A NOTE ON UNDISCOUNTED DYNAMIC PROGRAMMING[1]

By Ashok Maitra[2]

Mathematisch Centrum, Amsterdam

1. Introduction. We consider a system with a finite number of states $1, 2, \cdots, S$. Once a day, we observe the current state s of the system and choose an action a from an arbitrary set A of actions. As a result, two things happen: (1) we receive an immediate income $i(s, a)$, and (2) the system moves to a new state s' with probability $q(s' \mid s, a)$. Assume that the incomes are bounded, that is, there exists a positive number M such that $|i(s, a)| \leqq M$, $s = 1, 2, \cdots, S$, $a \varepsilon A$. The problem is to maximise the average rate of income (to be defined below).

Denote by F the set of all functions f on S into A. A *policy* $\pi = \{f_1, f_2, \cdots\}$ is a sequence of functions $f_n \varepsilon F$. Thus, to use policy π is to choose the action $f_n(s)$ on the nth day, if the system is in state s on that day. We shall call a policy $\pi = \{f_n\}$ *stationary* if $f_n = f$, $n = 1, 2, \cdots$, and denote it by $f^{(\infty)}$.

With each $f \varepsilon F$, associate (1) the $S \times 1$ vector $r(f)$, whose sth coordinate is $i(s, f(s))$ and (2) the $S \times S$ stochastic matrix $Q(f)$, whose (s, s') element is $q(s' \mid s, f(s))$. Hence, if we use the policy $\pi = \{f_n\}$, the n-step transition matrix of the system is $Q_n(\pi) = \prod_{k=1}^{n} Q(f_k)$. In particular, if our policy is stationary, the system becomes a discrete time-parameter Markov chain with stationary transition probabilities.

Given a policy π, let us denote by $W_n(\pi)$ the $S \times 1$ vector of incomes on the nth day, when the policy π is used. Set

$$x(\pi) = \lim_{N \to \infty} N^{-1} \sum_{n=1}^{N} W_n(\pi)$$

whenever the limit exists. Blackwell [1] has shown that the limit exists whenever π is a stationary policy. In the case of a stationary policy, $x(f^{(\infty)})$ is the vector of average rates of income, when the policy $f^{(\infty)}$ is used.

We shall say that a policy $f_0^{(\infty)}$ is *optimal* among stationary policies if $x(f_0^{(\infty)}) \geqq x(f^{(\infty)})$ for all $f \varepsilon F$ (for any two $S \times 1$ vectors w_1 and w_2, we shall write $w_1 \geqq w_2$ if every coordinate of w_1 is at least as large as the corresponding coordinate of w_2, and $w_1 > w_2$ if $w_1 \geqq w_2$ and $w_1 \neq w_2$).

Blackwell [1] showed that, if A is finite, there exists an optimal policy among stationary policies. When A is not finite, there may not exist an optimal policy. Consider, for instance, a system with a single state and $A = \{1, 2, \cdots\}$. Choice of action i brings an income of $1 - 1/i$ dollars. It is clear that there is no optimal stationary policy.

The purpose of this note is to prove:

Theorem. *Let A be arbitrary. Given $\epsilon > 0$, there exists a stationary policy $f_\epsilon^{(\infty)}$*

Received 24 November 1965.

[1] Report SP 89 of the Statistics Department, Mathematisch Centrum, Amsterdam.

[2] Now with Indian Statistical Institute, Calcutta.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Mathematical Statistics.* www.jstor.org