stichting
mathematisch
centrum

MC

AFDELING MATHEMATISCHE BESLISKUNDE BW 19/73 APRIL

A. HORDIJK and H.C. TIJMS

THE METHOD OF SUCCESSIVE APPROXIMATIONS AND MARKOVIAN DECISION PROBLEMS

BA

2e boerhaavestraat 49 amsterdam

Printed at the Mathematical Centre, 49, 2e Boerhaavestraat, Amsterdam.

The Mathematical Centre, founded the 11-th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications. It is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O), by the Municipality of Amsterdam, by the University of Amsterdam, by the Free University at Amsterdam, and by industries.

AMS (MOS) subject classification scheme (1970): 90C40

ABSTRACT. This paper is concerned with the finite state, discrete-time Markovian decision model. For the average return criterion it is shown that the method of successive approximations produces monotonic upper and lower bounds on the maximal average return. Also, it yields at each iteration a stationary policy whose average return is at least as good as the lower bound found at that iteration. These results are proved without making any assumption about the chain structure of the Markov chains associated with the stationary policies. Moreover, we extend MACQUEEN's results for the discounted reward model and we establish some relations between the finite time horizon model with a discount factor near 1 and the one with no discounting.

We are concerned with a dynamic system which at times t=1,2,... is observed in one of a finite number of states i, labeled by the integers 1,...,N. After observing state i, an action a must be chosen from a finite set A(i) of possible actions. If we choose at time t action a in state i, then we receive an (expected) reward $r(i,a)$, and at time t+1 the system will be in state j with probability $p_{ij}(a)$.

Let $\{X_t\}$ and $\{\Delta_t\}$ denote the sequences of states and actions. 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, to be denoted by f, is a policy which prescribes in each state i a single decision $f(i) \in A(i)$ whenever the system is in state i.

There is a considerable literature on this subject and much of the literature was stimulated by the basic work of HOWARD [5]. In his book, Howard gives both for the total expected return criterion and for the average return criterion a policy-iteration algorithm which leads after a finite number of iterations to an optimal policy. The policy-iteration algorithm has the drawback that each iteration involves the solution of a system of linear simultaneous equations whose order is the same as the number of states. In the literature there are published several alternatives which avoid this drawback [8, 9, 10, 12, 13]. An alternative which is very attractive from a computational point of view is the (modified) method of successive approximations. However, this method need not converge in a finite number of iterations to an optimal policy. This gives rise to the question how good an optimal policy and the optimal value of the criterion function can be approximated with this method. MACQUEEN [8] has investigated this question

for the discounted reward model. He has shown that monotonic upper and lower bounds on the maximal total expected discounted return are produced by the method of successive approximations at each iteration. Moreover, he has proved that the policy determined at each step achieves a total expected discounted return at least as good as the lower bound found at that step. In this paper we shall prove corresponding results for the undiscounted model with the long-run average return as criterion. It is important to note that we have not to make *any* assumption about the chain structure of the Markov chains $\{X_t\}$ associated with the stationary policies. The special case where for each stationary policy the associated Markov chain has a single ergodic class and has no periodic states was studied by ODONI [10] who obtained only monotonic upper and lower bounds on the maximal average return. Other related work was done by SCHWEITZER [12] who investigated for undiscounted single chain Markov renewal programming an algorithm in which each policy evaluation is followed by repeated policy improvements.

We shall give two approaches to analyze the undiscounted model. The first one treats the undiscounted model as a limiting case of the discounted model. In this approach we extend MACQUEEN's results and we prove some relations between the finite time horizon model with a discount factor near 1 and the one with no discounting. These results are also of interest in itself. The second approach is a direct one which does not use any result for the discounted model.

DEFINITIONS AND PRELIMINARIES

Let $V_0(i)$, $1 \le i \le N$, be an arbitrary function. For any α with $0 < \alpha < 1$, we define for n=0,1,...,

$$V_{n+1}(i,\alpha) = \max_{a\in A(i)}\{r(i,a) + \alpha\Sigma_{j=1}^{N} p_{ij}(a)V_n(j,\alpha)\}, \qquad (1\le i\le N), \quad (1)$$

where
$$V_0(i,\alpha) = V_0(i), \qquad (1\le i\le N).$$

If we interprete $V_0(j)$ as the terminal reward received when in the finite period problem the final state is j, then $V_n(i,\alpha)$ denotes the maximal expected discounted return over the periods 1,...,n when there is a discount factor α and the initial state is i. For n=0,1,..., let

$$V_{n+1}(i) = \max_{a\in A(i)}\{r(i,a) + \Sigma_{j=1}^{N} p_{ij}(a)V_n(j)\}, \qquad (1\le i\le N), \quad (2)$$

that is, $V_n(i)$ is the maximal expected return for the n-period problem when the rewards are not discounted.

Let $0 < \alpha < 1$, and let $\varepsilon > 0$. For any n=0,1,..., we introduce the following sets of stationary policies

$$F_n(\alpha,\varepsilon) = \{f \mid r(i,f(i)) + \alpha\Sigma_{j=1}^{N} p_{ij}(f(i))V_n(j,\alpha) \ge V_{n+1}(i,\alpha)-\varepsilon,\ 1\le i\le N\},$$

$$F_n(\alpha) = \{f \mid r(i,f(i)) + \alpha\Sigma_{j=1}^{N} p_{ij}(f(i))V_n(j,\alpha) = V_{n+1}(i,\alpha),\ 1\le i\le N\},$$

$$F_n = \{f \mid r(i,f(i)) + \Sigma_{j=1}^{N} p_{ij}(f(i))V_n(j) = V_{n+1}(i),\ 1\le i\le N\}.$$

For any $0 < \alpha < 1$ and any policy R, let

$$V_\alpha(i,R) = \Sigma_{t=1}^{\infty} \alpha^{t-1}E_R\{r(X_t,\Delta_t) \mid X_1 = i\}, \qquad (1\le i\le N)$$

and let,

$$g(i,R) = \lim_{n\to\infty} \sup (1/n)\Sigma_{t=1}^{n} E_R\{r(X_t,\Delta_t)|X_1 = i\}, \qquad (1\le i\le N),$$

where E_R denotes the expectation under policy R. That is, $V_\alpha(i,R)$ is the total expected discounted return for the infinite period problem when policy R is used and the initial state is i, and $g(i,R)$ is the long-run average return per unit time for policy R and initial state i. For any stationary policy f, we have [2,3]

$$\lim_{\alpha\to 1} (1-\alpha)V_\alpha(i,f) = g(i,f), \qquad (1\le i\le N). \quad (3)$$

Let

$$V_\alpha(i) = \sup_R V_\alpha(i,R) \text{ and } g(i) = \sup_R g(i,R), \qquad (1\le i\le N).$$

That is, for the infinite period problem, $V_\alpha(i)$ is the maximal total expected discounted return, and $g(i)$ is the maximal average return per unit time. Both $V_\alpha(i)$ and $g(i)$ are achieved by a stationary policy [2,3]. It is known that $V_n(i,\alpha)$ converges to $V_\alpha(i)$ as $n\to\infty$ for all i [3]. Moreover, $V_\alpha(i)$ is the unique solution to [2,3]

$$V_\alpha(i) = \max_{a\in A(i)}\{r(i,a) + \alpha\Sigma_{j=1}^{N} p_{ij}(a)V_\alpha(j)\}, \qquad (1\le i\le N). \quad (4)$$

It is known that there is a stationary policy f^* such that $V_\alpha(i,f^*) = V_\alpha(i)$, $1\le i\le N$, for all α near enough to 1 [2,3]. Moreover, $g(i,f^*) = g(i)$ for all i [2,3]. Together these facts and (3) imply

$$\lim_{\alpha\to 1} (1-\alpha)V_\alpha(i) = g(i), \qquad (1\le i\le N). \quad (5)$$

OPTIMALITY RELATIONS IN THE FINITE PERIOD MODEL.

We shall now prove that $V_n(i,\alpha)$ converges to $V_n(i)$ as $\alpha\to 1$ and further we establish inclusion relations between the sets $F_n(\alpha,\varepsilon)$, $F_n(\alpha)$ and F_n.

THEOREM 1. *For any* $n \geq 0$, $\lim_{\alpha\to 1} V_n(i,\alpha) = V_n(i)$, $1 \leq i \leq N$.

Proof. The proof is by induction on n. Since $V_0(i,\alpha) \equiv V_0(i)$ the theorem is true for n = 0. Assuming that the theorem has been proved for n = m, we shall show that $V_{m+1}(i,\alpha)$ converges to $V_{m+1}(i)$ as $\alpha\to 1$ for all i. To do this, we fix i and we observe that $V_{m+1}(i,\alpha)$ is bounded by (m+2)B for all α, where the constant B is such that r(j,a) and $V_0(j)$ are bounded by B. Hence it suffices to prove that $V_{m+1}(i,\alpha_k)$ converges to $V_{m+1}(i)$ as $k\to\infty$ for any sequence $\{\alpha_k,\ k\geq 1\}$ such that $\alpha_k \to 1$ as $k\to\infty$ and $V_{m+1}(i,\alpha_k)$ has a limit as $k\to\infty$. Let $\{\alpha_k\}$ be such a sequence. Since A(i) is finite, it follows from (1) that we can choose an action $a^* \in A(i)$ and a subsequence $\{\alpha_k'\}$ of $\{\alpha_k\}$ such that for all k,

$$V_{m+1}(i,\alpha_k') \geq r(i,a) + \alpha_k' \Sigma_{j=1}^{N} p_{ij}(a) V_m(j,\alpha_k') \qquad \text{for all } a \in A(i),$$

with equality for $a = a^*$. Letting $k\to\infty$ and using the induction hypothesis, we get

$$\lim_{k\to\infty} V_{m+1}(i,\alpha_k) \geq r(i,a) + \Sigma_{j=1}^{N} p_{ij}(a) V_m(j) \qquad \text{for all } a \in A(i),$$

with equality for $a = a^*$. Hence, by (2), $\lim_{k\to\infty} V_{m+1}(i,\alpha_k) = V_{m+1}(i)$, which proves the theorem.

THEOREM 2.

(a) *For any* $n \geq 0$ *and any* $\varepsilon > 0$, *there is a number* $\alpha_n(\varepsilon)$ *such that* $F_n \subseteq F_n(\alpha,\varepsilon)$ *for all* $\alpha_n(\varepsilon) \leq \alpha < 1$.

(b) *For any* $n \geq 0$, *there is a number* α_n *such that* $F_n \supseteq F_n(\alpha)$ *for all* $\alpha_n \leq \alpha < 1$.

Proof.

(a) Fix n and fix $\varepsilon > 0$. Assume to the contrary that there is a sequence $\{\alpha_k\}$ such that $\alpha_k \to 1$ as $k \to \infty$ and, for each k, $F_n \setminus F_n(\alpha_k, \varepsilon)$ is not empty. Since both the number of stationary policies and the number of states are finite, we can choose a stationary policy f^*, a state s and a subsequence $\{\alpha_k'\}$ of $\{\alpha_k\}$ such that, for all k,

$$r(s,f^*(s)) + \alpha_k' \Sigma_j\, p_{sj}(f^*(s))V_n(j,\alpha_k') < V_{n+1}(s,\alpha_k') - \varepsilon$$

and

$$r(s,f^*(s)) + \Sigma_j\, p_{sj}(f^*(s))V_n(j) = V_{n+1}(s).$$

Letting $k \to \infty$ and using theorem 1, we obtain a contradiction. This proves (a).

(b) The proof of (b) is very similar to that of (a) and is omitted.

Remark. The following example shows that $F_n \subseteq F_n(\alpha)$ for α near 1 need not hold. There are two states 1 and 2. In state 1 two actions a_1 and a_2 are possible with $p_{11}(a_1) = p_{12}(a_2) = 1$. Let $r(1,a_1) = 0$ and $r(1,a_2) = 1$, and let $V_0(1) = 1$ and $V_0(2) = 0$. Then, $r(1,a_1) + V_0(1) = r(1,a_2) + V_0(2)$ and $r(1,a_1) + \alpha V_0(1) < r(1,a_2) + \alpha V_0(2)$ for $0 < \alpha < 1$.

SUCCESSIVE APPROXIMATIONS AND THE DISCOUNTED MODEL.

We shall now extend MACQUEEN's results [8]. For given α with $0 < \alpha < 1$, the following transformations are introduced. Let $v(i)$, $1 \leq i \leq N$, be an arbitrary function, then the function $Tv(i)$ is defined by

$$Tv(i) = v(i) - [\max_{a \in A(i)} \{r(i,a) + \alpha \Sigma_{j=1}^{N}\, p_{ij}(a)v(j)\}], \quad (1 \leq i \leq N),$$

and, for any stationary policy f, the function $T_f v(i)$ is defined by

$$T_f v(i) = v(i) - [r(i,f(i)) + \alpha \Sigma_{j=1}^{N} p_{ij}(f(i))v(j)], \qquad (1 \le i \le N).$$

The following theorem has been proved by Macqueen [8].

THEOREM 3.

(a) $Tu(i) \le Tv(i)$ *for* $1 \le i \le N$ *implies* $u(i) \le v(i)$ *for* $1 \le i \le N$.

(b) $T_f u(i) \le T_f v(i)$ *for* $1 \le i \le N$ *implies* $u(i) \le v(i)$ *for* $1 \le i \le N$.

For any $0 < \alpha < 1$ and any $\varepsilon > 0$, we define for $i=1,\ldots,N$ and $n=0,1\ldots,$

$$u_n'(i,\alpha,\varepsilon) = V_n(i,\alpha) + (1-\alpha)^{-1} \min_{1 \le j \le N}\{V_{n+1}(j,\alpha) - V_n(j,\alpha)\} - \varepsilon(1-\alpha)^{-1},$$

$$u_n''(i,\alpha) = V_n(i,\alpha) + (1-\alpha)^{-1} \max_{1 \le j \le N}\{V_{n+1}(j,\alpha) - V_n(j,\alpha)\}.$$

We note that the function $V_n(i)$ and the function v_n defined on p. 41 of reference 8 are related by $v_n(i) = V_n(i) - V_n(s)$ for some fixed state s, which implies that the functions $u_n'(i,\alpha,\varepsilon)$ and $u_n''(i,\alpha)$ are related to the functions u_n' and u_n'' introduced by Macqueen [8] by $u_n'(i,\alpha,\varepsilon) = u_n'(i) - \varepsilon(1-\alpha)^{-1}$ and $u_n''(i,\alpha) = u_n''(i)$. The proof of the next theorem follows that of Macqueen's theorem 2 [8].

THEOREM 4. *For any* $0 < \alpha < 1$ *and any* $\varepsilon > 0$ *holds,*

(a) *For any* $n \ge 0$, $u_n'(i,\alpha,\varepsilon) \le V_\alpha(i,f) \le V_\alpha(i) \le u_n''(i,\alpha)$ for all $1 \le i \le N$ and all $f \in F_n(\alpha,\varepsilon)$.

(b) $u_n'(i,\alpha,\varepsilon)$ *is nondecreasing in* n *and* $u_n''(i,\alpha)$ *is nonincreasing in* n.

Proof.

(a) Let us first observe that, by (4), $TV_\alpha(i) = 0$ for $1 \le i \le N$. Further, for any stationary policy f, $V_\alpha(i,f) = r(i,f(i)) + \alpha \Sigma_j p_{ij}(f(i))V_\alpha(j,f)$, so $T_f V_\alpha(i,f) = 0$ for $1 \le i \le N$. Fix now n. Put for abbreviation

$w_n'(\alpha) = \min_j\{V_{n+1}(j,\alpha) - V_n(j,\alpha)\}$. Let $f\epsilon F_n(\alpha,\varepsilon)$. Then, by using the definition of $F_n(\alpha,\varepsilon)$, for each i,

$$T_f u_n'(i,\alpha,\varepsilon) = V_n(i,\alpha) + (1-\alpha)^{-1}w_n'(\alpha) - \varepsilon(1-\alpha)^{-1} - [r(i,f(i)) +$$

$$+ \alpha\Sigma_j p_{ij}(f(i))V_n(j,\alpha) + \alpha(1-\alpha)^{-1}w_n'(\alpha) - \alpha\varepsilon(1-\alpha)^{-1}] \leq$$

$$\leq V_n(i,\alpha) + w_n'(\alpha) - \varepsilon - V_{n+1}(i,\alpha) + \varepsilon \leq 0 = T_f V_\alpha(i,f).$$

Hence, by theorem 3(b), $u_n'(i,\alpha,\varepsilon) \leq V_\alpha(i,f)$ for all i. Similarly, by using (1), we find that $Tu_n''(i,\alpha) = V_n(i,\alpha) - V_{n+1}(i,\alpha) + \max_j\{V_{n+1}(j,\alpha) - V_n(j,\alpha)\} \geq$ $\geq 0 = TV_\alpha(i)$ for all i. Hence, by theorem 3(a), $V_\alpha(i) \leq u_n''(i,\alpha)$ for all i. This completes the proof of (a).

(b) This assertion trivially follows from theorem 2(ii) in reference 8, since $u_n'(i,\alpha,\varepsilon) = u_n'(i) - \varepsilon(1-\alpha)^{-1}$ and $u_n''(i,\alpha) = u_n''(i)$, where u_n' and u_n'' come from reference 8.

We note that $u_n'(i,\alpha,\varepsilon)$ and $u_n''(i,\alpha)$ converge to $V_\alpha(i) - \varepsilon(1-\alpha)^{-1}$ and $V_\alpha(i)$, respectively, as $n\to\infty$, since $V_n(i,\alpha)$ converges to $V_\alpha(i)$ as $n\to\infty$. Further $F_n(\alpha) \subseteq F_n(\alpha,\varepsilon)$ for all $\varepsilon>0$. From these facts and theorem 4 we obtain Macqueen's theorem 2 [8] as a corollary.

SUCCESSIVE APPROXIMATIONS AND THE UNDISCOUNTED MODEL.

We are now in a position to prove for the undiscounted model that the method of successive approximations produces monotonic upper and lower bounds on the maximal average return 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. These results will be proved without making *any* assumption about the chain structure of the Markov chains $\{X_t\}$ associated with the stationary policies.

THEOREM 5. *For any* $1 \le i \le N$ *and any* $n \ge 0$, *let* $u_n^* = \min_j\{V_{n+1}(j) - V_n(j)\}$, *and let* $u_n^{**} = \max_j\{V_{n+1}(j) - V_n(j)\}$. *Then,*

(a) For each n, $u_n^* \le g(i,f) \le g(i) \le u_n^{**}$ *for all* $i=1,\ldots,N$ *and all* $f \in F_n$.

(b) u_n^* *is nondecreasing in* n *and* u_n^{**} *is nonincreasing in* n.

Proof.

(a) Fix n and fix $\varepsilon>0$. Since $V_n(i,\alpha)$ is bounded by $(n+1)B$ for some finite constant B, we have by theorem 1 that, for $i=1,\ldots,N$,

$$\lim_{\alpha\to 1}(1-\alpha)u_n'(i,\alpha,\varepsilon) = u_n^* - \varepsilon \text{ and } \lim_{\alpha\to 1}(1-\alpha)u_n''(i,\alpha) = u_n^{**}. \quad (6)$$

By theorem 2(a), there is a number $\alpha_n(\varepsilon)$ such that $F_n \subseteq F_n(\alpha,\varepsilon)$ for all $\alpha_n(\varepsilon) \le \alpha < 1$. Let $f \in F_n$. Then, by theorem 4(a), for all $\alpha_n(\varepsilon) \le \alpha < 1$,

$$(1-\alpha)u_n'(i,\alpha,\varepsilon) \le (1-\alpha)V_\alpha(i,f) \le (1-\alpha)V_\alpha(i) \le (1-\alpha)u_n''(i,\alpha), \ (1 \le i \le N).$$

Letting $\alpha \to 1$ and using (3), (5) and (6), we get $u_n^* - \varepsilon \le g(i,f) \le g(i) \le u_n^{**}$ for all i. This proves (a), since ε was chosen arbitrarily.

(b) This part is an immediate consequence of theorem 4(b) and (6).

Theorem 5 can also be directly proved without using any result for the discounted model.

Alternative proof of theorem 5.

(a) Let us first note that $g(i) = \max_f g(i,f)$ for all i, since there is a stationary policy f^* such that $g(i,f^*) = g(i)$ for all i [2,3]. For any stationary policy f, denote by P(f) the $N \times N$ matrix whose (i,j) element is $p_{ij}(f(i))$. It is known that the sequence $\{n^{-1}\sum_{k=0}^{n-1}[P(f)]^k\}$ converges to a stochastic matrix $P^*(f)$ such that $P^*(f)P(f) = P^*(f)$ [6]. Denote by $p_{ij}^*(f)$ the (i,j) element of $P^*(f)$. Clearly, for any stationary policy f,

$$g(i,f) = \Sigma_{j=1}^{N} p_{ij}^{*}(f)r(j,f(j)), \qquad (1\leq i\leq N). \quad (7)$$

Fix n. Let f be any stationary policy. Since $V_{n+1}(i) - V_n(i) \leq u_n^{**}$ for all i, we have by (2),

$$r(i,f(i)) + \Sigma_j\, p_{ij}(f(i))V_n(j) \leq V_{n+1}(i) \leq V_n(i) + u_n^{**}, \quad (1\leq i\leq N).$$

Multiplying the extreme sides of this inequality by $p_{ki}^{*}(f)$, summing over i, and using (7) and $P^{*}(f)P(f) = P^{*}(f)$, we find $g(k,f) \leq u_n^{**}$ for $1\leq k\leq N$. Hence $g(i) \leq u_n^{**}$ for all i, since f was chosen arbitrarily. Choose now $f\epsilon F_n$. Since $V_{n+1}(i) - V_n(i) \geq u_n^{*}$ for all i, it then follows from (2) that

$$r(i,f(i)) + \Sigma_j\, p_{ij}(f(i))V_n(j) = V_{n+1}(i) \geq V_n(i) + u_n^{*}, \quad (1\leq i\leq N).$$

Multiplying the extreme sides of this relation by $p_{ki}^{*}(f)$ and summing over i, we find $g(k,f) \geq u_n^{*}$ for all k. This completes the proof of (a).

(b) A direct proof of this assertion can be found in reference 10.

We note that theorem 5(a) implies $g(i,f) \geq g(i) - (u_n^{**}-u_n^{*})$ for all $f\epsilon F_n$ and all i; this bound can also be deduced from theorem 6.1 in part II of reference 1. SCHWEITZER [12] has proposed for the undiscounted single chain Markov renewal program an algorithm in which each value-determination step is followed by repeated policy improvements. The bounds in Schweitzer's relations (13) and (14) with N = 1 for the unmodified policy improvement procedure can also be easily derived from theorem 5.

Specializing theorem 5 to the case of a single action in each state, we obtain the following corollary

COROLLARY. *Let* f *be any stationary policy, and let* $y_0(i)$, $1\leq i\leq N$, *be an arbitrary function. For* n=0,1,..., *we define the functions* $y_{n+1}(i)$ *by*

$$y_{n+1}(i) = r(i,f(i)) + \Sigma_{j=1}^{N} p_{ij}(f(i))y_n(j), \qquad (1 \le i \le N).$$

Then, $\min_j\{y_{n+1}(j) - y_n(j)\} \le g(i,f) \le \max_j\{y_{n+1}(j) - y_n(j)\}$ *for all* i *and all* n, *where the lower (upper) bound is nondecreasing (nonincreasing).*

This corollary may be helpful with regard to the procedure followed by MORTON to solve the systems of linear equations (3) and (4) in his paper [9].

Remark 1. Consider the special case where for every average-return optimal stationary policy the associated Markov chain $\{X_t\}$ has a single ergodic class and has no periodic states. Then, g(i) is constant (say g) and $\lim_{n\to\infty}\{V_n(i)-ng\}$ exists and is finite for all i [4,7,11]. Together this and theorem 5(b) imply that u_n^* is nondecreasing to g and u_n^{**} is nonincreasing to g. This result has been also found by ODONI [10]. The following example shows that u_n^* and u_n^{**} need not converge to g when there is an average return optimal policy with periodic states. There are two states 1 and 2. In each state there is a single action a_0. Let $p_{12}(a_0) = p_{21}(a_0) = 1$, and let $r(1,a_0) = 0$ and $r(2,a_0) = 1$. Choose $V_0(i) = 0$, i=1,2. Then $u_n^* = 0$ and $u_n^{**} = 1$ for all n, and $g(i) = \frac{1}{2}$ for i=1,2.

Remark 2. Since the sequence $\{V_n(i), n\ge1\}$ in general will not be bounded, it may be inconvenient to compute u_n^*, u_n^{**} and F_n by the recurrence relation (2). In case g(i) is constant (say g) this difficulty is avoided by applying WHITE's modified method of successive approximations [10,13], since in that case the sequence $\{V_n(i)-ng\}$ is bounded for all i.

REFERENCES

1. J. BATHER, "Optimal Decision Procedures for Finite Markov Chains", Technical Report No. 43, Department of Statistics, Stanford Univ., 1972 (to appear in Adv. Appl. Prob.).
2. D. BLACKWELL, "Discrete Dynamic Programming", *Ann. Math. Stat.* 33, 719-726 (1962).
3. C. DERMAN, *Finite State Markovian Decision Processes*, Academic Press, New York, 1970.
4. A. HORDIJK and H.C. TIJMS, "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, 1973.
5. R.A. HOWARD, *Dynamic Programming and Markov Processes*, The M.I.T. Press, Cambridge, 1960.
6. J.G. KEMENY and J.L. SNELL, *Finite Markov Chains*, Van Nostrand, New York, 1960.
7. E. LANERY, "Etude Asymptotic des Systèmes Markoviens a Commande, *R.I.R.O.* 1, No. 5, 3-57 (1967).
8. J.B. MACQUEEN, "A Modified Dynamic Programming Method for Markovian Decision Problems", *J. Math. Anal. and Appl.* 14, 38-43 (1966).
9. T.E. MORTON, "Undiscounted Markov Renewal Programming via Modified Successive Approximations", *Opns. Res.* 19, 1081-1089 (1971).
10. A.R. ODONI, "On Finding the Maximal Gain for Markov Decision Processes", *Opns. Res.* 17, 857-860 (1969).
11. P.J. SCHWEITZER, "Perturbation Theory and Markovian Decision Processes", M.I.T. Operations Research Center Technical Report No. 15, 1965.
12. P.J. SCHWEITZER, "Multiple Policy Improvements in Undiscounted Markov Renewal Programming", *Opns. Res.* 19, 784-793 (1971).
13. D.J. WHITE, "Dynamic Programming, Markov Chains, and the Method of Successive Approximations", *J. Math. Anal. and Appl.* 6, 373-376 (1963).