stichting
mathematisch
centrum

MC

AFDELING MATHEMATISCHE BESLISKUNDE BN 18/73 JULY

A. HORDIJK and H.C. TIJMS

A NOTE ON HOWARD'S VALUE DETERMINATION STEP

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

1. INTRODUCTION

Let P be an N×N Markov matrix whose (i,j) element is p_{ij} (i,j=1,...,N), i.e., $p_{ij} \geq 0$ and $\sum_j p_{ij}=1$. Let T be an N component column vector whose ith element is T_i where $T_i > 0$ for i=1,...,N, and let q be an N component column vector whose ith element is q_i (i=1,...,N). The triple (P,T,q) can be thought of as a semi-Markov reward process with transition probabilities p_{ij}, expected transition times T_i and one-transition rewards q_i. It is assumed that the Markov matrix P has a single recurrent chain. Let state N be a recurrent state of the Markov matrix P.

In each iteration of Howard's [2] well known policy-iteration algorithm a set of linear simultaneous equations must be solved. For the single chain case this set of equations is of the following type:

$$gT + v = q + Pv \ , \tag{1}$$

where g is an unknown scalar and v is an unknown N component column vector whose ith element is v_i (i=1,...,N). It is important to have an efficient method for solving (1). For the case where P is an aperiodic Markov matrix Morton [4] has given a simple iterative scheme to solve (1).

The purpose of this note is to demonstrate that a solution of (1) can be found by solving two sets of linear simultaneous equations which are more easy to tackle than (1). In our approach we need not require that P is aperiodic. Despite the fact that our approach is implied in the paper of Derman and Veinott[1], the theorem below seems to have passed unnoticed.

2. RESULTS

We first introduce some notation. Let T^* be the N-1 component column vector whose ith element is T_i, let q^* be the N-1 component column vector whose ith element is q_i, and let R be the N-1 component row vector whose ith element is p_{Ni} $(i=1,\ldots,N-1)$. Denote by Q the $(N-1)\times(N-1)$ matrix whose (i,j) element is p_{ij} $(i,j=1,\ldots,N-1)$. Observe that $Q^n \to 0$ as $n \to \infty$, since N is a recurrent state of the Markov matrix P.

We have the following theorem (cf. Derman and Veinott [1] and Theorem 1 of Morton [4])

THEOREM. *Let the column vector $x=(x_1,\ldots,x_{N-1})$ be the unique solution to*

$$x = q^* + Qx \ , \tag{2}$$

and let the column vector $y=(y_1,\ldots,y_{N-1})$ be the unique solution to

$$y = T^* + Qy \ . \tag{3}$$

Define the scalar g by

$$g=(q_N+Rx)/(T_N+Ry) \ , \tag{4}$$

and define the N component column vector $v=(v_1,\ldots,v_N)$ by

$$v_i = x_i - gy_i \quad \text{for } i=1,\ldots,N-1 \ , \qquad v_N = 0 \ . \tag{5}$$

Then g,v satisfy equation (1).

Proof. Let us first observe that both (2) and (3) have a unique solution, since $Q^n \to 0$ as $n \to \infty$. Denote by v^* the N-1 component column vector whose ith element is v_i (i=1,...,N-1). From (2), (3) and (5),

$$gT^* + v^* = gT^* + q^* + Qx-g(T^*+Qy) = q^* + Q(x-gy) = q^* + Qv^* ,$$

while from (4) and (5) it follows that

$$gT_N + v_N = q_N + Rx - gRy = q_N + R(x-gy) = q_N + Rv^* .$$

Using $v_N = 0$ the theorem now follows.

Observe that g in (4) can be interpreted as the ratio of the expected return earned during a cycle and the expected length of a cycle, where a cycle is defined as the time interval between two successive visits to the recurrent state N. It is well-known that this ratio equals the long-run average return.

Remark. Suppose that $p_{iN}=1-\alpha_i > 0$ for i=1,...,N-1. Let z_0 be an arbitrary N-1 component column vector, and for $n\geq 1$ define z_n by $z_n = b + Qz_{n-1}$, where b is a given N-1 component column vector. Let z be the unique solution to z=b+Qz. Define for any $n\geq 1$,

$$u_n'(i)=z_n(i)+(1-\alpha_i)^{-1} \min_j\{z_n(j)-z_{n-1}(j)\} \qquad \text{for } i=1,\dots,N-1 ,$$

and

$$u_n''(i)=z_n(i)+(1-\alpha_i)^{-1} \max_j\{z_n(j)-z_{n-1}(j)\} \qquad \text{for } i=1,\dots,N-1 .$$

Then, for any $n\geq 1$, $u_n'(i)\leq z(i)\leq u_n''(i)$ for i=1,...,N-1, where $u_n'(i)$ is nondecreasing in n to z(i) and $u_n''(i)$ is nonincreasing in n to z(i) for all i. The proof of this assertion is a slight modification of proofs given by Macqueen [3] and is based on the following fact: If $Tu\leq Tw$ then $u\leq w$, where the transformation T is defined by Tu=u-(b+Qu) for any N-1 component column vector u.

Remark. It is straightforward to extend the analysis above to the case of a general Markov matrix P; in this case the set of simultaneous equations g=Pg and gT+v=q+Pv has to be solved where g and v are unknown N component column vectors.

REFERENCES.

1. C. DERMAN AND A.F. VEINOTT Jr., A solution to a countable system of equations arising in Markovian decision processes, *Ann. Math. Statist.* 38 (1967), 582-584.

2. R.A. HOWARD, "Dynamic programming and Markov Processes", Wiley, New York, 1960.

3. J.B. MACQUEEN, A modified dynamic programming method for Markovian decision problems, *J. Math. Anal. and Appl.* 14 (1966), 38-43.

4. T.E. MORTON, Undiscounted Markov renewal programming via modified successive approximations, *Opns. Res.* 19 (1971), 1081-1089.