority is given to finding an average-optimal policy, using additional criteria for further selections within the class of such policies. This approach is useful in the situations where the average return criterion is of primary interest, but one wants to select average-optimal policies which make *reasonable* early decisions. Note, however, that under this approach, the performance of average-optimal policies in a *particular* discounted process are not being compared to each other, since we let α approach 1.

Our development is motivated by the desire to treat situations where the discount factor α is *fixed* (e.g., it might correspond to the prevailing inflation rate) and is not necessarily very high. As an alternative objective, we consider a convex combination of the discounted and the long-run average return criteria (the weighted reward criterion). The following example motivates this choice (see Krass 1989 for a more complete discussion of the motivation for the weighted reward criterion).

Example 1. Suppose that the owner of a family business relies on a supplier of a rare raw material which

Subject classifications: Decision analysis, sequential: tradeoffs between discounted and long-term average objectives. Dynamic programming, Markov, finite state: new reward criteria for Markov decision processes.
Area of review: STOCHASTIC PROCESSES AND THEIR APPLICATIONS.

0030-364X/92/4006-1180 $01.25

is essential for his business. Currently he is using a very dependable supplier, however, a new supplier proposes to meet his demand at a lower price. The owner estimates that with probability 0.1 the new supplier may fail to deliver the raw material in any given year, and that such a failure would result in bankruptcy of the business because the former (dependable) supplier will no longer be available for that year. Assuming that the annual profits with the old (new) supplier will be \$100,000 (\$142,500), respectively, and that the discount factor is $\alpha = 0.8$, we now have a discounted Markov decision model with the data:

$i = 1$ (business operating) $\quad i = 2$ (bankruptcy)

$$\begin{pmatrix} 142.5 \\ \\ 100 \end{pmatrix} \begin{matrix} \rightarrow (.9, .1) \\ \\ \rightarrow (1, 0) \end{matrix} \qquad (0) \rightarrow (0, 1)$$

The notation $(r_{ia}) \rightarrow (p_{ia1}, p_{ia2}, \ldots, p_{ian})$ gives the reward and the transition probabilities resulting from the choice of an action a in state i. For instance, action 2 in state 1 corresponds to the decision to purchase the raw material from the old supplier. There are only two deterministic policies in this model: policy f^0 which corresponds to signing a contract with the old supplier at the beginning of each year (i.e., action 2 is chosen in state 1), and f^* which corresponds to signing a contract with the new supplier each year (i.e., action 1 is chosen in state 1). These result in the discounted overall rewards: $\psi_1(f^0) = 500{,}000$, $\psi_1(f^*) = 509{,}000$, and $\psi_2(f^0) = \psi_2(f^*) = 0$.

An optimal policy is f^* which always selects the cheaper but less reliable supplier. Indeed, in this example f^* is the unique optimal policy. However, the use of this policy results in the ultimate loss of the business. In fact, the owner should expect bankruptcy within ten years under this policy. Note that the conservative policy f^0 is optimal in the corresponding average reward model and has zero risk of bankruptcy. Furthermore, policies of the form $f_\tau = (f^1, \ldots, f^t, \ldots)$, where $f^t = f^*$ for $t \leq \tau$ and $f^t = f^0$ otherwise, will yield the overall discounted reward $\psi_1(f_\tau) \in (500{,}000, 509{,}000)$, and a risk of bankruptcy of $[1 - (0.9)^\tau]$. Hence, policies of this form will enable the owner to balance the increased profits against the risk of bankruptcy. Inasmuch as bankruptcy carries with it a virtually permanent stigma, the owner may prefer one of the policies f_τ because these policies are more "sensitive to the future" than the discounted optimal policy f^*.

The weighted reward criterion is formally defined in Section 1. There we also demonstrate that the seemingly innocuous step of taking a convex combination of two "well-behaved" MDPs results in a breakdown of the standard properties of MDPs (see (7) and the following discussion). Indeed, Example 2 demonstrates that optimal policies need not exist in the weighted MDP. In Section 2, we analyze the important special case which arises when the maximal achievable value (or gain-rate) in the average return process is independent of the starting state. Theorem 1 shows that under this assumption the weighted MDP possesses "ultimately deterministic" ϵ-optimal policies that consist of first following a deterministic policy which is optimal in the discounted reward process for a certain period of time, and then switching (permanently) to a deterministic policy which is optimal in the average return process. This particularly simple structure of the ϵ-optimal policies is not preserved when the gain-rate is allowed to depend on the initial state. Section 3 is devoted to developing an iterative algorithm for the construction of ϵ-optimal policies for general MDPs. Lemmas 2 and 3 construct the optimality equation for the weighted MDP, whose principal difference from its counterpart in the classical MDP theory, is that the form of the payoff criterion depends on time. The existence of ϵ-optimal policies independent of the starting state is also established in Section 3. Finally, Theorem 3 describes an algorithm for constructing "ultimately stationary" ϵ-optimal policies.

The approach taken here is somewhat similar in spirit to the concept of "forecast horizon optimality" (see Bean and Smith 1984, and Hopp, Bean and Smith 1987). However, the latter approach is usually concerned with approximating an early part of the solution to an infinite horizon MDP by the solution to an associated finite horizon MDP.

1. DEFINITIONS AND PRELIMINARIES

A finite discrete Markov decision process Γ is observed at discrete time points $t = 1, 2, \ldots,$. The state space is denoted by $E = \{1, 2, \ldots, N\}$. With each state $i \in E$, we associate a finite action set $A(i)$. At any time point t, 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_{ia} is earned and the process moves to a state $j \in E$ with transition probability p_{iaj}, where $p_{iaj} \geq 0$ and $\sum_{j=1}^{N} p_{iaj} = 1$.

A *decision rule* f^t at time t is a function that assigns a probability to the event that action a is taken at time

t. In general, f^t may depend on all realized states up to and including time t, and on all realized actions up to time t. A *policy* f is a sequence of decision rules: $f = (f^1, f^2, \ldots, f^t \ldots)$. A *Markov policy* is one in which f^t depends only on the current state at time t. A Markov policy is called *stationary* if all its decision rules are identical. A *deterministic policy* is a stationary policy with nonrandomized decision rules.

Let C, $C(M)$, $C(S)$ and $C(D)$ be the sets of all policies, the Markov policies, the stationary policies, and the deterministic policies, respectively. Let X_t be the state at time t, Y_t be the action at time t, and $P_f(X_t = j, Y_t = a \mid X_1 = i)$ be the conditional probability that at time t the state is j and the action taken is a, given that the initial state is i and the decision maker uses policy f. Now, if R_t denotes the reward at time t, then for any policy f and initial state i, the expectation of R_t is given by

$$E_f(R_t, i) = \sum_{j \in E} \sum_{a \in A(j)} P_f(X_t = j, Y_t = a \mid X_1 = i) r_{ja}. \tag{1}$$

The manner in which we aggregate the resulting stream of expected rewards $\{E_f(R_t, i): t = 1, 2, \ldots\}$ defines the Markov decision processes discussed in the sequel.

Discounted MDP. Here the corresponding overall reward criterion is defined by

$$\psi_i(f) = \sum_{t=1}^{\infty} \alpha^{t-1} E_f(R_t, i), \tag{2}$$

where $\alpha \in [0, 1)$ is a fixed discount factor. A policy f^* is called *discounted-optimal* if for all $i \in E$, $\psi_i(f^*) = \max_f \psi_i(f)$. We will denote this process by $\Gamma(\alpha)$.

Average Reward MDP. Here the overall reward criterion is defined by

$$\phi_i(f) = \lim_{T \to \infty} \inf \frac{1}{T} \sum_{t=1}^{T} E_f(R_t, i); \quad i \in E. \tag{3}$$

A policy f^* is called *average-optimal* if for all $i \in E$, $\phi_i(f^*) = \max_f \phi_i(f)$. We will denote this process by Γ_A.

Note that both (2) and (3) are well defined. It is well known (e.g, see Derman 1970) that for both $\Gamma(\alpha)$ and Γ_A, deterministic optimal policies exist. Let

$$v(\alpha, i) = \max_f \psi_i(f), \quad i \in E, \tag{4}$$

$$v(i) = \max_f \phi_i(f), \quad i \in E \tag{5}$$

($v(i)$ is usually referred to as the *gain-rate* of Γ_A).

We now introduce the weighted reward criterion.

Weighted Reward Criterion. Here the overall reward criterion is defined by

$$\omega_i(f) = \lambda(1 - \alpha)\psi_i(f) + (1 - \lambda)\phi_i(f); \quad i \in E, \tag{6}$$

where $\lambda \in [0, 1]$ is a fixed weight parameter, and α is the discount factor in the process $\Gamma(\alpha)$. We denote this weighted process by $\Gamma(\alpha, \lambda)$.

Remark 1. The term "reward" for r_{ia} is somewhat of a misnomer as there are many potential applications of Markov decision processes where the r_{ia}'s might actually be "outputs" of some process (e.g., water levels at a dam, or volumes of production). Consequently, it is not hard to conceive of a decision maker who has to cater to two distinct constituencies: one that wishes to discount the future outputs, and another that is interested only in the long-run average performance. In this context a policy f^0 that optimizes the weighted criterion (6) with $\lambda \in (0, 1)$, is clearly Pareto-optimal in a bi-objective optimization problem in which the decision maker wishes to "simultaneously optimize" both the discounted and average reward criteria.

Also, note that it is possible to define an overall reward criterion that is the weighted sum of two or more discounted reward criteria with distinct discount factors. While much of the analysis of the subsequent section could be extended to this case, we find it a less natural model. In the remainder of this section we shall attempt to analyze some of the properties of the weighted MDP $\Gamma(\alpha, \lambda)$.

Remark 2. Note that the results of Derman and Strauch (1966) imply that $\sup_{C(M)} \omega_i(f) = \sup_C \omega_i(f)$ for all $i \in E$, and so we need only consider Markov policies. From now on, $C(M)$ will be the largest class of policies under consideration.

Observe that the following inequalities hold since $C(D) \subset C(S) \subset C(M)$ and by definitions of $v(i)$, $v(\alpha, i)$:

$$\begin{aligned} \sup_{C(D)} \omega_i(f) &\leq \sup_{C(S)} \omega_i(f) \leq \sup_{C(M)} \omega_i(f) \\ &\leq \lambda(1 - \alpha)v(\alpha, i) + (1 - \lambda)v(i). \end{aligned} \tag{7}$$

The upper bound on the left-hand side of (7) will be referred to as the *utopian bound.*

Remark 3. Note that in the discounted MDP $\Gamma(\alpha)$ there always exists $\alpha_0 \in [0, 1)$, and a deterministic policy f^0 such that for all $\alpha \geq \alpha_0$, f^0 is optimal in both $\Gamma(\alpha)$ and Γ_A (see Blackwell 1962). Such a policy clearly attains the utopian bound of (7) and is therefore

optimal in the weighted process $\Gamma(\alpha, \lambda)$ for all $\lambda \in [0, 1]$ and $\alpha \geq \alpha_0$.

The "classical" analysis of MDPs is based on the ability to restrict consideration to the simple class of deterministic policies $C(D)$. In the present context that would mean that the first two inequalities in (7) are, in fact, equalities. However, Example 2 will demonstrate that the first two inequalities in (7) can be strict. The example also shows that optimal policies do not generally exist in weighted MDPs.

Example 2. Consider a two-state MDP with the data

$$\begin{array}{ll} i = 1 & i = 2 \\ \begin{pmatrix} 0 \\ -10 \end{pmatrix} \begin{matrix} \rightarrow (1, 0) \\ \rightarrow (0, 1) \end{matrix} & (12) \rightarrow (1, 0) \end{array}$$

the discount factor $\alpha = \frac{1}{2}$, and the weight parameter $\lambda = \frac{1}{4}$. There are only two deterministic policies here: f_1 which chooses action 1 in state 1, and f_2 which chooses action 2 in state 1. Note that $\phi_1(f_1) = (1 - \alpha)\psi_1(f_1) = \omega_1(f_1) = 0$, while $\phi_1(f_2) = 1$, $(1 - \alpha)\psi_1(f_2) = -\frac{8}{3}$, $\omega_1(f_2) = 0.08(3)$. Clearly f_1 is discounted-optimal, f_2 is average-optimal, and f_2 is the best deterministic policy in the weighted process $\Gamma(\frac{1}{2}, \frac{1}{4})$. Let f_p be a randomized stationary policy which chooses action 1 in state 1 with probability p. Any stationary policy is of the form f_p. It can be verified that

$$\omega_1(f_p) = (3/2)\frac{1 - p}{2 - p} - (2)\frac{1 - p}{3 - p}.$$

The above attains its maximum over $[0, 1]$ at $p^* \sim 0.42$, yielding

$$\omega_1(f_{p^*}) \sim 0.10102 > \max\{\omega_1(f_1), \omega_1(f_2)\},$$

thereby demonstrating that the first inequality in (7) is strict in this case. Now consider the Markov policy f^τ which uses f_1 for the first τ stages, and then switches to f_2 permanently. Then,

$$\psi_1(f^\tau) = -\frac{16}{3}\left(\frac{1^\tau}{2}\right), \quad \phi_1(f^\tau) = 1,$$

and thus

$$\omega_1(f^\tau) = \frac{3}{4} - \frac{2}{3}\left(\frac{1^\tau}{2}\right) \leq \frac{3}{4}.$$

Thus, f^τ approaches the "utopian" upper bound of 0.75 arbitrarily closely as $\tau \rightarrow \infty$, but never reaches it. This immediately implies that the second inequality in (7) is also strict.

We now show that optimal policies do not exist in this example. Indeed, suppose that f^* is the optimal policy. Without loss of generality, we can assume that $f^* = (f_1^*, f_2^*, \ldots)$ is Markov. We must have $\omega_1(f^*) = 0.75$ because f^* is better than f^τ for any τ. Clearly, f^* cannot equal f_1 because then $\omega_1(f^*) = 0$. Therefore, there exists some t such that f_t^* is the first decision rule which calls for the choice of action 2 in state 1 with positive probability, say p. The discounted payoff of f^* can be bounded as follows: the payoff for the first $t - 1$ stages is 0; the payoff at stage t is $-10p(\frac{1}{2})^{t-1}$; the payoff at stage $t + 1$ is $\leq 12p(\frac{1}{2})^t$; the payoff for all stages after $t + 2$ is $\leq (\frac{1}{2})^{t+1}\psi_1(f_1) = 0$. (The last statement follows because f_1 is discount-optimal.) Thus,

$$\psi_1(f^*) \leq -10p\left(\frac{1}{2}\right)^{t-1} + 12p\left(\frac{1}{2}\right)^t = -4p\left(\frac{1}{2}\right)^t < 0.$$

Since $\phi_1(f^*) \leq \phi_1(f_2) = 1$, we have:

$$\omega_1(f^*) < (1 - \lambda) = \frac{3}{4},$$

thus contradicting maximality of f^*.

In view of the preceding example it is evident that we must focus on policies that *nearly* achieve the $\sup_{C(M)} \omega_i(f)$ for all $i \in E$. Toward this end we define an *ϵ-i-optimal policy* f to be one for which $\omega_i(f) \geq \sup_{C(M)} \omega_i(f) - \epsilon$, where $\epsilon > 0$ and $i \in E$ is given. Policy f is *ϵ-optimal* if $\omega_i(f) \geq \sup_{C(M)} \omega_i(f) - \epsilon$ for all $i \in E$. Note that while the existence of *ϵ-i-optimal* policies follows immediately from the definition, the existence of *ϵ-optimal* policies is not obvious. We will settle this issue in Section 3.

We shall say that a Markov policy $f = (f_1, f_2, \ldots)$ is *ultimately deterministic* if there exists an integer τ_0 and a deterministic policy g such that $f_\tau = g$ for all $\tau \geq \tau_0$. Let $C(UD)$ denote the set of ultimately deterministic policies.

In Section 3 we shall demonstrate that for each fixed $i \in E$

$$\sup_{C(UD)} \omega_i(f) = \sup_{C(M)} \omega_i(f), \tag{8}$$

and, therefore, that the relatively simple class of ultimately deterministic policies performs the role played by the class $C(D)$ in the classical theory of MDPs.

2. THE CASE OF IDENTICAL GAIN-RATE IN THE AVERAGE REWARD PROCESS

Average reward MDPs with the following special structures of the transition probabilities have received

considerable attention in the literature: *unichain* MDPs, where for any deterministic policy f the Markov chain $P(f)$ induced by that policy has exactly one recurrent class, and *communicating* MDPs, where for any two states i and j there exists a deterministic policy f such that i is accessible from j in the Markov chain $P(f)$. The reasons for the importance of these special cases are twofold: first, in many practical applications of MDPs the resulting models possess either unichain or communicating structure, and second, very efficient iterative algorithms have been developed for MDPs with these structures.

One of the key properties of unichain and communicating MDPs is that the gain-rate in Γ_A is independent of the starting state, i.e.,

$$v(i) \equiv u \quad \text{for all } i \in E.$$

The next theorem shows that for MDPs with this property the utopian bound in (7) is achieved and that ϵ-optimal ultimately deterministic policies with a simple structure can be constructed easily.

Theorem 1. *Consider the weighted process $\Gamma(\alpha, \lambda)$, with $\lambda > 0$. Suppose that $v(i) \equiv u$ for all $i \in E$. Then:*

i. $\sup_{C(M)} \omega_i(f) = \lambda(1-\alpha)v(\alpha, i) + (1-\lambda)u$ *for all* $i \in E$.
ii. *Given any $\epsilon > 0$, there exists $f^\epsilon \in C(UD)$ which is ϵ-optimal in $\Gamma(\alpha, \lambda)$.*
iii. *Let $f^0 \in C(D)$ be optimal in $\Gamma(\alpha)$ and $g^0 \in C(D)$ be optimal in Γ_A, and $\epsilon > 0$ be given. Let $R = \max_{i \in E, a \in A(i)} r_{ia}$ and take $\tau(\epsilon) = \lfloor \log(\epsilon/\lambda R)/\log \alpha \rfloor + 1$. Define $f^\epsilon = (f_1, f_2, \ldots)$, with $f_\tau = f^0$ for $\tau \leq \tau(\epsilon)$ and $f_\tau = g^0$ for $\tau > \tau(\epsilon)$. Then f^ϵ is ϵ-optimal in $\Gamma(\alpha, \lambda)$.*

Proof. Fix $\epsilon > 0$. For every $i \in E$

$$v(\alpha, i) = \sum_{t=1}^{\infty} \alpha^{t-1} E_{f^0}(R_t, i).$$

Note that for any integer $\tau > 0$,

$$\sum_{t=\tau+1}^{\infty} \alpha^{t-1} E_{f^0}(R_t, i) \leq \sum_{t=\tau+1}^{\infty} \alpha^{t-1} R = \frac{R\alpha^\tau}{1-\alpha}.$$

Thus to assure that

$$\frac{R\alpha^\tau}{1-\alpha} < \frac{\epsilon}{\lambda(1-\alpha)}$$

it suffices to take $\tau > \log(\epsilon/\lambda R)/\log \alpha$. Let

$$\tau(\epsilon) = \left\lfloor \log \frac{\epsilon}{\lambda R} \Big/ \log \alpha \right\rfloor + 1.$$

Then for all $\tau \geq \tau(\epsilon)$

$$\sum_{t=1}^{\tau} \alpha^{t-1} E_{f^0}(R_t, i) \geq v(\alpha, i) - \frac{R\alpha^\tau}{1-\alpha} > v(\alpha, i) - \frac{\epsilon}{\lambda(1-\alpha)}. \tag{9}$$

Define f^ϵ as in part iii of the theorem. From (2), (9) and the structure of f^ϵ we now have

$$\lambda(1-\alpha)\psi_i(f^\epsilon) > \lambda(1-\alpha)v(\alpha, i) - \epsilon \tag{10}$$

for every $i \in E$. Furthermore, since $\phi_i(f^\epsilon) = \phi_i(g^0) = u$ for all $i \in E$, we have that

$$\omega_i(f^\epsilon) > \lambda(1-\alpha)v(\alpha, i) + (1-\lambda)u - \epsilon$$

for all $i \in E$.

3. COMPUTATION OF ϵ-OPTIMAL POLICIES FOR GENERAL MDPs

In this section, we turn our attention back to the general case, that is, the gain-rate is now allowed to depend on the starting state. It is easy to construct examples showing that parts i and iii of Theorem 1 do not hold for the nonidentical gain-rate case. However, part ii of the theorem is extended below to an arbitrary weighted process $\Gamma(\alpha, \lambda)$.

We will first need the following technical result whose proof can be found in Krass.

Lemma 1. *Let $f = (f^1, f^2, \ldots)$ be a Markov policy in Γ_A. Let g be a stationary average-optimal policy. For some $t \in \{1, 2, \ldots\}$ define $\hat{f} = (\hat{f}^1, \hat{f}^2, \ldots)$, where*

$$\hat{f}^k = \begin{cases} f^k & \text{if } k < t \\ g & \text{if } k \geq t \end{cases}.$$

Then $\phi_i(f) \leq \phi_i(\hat{f})$ for all $i \in E$.

Theorem 2. *Consider the weighted process $\Gamma(\alpha, \lambda)$ with $\lambda > 0$. Let $g^0 \in C(D)$ be optimal in Γ_A. Given a fixed initial state i and arbitrary $\epsilon > 0$, there exists a positive $\tau(\epsilon)$ and a policy $f^\epsilon = (f_1, f_2, \ldots) \in C(UD)$ with $f_\tau = g^0$ for all $\tau > \tau(\epsilon)$ such that f^ϵ is ϵ-i-optimal in $\Gamma(\alpha, \lambda)$.*

Proof. By Remark 2 there exists a Markov policy $\hat{f} = (\hat{f}_1, \hat{f}_2, \ldots)$, which is $(\epsilon/2)$-i-optimal in $\Gamma(\alpha, \lambda)$. As in (9), there exists $\tau(\epsilon)$ such that for any $\tau \geq \tau(\epsilon)$

$$\sum_{t=1}^{\tau} \alpha^{t-1} E_{\hat{f}}(R_t, i) \geq \psi_i(\hat{f}) - \frac{\epsilon}{2\lambda(1-\alpha)} \tag{11}$$

for the fixed $i \in E$. Now define $f^\epsilon = (f_1, f_2, \ldots)$ with $f_t = \hat{f}_t$ for $t \leq \tau(\epsilon)$ and $f_t = g^0$ for $t > \tau(\epsilon)$. Similarly to

(10), we have for the fixed $i \in E$

$$\lambda(1-\alpha)\psi_i(f^\epsilon) \geq \lambda(1-\alpha)\psi_i(\hat{f}) - \frac{\epsilon}{2}. \tag{12}$$

By the construction of f^ϵ and optimality of g^0 in the average-return process Γ_A it follows by Lemma 1 that for all $i \in E$

$$\phi_i(f^\epsilon) \geq \phi_i(\hat{f}). \tag{13}$$

Combining (12) and (13) we obtain that for the fixed $i \in E$

$$\omega_i(f^\epsilon) \geq \lambda(1-\alpha)\psi_i(\hat{f}) + (1-\lambda)\phi_i(\hat{f}) - \frac{\epsilon}{2}$$

$$= \omega_i(\hat{f}) - \frac{\epsilon}{2} \geq \sup_{C(M)} \omega_i(f) - \epsilon.$$

In the remainder of this section we will construct an algorithm to compute ϵ-optimal ultimately deterministic policies in the weighted process $\Gamma(\alpha, \lambda)$. To that end, we will use Theorem 2, which guarantees the existence of ϵ-*i*-optimal ultimately deterministic policies whose "tail" consists of an average optimal deterministic policy. The algorithm below will be used to compute the "head" of such policies.

To simplify the notation in the rest of this section, we slightly generalize the definition of the weighted reward process: The $\Gamma(\alpha, \lambda_1, \lambda_2)$ process is the MDP with overall reward criterion defined by

$$\omega_i[\lambda_{,1}, \lambda_2](f) := \lambda_1(1-\alpha)\psi_i(f) + \lambda_2\phi_i(f); \quad i \in E,$$

where $\lambda_1, \lambda_2 \in [0, 1]$, and α is the discount factor in the process $\Gamma(\alpha)$. Note that the weighted reward process defined in Section 1 is simply $\Gamma(\alpha, \lambda, (1-\lambda))$.

The next result allows a first-step decomposition of payoffs. While the result is stated for ultimately deterministic policies only, it holds for any policy for which the limit (rather than lim inf) is achieved in (3).

Lemma 2. *Let $f = (f^1, f^2, \ldots) \in C(UD)$. Let $\hat{f} = (f^2, f^3, \ldots)$, that is $\hat{f}$ uses the decision rule f^{t+1} at time t. Then*

$$\omega_i[\lambda_1, \lambda_2](f) = (1-\alpha)\lambda_1 \sum_{a\in A(i)} r_{ia} f_i^1(a) + \sum_{a\in A(i)} \sum_j p_{iaj} f_i^1(a)\omega_j[\alpha\lambda_1, \lambda_2](\hat{f})$$

for any $i \in E$, where $f_i^1(a)$ is the probability of taking action a in state i under the decision rule f^1.

Proof. For any $i \in E$ we have by definition of $\hat{f}$,

$$\psi_i(f) = \sum_{a\in A(i)} r_{ia} f_i^1(a) + \alpha \sum_{a\in A(i)} \sum_j p_{iaj} f_i^1(a)\psi_j(\hat{f}).$$

Also, since the limit holds in (3) for any simple ultimately deterministic policy, we have:

$$\phi_i(f) = \sum_{a\in A(i)} \sum_j p_{iaj} f_i^1(a)\phi_j(\hat{f}).$$

Thus,

$$\omega_i[\lambda_1, \lambda_2](f)$$

$$= \lambda_1(1-\alpha)\psi_i(f) + \lambda_2\phi_i(f)$$

$$= \lambda_1(1-\alpha) \sum_{a\in A(i)} r_{ia} f_i^1(a)$$

$$+ \sum_{a\in A(i)} \sum_j p_{iaj} f_i^1(a)[\psi_j(\hat{f})\lambda_1(1-\alpha)\alpha + \phi_j(\hat{f})\lambda_2]$$

$$= \lambda_1(1-\alpha) \sum_{a\in A(i)} r_{ia} f_i^1(a)$$

$$+ \sum_{a\in A(i)} \sum_j p_{iaj} f_i^1(a)\omega_j[\alpha\lambda_1, \lambda_2](\hat{f}).$$

The next lemma allows us to construct an ϵ-optimal ultimately stationary policy for the process $\Gamma(\alpha, \lambda_1, \lambda_2)$ from an ϵ-optimal ultimately stationary policy for $\Gamma(\alpha, \alpha\lambda_1, \lambda_2)$. Note, however, that at the moment we have not shown that any ϵ-optimal policies exist.

Lemma 3. *Suppose that $f \in C(UD)$ is an ϵ-optimal policy in $\Gamma(\alpha, \alpha\lambda_1, \lambda_2)$. For each $i \in E$, let a_i be the action which achieves*

$$\max_{a\in A(i)} \left[r_{ia}(1-\alpha)\lambda_1 + \sum_j p_{iaj}\omega_j[\alpha\lambda_1, \lambda_2](f) \right].$$

Let $\hat{f}^1$ be the deterministic decision rule which uses action a_i in state i. Let $\hat{f} = (\hat{f}^1, f)$, that is, $\hat{f}$ uses the decision rule $\hat{f}^1$ at time 1, and follows f thereafter. Clearly, $\hat{f}$ is ultimately deterministic. Then $\hat{f}$ is ϵ-optimal in $\Gamma(\alpha, \lambda_1, \lambda_2)$.

Proof. Take any $i \in E$ and some $\delta > 0$. Let $g \in C(UD)$ be an δ-*i*-optimal policy in $\Gamma(\alpha, \lambda_1, \lambda_2)$, $g = (g^1, g^2, \ldots)$, and $\rho = (g^2, g^3, \ldots)$. Note that the

existence of g is assured by Theorem 2. We have:

$$\omega_i[\lambda_1, \lambda_2](\hat{f})$$

(by Lemma 2)

$$= r_{ia_i}(1-\alpha)\lambda_1 + \sum_j p_{ia_ij}\omega_j[\alpha\lambda_1, \lambda_2](f)$$

(by definition of a_i)

$$= \max_a [r_{ia}(1-\alpha)\lambda_1 + \sum_j p_{iaj}\omega_j[\alpha\lambda_1, \lambda_2](f)]$$

(by ϵ-optimality of f in $\Gamma[\alpha\lambda_1, \lambda_2]$)

$$\geqslant \max_a [r_{ia}(1-\alpha)\lambda_1 + \sum_j p_{iaj}(\omega_j[\alpha\lambda_1, \lambda_2](\rho) - \epsilon)]$$

$$= \max_a [r_{ia}(1-\alpha)\lambda_1 + \sum_j p_{iaj}\omega_j[\alpha\lambda_1, \lambda_2](\rho)] - \epsilon$$

$$\geqslant \sum_{a\in A(i)} r_{ia}(1-\alpha)\lambda_1 g_i^1(a) + \sum_{a\in A(i)}\sum_j p_{iaj}g_i^1(a)\omega_j[\alpha\lambda_1, \lambda_2](\rho) - \epsilon$$

(by Lemma 2)

$$= \omega_i[\lambda_1, \lambda_2](g) - \epsilon.$$

Now let h be an arbitrary policy. Using the above result together with δ-i-optimality of g, we get:

$$\omega_i[\lambda_1, \lambda_2](\hat{f}) \geqslant \omega_i[\lambda_1, \lambda_2](g) - \epsilon \geqslant \omega_i[\lambda_1, \lambda_2](h) - \epsilon - \delta.$$

Since δ is arbitrarily small, this implies:

$$\omega_i[\lambda_1, \lambda_2](\hat{f}) \geqslant \omega_i[\lambda_1, \lambda_2](h) - \epsilon.$$

This completes the proof since i was arbitrary.

Remark 4. Lemmas 2 and 3 are very similar to well known results in the classical theory of MDPs. The main difference is that, in our case, the form of the reward criterion changes at every stage.

We next make two simple observations: For any n,

$$\omega_i[\alpha^n\lambda_1, \lambda_2](f) = \alpha^n\lambda_1(1-\alpha)\psi_i(f) + \lambda_2\phi_i(f) \leqslant \alpha^n\lambda_1(1-\alpha)v(i,\alpha) + \lambda_2 v(i) \quad \text{for any } f \in C.$$

For any $\epsilon > 0$, there exists N_i such that

$$\alpha^n\lambda_1(1-\alpha)v(i,\alpha) \leqslant \epsilon \quad \text{wherever } n \geqslant N_i.$$

We now define $N := \max_{i\in E} N_i$.

Let f_A be an optimal deterministic policy in Γ_A. Then f_A is ϵ-optimal in $\Gamma[\alpha, \alpha^N\lambda_1, \lambda_2]$. Thus, ultimately stationary ϵ-optimal policies exist in $\Gamma[\alpha, \alpha^n\lambda_1, \lambda_2]$ for n large enough. Repeated applications of Lemma 3 now assure the existence of ϵ-optimal policies in $\Gamma[\alpha, \lambda_1, \lambda_2]$ for any λ_1, λ_2. This observation together with Lemma 3 immediately establishes the validity of the algorithm below.

Theorem 3. (The Algorithm) *Fix $\epsilon > 0$. The following algorithm constructs an ϵ-optimal ultimately deterministic policy.*

Step 0. *Choose integer $N = \max_{i\in E} N_i$, where N_i is the smallest positive integer such that*

$$\alpha^{N_i}\lambda_1(1-\alpha)v(i,\alpha) \leqslant \epsilon.$$

Let f_A be an average-optimal deterministic policy. Set $f = f_A$.

Step 1. *For $k = N$ down to* 1. *For each $i \in E$ select a_i that achieves*

$$\max_{a\in A(i)}\{r_{ia}1-\alpha)\lambda_1\alpha^{k-1} + \sum_j p_{iaj}\omega_j[\alpha^k\lambda_1, \lambda_2](f)\}.$$

Let f^k be the nonrandomized decision rule taking action a_i in state i. Set $f = (f^k, f)$ (i.e., use f^k at time 1 *and then follow the original f). Increment k, and repeat.*

The ultimately deterministic policy $f = (f^1, f^2, \ldots f^N, f_A, f_A, \ldots)$ constructed above is now ϵ-optimal in $\Gamma(\alpha, \lambda_1, \lambda_2)$.

Proof. After Step 0, f is ϵ-optimal in $\Gamma(\alpha, \alpha^N\lambda_1, \lambda_2)$. After each iteration in Step 1, f is ϵ-optimal in $\Gamma(\alpha, \alpha^{k-1}\lambda_1, \lambda_2)$ by Lemma 3.

Remark 5. The amount of work required for the above algorithm is determined by N defined in Step 0. This definition is independent of f_A—the average optimal policy used as the "tail" of the ϵ-optimal policy constructed by the algorithm. However, if a particular f_A performs well in the discounted process $\Gamma(\alpha)$, this should make the calculation of an ϵ-optimal policy easier. This idea can be incorporated into our algorithm by replacing Step 0 with the following step.

Modified Step 0

Choose some average optimal f_A. Set $f = f_A$. Let $\delta_i(f_A) = v(i, \alpha) - \psi_i(f_A)$. Choose integer $N = \max_{i\in E} N_i$, where N_i is the smallest positive integer such that

$$\alpha^{N_i}\lambda_1(1-\alpha)\delta_i(f_A) \leqslant \epsilon.$$

Note that if f_A is optimal in both Γ_A and $\Gamma(\alpha)$, then the algorithm would terminate immediately (see also

Remark 3). The proposed modification raises the problem of selecting an average-optimal policy whose α-discounted payoff is the best among all average-optimal policies. We feel this problem deserves further study (for α large enough the problem was treated in Veinott 1969). Finally, we reiterate that to apply the results of Section 3 to the weighted reward process of Sections 1 and 2, we merely set $\lambda_1 = \lambda$, $\lambda_2 = 1 - \lambda$.

ACKNOWLEDGMENT

This research was supported in part by the AFOSR and the National Science Foundation under grant ECS-8704954. We are indebted to A. J. Goldman for reading the manuscript and for his valuable comments.

REFERENCES

BEAN, J., AND R. SMITH. 1984. Conditions for the Existence of Planning Horizons. *Math. Opns. Res.* **9,** 391–401.

BERTSEKAS, D. 1976. *Dynamic Programming and Stochastic Control.* Academic Press, New York.

BLACKWELL, D. 1962. Discrete Dynamic Programming. *Ann. Math. Stat.* **33,** 719–726.

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

DERMAN, C., AND R. STRAUCH. 1966. A Note on Memoryless Rules for Controlling Sequential Control Problems. *Ann. Math. Stat.* **37,** 276–278.

HOPP, W., J. BEAN AND R. SMITH. 1987. A New Optimality Criterion for Nonhomogeneous Markov Decision Processes. *Opns. Res.* **35,** 875–883.

KRASS, D. 1989. Contributions to the Theory and Applications of Markov Decision Processes. Ph.D. Thesis, Johns Hopkins University, Baltimore.

MILLER, B. L., AND A. F. VEINOTT. 1969. Discrete Dynamic Programming With Small Interest Rate. *Ann. Math. Stat.* **37,** 1284–1294.

TIJMS, H. C. 1986. *Stochastic Modeling and Analysis: A Computational Approach.* John Wiley, New York.

VEINOTT, A. F. 1966. On Finding Optimal Policies in Discrete Dynamic Programming With No Discounting. *Ann. Math. Stat.* **37,** 1284–1294.

VEINOTT, A. F. 1969. Discrete Dynamic Programming With Sensitive Discount Optimality Criteria. *Ann. Math. Stat.* **40,** 1635–1660.

WHITE, D. J. 1985. Real Applications of Markov Decision Processes. *Interfaces* **15**(6), 73–83.