The Annals of Mathematical Statistics
1971, Vol. 42, No. 5, 1767–1768

ON THE NONEXISTENCE OF ε-OPTIMAL RANDOMIZED STATIONARY POLICIES IN AVERAGE COST MARKOV DECISION MODELS

BY SHELDON M. ROSS

University of California, Berkeley

1. Introduction. Consider a Markov decision process [see Derman (1966) or Ross (1968)] having a countable state space I and a finite action space A. If action a is chosen when in state i, then

(i) a cost $c[i, a]$ is incurred, and

(ii) the next state is determined according to the transition probabilities $\{P_{ij}(a), j \in I\}$.

A policy is any measurable rule for choosing actions, and is called stationary if the (possibly randomized) action the policy chooses at any time depends only on the state of the process at that time. In Maitra (1966), the question is asked if whether or not there always exists an ε-optimal stationary policy under the average expected cost criterion. That is, is there a stationary policy whose average expected cost is within ε of the infimum over all policies? We answer this in the negative by the following counterexample.

2. The counterexample. Let the states by given by $1, 1', 2, 2', \cdots, n, n', \cdots, \infty$. In state n, $1 \leqq n < \infty$, there are two actions, with transition probabilities given by

$$P_{n,n+1}(1) = 1$$

$$P_{n,n'}(2) = \alpha_n = 1 - P_{n,\infty}(2).$$

In state n', there is a single action, having transition probabilities

$$P_{n',(n-1)'} = 1, \qquad n \geqq 2$$

$$P_{1',1} = 1.$$

State ∞ is an absorbing state and once entered is never left, i.e.,

$$P_{\infty\infty} = 1.$$

The costs depend only on the state and are given by

$$c[n, a] = 2 \qquad \text{all } n = 1, 2, \cdots, \infty, \text{all actions } a$$

$$c[n', a] = 0 \qquad \text{all } n \geqq 1, \text{ all } a.$$

The values α_n are chosen to satisfy

(i) $\alpha_n < 1$

(ii) $\prod_{n=1}^{\infty} \alpha_n = \frac{3}{4}$.

Received December 22, 1970.

Suppose the initial state is state 1. If a stationary policy is employed, then, with probability 1, a cost of 2 will be incurred in all but a finite number of time periods. This follows, since, under a stationary policy, each time the process enters state 1 there is a fixed positive probability that the process will never again re-enter that state. Therefore, under a stationary policy, the average cost will, with probability 1, equal 2. Hence, by the bounded convergence theorem, the average expected cost will also equal 2.

Now let R be the nonstationary policy which initially chooses action 2, and then on its nth return to state 1, chooses action 1 n times and then chooses action 2. The average cost under this policy will equal

$$\begin{aligned} &2 \text{ with probability } 1-\textstyle\prod_{n=1}^{\infty}\alpha_n \\ &1 \text{ with probability } \textstyle\prod_{n=1}^{\infty}\alpha_n. \end{aligned}$$

This is true since $\prod_{n=1}^{\infty}\alpha_n$ represents the probability that, under R, the process will never enter state ∞. Hence, by the bounded convergence theorem, the average expected cost under R is $\frac{3}{4}+\frac{2}{4} = \frac{5}{4}$.

Hence, there is no ε-optimal randomized stationary policy for $\varepsilon < \frac{3}{4}$.

Acknowledgment. This research has been partially supported by the Office of Naval Research under Contract N00014-69-A-0200-1010 and the U.S. Army Research Office, Durham under contract DA-31-124-ARO-D-331 with the University of California.

REFERENCES

Derman, Cyrus (1966). Denumerable state Markovian decision processes-average cost criterion. *Ann. Math. Statist.* **37** 1545–1556.

Fisher, L. and Ross, S. (1968). An example in denumerable decision processes. *Ann. Math. Statist.* **39** 674–675.

Maitra, Ashok (1966). A note on undiscounted dynamic programming. *Ann. Math. Statist.* **37** 1042–1044.

Ross, Sheldon (1968). Nondiscounted denumerable Markovian decision models. *Ann. Math. Statist.* **39** 412–423.