DISCOUNTED DYNAMIC PROGRAMMING

By David Blackwell[1]

University of California, Berkeley

1. Introduction. Soon after the appearance of Wald's work in sequential analysis, Richard Bellman recognized the broad applicability of the methods of sequential analysis, named this body of methods dynamic programming, and applied the methods to many problems (see [1] and papers cited there). The first development of a general theory underlying these methods is due to Karlin [6], and a rather complete analysis of the finite case was given by Howard [5]. Dubins and Savage [3] have recently developed a general theory of gambling; the relation of gambling to dynamic programming is not completely clear, but it is certainly close.

Our formulation of a dynamic programming problem is somewhat narrower than Bellman's. For us, a dynamic programming problem is specified by four objects S, A, q, r, where S, A are any non-empty Borel sets, q associates with each pair $(s, a) \in S \times A$ a probability distribution $q(\cdot \mid s, a)$ on S, and r is a bounded Baire function on $S \times A \times S$. We think of S as the set of possible states of some system, and A as the set of acts available to you. Periodically, say once a day, you observe the current state s of the system, then choose an act $a \in A$. Then the system moves to a new state s' (which will be the state you observe tomorrow), selected according to $q(\cdot \mid s, a)$, and you receive a reward $r(s, a, s')$. Your problem is, given the initial state of the system, to maximize your total expected reward over the infinite future.

This total expected reward may well be infinite, for example, if $r \equiv 1$. Or it may well be undefined. For example, if S has two elements 0, 1, A has only a single element, q is deterministic with $0 \to 1$, $1 \to 0$, and the transition $0 \to 1$ yields \$1, while $1 \to 0$ costs \$1, the series of rewards, starting in state 0, is $1 - 1 + 1 - 1 + \cdots$. We shall avoid this problem by introducing a discount factor β, $0 \leqq \beta < 1$, so that unit reward on the nth day is worth only β^{n-1}, and shall try to maximize the total discounted expected reward.

A *plan* π specifies for each $n \geqq 1$ what act to choose on the nth day as a Borel measurable function of the history $h = (s_1, a_1, \cdots, s_n)$ of the system to date or, more generally, π specifies for each h a probability distribution over A. Associated with each π is a bounded function $I(\pi)$ on S, the total expected discounted reward from π, as a function of the initial state of the system. We shall be especially interested in the (non-randomized) *stationary* plans π. A stationary π is defined by a single function f mapping S into A: whenever the system is in state s, you choose act $f(s)$.

Received 24 September 1964.

[1] Prepared with the partial support of the National Science Foundation, Grant GP-2593.