NEGATIVE DYNAMIC PROGRAMMING[1]

By Ralph E. Strauch

University of California, Berkeley

1. Introduction. A dynamic programming problem is determined by four objects, S, A, q, and r. S and A are non-empty Borel sets, q is a regular conditional probability on S given $S \times A$, and r is a Baire function on $S \times A \times S$. We interpret S as the set of states of some system, and A as the set of actions available at each state. (The set of actions available is assumed to be independent of the state.) When the system is in state s and we take action a, we move to a new state s' selected according to $q(\cdot \mid s, a)$, and we receive a return $r(s, a, s')$. The process is then repeated from the new state s', and we wish to maximize the total expected return over the infinite future.

A *policy* π is a sequence $\pi_1, \pi_2, \cdots$, where π_n is a regular conditional probability on A given $h = (s_1, a_1, \cdots, a_{n-1}, s_n)$, the history of the system up to the nth stage. Given that we have experienced history h up to the nth stage, we choose the nth action according to $\pi_n(\cdot \mid h)$. Certain types of policies are of special interest. A *random semi-Markov policy* is one in which π_n depends only on s_1 and s_n, and a *random Markov policy* is one in which π_n depends only on s_n. A *non-random* policy is one in which each π_n is degenerate, i.e. is a measurable function from histories to actions. A *semi-Markov policy* is a sequence $f_1, f_2, \cdots$, where each f_n is a measurable function from $S \times S$ to A, and $f_n(s_1, s_n)$ is the action we take at the nth stage if we start in state s_1 and the nth state is s_n. A *Markov policy* is a sequence $f_1, f_2, \cdots$ where each f_n is a measurable function from S to A and $f_n(s)$ is the action we choose at the nth stage if the nth state is s. (Notice that the term *Markov policy* and *semi-Markov policy* refer to non-random policies, and are modified by the adjective *random* if the elements of the policies are probability distributions.) A *stationary policy* is a Markov policy in which $f_n = f$ for some measurable f from S to A and all n.

If $\pi = \{f_1, f_2, \cdots\}$ is a Markov policy, the function of g is *π-generated* if there exists a measurable partition $S_1, S_2, \cdots$ of S such that $g = f_n$ on S_n. A Markov policy $\pi' = \{g_1, g_2, \cdots\}$ is *π-generated* if each g_n is *π-generated*.

Associated with each π is a Baire function on S, $I(\pi)(s)$, the total expected return starting from s and using π. This total return may well be infinite, or may be undefined. There are, however, three cases in which the problem is well defined, which may be described as follows:

(a) *The discounted case.* If the return function r is bounded, and we discount our future return with a discount factor β, $0 \leqq \beta < 1$, so that a return of one

Received 17 May 1965.

[1] This paper is the author's doctoral dissertation at the University of California, Berkeley, and was written with the partial support of a National Science Foundation Cooperative Graduate Fellowship, and the Office of Naval Research, contract NONR 222-43.

[2] Now at The RAND Corporation.