BA

stichting
mathematisch
centrum

$\sum$
MC

AFDELING MATHEMATISCHE BESLISKUNDE   BW 21/73   MAY

A. HORDIJK and H.C. TIJMS   BA
A MODIFIED FORM OF THE ITERATIVE METHOD
OF UNDISCOUNTED DYNAMIC PROGRAMMING

2e boerhaavestraat 49 amsterdam

A modified form of the iterative method of undiscounted dynamic programming

by

A. Hordijk   and   H.C. Tijms

ABSTRACT.   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, ... is observed to be in one of S states labeled 1, ..., 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, ..., 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) $P\{X_{t+1} = j \mid X_1, \Delta_1, \ldots, X_t = i, \Delta_t = a\} = p_{ij}(a)$, 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 $(X_1, \Delta_1, \ldots, X_{t-1}, \Delta_{t-1})$. A stationary policy f is a rule which for each i selects an action $f(i) \in A(i)$ such that always action f(i) is 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 × 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 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

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

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

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

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 [5]).

Let $\{\alpha_n,\ n = 1,2,\ldots\}$ be an arbitrary sequence of finite numbers, and

let $y_0(i)$ be an arbitrary function. Define for $n = 1, 2, \ldots,$

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

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

$\{y_n(i),\ n \geq 0\}$ was studied by Bather [1], Brown [4], Hordijk and Tijms [6],

Lanery [9] and Schweitzer [10]. Bather [1] investigated also the sequence

$\{y_n(i)\}$ for the case $\alpha_n = 1 - 1/n$. For a Markov decision model with a fini-

te number of communicating states and a convex decision space, he used this

sequence to prove the existence of an optimal policy. Also, under the as-

sumption that for each stationary policy the associated Markov chain $\{X_t\}$

is irreducible, Bather [1] determined the asymptotic behaviour of the se-

quence $\{y_n(i)\}$. 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 as-

sumption that for each optimal stationary policy the associated Markov

chain $\{X_t\}$ is aperiodic.

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 sufficient 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 insensible to possible periodicity of the Markov chains $\{X_t\}$ associated with the stationary policies. Moreover, it is our conjecture that the convergence of the modified method will in general be faster than that of the standard iterative method.

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, ..., 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

$$(3) \qquad \gamma_0 = 0 \quad \text{and} \quad \gamma_n = 1 + \alpha_n \gamma_{n-1} \qquad \text{for } n = 1, 2, \ldots .$$

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, ...,

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

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

(5) $\hat{y}_n(i) = y_n(i) - \gamma_n g$ for $i = 1,\ldots,S$ and $n = 0,1,\ldots$,

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

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

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 [5])

(6) $\hat{V}_\alpha(i) = \max_a \{\hat{r}(i,a) + \alpha \sum_{j=1}^{N} p_{ij}(a)\hat{V}_\alpha(j)\}$ for $i = 1,\ldots,S$.

We shall need the following lemma.

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

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

PROOF. Choose 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. Blackwell [2] proved that such a policy exists. Moreover, it follows from part (a) of Theorem 4 in [2] that each element of $P^*(f)r(f)$ equals g. See also [5]. Let $\hat{V}_\alpha$ be the S component vector whose ith element is $\hat{V}_\alpha(i,f)$, and let $Q = P(f) - P^*(f)$. Using $[P(f)]^n - P^*(f) = Q^n$ for $n \ge 1$, we find $\hat{V}_\alpha(f) = \sum_{n=0}^{\infty} \alpha^n Q^n r(f) - P^*(f)r(f)$. The series $\sum_0^{\infty} \alpha^n Q^n$ is bounded in $\alpha \in (0,1)$, since this series converges absolutely for $0 < \alpha < 1$ and has a finite limit as $\alpha \to 1$ (see [2] and [11]). It can be directly verified that (see also the resolvent equation (6) in Veinott [11, p.1640])

$$\sum_0^{\infty}(\alpha^n-\beta^n)Q^n r(f) = (\alpha-\beta)Q(\sum_0^{\infty}\alpha^n Q^n)(\sum_0^{\infty}\beta^n Q^n)r(f)$$

for all α, $\beta \in (0,1)$. Now it is readily seen that there is a finite constant B such that $|\hat{V}_\alpha(i,f) - \hat{V}_\beta(i,f)| \leq |\alpha-\beta|B$ for all i and all α, $\beta \in (0,1)$. This completes the proof.

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

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

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

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

(7) $\qquad ||\hat{y}_n - \hat{V}_n|| \leq \alpha_n ||\hat{y}_{n-1} - \hat{V}_n|| \qquad\qquad$ for n = 2,3,... .

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

(8) $\qquad ||\hat{V}_n - \hat{V}_m|| \leq |\alpha_n - \alpha_m|B \qquad\qquad$ for all n, m $\geq n_0$.

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

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

Iterating this inequality, we find for n = 1,2,...

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

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

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 associa- Markov chain $\{X_t\}$ is irreducible.

3. <u>Bounds on the maximal average return and ε-optimal policies.</u> The next theorem deals with the question how good 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 1. *For any $n \ge 1$, 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 $n \ge 1$, $L_n \le \phi(i,f) \le g(i) \le U_n$ for all $i=1,\ldots,S$ 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 $n \ge N$ 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), $r(f) + \alpha_n P(f)y_{n-1} \leq y_n$, so

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

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 $\phi(f) \leq U_n e$. Hence $g(i) \leq U_n$ 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

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

Multiplying both sides of this inequality by $P^*(f)$, we find $\phi(f) \geq L_n e$. 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 $m \geq N$, then $f \in \Gamma_n$ for infinitely many values of n. Let $f \in \Gamma_n$ for some $n \geq N$. 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),

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

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

It follows from Theorem 2 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 [8] policy-iteration algorithm the iterative method (2) does not involve the solution of a set of linear simultaneous equations at each iteration.

REMARK. Let f be any stationary policy such that for some constant ϕ the average expected return $\phi(i,f) = \phi$ for all i. Theorem 2 provides us with a way to compute ϕ without solving a set of S linear simultaneous equations. Choose a sequence $\{\alpha_n\}$ that satisfies the conditions (i)-(iv) of Theorem 1, and choose a function $w_0(i)$. For any $n \geq 1$, define the function w_n by $w_n = r(f) + \alpha_n P(f)w_{n-1}$. Specializing Theorem 2 to the case of a single action in each state, we find

$$\min_i\{w_n(i) - \alpha_n w_{n-1}(i)\} \leq \phi \leq \max_i\{w_n(i) - \alpha_n w_{n-1}(i)\} \quad \text{for all } n \geq 1,$$

As $\{\alpha_n\}$ satisfies the conditions (i)-(iv) of Theorem 1, these bounds have ϕ as limit, and $w = \lim_{n\to\infty}\{w_n - \gamma_n \phi e\}$ satisfies $\phi e + w = r(f) + P(f)w$.

REFERENCES

[1] BATHER, J. (1972). Optimal decision procedures for finite Markov chains. Technical Report No.43. Dept. of Statistics, Stanford Univ. To appear in *Adv. Appl. Prob.*

[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] DERMAN, C. (1970). *Finite State Markovian Decision Processes*. Acade-
      mic Press, New York.

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

[7] HORDIJK, A. (1973). On the convergence of the average expected re-
      turn in dynamic programming. Report BW 20/73. Mathematisch Centrum,
      Amsterdam.

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

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

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

[11] VEINOTT, A.F., Jr. (1969). Discrete dynamic programming with sensi-
      tive discount optimality criteria. *Ann. Math. Statist.* 40
      1635-1660.