The Annals of Statistics
1974, Vol. 2, No. 2, 411-413

AVERAGING VS. DISCOUNTING IN DYNAMIC PROGRAMMING: A COUNTEREXAMPLE

BY JAMES FLYNN

University of Chicago

We consider countable state, finite action dynamic programming problems with bounded rewards. Under Blackwell's optimality criterion, a policy is optimal if it maximizes the expected discounted total return for all values of the discount factor sufficiently close to 1. We give an example where a policy meets that optimality criterion, but is not optimal with respect to Derman's average cost criterion. We also give conditions under which this pathology cannot occur.

1. Introduction. We consider a dynamic programming problem with a countable state space S (see Blackwell (1962), (1965), Derman (1965), (1966) and Maitra (1965)). Each day we observe the current state s of some system and choose an action a from a finite action space A. This selection results in (1) an immediate income $i(s, a)$ and (2) a transition to a new state s' with probability $q(s' \mid s, a)$. We assume that the incomes are bounded. The problem is to control the system in the most effective manner over an infinite future.

A rule or policy π for controlling the system specifies for each $n \geqq 1$ what act to choose on the nth day as a function of the system's current history $h = (s_1, a_1, \cdots, s_n)$ or, more generally, π specifies for each h a probability distribution on A. A (nonrandomized) stationary policy is a policy which is specified by a single function f mapping S into A: under it, you select act $f(s)$ whenever the system is in state s.

There are different ways of measuring the effectiveness of a policy. Blackwell's (1962) approach is to favor policies which maximize the expected value of discounted total return for all values of the discount factor β sufficiently close to 1, while Derman's (1966) is to favor policies which minimize the expected value of the long-run average cost. To be more specific, we need some notation and definitions.

Let $r_j(s, \pi)$ denote the expected return on the jth day under the policy π when the initial state is s ($j = 1, 2, \cdots$). For each $\beta \in (0, 1)$, let

$$V_\beta(s, \pi) = \sum_{j=1}^{\infty} \beta^{j-1} r_j(s, \pi) \qquad (s \in S) \tag{1}$$

and

$$x(s, \pi) = \liminf_n \left(\sum_{j=1}^{n} r_j(s, \pi)\right)/n \qquad (s \in S). \tag{2}$$

Received October 1972; revised May 1973.

AMS 1970 *subject classifications.* Primary 49C15, 62L99, 90C40, 93C55; Secondary 60J10, 60J20.

Key words and phrases. Dynamic programming, average cost criteria, discounting, Markov decision process.

DEFINITION 1. A policy π_* is B-optimal if there exists a $\beta_0 \in (0, 1)$ such that

$$V_\beta(s, \pi_*) \geqq V_\beta(s, \pi) \qquad (s \in S, \beta \in (\beta_0, 1)) \tag{3}$$

for any policy π.

Blackwell (1962) and Derman (1965) established the existence of a (non-randomized) stationary B-optimal policy for finite S, while Maitra (1965) constructed a countable state system for which there was no B-optimal policy.

DEFINITION 2. A policy π_* is D-optimal if

$$x(s, \pi_*) \geqq x(s, \pi) \qquad (s \in S) . \tag{4}$$

for any policy π.

Intuitively, one would expect D-optimality to be weaker than B-optimality. Certainly, it is easy to construct D-optimal policies which are not B-optimal. Also, it is natural to conjecture that B-optimal policies are always D-optimal. This conjecture, however, turns out to be false. We will provide a counterexample. We will also show that the conjecture is true when S is finite.

2. A counterexample. Liggett and Lippman (1969) established the existence of a bounded sequence of real numbers $\{r_n\}_{n=1}^\infty$ satisfying

$$r^* \equiv \liminf_{\beta \to 1^-} (1 - \beta) \textstyle\sum_{j=1}^\infty \beta^{j-1} r_j > \liminf_n (\sum_{j=1}^n r_j)/n \equiv r_* . \tag{5}$$

Let the state space S consist of $0, r_1, r_2, \cdots$. To each state there corresponds two actions, 0 and 1. Transitions are deterministic:

$$q(r_{j+1} | r_j, 0) = q(r_{j+1} | r_j, 1) = 1 \qquad (j = 1, 2, \cdots)$$
$$q(0 | 0, 0) = q(r_1 | 0, 1) = 1 .$$

The immediate income depends only on the state:

$$i(r_j, 0) = i(r_j, 1) = r_j \qquad (j = 1, 2, \cdots)$$
$$i(0, 0) = i(0, 1) = (r^* + 2r_*)/3 .$$

Let π_j denote the policy which always selects action j ($j = 0, 1$). Clearly, π_1 is B-optimal. One can show that π_0 is D-optimal by establishing

$$x(0, \pi_0) = (r^* + 2r_*)/3 > r_* = x(0, \pi_1) . \tag{6}$$

It follows that π_1 is not D-optimal.

3. Sufficient conditions. A sufficient condition for a B-optimal policy π_* to be D-optimal is

$$\liminf_{\beta \to 1^-} (1 - \beta) V_\beta(s, \pi_*) = x(s, \pi_*) \qquad (S \in S) . \tag{7}$$

This follows immediately from the fact (Hobson (1926)) that

$$\liminf_{\beta \to 1^-} (1 - \beta) V_\beta(s, \pi) \geqq x(s, \pi) \qquad (S \in S) .$$

In particular, any B-optimal policy π_* is D-optimal when S is finite since (7) always holds in that case. We establish this result as follows: By Blackwell (1962) and Derman (1965), there exists a stationary policy $\hat{\pi}$ which is B-optimal. Hence for some $\beta_0 \in (0, 1)$, we have $V_\beta(s, \pi_*) = V_\beta(s, \hat{\pi})$ for all s and all $\beta \in (\beta_0, 1)$. Moreover, $V_\beta(s, \hat{\pi})$ is a rational function (Blackwell (1962)). Hence $\lim_{\beta \to 1^-} (1 - \beta) V_\beta(s, \pi_*)$ exists. The existence of this limit and the Hardy–Littlewood theorem (see Liggett and Lippman (1969)) give us (7).

4. Remarks. The case where D-optimality is defined in terms of the lim sup instead of the lim inf is similar. Using the approach of Section 2, one can construct an example where a B-optimal policy does not maximize the lim sup of the average returns. Results analogous to those of Section 3 are easy to establish for the lim sup case.

5. Acknowledgments. We wish to thank David Blackwell and Donald Iglehart for useful discussions. We also wish to thank Bennett Fox and the referee for their helpful comments on an earlier version of this note.

REFERENCES

BLACKWELL, D. (1962). Discrete dynamic programming. *Ann. Math. Statist.* **33** 719–726.

BLACKWELL, D. (1965). Discounted dynamic programming. *Ann. Math. Statist.* **36** 226–235.

DERMAN, C. (1965). Markovian sequential control processes—denumerable state space. *J. Math. Anal. Appl.* **10** 295–302.

DERMAN, C. (1966). Denumerable state Markovian decision processes—average cost criterion. *Ann. Math. Statist.* **37** 1545–1553.

HOBSON, E. (1926). *The Theory of Functions of a Real Variable and the Theory of Fourier's Series.* Cambridge Univ. Press.

LIGGETT, T. and LIPPMAN, S. (1969). Stochastic games with perfect information and time average payoff. *SIAM Rev.* **11** 604–607.

MAITRA, A. (1965). Dynamic programming for countable state systems. *Sankhyā Ser. A* **27** 259–266.

GRADUATE SCHOOL OF BUSINESS
UNIVERSITY OF CHICAGO
5836 GREENWOOD AVENUE
CHICAGO, ILLINOIS 60637