ON FINDING OPTIMAL POLICIES IN DISCRETE DYNAMIC PROGRAMMING WITH NO DISCOUNTING[1]

By Arthur F. Veinott, Jr.

Stanford University

1. Introduction. In an elegant paper [1] Blackwell has studied the infinite horizon discrete time parameter Markovian sequential decision problem with finitely many states and actions. He focuses initially on the case where there is a discount factor β, $0 \leqq \beta < 1$. The problem is to choose a policy, termed *β-optimal*, that maximizes the total expected discounted return over an unbounded time horizon. He shows that there is a β-optimal policy which is stationary. He also gives a neat proof that Howard's [2], p. 84, policy improvement method yields a β-optimal stationary policy in finitely many steps.

For the case $\beta = 1$, a policy is called 1-*optimal* if the difference between the total expected discounted return with that policy and the β-optimal policy for $0 \leqq \beta < 1$ tends to 0 as $\beta \nearrow 1$.[2] Blackwell established the existence of a 1-optimal policy that is stationary. He also shows that Howard's [2], p. 64, policy improvement method yields an element of the set of stationary policies that maximize the long run average return per unit time. Blackwell shows that this set contains the set of stationary 1-optimal policies. Thus if there is only one stationary policy with maximal average return, then that policy is 1-optimal and will be found by the policy improvement method. If there are two or more stationary policies having maximal average return, the method still yields a 1-optimal policy in certain special cases—e.g., where the chains associated with the stationary policies have a common absorbing state and transient states elsewhere. However, there are situations, e.g., example 2 in [1], p. 726, in which the policy improvement method fails to produce a 1-optimal policy.

Blackwell does not give an algorithm that will always find a 1-optimal policy. The principal purpose of this paper is to fill this gap by generalizing the policy improvement method to solve this problem (Theorem 14 below).

2. Review and extension of Blackwell's results. Following Blackwell [1] consider a system which is observed at each of a sequence of points in time labeled $1, 2, \cdots$. At those points the system is observed to be in one of S states labeled $1, 2, \cdots, S$. Each time the system is observed in state s, an action a is chosen from

Received 8 February 1966.

[1] This work was supported by the National Science Foundation under Grant GP-3739.

[2] Actually Blackwell uses the term "nearly optimal" for what we call 1-optimal policies for the case $\beta = 1$. He reserves the term "optimal" for $\beta = 1$ for a policy that is α-optimal (in our sense) for all α, $0 \leqq \alpha \leqq \beta$, sufficiently near β $(= 1)$. He does not consider this latter concept for $0 \leqq \beta < 1$ even though it is also meaningful (see Theorem 2 below) in that case. We have changed his terminology to establish what seems to us to be a more natural relationship between the definitions for the case $0 \leqq \beta < 1$ and $\beta = 1$.