AD619764

COPY 2 OF 3 ae
HARD COPY $. 1.00
MICROFICHE $. 0.50
11 p

ARCHIVE COPY

DDC
RECEIVED
AUG 30 1965
DDC-IRA E

Western Management Science Institute
University of California • Los Angeles

University of California

Los Angeles

Western Management Science Institute

Working Paper No. 81

A MODIFIED DYNAMIC PROGRAMMING METHOD FOR MARKOVIAN DECISION PROBLEMS

by

James B. MacQueen

June 1965

This work was supported by the Western Management Science Institute under a grant from the Ford Foundation, and by the Office of Naval Research under Contract No. 233(75), Task No. 047-041. Reproduction in whole or in part is permitted for any purpose of the United States Government.

A MODIFIED DYNAMIC PROGRAMMING METHOD FOR MARKOVIAN DECISION PROBLEMS[1]

J. MacQueen

University of California, Los Angeles

1. Introduction. Let $X_1, X_2, \ldots$ be a sequence of random variables taking values in a finite set S, and controlled by a decision maker who at each time $t = 1, 2, \ldots,$ observes X_t and then picks an action a belonging to a finite set A; then if $X_t = x$, the probability that $X_{t+1} = y$ becomes $p(y; x, a)$, where p is a known function. Also, choice of action a when $X_t = x$ earns a known amount $g(x, a)$ immediately. Future income is discounted by a constant factor $\alpha < 1$. Thus, if a_t is the action chosen after observing X_t, $t = 1, 2, \ldots,$ the discounted return is defined to be $g(X_1, a_1) + \alpha g(X_2, a_2) + \alpha^2 g(X_3, a_3) + \ldots.$ A policy r is a rule for determining each of the actions a_t as a function of X_t and (possibly) the sequences $X_1, X_2, \ldots X_{t-1}$ and $a_1, a_2, \ldots a_{t-1}$. If the policy r is used and $X_1 = x$, the expected discounted return is given by $u_r(x)$, say, and we are interested in maximizing $u_r(x)$ by an appropriate choice of r. Let $u^*(x) = \sup_r u_r(x)$.

This paper describes a simple algorithm for this problem that is basically an improved version of the standard dynamic programming iterative scheme (see below). Upper and lower bounds on the optimal return are produced by the algorithm at each iteration. These both converge monotonely to the optimal return. Also, the policy determined at each stage achieves a return at least as good as the corresponding lower bound. The sequence of policies produced is actually the same sequence produced by the dynamic programming method; the improvement consists of both better

[1]This research was supported in part by a grant from the Ford Foundation, and in part by the Office of Naval Research (Contract 233(75)).

information about convergence of the sequence of policies, and the fact that as regards computing u^*, the algorithm is apparently much faster. Thus, when the algorithm was applied to the automobile replacement problem described by Howard [5, p. 89], the upper and lower bounds were within 1.3% of u^* after 25 iterations, at which time the optimal policy was reached. The mean of the upper and lower bounds was within .08% of u^* at this point. After 50 iterations the upper and lower bounds were within .05% of u^* and their mean was within .0005% of u^*. The estimate of u^* produced by the standard dynamic programming method was 40.5% below u^* after 25 iterations; in fact, after 160 iterations, this estimate was still below u^* by 1.1%. Both methods require essentially the same computations.[2]

The method of policy iteration required only 9 iterations for the automobile replacement problem. However, while otherwise comparable, each iteration using this method involves the "value determination" operation, which amounts to solving N equations in N unknowns, N being the number of states. Because of this, it is not clear which method is superior from a computational point of view. The proposed method may have an important relative advantage in problems with a large number of states, where the value determination operation presents computational difficulties.

The main properties of the algorithm are described in Theorem 2 of Section 3. A key part of this theorem is based on the very simple but useful relationship contained in Theorem 1 of Section 2. Theorem 1 may be of independent interest.[3] The error bounds provided by parts (i) and (iv)

---

[2]In this comparison, the initial function used by both methods was set at zero, and the percentage errors given are based on the state where this error was maximal using the proposed method.

[3]Theorem 1 derives from some joint work [6] of R. M. Redheffer and the author.

of Theorem 2 can be applied to the policies and estimates of the optimal return produced by other methods.

For further relevant discussion of Markovian decision problems, the reader is referred to papers by d'Epenoux [3], Mann [7], Scarf [8], and Wagner [9].

2. Notation and preliminaries. For dealing with a sequence of real-valued functions on S, $v_1, v_2, \ldots$, it is convenient to associate with each v_n another function r_n on S into A, such that

$$g(x, r_n(x)) + \alpha \Sigma_y v_n(y) p(y; x, r_n(x)) = \max_a [g(x,a) + \alpha \Sigma_y v_n(y) p(y; x, a)],$$

and then define the function g_n by $g_n(x) = g(x, r_n(x))$ and the transformation P_n by $(P_n f)(x) = \Sigma_y f(y) p(y; x, r_n(x))$. In these terms the dynamic programming algorithm is defined by an initial function v_1 and the rule $v_{n+1} = g_n + \alpha P_n v_n$, $n = 1, 2, \ldots$. A function r on S into A is termed a stationary policy. For such a function, define the transformation T_r by

$$(T_r f)(f) = f(x) - g(x, r(x)) - \alpha \Sigma_y f(y) p(y; x, r(x)).$$

The expected return u_r for a stationary policy satisfies the equation $T_r u = 0$.

Now define the transformation T^* by

$$(T^* f)(x) = f(x) - \max_a [g(x, a) + \alpha \Sigma_y f(y) p(y; x, a)].$$

Thus $T^* v_n = v_n - (g_n + \alpha P_n v_n)$.

Using the principle of optimality [2], we can easily convince[4] ourselves that u^* satisfies the equation $T^* u = 0$.

Theorem 1. $T^* u \leq T^* v$ implies $u \leq v$.

---

[4] For rigorous treatment of this and related questions see [1] and [4].

Proof. Translated, the hypothesis $T^*u \le T^*v$ becomes

$$u(x) - v(x) \le \max_a [g(x;a) + \alpha\Sigma_y u(y)p(y;x,a)]$$
$$- \max_a [g(x,a) + \alpha\Sigma_y v(y)p(y;x,a)] \le \max_a \alpha\Sigma_y (u(y)-v(y))p(y;x,a).$$

Suppose the maximum of the left side is $m > 0$. The maximum will be achieved at a point x_0. Replacing $u - v$ with m on the right we get the contradiction,

$$u(x_0) - v(x_0) = m \le \max_a \alpha\Sigma_y \; mp(y;x_0,a) = \alpha m,$$

and the proof is complete.

If there is only one action for each state, T^* is of the same form as T_r. Thus, we have

Corollary 1. $T_r u \le T_r v$ implies $u \le v$.

An immediate application of Theorem 1 is

Corollary 2. The dynamic programming equation $T^*u = 0$ has at most one (finite) solution.

Proof. If $T^*u = T^*v = 0$, then $u \le v$ and $v \le u$ by Theorem 1. Hence, $u = v$.

3. The algorithm. Let v_1 be an arbitrary function with $v_1(s) = 0$ where s is a conveniently selected state, and define the sequence of functions $\{v_n\}$ and the sequences of constants $\{L_n'\}$ and $\{L_n''\}$, by,

$$v_{n+1} = g_n + P_n v_n - (g_n + \alpha\, P_n v_n)(s),$$
$$L_n' = \min_x \, (g_n + \alpha\, P_n v_n - v_n)(x),$$
$$L_n'' = \max_x \, (g_n + \alpha\, P_n v_n - v_n)(x).$$

Notice each function v_n is zero at S. Now let $t = (1-\alpha)^{-1}$, and define the sequence of functions $\{u_n'\}$ and $\{u_n''\}$ by

$$u_n' = v_n + tL_n',$$
$$u_n'' = v_n + tL_n''.$$

Theorem 2. (i) The optimal return u^* satisfies $u_n' \le u^* \le u_n''$.

(ii) $u_n' \leq u_{n+1}'$, $u_n'' \geq u_{n+1}''$. (iii) $u_n' \to u^*$, $u_n'' \to u^*$. (iv) Let u^*_n be the expected discounted return for the stationary policy r_n. Then $u^*_n \geq u_n'$.

<u>Proof</u>. In the following, let $v_n' = g_n + \alpha P_n v_n - L_n'$, so that $v_n' \geq v_n$, and let $v_n'' = g_n + \alpha P_n v_n - L_n''$, so that $v_n'' \leq v_n$. Also, $v_{n+1} = v_n' - v_n'(s) = v_n'' - v_n''(s)$.

(i) $u_n' \leq u^* \leq u_n''$. As was pointed out above, u^* satisfies $T^*u^* = 0$. From the definition of T^* we get,

$$\begin{aligned} T^*u_n' &= v_n + tL_n' - [g_n + \alpha P_n v_n + \alpha tL_n'] \\ &= v_n + tL_n' - [v_n' + L_n' + \alpha tL_n'] \\ &= v_n - v_n' \leq 0 = T^*u^*. \end{aligned}$$

Therefore $u_n' \leq v^*$ by Theorem 1. Similarly,

$$\begin{aligned} T^*u_n'' &= v_n + tL_n'' - [g_n + \alpha P_n v_n + \alpha tL_n''] \\ &= v_n + tL_n'' - [v_n'' + L_n'' + \alpha tL_n''] \\ &= v_n - v_n'' \geq 0 = T^*u^*, \end{aligned}$$

and $u_n'' \geq u^*$ again by Theorem 1.

(ii) $u_n' \leq u_{n+1}'$, $u_n'' \geq u_{n+1}''$. For convenience we use 1 and 2 in place of n and n+1. We have

$$\begin{aligned} u_2' &= v_2 + tL_2' = v_2 + t \min_x [g_2 + \alpha P_2 v_2 - v_2] \\ &\geq v_2 + t \min_x [g_1 + \alpha P_1 v_2 - v_2] \\ &= v_2 + t \min_x [g_1 + \alpha P_1 v_1' - \alpha v_1'(s) - v_1' + v_1'(s)] \\ &\geq v_2 + t \min_x [g_1 + \alpha P_1 v_1 - v_1' + (1-\alpha) v_1'(s)] \\ &= v_2 + tL_1' + v_1(s) = v_1' + tL_1' \geq v_1 + tL_1' = u_1'. \end{aligned}$$

Similarly,

$$\begin{aligned} u_2'' &= v_2 + tL_2'' = v_2 + t \max_x [g_2 + \alpha P_2 v_2 - v_2] \\ &= v_2 + t \max_x [g_2 + \alpha P_2 v_1'' - v_1'' + (1-\alpha) v_1''(s)] \\ &\leq v_2 + t \max_x [g_2 + \alpha P_2 v_1 - v_1'' + (1-\alpha) v_1''(s)] \end{aligned}$$

$$\leq v_2 + t \max_x [g_1 + \alpha P_1 v_1 - v_1'' + (1-\alpha)\, v_1''(s)]$$

$$= v_1'' + t\, L_1'' \leq v_1 + tL_1'' = u_1''.$$

(iii) Convergence of u_n' and u_n'' to u^*. Convergence itself is immediate from the monotonicity and the fact that u^* is an upper bound for u_n' and a lower bound for u_n''. Let $u_\infty = \lim u_n'$. We show that u_∞ satisfies $T^*u = 0$, and hence $u_\infty = u^*$ by Corollary 2. The argument is similar for u_n''. Since $L_n' = u_n'(s)/t \leq u^*(s)/t$, $\lim L_n' = L_\infty$ is finite. Let $\lim_n v_n = v_\infty = u_\infty - L_\infty t$.

First we establish that $v_n'(s) \to 0$; in fact $\Sigma\, v_n'(s)$ converges. Considering the proof of (ii) at the point $x = s$ yields $L_2' \geq L_1' + (1-\alpha)\, v_1'(s)$. Proceeding inductively gives $L_n' \geq L_1' + (1-\alpha) \Sigma_1^{r-1} v_1'(s)$. Since L_n' is bounded and since $v_n'(s) \geq 0$, $\Sigma\, v_n'(s)$ converges. Now, $v_n' = v_{n+1} - v_n'(s) = g_n + \alpha P_n v_n - L_n'$, so we write

$$v_{n+1}(x) - v_n'(s) = \max_a \Big[g(x;a) + \alpha \Sigma_y\, v_\infty(y) p(y;x,a)$$

$$+ \alpha \Sigma_y\, (v_n(y) - v_\infty(y)) p(y;x,a) \Big] - L_n'$$

$$\leq \max_a \Big[g(x,a) + \alpha \Sigma_y\, v_\infty p(y;x,a) \Big] - L_\infty$$

$$+ \max_x \max_a \Big[\alpha \Sigma_y\, (v_n(y) - v_\infty(y)) p(y;x,a) \Big] + L_\infty - L_n'.$$

Taking limits gives

$$v_\infty(x) \leq \max_a [g(x,a) + \alpha\, \Sigma\, v_\infty(y) p(y;x,a)] - L_\infty.$$

With $\min_x \min_a$ replacing $\max_x \max_a$ in the preceding, the inequality is reversed so that we get equality. Substitution of $v_\infty = u_\infty - tL_\infty$ gives $u_\infty(x) = \max_a [g(x,a) + \alpha \sum_{y \in S} u_\infty(y) p(y;x,a)]$, that is, $T^*u_\infty = 0$.

(iv) $u^*_n \geq u_n'$. Define T_{r_n} as indicated in Section 2, by $T_{r_n} f = f - (g_n + \alpha\, P_n f)$. Now, $u^*_n = g_n + \alpha P_n u^*_n$, that is, $T_{r_n} u^*_n = 0$. But $T_{r_n} u_n' \leq 0$ as was seen in this proof of (i). Application of Corollary 1 gives $u_n' \leq u^*_n$. This completes the proof.

BIBLIOGRAPHY

[1] Blackwell, David, "Discounted Dynamic Programming," Ann. Math. Stat. 36, 226-235 (1965).

[2] Bellman, Richard, "Dynamic Programming," Princeton University Press, 1957.

[3] d'Epenoux, F., "A Probabilistic Production and Inventory Problem," Management Science, 10, 98-108 (1963).

[4] Derman, C., "On Sequential Control Processes," Ann. Math. Stat. 35, 341-349 (1964).

[5] Howard, Ronald A., Dynamic Programming and Markov Processes, Wiley, New York (1960).

[6] MacQueen, J. and R. M. Redheffer, "Uniqueness and Monotone Operators in Markov Processes," (Abstract), Ann. Math. Stat. 35, 939 (1964).

[7] Manne, A. S., "Linear Programming and Sequential Decisions," Management Science 6, 259-267 (1960).

[8] Scarf, Herbert E., "A Survey of Analytic Techniques in Inventory Theory," in Multistage Inventory Models and Techniques, edited by Herbert E. Scarf, Dorothy M. Gildord, and Maynard W. Shelly, Stanford University Press, 1963.

[9] Wagner, Harvey M., Michael O'Hagen, and Bertil Lundh, "An Empirical Study of Exactly and Approximately Optimal Inventory Policies," Technical Report No. 5, Institute for Mathematical Studies in the Social Sciences, Stanford University.