The Annals of Statistics
1975, Vol. 3, No. 1, 203-208

A MODIFIED FORM OF THE ITERATIVE METHOD OF DYNAMIC PROGRAMMING[1]

BY ARIE HORDIJK AND HENK TIJMS

Mathematisch Centrum, Amsterdam

This paper considers the discrete time finite state Markovian decision problem with the average return criterion. A modified form of the iterative method of dynamic programming is studied. Under the assumption that the maximal average return is independent of the initial state the asymptotic behaviour of the sequence of functions generated by this modified method is found. It is shown that the modified iterative method supplies both upper and lower bounds on the maximal average return and ε-optimal policies. Moreover, a convergence result is proved for the policies produced by the modified iterative method.

1. Introduction. We are concerned with a dynamic system which at times $t = 1, 2, \cdots$ is observed to be in one of S states labeled $1, \cdots, S$. After observing state i, an action a must be chosen from a finite set $A(i)$ of possible actions. Let X_t and Δ_t, $t = 1, 2, \cdots$, denote the sequences of states and actions. If the system is in state i at time t and action a is chosen, then two things happen:

(i) We receive an immediate (expected) reward $r(i, a)$ and

(ii) $P\{X_{t+1} = j \mid X_1, \Delta_1, \cdots, X_t = i, \Delta_t = a\} = p_{ij}(a)$, where both the rewards $r(i, a)$ and the transition probabilities $p_{ij}(a)$ are assumed to be known.

A policy R for controlling the system is any (possibly randomized) rule which for each t specifies which action to take at time t given the current state X_t and the history $(X_1, \Delta_1, \cdots, X_{t-1}, \Delta_{t-1})$. A stationary policy f is a rule which for each i selects an action $f(i) \in A(i)$ such that action $f(i)$ is always taken whenever the system is in state i. For any stationary policy f, let $r(f)$ be the S component column vector whose ith element is $r(i, f(i))$, and let $P(f)$ be the $S \times S$ Markov matrix whose (i, j) element is $p_{ij}(f(i))$. It is known that the sequence $(n + 1)^{-1} \sum_{k=0}^{n} [P(f)]^k$ converges as $n \to \infty$ to a Markov matrix $P^*(f)$ such that $P^*(f)P(f) = P^*(f)$.

We shall be concerned in this paper with the average return criterion. For any policy R, let

$$\phi(i, R) = \limsup_{n\to\infty} n^{-1} \sum_{t=1}^{n} E_R\{r(X_t, \Delta_t) \mid X_1 = i\} \quad \text{for } i = 1, \cdots, S. \tag{1}$$

Thus $\phi(i, R)$ is the long run average expected return per unit time when the

Received May 1973; revised April 1974.

[1] This paper is registered as Mathematical Centre Report BW 21/73.

AMS 1970 *subject classifications*. 90C40.

Key words and phrases. Markov decision theory, average return, dynamic programming, modified iterative method, convergence results.

initial state is i and policy R is used. Clearly, for any stationary policy f, $\phi(f) = P^*(f)r(f)$, where $\phi(f)$ is the S component column vector whose ith element is $\phi(i, f)$. Let

$$g(i) = \sup_R \phi(i, R) \qquad \text{for} \quad i = 1, \cdots, S.$$

A policy R is called optimal if $\phi(i, R) = g(i)$ for all i. It is known that there is a stationary policy which is optimal (cf. Derman [6]).

Let $\{\alpha_n, n = 1, 2, \cdots\}$ be an arbitrary sequence of finite numbers, and let $y_0(i)$ be an arbitrary function. Define for $n = 1, 2, \cdots$,

$$(2) \qquad y_n(i) = \max_{a \in A(i)} \{r(i, a) + \alpha_n \textstyle\sum_{j=1}^{S} p_{ij}(a) y_{n-1}(j)\} \qquad \text{for} \quad i = 1, \cdots, S.$$

The purpose of this paper is to investigate the iterative method given by (2). If $\alpha_n \equiv 1$, then (2) reduces to the standard iterative method of dynamic programming. For the case $\alpha_n \equiv 1$ the asymptotic behaviour of the sequence $\{y_n(i), n \geqq 0\}$ was studied by Bather [1], Brown [4], Denardo [5], Hordijk and Tijms [8], Lanery [10], and Schweitzer [12]. Bather [1] investigated also the sequence $\{y_n(i)\}$ for the case $\alpha_n = 1 - 1/n$. For a Markov decision model with a finite number of communicating states and a convex decision space, he used this sequence to prove the existence of an optimal policy. Also, under the assumption that for each stationary policy the associated Markov chain $\{X_t\}$ is irreducible, Bather [1] determined the asymptotic behaviour of the sequence $\{y_n(i)\}$. This latter result will be generalized in Section 2. Under the assumption that $g(i) \equiv g$ for some constant g, we shall prove that for each sequence $\{\alpha_n\}$ which satisfies certain conditions a sequence $\{\gamma_n\}$ can be found such that $y_n(i) - \gamma_n g$ has a finite limit as $n \to \infty$ for all i. This limit result was established for the case $\alpha_n \equiv 1$ under the additional assumption that for each optimal stationary policy the associated Markov chain $\{X_t\}$ is aperiodic (see [1], [4], [5], [8], [10], and [12]). Further, we find in Section 2 a result concerning the second term of the Laurent series expansion of the total expected discounted rewards of a stationary policy that is α-optimal for all α close enough to 1 (see [2] and [11]).

In Section 3 we shall show that the iterative method (2) supplies upper and lower bounds on the maximal average return $g(i)$ and, moreover, yields at each iteration a stationary policy whose average return is at least as good as the lower bound found at that iteration. If $g(i)$ is independent of i and if $\{\alpha_n\}$ satisfies certain conditions, then for all n sufficiently large the policy found at the nth iteration is optimal.

As compared with the standard iterative method of dynamic programming the modified method (2) has the advantage that it is insensitive to possible periodicity of the Markov chains $\{X_t\}$ associated with the stationary policies.

2. Asymptotic behaviour of $\{y_n(i)\}$. The discussion in this section will be based on the next assumption.

ASSUMPTION. For some constant g, $g(i) = g$ for $i = 1, \cdots, S$.

This assumption is satisfied if there is an optimal stationary policy such that the associated Markov chain $\{X_t\}$ has a single recurrent class.

Given the sequence $\{\alpha_n\}$, we define the sequence $\{\gamma_n\}$ by

$$\gamma_0 = 0 \quad \text{and} \quad \gamma_n = 1 + \alpha_n \gamma_{n-1} \quad \text{for} \quad n = 1, 2, \cdots . \tag{3}$$

Under certain conditions on $\{\alpha_n\}$ we shall prove that $\lim_{n\to\infty} \{y_n(i) - \gamma_n g\}$ exists and is finite for all i. To do this, let $\hat{r}(i, a) = r(i, a) - g$, and define for $n = 1, 2, \cdots$,

$$\hat{y}_n(i) = \max_{a \in A(i)} \{\hat{r}(i, a) + \alpha_n \textstyle\sum_{j=1}^{S} p_{ij}(a)\hat{y}_{n-1}(j)\} \quad \text{for} \quad i = 1, \cdots, S, \tag{4}$$

where $\hat{y}_0(i) = y_0(i)$ for all i. By induction on n, we obtain from (2) and (4) that

$$\hat{y}_n(i) = y_n(i) - \gamma_n g \quad \text{for} \quad i = 1, \cdots, S \quad \text{and} \quad n = 0, 1, \cdots . \tag{5}$$

Further, for any α with $0 < \alpha < 1$ and any policy R, let

$$\hat{V}_\alpha(i, R) = \textstyle\sum_{t=1}^{\infty} \alpha^{t-1} E_R\{\hat{r}(X_t, \Delta_t) \mid X_1 = i\} \quad \text{and} \quad \hat{V}_\alpha(i) = \sup_R \hat{V}_\alpha(i, R) .$$

That is, given a reward function $\hat{r}(i, a)$ and a discount factor α, $\hat{V}_\alpha(i, R)$ is the total expected discounted return when the initial state is i and policy R is used, while $\hat{V}_\alpha(i)$ is the maximal expected discounted return. It is known that $\hat{V}_\alpha(i)$ is the unique solution to (Blackwell [3] and Derman [6])

$$\hat{V}_\alpha(i) = \max_a \{\hat{r}(i, a) + \alpha \textstyle\sum_{j=1}^{S} p_{ij}(a)\hat{V}_\alpha(j)\} \quad \text{for} \quad i = 1, \cdots, S . \tag{6}$$

Using the fact that there is a stationary policy f such that $\hat{V}_\alpha(i, f) = \hat{V}_\alpha(i)$ for all i and all α close enough to 1 (see Blackwell [2]), the next lemma follows easily from the Laurent series expansion of $\hat{V}_\alpha(i, f)$ for α near 1 (see Miller and Veinott [11], page 367).

LEMMA 1. *There is a number α^*, $0 < \alpha^* < 1$, and a finite constant B such that, for all i,*

$$|\hat{V}_\alpha(i) - \hat{V}_\beta(i)| \leqq |\alpha - \beta| B \quad \text{for all} \quad \alpha, \beta \in (\alpha^*, 1) .$$

The next theorem gives the asymptotic behaviour of the sequence $\{y_n(i)\}$.

THEOREM 1. *Let the sequence $\{\alpha_n, n = 1, 2, \cdots\}$ be such that* (i) $0 < \alpha_n < 1$ *for all* $n \geqq 2$; (ii) $\alpha_n \to 1$ *as* $n \to \infty$; (iii) $\alpha_2\alpha_3 \cdots \alpha_n \to 0$ *as* $n \to \infty$; (iv) $\sum_{j=2}^{n} (a_n \alpha_{n-1} \cdots \alpha_j)|\alpha_j - \alpha_{j-1}| \to 0$ *as* $n \to \infty$. *Then,*

$$\lim_{n\to\infty} \{y_n(i) - \gamma_n g\} \quad \textit{exists and is finite for all} \quad i .$$

PROOF. The proof is a generalization of one given in [7]. For any S component vector x, let $||x|| = \max_i |x_i|$. For any $n \geqq 2$, denote by $\hat{V}_n$ the S component vector whose ith element is $\hat{V}_{\alpha_n}(i)$, and for $n \geqq 1$, let $\hat{y}_n$ be the S component vector whose ith element is $\hat{y}_n(i)$. From (4) and (6) we easily deduce

$$||\hat{y}_n - \hat{V}_n|| \leqq \alpha_n ||\hat{y}_{n-1} - \hat{V}_n|| \quad \text{for} \quad n = 2, 3, \cdots . \tag{7}$$

Since $\alpha_n \to 1$ as $n \to \infty$, we have by Lemma 1 that there is an integer $n_0 \geqq 2$ and a finite constant B such that

$$||\hat{V}_n - \hat{V}_m|| \leqq |\alpha_n - \alpha_m| B \quad \text{for all} \quad n, m \geqq n_0 . \tag{8}$$

Fix now an integer $K \geqq n_0$. By the triangle inequality, we obtain from (7) and (8) that for $n = 1, 2, \cdots$,

$$||\hat{y}_{n+K} - \hat{V}_{n+K}|| \leqq \alpha_{n+K}||\hat{y}_{n+K-1} - \hat{V}_{n+K-1}|| + \alpha_{n+K}|\alpha_{n+K} - \alpha_{n+K-1}|B .$$

Iterating this inequality, we find for $n = 1, 2, \cdots$

$$||\hat{y}_{n+K} - \hat{V}_{n+K}|| \leqq (\alpha_{n+K} \cdots \alpha_{K+1})||\hat{y}_K - \hat{V}_K|| + \sum_{j=K+1}^{n+K} (\alpha_{n+K} \cdots \alpha_j)|\alpha_j - \alpha_{j-1}|B ,$$

from which we get $\lim_{n\to\infty} \{\hat{y}_n(i) - \hat{V}_n(i)\} = 0$ for all i. It follows from relation (8) that $\hat{V}_n(i)$ has a finite limit as $n \to \infty$ for all i, so, by (5), the proof is complete.

REMARK. Using the fact that $n^c - (n-1)^c \leqq 1$ for $n \geqq 1$ and $\sum_{k=1}^{n} k^{-c} - \int_1^n x^{-c}\,dx$ is bounded in n when $0 < c \leqq 1$, it is readily verified that the conditions (i)—(iv) of Theorem 1 are satisfied for any choice $\alpha_n = 1 - n^{-b}$ with $\frac{1}{2} < b \leqq 1$. In case $\alpha_n = 1 - 1/n$ for all n, then $\gamma_n = (n+1)/2$ for $n \geqq 1$.

For the choice $\alpha_n = 1 - 1/n$ Theorem 1 was proved in a different way by Bather [1] under the assumption that for each stationary policy the associate-Markov chain $\{X_t\}$ is irreducible.

In general the sequence $\{y_n(i), n \geqq 0\}$ will diverge. This numerical difficulty can be circumvented as follows (cf. White [14]). Fix some state s. For any $n \geqq 0$, let $v_n(i) = y_n(i) - y_n(s)$ for $i = 1, \cdots, S$, and let $g_n = y_n(s) - \alpha_n y_{n-1}(s)$ for $n \geqq 1$. By (2) we can compute for any $n \geqq 1$ these quantities from

$$g_n = \max_a \{r(s, a) + \alpha_n \sum_{j=1}^{S} p_{sj}(a)v_{n-1}(j)\} ,$$
$$v_n(i) = \max_a \{r(i, a) + \alpha_n \sum_{j=1}^{S} p_{ij}(a)v_{n-1}(j)\} - g_n \qquad \text{for} \quad i = 1, \cdots, S .$$

THEOREM 2. *Suppose that* $\{\alpha_n\}$ *satisfies the conditions* (i)—(iv) *in Theorem* 1. *Then,* g_n *converges to* g *as* $n \to \infty$ *and* $v_n(i)$ *has a finite limit* $v(i)$ *as* $n \to \infty$ *for all* i, *where*

$$g + v(i) = \max_a \{r(i, a) + \sum_{j=1}^{S} p_{ij}(a)v(j)\} \qquad \textit{for} \quad i = 1, \cdots, S .$$

Let the stationary policy f *be* α*-optimal for all* α *close enough to* 1, *and let* $u(i)$ *be the* ith *element of the vector* $H(f)r(f)$ *where* $H(f) = [I - P(f) + P^*(f)]^{-1} - P^*(f)$. *Then, for all* $i = 1, \cdots, S$, $y_n(i) - \gamma_n g$ *converges to* $u(i)$ *as* $n \to \infty$ *and* $v(i) = u(i) - u(s)$.

PROOF. The first assertion is an immediate consequence of Theorem 1. Using the relation $H(f)ge = H(f)P^*(f)r(f) = 0$ (see [2]), and the partial Laurent series expansion given in [2], it follows that $\hat{V}_\alpha(i)$ converges to $u(i)$ as $\alpha \to 1$ for all i. In the proof of Theorem 1 it was shown that $\hat{y}_n(i) - \hat{V}_n(i)$ converges to zero as $n \to \infty$ for all i. The theorem now follows.

REMARK. The bias of any 1-optimal policy (cf. [2], [5], and [13]) is given by the vector whose ith component is $\lim_{n\to\infty} y_n(i) - \gamma_n g$.

3. Bounds on the maximal average return and ε-optimal policies. The next

theorem deals with the question how well the maximal average return and an optimal policy can be approximated with the iterative method (2). The first part of the theorem below involves no assumption about the chain structure of the Markov chains $\{X_t\}$ associated with the stationary policies or about the sequence $\{\alpha_n\}$.

THEOREM 3. *For any $n \geqq 1$, denote by Γ_n the set of the stationary policies f such that $f(i)$ maximizes the right-hand side of* (2) *for all i. Let $L_n = \min_i \{y_n(i) - \alpha_n y_{n-1}(i)\}$, and let $U_n = \max_i \{y_n(i) - \alpha_n y_{n-1}(i)\}$. Then,* (a) *For each $n \geqq 1$, $L_n \leqq \phi(i, f) \leqq g(i) \leqq U_n$ for all $i = 1, \cdots, S$ and $f \in \Gamma_n$.* (b) *If $g(i) = g$ for all i for some constant g and if the sequence $\{\alpha_n\}$ satisfies the conditions* (i)—(iv) *of Theorem* 1, *then both L_n and U_n converge as $n \to \infty$ to g, and, moreover, there is a finite integer N such that for all $n \geqq N$ each policy from Γ_n is optimal.*

PROOF. (a) Let y_n be the S component column vector whose ith element is $y_n(i)$, and let e be the S component column vector of ones. Fix n. Let f be any stationary policy. Then, by (2), $r(f) + \alpha_n P(f) y_{n-1} \leqq y_n$, so

$$r(f) + \alpha_n P(f) y_{n-1} \leqq \alpha_n y_{n-1} + U_n e .$$

Multiplying both sides of this inequality by $P^*(f)$ and using the relations $\phi(f) = P^*(f) r(f)$ and $P^*(f) P(f) = P^*(f)$, we find $\phi(f) \leqq U_n e$. Hence $g(i) \leqq U_n$ for all i, since $g(i) = \max_f \phi(i, f)$ for all i. Choose now $f \in \Gamma_n$. Then, by (2), $r(f) + \alpha_n P(f) y_{n-1} = y_n$, so

$$r(f) + \alpha_n P(f) y_{n-1} \geqq \alpha_n y_{n-1} + L_n e .$$

Multiplying both sides of this inequality by $P^*(f)$, we find $\phi(f) \geqq L_n e$. This completes the proof of (a).

(b) By Theorem 1 and (3) we have that both L_n and U_n converge as $n \to \infty$ to g. Since the number of stationary policies is finite, it follows that a finite integer N exists with the following property: if $f \in \Gamma_m$ for some $m \geqq N$, then $f \in \Gamma_n$ for infinitely many values of n. Let $f \in \Gamma_n$ for some $n \geqq N$. Choose a sequence $\{n_k\}$ with $n_k \to \infty$ as $k \to \infty$ such that $f \in \Gamma_{n_k}$ for all k. By (2) and (3),

$$y_{n_k} - \gamma_{n_k} g e = r(f) + \alpha_{n_k} P(f)[y_{n_k - 1} - \gamma_{n_k - 1} g e] - g e \qquad \text{for all } k .$$

Letting $k \to \infty$ and using Theorem 1, we find $v = r(f) + P(f) v - g e$ for some S component column vector v. Multiplying both sides of the latter equality by $P^*(f)$, we find $P^*(f) r(f) = g e$, so policy f is optimal. This ends the proof.

It follows from Theorem 3 that if $g(i)$ is independent of i, then, by an appropriate choice of α_n, we can determine by (2) for each $\varepsilon > 0$ a stationary policy whose average return differs at most ε from the maximal average return. In contrast with Howard's [9] policy-iteration algorithm the iterative method (2) does not involve the solution of a set of linear simultaneous equations at each iteration.

REFERENCES

[1] BATHER, J. (1973). Optimal decision procedures for finite Markov chains. *Adv. Appl. Prob.* **5** 328-339, 521-540, 541-553.

[2] BLACKWELL, D. (1962). Discrete dynamic programming. *Ann. Math. Statist.* **33** 719-726.

[3] BLACKWELL, D. (1965). Discounted dynamic programming. *Ann. Math. Statist.* **36** 226-235.

[4] BROWN, B. W. (1965). On the iterative method of dynamic programming on a finite space discrete time Markov process. *Ann. Math. Statist.* **36** 1279-1285.

[5] DENARDO, E. V. (1972). A Markov decision problem, pp. 33-68 in T. C. Hu and S. M. Robinson, *Mathematical Programming*, Academic Press, New York.

[6] DERMAN, C. (1970). *Finite State Markovian Decision Processes*. Academic Press, New York.

[7] HORDIJK, A. (1974). On the convergence of the average expected return in dynamic programming. *J. Math. Anal. Appl.* **46** 542-544.

[8] HORDIJK, A. and TIJMS, H. C. (1973). The asymptotic behaviour of the minimal total expected cost in denumerable state dynamic programming and an application in inventory theory. Report BW 17/73. Mathematisch Centrum, Amsterdam.

[9] HOWARD, R. A. (1960). *Dynamic Programming and Markov Processes*. Wiley, New York.

[10] LANERY, E. (1967). Étude asymptotique des systèmes Markoviens a commande. *Rev. Inf. Rech. Op.* **1** No. 5, 3-57.

[11] MILLER, B. L. and VEINOTT, A. F. JR. (1969). Discrete dynamic programming with a small interest rate. *Ann. Math. Statist.* **40** 366-370.

[12] SCHWEITZER, P. J. (1965). Perturbation theory and Markovian decision processes. M.I.T. Operations Research Center Technical Report No. 15.

[13] VEINOTT, A. F. JR. (1969). Discrete dynamic programming with sensitive discount optimality criteria. *Ann. Math. Statist.* **40** 1635-1660.

[14] WHITE D. J. (1963). Dynamic programming, Markov chains, and the method of successive approximations. *J. Math. Anal. Appl.* **6** 373-376.

MATHEMATISCH CENTRUM
2E BOERHAAVESTRAAT 49
AMSTERDAM, NETHERLANDS