*The Annals of Mathematical Statistics*
1969, Vol. 40, No. 2, 366–370

# DISCRETE DYNAMIC PROGRAMMING WITH A SMALL INTEREST RATE[1]

By Bruce L. Miller and Arthur F. Veinott, Jr.

*The RAND Corporation and Stanford University*

**1. Introduction.** In a fundamental paper on stationary finite state and action Markovian decision processes, Blackwell [1] defines an optimal policy to be one that maximizes the expected total discounted rewards for all sufficiently small interest rates $\rho > 0$. He also establishes the existence of a stationary optimal policy by a limit process that does not give a finite algorithm. The purpose of this paper is to prove this result constructively by devising a finite policy improvement method for finding stationary optimal policies. The algorithm is based on the representation of the vector of expected discounted returns under a stationary policy as a Laurent series in the interest rate for all small enough $\rho > 0$.

**2. Preliminaries.** Consider a system which is observed at each of a sequence of points in time labeled $1, 2, \cdots$. At each of these points the system is found to be in one of $S$ states labeled $1, \cdots, S$. Each time the system is observed in state $s$, an action $a$ is chosen from a finite set $A_s$ of possible actions and a reward $r(s, a)$ is received. The conditional probability that the system is observed in state $t$ at time $N + 1$ given that it is found in state $s$ at time $N$, that action $a$ is taken at that time, and given the observed states and actions taken at times $1, 2, \cdots, N - 1$ is assumed to be a function $p(t \mid s, a)$ depending only on $t$, $s$, and $a$.

Let $F = \times_{s=1}^{S} A_s$. A *policy* is a sequence $\pi = (f_1, f_2, \cdots)$ of elements $f_N$ of $F$. Using the policy $\pi$ means that if the system is observed in state $s$ at time $N$, the action chosen at that time is $f_N(s)$, the $s$th component of $f_N$. We write $f^\infty$ for the *stationary policy* $(f, f, \cdots)$ and $(g, f^\infty)$ for the policy $(g, f, f, \cdots)$.

For any $f \varepsilon F$, let $r(f)$ be the $S$ component column vector whose $s$th component is $r(s, f(s))$, and let $P(f)$ be the $S \times S$ Markov matrix whose $st$th element is $p(t \mid s, f(s))$. If $\pi = (f_1, f_2, \cdots)$, let $P^N(\pi) = P(f_1) \cdots P(f_N)$ for $N > 0$ and $P^0(\pi) = I$.

Denote by $\rho > 0$ the rate of interest and let $\beta = (1 + \rho)^{-1}$ be the associated discount factor. If $\rho = \infty$, $\beta \equiv 0$. We suppress the dependence of $\beta$ on $\rho$ in the sequel for simplicity.

The vector of expected total discounted rewards starting from each state and using the policy $\pi$ is

$$V_\rho(\pi) = \sum_{N=0}^{\infty} \beta^N P^N(\pi) r(f_{N+1}).$$

A policy $\pi^*$ is called *$\rho$-optimal* if $V_\rho(\pi^*) \geqq V_\rho(\pi)$ for all $\pi$, and *optimal* if it is $\rho$-optimal for all sufficiently small $\rho > 0$.

---

Received 6 June 1968.

[1] This research was supported by the United States Air Force under project RAND Contract No. F44620-67-C-0045, the National Science Foundation under Grant GK-1420, and the Office of Naval Research under Contract Nonr 225(77) (NR-347-010).