The Annals of Statistics
1973, Vol. 1, No. 3, 496-505

MARKOV DECISION PROCESSES WITH A NEW OPTIMALITY CRITERION: DISCRETE TIME[1]

BY STRATTON C. JAQUETTE

Cornell University

Standard finite state and action discrete time Markov decision processes with discounting are studied using a new optimality criterion called moment optimality. A policy is moment optimal if it lexicographically maximizes the sequence of signed moments of total discounted return with a positive (negative) sign if the moment is odd (even). This criterion is equivalent to being a little risk adverse. It is shown that a stationary policy is moment optimal by examining the negative of the Laplace transform of the total return random variable. An algorithm to construct all stationary moment optimal policies is developed. The algorithm is shown to be finite.

1. Introduction. This paper is concerned with finite state and action discrete time Markov decision processes where future returns are discounted. We define and study a new optimality criterion which we call moment optimality.

The Markov decision process as well as most of the notation needed to develop the results in this paper are defined in Section 2. We also define and discuss moment optimality. Section 3 contains the result that there is a stationary policy that is optimal under the criterion of moment optimality. As the methods of Section 3 do not yield a workable algorithm to construct optimal policies, a slightly different approach is given in Section 4 to obtain an algorithm to construct all stationary moment optimal policies. Section 5 is devoted to showing that the algorithm is finite.

2. Preliminaries. We consider a standard Markov decision process. We assume that the stochastic process has a finite state space denoted by S, and without loss of generality assume that $S = \{1, 2, \cdots, s\}$. We assume that the process starts at time $t = 0$ and that the process can jump from state to state at discrete points in time $t = 0, 1, 2, \cdots$. At each point in time an action is selected and applied to the Markov process. We assume that there is a single finite set of actions, which we denote by A. Any element a in A is an action which may be applied to the process in any state and at any point in time. We can just as easily assume that there are perhaps distinct finite action sets A_i available when the process is in state i and obtain the same results, but we make the more restrictive assumption to simplify the exposition.

If the process is in state i at some point in time, applying action a determines the return and transition probabilities for that period. The return obtained is

Received May 1971; revised September 1972.

[1] The original work for this paper was partially supported by Office of Naval Research Contract N00014-67-A-0112-0031 (NR-042-265) and National Science Foundation Grant GP-8790.

denoted by $r(i, a)$. The returns are bounded since both state space and action sets are finite. We denote the transition probabilities by $p_{ij}(a)$ $(j = 1, 2, \cdots, s)$. The transition probability $p_{ij}(a)$ is the probability that the process will be in state j at time $t + 1$ given that the process is in state i and action a is applied at time t. We assume that $p_{ij}(a) \geqq 0$ for all i, j, and a, and that $\sum_{j=1}^{s} p_{ij}(a) = 1$ for all i and a.

We define the set F by $F = \bigtimes_{i=1}^{s} A$. The elements of F, denoted $\mathbf{f}$, are called action vectors. The ith component of $\mathbf{f}$, denoted $f(i)$, is the action taken if the process is in state i.

We denote a policy by π and can write this policy as a sequence of action vectors, e.g. $\pi = \{\mathbf{f}_0, \mathbf{f}_1, \cdots, \mathbf{f}_t, \cdots\}$. Here $\mathbf{f}_t$ is the action vector applied at time t using policy π, and we can write $\pi(t) = \mathbf{f}_t$. If π is stationary we have $\pi(0) = \pi(t)$ $(t = 1, 2, \cdots)$, and if $\pi(0) = \mathbf{f}$, we usually write f^{∞} to denote this stationary policy.

The state of the process is a random variable which depends on the starting state of the process, the policy used, and the time in question. We denote by $X_\pi(t)$ the random state of the system at time t when policy π is used and suppress the dependence of $X_\pi(t)$ on the starting state.

A return is obtained in each period. This return is a function of the random variable $X_\pi(t)$ and is itself therefore a random variable. In explicit notation the return at time t when policy π is used is $r(X_\pi(t), \mathbf{f}(X_\pi(t)))$ where $\mathbf{f} = \pi(t)$. We shall simplify this to $r(X_\pi(t), \pi)$. Future returns are discounted by a constant discount factor β $(0 \leqq \beta < 1)$ per period.

We denote by $\mathbf{r}(f)$ the column vector of returns associated with the action vector $\mathbf{f}$. We denote by $P(f)$ the one period transition probability matrix associated with $\mathbf{f}$. The vector $\mathbf{R}(\pi)$ will denote the total discounted return using policy π. The total discounted return random variable given that $X_\pi(0) = i$ is:

$$
\begin{aligned}
(2.1) \qquad [\mathbf{R}(\pi)]_i &= r(i, \pi) + \beta r(X_\pi(1), \pi) + \beta^2 r(X_\pi(2), \pi) + \cdots \\
&= \sum_{t=0}^{\infty} \beta^t r(X_\pi(t), \pi) \qquad (X_\pi(0) = i) .
\end{aligned}
$$

Let $\mathbf{X}_\pi(t)$ be the vector whose ith component is 1 if $X_\pi(t) = i$ and is 0 otherwise and let $\circ$ denote componentwise multiplication i.e., $[\mathbf{u} \circ \mathbf{v}]_i \equiv [\mathbf{u}]_i[\mathbf{v}]_i$ if $\mathbf{u}$ and $\mathbf{v}$ have the same dimension. Then we can rewrite (2.1) as

$$
(2.2) \qquad \mathbf{R}(\pi) = \sum_{t=0}^{\infty} \beta^t \mathbf{r}(\pi(t)) \circ \mathbf{X}_\pi(t) .
$$

We can characterize the total discounted return random vector $\mathbf{R}(\pi)$ by its moments. We denote by $\mathbf{M}_n(\pi)$ the nth moment of $\mathbf{R}(\pi)$. Noting that $\mathbf{R}(\pi)^n = \mathbf{R}(\pi) \circ \mathbf{R}(\pi) \circ \cdots \circ \mathbf{R}(\pi)$, we define $\mathbf{M}_n(\pi)$ as follows:

$$
\begin{aligned}
(2.3) \qquad \mathbf{M}_n(\pi) &\equiv E[\mathbf{R}(\pi)^n] \qquad (n = 1, 2, \cdots) \\
\mathbf{M}_0(\pi) &\equiv \mathbf{1} .
\end{aligned}
$$

We also consider $\mathbf{M}_n(\pi)$ multiplied by $+1$ or -1 depending on whether n is odd

or even. For this reason we define the vectors $\mathbf{N}_n(\pi)$ as follows:

$$\mathbf{N}_n(\pi) \equiv (-1)^{n+1}\mathbf{M}_n(\pi) \qquad (n = 0, 1, 2, \cdots). \tag{2.4}$$

We also use vectors $\mathbf{N}^m(\pi)$ whose $m + 1$ components are themselves vectors. We define $\mathbf{N}^m(\pi)$ as follows:

$$\mathbf{N}^m(\pi) \equiv (\mathbf{N}_0(\pi), \mathbf{N}_1(\pi), \cdots, \mathbf{N}_m(\pi)) \qquad (m = 0, 1, 2, \cdots). \tag{2.5}$$

By $\mathbf{N}(\pi)$ we mean $\mathbf{N}^\infty(\pi)$. We use the negative of the Laplace transform of the total discounted return random vector extensively and denote this by $\mathbf{U}_\pi(\lambda)$:

$$\mathbf{U}_\pi(\lambda) \equiv -E[\exp(-\lambda\mathbf{R}(\pi))]. \tag{2.6}$$

By expression (2.6) we mean that $[\mathbf{U}_\pi(\lambda)]_i = -E[\exp(-\lambda[\mathbf{R}(\pi)]_i)]$.

Suppose $\mathbf{u}$ and $\mathbf{v}$ are two vectors of the same dimension with components $u(i)$ and $v(i)$ respectively. We write $\mathbf{u} \succ \mathbf{v}$, that $\mathbf{u}$ is lexicographically greater than $\mathbf{v}$, if there is an integer n such that $u(i) = v(i)$ for $i < n$ and $u(n) > v(n)$. If $\mathbf{N}$ and $\mathbf{M}$ are vectors whose components are themselves vectors, then we write $\mathbf{N} \succ \mathbf{M}$ to signify that there is an integer n such that $\mathbf{N}(i) = \mathbf{M}(i)$ for $i < n$ and $\mathbf{N}(n) \succ \mathbf{M}(n)$. The relationship $\prec$, $\preceq$, and $\succeq$ are defined analogously. In this paper we shall write $\mathbf{u} < \mathbf{v}$ for vectors if $\mathbf{u} \leqq \mathbf{v}$ and $\mathbf{u} \neq \mathbf{v}$.

The optimality criterion which has been used almost exclusively is maximization of the expected value of total return. Exceptions to this are Derman [1] and Whitt [4], both for the case of infinite horizon and no discounting: Derman considers the sample path average return and Whitt considers weak convergence of the total return suitably normalized. We choose instead to retain the fact that the total return obtained from using a given policy is a random variable and pose the question of what sort of random return would be desirable. Using the notation above we say that a policy π^* is *moment optimal* if $\mathbf{N}(\pi^*) \succeq \mathbf{N}(\pi)$ for all policies π.

Moment optimality measures the desirability of a policy by examining the moments of the associated return random variable. A policy is good if its odd moments are large and even moments small, where we consider the moments lexicographically. Thus if two policies have unequal means, the one with the greater mean is better, as in the usual expected value case. If they have equal means, the one with the smaller variance is better. If they have equal means and variances, the one with the greater third moment is better, and so forth. In some sense a return distribution is good if it has large expected value with small risk of falling far below the expected value.

Moment optimality will not differ from the usual expected value case if there is a unique stationary policy which maximizes the expected return. In general there may be several stationary policies which attain the maximum expected return and many more nonstationary policies which do as well. Examples are easy to construct; a simple two state example can be found in Jaquette ([3], page 15).

Moment optimality can also be viewed as being a "little bit risk adverse"

when an exponential utility function is used. If the utility function $-e^{-\lambda x}$ is assumed for the utility of having wealth x, then λ is the constant aversion to risk implicit in the utility function. The problem of finding a policy which yields maximum expected utility of total discounted return for this utility function when the aversion to risk is small enough is equivalent to the problem of finding a moment optimal policy. This will be clear in the treatment which follows, although it will not be stated explicitly. For a discussion of this and other implications for utility theory and Markov decision processes, the reader should see Jaquette ([3] Chapter 5), and Howard and Matheson ([2]). The approach here and in Jaquette ([3]) is different from that used by Howard and Matheson; however the results are quite similar.

3. A stationary policy is moment optimal. The main result of this section is:

THEOREM 1. *For the finite state and action discrete time Markov decision process defined above, there exists a stationary policy which is moment optimal.*

We dispense with some preliminaries before proving this theorem. We know that $r(\cdot, \cdot)$ is bounded; assume that $r(i, a) \leqq B < \infty$ for all $i \in S$ and $a \in A$. From (2.1) we conclude that $|[\mathbf{R}(\pi)]_i| \leqq (1 - \beta)^{-1}B < \infty$. Thus we can conclude that $\mathbf{R}(\pi)$ is bounded everywhere and that its moments grow at most geometrically since $|\mathbf{M}_n(\pi)| \leqq B^n\mathbf{1}$. This is sufficient to ensure that the Laplace transform of $\mathbf{R}(\pi)$, and hence $\mathbf{U}_\pi(\lambda)$, exists and is finite on the whole real line for any policy π. We are thus justified in using the Taylor series expansion for $\mathbf{U}_\pi(\lambda)$.

From the definition of the function $\mathbf{U}$ we have

$$[\mathbf{U}_{f\pi}(\lambda)]_i = -E[\exp(-\lambda[\mathbf{R}(f\pi)]_i)] ,$$

where by the policy $f\pi$ we mean that $\mathbf{f}$ is used at time zero and then the elements of π are used in sequence. The Markov property of the process allows us to conclude that

$$[\mathbf{U}_{f\pi}(\lambda)]_i = \exp(-\lambda[\mathbf{r}(f)]_i) \textstyle\sum_{j=1}^{s} p_{ij}(f(i))[\mathbf{U}_\pi(\beta\lambda)]_j . \tag{3.1}$$

This result follows directly by using the Markov property and taking suitable conditional expectations. Equation (3.1) can be written in vector and matrix form as follows:

$$\mathbf{U}_{f\pi}(\lambda) = \{\exp(-\lambda\mathbf{r}(f))\} \circ \{P(f)\mathbf{U}_\pi(\beta\lambda)\} . \tag{3.2}$$

We can simplify the calculations by defining the operator L_f and restating (3.2) in terms of this operator. Let $\mathscr{L}$ be the space of functions of a real variable which are the negative of Laplace transforms of bounded random variables. Let $\mathscr{L}^s$ be the s-fold direct product of $\mathscr{L}$: $\mathscr{L}^s = \bigtimes_{i=1}^{s} \mathscr{L}$. We define the operator L_f as follows:

$$(L_f\mathbf{u})(\lambda) \equiv \{\exp(-\lambda\mathbf{r}(f))\} \circ \{P(f)\mathbf{u}(\beta\lambda)\} , \qquad \mathbf{u}(\lambda) \in \mathscr{L}^s . \tag{3.3}$$

It is easy to verify that $L_f\mathbf{u}$ is in $\mathscr{L}^s$ if $\mathbf{u}$ is in $\mathscr{L}^s$ and thus that $L_f\colon \mathscr{L}^s \to \mathscr{L}^s$. We can then restate (3.2) as:

$$\mathbf{U}_{f\pi}(\lambda) = L_f\mathbf{U}_\pi(\lambda)\,. \tag{3.4}$$

The following lemma establishes a sense in which L_f is monotone.

LEMMA 3.1. *Let* $\mathbf{f}$ *be any action vector,* $\lambda_0 > 0$, *and* $\mathbf{u}$ *in* $\mathscr{L}^s$. *If*

$$L_f(\mathbf{u})(\lambda) \geqq \mathbf{u}(\lambda) \tag{3.5}$$

for all λ *in* $[0, \lambda_0]$, *then for all* λ *in* $[0, \lambda_0]$

$$\mathbf{U}_{f^\infty}(\lambda) \geqq \mathbf{u}(\lambda)\,.$$

PROOF. Equations (3.3) and (3.5) suffice to establish inductively that for λ in $[0, \lambda_0]$

$$(L_f)^n(\mathbf{u})(\lambda) \geqq \mathbf{u}(\lambda)\,.$$

It suffices to show that $(L_f)^n(\mathbf{u})(\lambda) \to U_{f^\infty}(\lambda)$. To show this define a policy $\pi^n\pi^*$ to be the policy that uses the action vectors from π until time n and then uses the action vectors from π^* in their proper order. If $\pi = \{\mathbf{f}_0, \mathbf{f}_1, \cdots\}$, define $L_\pi{}^n$ as follows:

$$L_\pi{}^n(\mathbf{u}) \equiv L_\pi{}^{n-1}(L_{f_n}(\mathbf{u}))\,, \qquad \text{and} \qquad L^0(\mathbf{u}) \equiv L_{f_0}(\mathbf{u})\,.$$

It is elementary to show that $\mathbf{R}(\pi^n\pi^*) \to \mathbf{R}(\pi)$ a.e. $(n \to \infty)$, that $\mathbf{U}_\pi n_{\pi^*}(\lambda) \to \mathbf{U}_\pi(\lambda)$ $(n \to \infty)$, and that

$$L_\pi{}^n(\mathbf{u}) \to \mathbf{U}_\pi \qquad (n \to \infty) \quad \text{on the real line.} \tag{3.6}$$

We note that Lemma 3.1 and its proof hold with all inequalities reversed. It also holds with a strict inequality. We shall use these generalizations in what follows.

LEMMA 3.2. *Let* $\mathbf{f}$ *and* $\mathbf{g}$ *be any action vectors in* F. *Define the function* $h(\cdot)$ *as follows*:

$$h(\lambda) \equiv [\mathbf{U}_{f^\infty}(\lambda)]_i - [\mathbf{U}_{g^\infty}(\lambda)]_i\,. \tag{3.7}$$

There exists a positive number λ_0 *such that* $h(\cdot)$ *does not change sign on the interval* $[0, \lambda_0]$.

PROOF. Laplace transforms are analytic on the interior of their region of convergence. Thus $h(\cdot)$ is analytic on the real line. If the lemma were false, then h would cross zero infinitely often as λ approaches zero, and there would exist a sequence $\{\lambda_n\} \to 0^+$ such that $h(\lambda_n) = 0$ for all n. This would imply that h is identically zero, since h is analytic. Thus the lemma must hold.

For convenience in notation for the following lemma, we write $h > 0$ $(\lambda \to 0^+)$ if a function as defined in Lemma 3.2 does not change sign on some interval $[0, \lambda_0]$ and if $h(\lambda) > 0$ on $(0, \lambda_0)$.

LEMMA 3.3. *Let* **f** *and* **g** *be any action vectors in F and let h be as in Lemma* 3.2. *Define the action vector* **e** *as follows*:

$$e(i) \equiv f(i) \quad \text{if} \quad h > 0 \quad (\lambda \to 0^+), \tag{3.8}$$
$$\equiv g(i) \quad \text{otherwise.}$$

There exists a scalar $\lambda_0 > 0$ *such that*

$$\mathbf{U}_{e^\infty}(\lambda) \geqq \max\{\mathbf{U}_{f^\infty}(\lambda), \mathbf{U}_{g^\infty}(\lambda)\} \qquad \text{for all} \quad \lambda \in [0, \lambda_0].$$

PROOF. Associated with each index i there is a positive $\lambda_0{}^i$ guaranteed by Lemma 3.2 such that the appropriate $h(\cdot)$ does not change sign. Take the λ_0 for Lemma 3.3 to be the minimum of these $\lambda_0{}^i$, which must be positive since S is finite. The definition for **e**, (3.8), leads directly to the statements that for $\lambda \in [0, \lambda_0]$

$$L_e(\mathbf{U}_{f^\infty})(\lambda) \geqq \mathbf{U}_{f^\infty}(\lambda), \qquad \text{and} \qquad L_e(\mathbf{U}_{g^\infty})(\lambda) \geqq \mathbf{U}_{g^\infty}(\lambda).$$

We can now apply Lemma 3.1 to obtain the conclusion of the lemma.

We now define an alternative optimality criterion. A policy π^* is called *U-optimal* if there exists a positive scalar λ_0 such that $\mathbf{U}_{\pi^*}(\lambda) \geqq \mathbf{U}_\pi(\lambda)$ for every π and all λ in $[0, \lambda_0]$. We remark that *U*-optimality is optimality being a "little bit risk adverse."

LEMMA 3.4. *There exists a stationary policy that is U-optimal.*

PROOF. Restrict consideration first to stationary policies. Starting with any stationary policy and applying Lemma 3.3 repeatedly, we are assured of the existence of a positive λ_0 and an action vector **f** such that $\mathbf{U}_{f^\infty}(\lambda) \geqq \mathbf{U}_{g^\infty}(\lambda)$ for all $\lambda \in [0, \lambda_0]$ and for every $\mathbf{g} \in F$. This follows since there are only a finite number of distinct elements in F. We can then conclude immediately that every action vector **g** satisfies

$$L_g(\mathbf{U}_{f^\infty})(\lambda) \leqq \mathbf{U}_{f^\infty}(\lambda) \qquad \text{for all} \quad \lambda \in [0, \lambda_0]. \tag{3.9}$$

If this conclusion did not follow, then Lemma 3.3 would ensure an improvement on **f**, which cannot obtain. By suitable choices of **g** in (3.9) and applying these L_g repeatedly, we can construct $L_\pi{}^n$ for any choice of policies π. This application ensures that for all $\lambda \in [0, \lambda_0]$

$$L_\pi{}^n(\mathbf{U}_{f^\infty})(\lambda) \leqq \mathbf{U}_{f^\infty}(\lambda).$$

Equation (3.6) allows us to conclude that $\mathbf{U}_\pi(\lambda) \leqq \mathbf{U}_{f^\infty}(\lambda)$ on $[0, \lambda_0]$, which completes the proof.

LEMMA 3.6. *A policy is moment optimal if and only if it is U-optimal.*

PROOF. This follows directly from the definitions of moment optimality and *U*-optimality, Lemma 3.5, and the Taylor series expansion for $\mathbf{U}_\pi(\lambda)$, which converges for all λ,

$$\mathbf{U}_\pi(\lambda) = \textstyle\sum_{n=0}^{\infty} \mathbf{N}_n(\pi)\lambda^n/n!. \tag{3.10}$$

It is clear the $\mathbf{N}(\pi^*) \succeq \mathbf{N}(\pi)$ if and only if $\mathbf{U}_{\pi^*}(\lambda) \geqq \mathbf{U}_{\pi}(\lambda)$ for all λ is some suitable interval $[0, \lambda_0]$ $(\lambda_0 > 0)$.

Theorem 1 can now be verified very easily. Lemma 3.5 establishes the existence of a stationary policy that is U-optimal. Lemma 3.6 establishes that the moment optimal policies are exactly the U-optimal policies.

It should also be clear from the proof that the assumption of a single action set A is merely for convenience and that the more general action sets A_i with $F = \bigtimes_{i=1}^{s} A_i$ admits the identical proofs.

It should also be clear that the particular selection of alternating signs in the definition of moment optimality is not critical in the proofs. In fact the same results obtain if we choose any of the following sequence of signs for the moments of return in the definition of $\mathbf{N}(\pi)$ in (2.4): $+1, -1, +1, \cdots$ or $-1, +1, -1, \cdots$ or $+1, +1, +1, \cdots$ or $-1, -1, -1, \cdots$.

4. An algorithm to construct a moment optimal policy. To facilitate construction of all stationary policies which are moment optimal, we define additional optimality criteria which are related to moment optimality. We define the criterion of (m) *moment optimality* to be moment optimality ignoring all moments higher than the mth moment. Recalling the definition of $\mathbf{N}^m(\pi)$ in (2.5), we call a policy π^* (m) moment optimal if $\mathbf{N}^m(\pi^*) \succeq \mathbf{N}^m(\pi)$ for all policies π.

It should be evident that a policy is moment optimal if and only if it is (m) moment optimal for all m. We base an algorithm on this observation. From Theorem 1 we know that there is a stationary policy which is moment optimal. A stationary moment optimal policy is (m) moment optimal for all m. We can, therefore, restrict our attention to stationary policies.

Define the sets of action vectors, F^m, as follows:

$$F^m \equiv \{\mathbf{f}: \mathbf{f} \in F \text{ and } f^\infty \text{ is } (m) \text{ moment optimal}\}. \tag{4.1}$$

From our previous observations $F^m \supseteq F^{m+1}$ and $F^\infty = \lim_{m\to\infty} F^m \neq \emptyset$. The set F^∞ is the set of action vectors, $\mathbf{f}$, such that f^∞ is moment optimal.

In this section we shall use some additional notation. If $\mathbf{N}$ and $\mathbf{M}$ are vectors with vector components $\mathbf{N}_i$ and $\mathbf{M}_i$ respectively, we write $\mathbf{N} =_n \mathbf{M}$ if $\mathbf{N}_i = \mathbf{M}_i$ for $i \leqq n$. Similarly we write $\mathbf{N} \succeq_n \mathbf{M}$ if $\mathbf{N} \succ \mathbf{M}$ and $\mathbf{N} \neq_n \mathbf{M}$. The relation $\succeq_n$, $\preceq_n$, and $\prec_n$ are similarly defined.

Suppose that $\mathbf{U}$ is in $\mathscr{L}^s$; then we can write the Taylor expansion as $\mathbf{U}(\lambda) = \sum_{n=0}^{\infty} \mathbf{u}_n \lambda^n/n!$. By the notation $\mathbf{U}$ we mean the vector whose vector components are given by this Taylor expansion, e.g., $\mathbf{U} = \{\mathbf{u}_0, \mathbf{u}_1, \cdots\}$. If $\mathbf{U}(\lambda)$ in $\mathscr{L}^s$ is viewed as the negative of the Laplace transform of some bounded random vector $\mathbf{R}$, then $E[\mathbf{R}^n] = (-1)^{n+1}\mathbf{u}_n$.

LEMMA 4.1. *Suppose* $\mathbf{U}(\lambda)$ *is in* $\mathscr{L}^s$ *and let* $\mathbf{f}$ *be any action vector. Then*

$$\mathbf{U}_{f^\infty} =_m \mathbf{U} \tag{4.2}$$

if and only if

$$L_f(\mathbf{U}) =_m \mathbf{U}. \tag{4.3}$$

Suppose that $L_f(\mathbf{U}) \succ_m \mathbf{U}$. *Then*

$$\mathbf{U}_{f^\infty} \succ_m \mathbf{U}\,.$$

PROOF. Suppose that (4.3) holds. Then a simple induction argument leads to $(L_f)^n(\mathbf{U}) =_m \mathbf{U}$. The result in (3.6) concerning convergence of the transform implies $(L_f)^n(\mathbf{U}) \to \mathbf{U}_{f^\infty}$, which yields (4.2).

Suppose that (4.2) holds. Using (3.4) establishes that $L_f(\mathbf{U}_{f^\infty}) = \mathbf{U}_{f^\infty}$, therefore if we can show that $L_f(\mathbf{U}) =_m L_f(\mathbf{U}_{f^\infty})$ follows from (4.2) we could establish that

$$L_f(\mathbf{U}) =_m L_f(\mathbf{U}_{f^\infty}) = \mathbf{U}_{f^\infty} =_m \mathbf{U}\,,$$

which is just (4.3). To fill the gap expand the components of (3.3) as a power series in λ and equate coefficients of λ^n:

$$[L_f(\mathbf{U})]_n = \textstyle\sum_{i=0}^{n} \beta^i \binom{n}{i} (-\mathbf{r}(f))^{n-i} \circ P(f)\mathbf{u}_i\,. \tag{4.4}$$

Since $[L_f(\mathbf{U})]_n$ is a linear combination of $\mathbf{u}_i$ for $i \leqq n$ only and since $\mathbf{U}_{f^\infty} =_m \mathbf{U}$ by (4.2), we can conclude that $[L_f(\mathbf{U})]_n = [L_f(\mathbf{U}_{f^\infty})]_n$ for $n \leqq m$. This completes this portion of the proof.

If $L_f(\mathbf{U}) \succ_m \mathbf{U}$, then using (3.10) it is clear that the hypotheses of Lemma 3.1 are satisfied and hence that $\mathbf{U}_{f^\infty} \succeq \mathbf{U}$. Applying L_f to both sides of this relationship yields

$$\mathbf{U}_{f^\infty} = L_f(\mathbf{U}_{f^\infty}) \succeq L_f(\mathbf{U}) \succ_m \mathbf{U}\,.$$

LEMMA 4.2. *The set* F^m *is a direct product. If* $f \in F^m$ *and* $F^m = \bigtimes_{i=1}^{s} A_i^m$, *then*

$$A_i^m = \{a \in A_i \colon L_g(\mathbf{U}_{f^\infty}) =_m \mathbf{U}_{f^\infty} \text{ for } g(i) = a \text{ and } g(j) = f(j)\ (i \neq j)\}\,. \tag{4.5}$$

PROOF. Let $\mathbf{U}$ in Lemma 4.1 be $\mathbf{U}_{f^\infty}$. Since $\mathbf{f} \in F^m$, Lemma 4.1 implies that $\mathbf{g} \in F^m$ if and only if $L_g(\mathbf{U}_{f^\infty}) =_m \mathbf{U}_{f^\infty}$. Clearly $F^{m-1} \supseteq F^m$, so that to define F^m we need only consider $\mathbf{g} \in F^{m-1}$. This eliminates consideration of all components of $L_g(\mathbf{U}_{f^\infty})$ except the mth. From (4.4) it is evident that the ith component of $[L_g(\mathbf{U})]_m$ depends only on $g(i)$ and not on $g(j)$ for $j \neq i$. It thus follows directly that F^m is a direct product and that A_i^m can be given by (4.5).

We remark that $A_i^{m-1} \supseteq A_i^m$ follows directly from $F^{m-1} \supseteq F^m$. Since F^∞ is nonempty by Theorem 1 so are the A_i^m. This indicates that F^m can easily be constructed from F^{m-1} if an element $\mathbf{f}$ in F^m can be found. This construction can be accomplished by using (4.5) and (4.4) and noting that $[\mathbf{U}_{f^\infty}]_m = \mathbf{N}_m(f^\infty)$.

LEMMA 4.3. *Pick any* $\mathbf{f}$ *in* F^{m-1}. *Either*

(a) $\mathbf{f} \in F^m$ *or*

(b) *there exists a* $\mathbf{g} \in F^{m-1}$ *such that*

$$L_g(\mathbf{U}_{f^\infty}) \succ_m \mathbf{U}_{g^\infty}\,. \tag{4.6}$$

If (4.6) *holds, then* $\mathbf{U}_{g^\infty} \succ_m \mathbf{U}_{f^\infty}$.

PROOF. First assume that (4.6) holds. Lemma 4.1 can be applied to show directly that $\mathbf{U}_{g^\infty} \succ_m \mathbf{U}_{f^\infty}$.

Suppose then that (4.6) holds for no $\mathbf{g}$ in F^{m-1}. For every $\mathbf{g}$ in F^{m-1} we then have $L_g(\mathbf{U}_{f^\infty}) \preceq_m \mathbf{U}_{f^\infty}$. A direct analog of Lemma 4.1 for the inequality $\prec$ and that lemma as stated combine to show that then $\mathbf{U}_{g^\infty} \preceq_m \mathbf{U}_{f^\infty}$ for all $\mathbf{g}$ in F^{m-1}. This is simply the statement that $\mathbf{f}$ is in F^m, which is part (a) of the lemma.

Lemma 4.3 indicates a policy improvement algorithm. An action vector $\mathbf{g}$ is sought which improves upon all others in F^{m-1} in the sense that it is in F^m. Lemma 4.1, especially (4.4) indicates how to find a $\mathbf{g}$ in F^{m-1} satisfying (4.6). Since $L_g(\mathbf{U}_{f^\infty}) =_{m-1} \mathbf{U}_{f^\infty}$, we seek an improvement $\mathbf{g}$ satisfying $[L_g(\mathbf{U}_{f^\infty})]_m > [\mathbf{U}_{f^\infty}]_m$ or

$$\sum_{i=0}^{m} \beta^i \binom{m}{i} (-\mathbf{r}(g))^{m-i} \circ P(g)\mathbf{N}_i(f^\infty) > \mathbf{N}_m(f^\infty) . \tag{4.7}$$

The ith component of the left side of (4.7) depends only on $g(i)$ and not on $g(j)$ $(j \neq i)$. Thus an improvement on $\mathbf{f}$ can be obtained by $\mathbf{g}$ defined by

$$g(i) = \{a \in A_i^{m-1} : a \text{ maximizes the } i\text{th component of the left side of (4.7)}\} .$$

If $\mathbf{g}$ obtained equals $\mathbf{f}$, then $\mathbf{f}$ is in F^m and no further improvement is possible. This $\mathbf{f}$ then allows direct construction of the A_i^m and F^m as indicated by Lemma 4.2 and the following remarks.

THEOREM 2. *The set F^m can be constructed in a finite number of steps.*

PROOF. The theorem follows by an inductive argument using Lemmas 4.2 and 4.3 and the following remarks. The finiteness of each step, generating F^m from F^{m-1}, follows since the policy improvement can be performed only a finite number of times. This is a result of the finiteness of F.

5. The algorithm to construct moment optimal policies is finite. Theorem 2 assures that (m) moment optimal policies can be constructed a finite algorithm. Moment optimality requires consideration of all moments, and construction of (∞) moment optimal policies would require a countable algorithm. In fact only a finite algorithm is needed, as indicated in the following theorem.

THEOREM 3. *Moment optimality is equivalent to (n^*) moment optimality for some finite n^*. All stationary moment optimal policies can be constructed by a finite algorithm.*

PROOF. Since there are only a finite number of stationary policies, there must be a finite number n^* such that for all pairs of action vectors $\mathbf{f}$ and $\mathbf{g}$, $\mathbf{U}_{f^\infty} \succ \mathbf{U}_{g^\infty}$ if and only if $\mathbf{U}_{f^\infty} \succ_{n^*} \mathbf{U}_{g^\infty}$.

Under certain circumstances n^* can be determined beforehand to be no greater than s; however in general this is not the case. If the algorithm is performed iteratively, a stopping rule can be given: if $\mathbf{R}(f^\infty) =_{\mathscr{D}} \mathbf{R}(g^\infty)$ for all $\mathbf{f}$ and $\mathbf{g}$ in F^m for some m, then $n^* \leqq m$. This is true if F^m has a single element; however there may be several elements in F^∞ with the same return distributions, in which case verifying that $\mathbf{R}(f^\infty) =_{\mathscr{D}} \mathbf{R}(g^\infty)$ is equivalent to calculating a countable number of moments.

6. Acknowledgments. The author would like to acknowledge the assistance of his advisor, Professor Donald L. Iglehart, throughout the preparation of his dissertation while at Stanford University. These results are given in much greater detail in the dissertation [3]. The author also appreciates the suggestions of the referee, which were very helpful in indicating a more compact method of proof.

REFERENCES

[1] DERMAN, C. (1964). On sequential control processes. *Ann. Math. Statist.* **35** 341–349.

[2] HOWARD, R. A. and MATHESON, J. E. (1972). Risk-sensitive Markov decision processes. *Management Sci.* **18** 356–369.

[3] JAQUETTE, S. C. (1971). Markov decision processes with a new optimality criterion. Technical Report No. 15, Department of Operations Research, Stanford Univ.

[4] WHITT, W. (1972). Stochastic abelian and tauberian theorems. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **22** 251–267.

CORNELL UNIVERSITY
DEPARTMENT OF OPERATIONS RESEARCH
UPSON HALL
ITHACA, NEW YORK 14850