stichting
mathematisch
centrum

DEPARTMENT OF OPERATIONS RESEARCH BW 59/76 FEBRUARY

P.J. WEEDA

SENSITIVE TIME AND DISCOUNT OPTIMALITY IN MARKOV RENEWAL DECISION PROBLEMS WITH INSTANTANEOUS ACTIONS

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): 90C45

Sensitive time and discount optimality in Markov renewal decision problems with instantaneous actions.

by

P.J. Weeda

ABSTRACT

Finite state Markov renewal decision problems are considered in which some of the feasible actions are instantaneous. Because these actions take a zero time they are not distinctive with respect to sensitive discount criteria. On the other hand the fact that they take a zero time is usually a simplification of reality because in practice actions take some time in most cases. Here a method is developed to obtain policies which are optimal with respect to sensitive discount criteria as well as optimal for sufficiently small non-negative action times. To this end instantaneous actions are replaced by actions taking time $\varepsilon \geq 0$ and specifying a direct income which is a function of ε, analytic at $\varepsilon = 0$. A partial Laurent expansion in two variables s(discount rate) and ε is derived for a fixed policy. Based on this expansion it is shown that policies which are optimal with respect to the above criteria can be computed by solving a sequence of Markov renewal decision problems with policy iteration or linear programming.

KEY WORDS & PHRASES: *Markov renewal decision problems, instantaneous actions, sensitive time and discount optimality, Laurent expansion.*

1. INTRODUCTION

Sensitive discount criteria were introduced originally by VEINOTT and MILLER [15] and VEINOTT [14] in discrete and continuous time Markov decision problems. Extensions of these results are given by SLADKÝ [13] to the set of history remembering policies, by ROTHBLUM [12] to non-negative matrices with spectral radius not exceeding one and by HORDYK and SLADKÝ [7] to a countable state space. The extensions of [14] and [15] to the finite Markov renewal case have been given by DENARDO [4] who also observed that an n-discount optimal policy can be computed (under certain conditions about the existence of the moments) by solving each of a sequence of n + 2 Markov renewal decision problems by means of policy iteration.

This report devotes a special attention to finite state Markov renewal decision problems in which some of the feasible actions in certain states are instantaneous. Because these actions take a zero time they are not distinctive if sensitive discount criteria are used. On the other hand the fact that they take a zero time is usually a simplification of reality because in practice actions always take some time. It is therefore of interest to find a policy in these problems which not only satisfies sensitive discount optimal criteria but is at the same time optimal for sufficiently small non-negative action times in those states in which instantaneous actions are applied by this policy. Actually the model treated here assumes moreover that each action with a small action time ε specifies a direct income or cost which is a function of ε, analytic at the origin $\varepsilon = 0$. Roughly speaking one seeks then an optimal policy which allows "hesitation".

For illustration we present a two-state numerical example with two policies. Let P denote the matrix of transition probabilities, h the vector of direct income (in this example *not* a function of ε) and t the vector of intertransition times. The numerical data are

$$\underline{\text{policy 1}} \quad P = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ 1 & 0 \end{bmatrix} \quad h = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad t = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

<u>policy 2</u> $P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ $h = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$ $t = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$

For policy 1 a simple calculation yields

$$P^* = \begin{bmatrix} 2/3 & 1/3 \\ 2/3 & 1/3 \end{bmatrix} \quad P^* h = \begin{bmatrix} 2/3 \\ 2/3 \end{bmatrix} \quad P^* t = \begin{bmatrix} 1/3 \\ 1/3 \end{bmatrix}$$

and the average income vector y_1 equals $y_1 = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$.
For policy 2 we have

$$P^* = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} \quad P^* h = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad P^* t = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}$$

and $y_2 = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$. Thus both policies are gain-optimal. If we replace the t-vectors by $\begin{bmatrix} \varepsilon \\ 1 \end{bmatrix}$, ε non-negative and real, then

$$y_1(\varepsilon) = \begin{bmatrix} \frac{2}{2\varepsilon+1} \\ \frac{2}{2\varepsilon+1} \end{bmatrix} \quad \text{and } y_2(\varepsilon) = \begin{bmatrix} \frac{2}{\varepsilon+1} \\ \frac{2}{\varepsilon+1} \end{bmatrix}$$

and obviously $y_1(\varepsilon) < y_2(\varepsilon) < \begin{bmatrix} 2 \\ 2 \end{bmatrix}$ for $\varepsilon > 0$.
Hence policy 2 has to be preferred if the instantaneous actions take a small positive time. Policy iteration starting with policy 1 however terminates with policy 1. This report derives a general computational procedure which converges in the numerical example to policy 2, irrespective of the initial policy.

2. MODEL FORMULATION AND PRELIMINARIES

In a *Markov renewal decision problem* a system is observed at stochastic epochs given by the sequence $\{\underline{t}_n, n = 0,1,2,\ldots\}$ of non-negative random variables satisfying $0 = \underline{t}_0 \leq \underline{t}_1 \leq \underline{t}_2 \ldots$. At each epoch the system is in one of a finite set of states J. Let N be the number of states. In each state $i \in J$ there is a finite set $K(i)$ of feasible actions available to the decision maker. Let $\{\underline{i}_n,\ n = 0,1,2,\ldots\}$ be the sequence of states and let $\{\underline{k}_n,\ n = 0,1,2,\ldots\}$ be the sequence of actions chosen at the epochs $\underline{t}_n$, $n = 0,1,2,\ldots$.

2.1. ASSUMPTION. The joint probability

$$P\{\underline{i}_{n+1} = j,\ \underline{t}_{n+1} - \underline{t}_n \leq t \mid \underline{i}_0,\ldots,\underline{i}_n;\ \underline{t}_0,\ldots,\underline{t}_n;\ \underline{k}_0,\ldots,\underline{k}_n\}$$

depends only on $\underline{i}_n$, $\underline{k}_n$, j and t. Moreover we assume that this probability is independent of n and define

$$Q_{ij}^{k}(t) \stackrel{\text{def}}{=} P\{\underline{i}_{n+1} = j,\ \underline{t}_{n+1} - \underline{t}_n \leq t \mid \underline{i}_n = i,\ \underline{k}_n = k\}$$

for $t \in [0,\infty)$ and $Q_{ij}^{k}(t) \stackrel{\text{def}}{=} 0$ for $t < 0$.

Let $F \stackrel{\text{def}}{=} \bigtimes_{i \in J} K(i)$. F is the set of functions f having J as domain and assuming a value $f(i) \in K(i)$ for each $i \in J$. Such a function will be called a *policy*. The following definitions are relevant for each policy $f \in F$.

2.2. DEFINITION. An $N \times N$ matrix Q is called a *semi-Markov matrix* if each entry $Q_{ij}(t)$, $i, j \in J$, is a non-decreasing, right continuous, Borel measurable, real-valued function of t satisfying $Q_{ij}(t) = 0$ for $t < 0$ and $Q_{ij}(t) \leq 1$ for $t \geq 0$ such that

$$S_i(t) \stackrel{\text{def}}{=} \sum_{j \in J} Q_{ij}(t) \qquad \text{for } t \in \mathbb{R}$$

satisfies $S_i(0^-) = 0$ and $S_i(\infty) = 1$.
We assume that the probabilities $Q_{ij}^{f(i)}(t)$ constitute a semi-Markov matrix for all $f \in F$. In the sequel we drop the dependence on f in the notation as long as we consider a fixed $f \in F$.

2.3. DEFINITION. The sequence of matrices $Q^n(t)$, $n = 0,1,2,\ldots$ is for $t \in [0,\infty)$ defined by the recurrence relations

$$Q^{(0)}(t) \overset{\mathrm{def}}{=} I$$

$$Q^{(n)}(t) \overset{\mathrm{def}}{=} \int_{y\in[0,t)} Q(dy)Q^{(n-1)}(t-y) \qquad n \geq 1$$

2.4. DEFINITION. For each $t \in [0,\infty)$ the matrix $R(t)$ is defined by

$$R(t) \overset{\mathrm{def}}{=} \sum_{n=0}^{\infty} Q^{(n)}(t)$$

if the series in the right hand member converges. The matrix $R(t)$ is called the *Markov-renewal matrix* corresponding to $Q(t)$.

2.5. DEFINITION. An action $k \in K(i)$ is called instantaneous if

$$Q^k_{ij}(t) = \begin{cases} Q^k_{ij}(\infty) & \text{for } t \geq 0 \\ 0 & \text{for } t < 0 \end{cases}$$

Note that for each instantaneous action $k \in K(i)$: $S^k_i(t) = 1$ for $t \in [0,\infty)$.

In the Markov renewal decision model considered here we allow that for any policy in a subset of states instantaneous actions can be taken provided that one statement of the following theorem is valid.

2.6. THEOREM. (c.f. CINLAR [2], p.132) *The following statements are equivalent*

(i) $R(0) < \infty$

(ii) $R(t) < \infty$ *for each* $t \in [0,\infty)$

(iii) *Each simple ergodic set (= irreducible closed set of persistent states)* $E \subseteq J$ *contains at least one state* $i \in E$ *for which* $S_i(0) < 1$.

Under the conditions of definition 2.2 and one of the statements of theorem 2.6 the Laplace transforms of $Q(t)$ and $R(t)$ exist and are defined by

$$q(s) \overset{\mathrm{def}}{=} \int_{t\in[0,\infty)} e^{-st}Q(dt) \qquad \textit{for } s \geq 0$$

and

$$r(s) \overset{\text{def}}{=} \int_{t\in[0,\infty)} e^{-st} R(dt) \quad \textit{for } s > 0$$

By taking the Laplace transform of definiton 2.4 we have by theorem 2.6

$$(2.1) \qquad r(s) = \sum_{n=0}^{\infty} [q(s)]^n = [I - q(s)]^{-1}.$$

By the elementary renewal theorem (c.f. Ross [9], p.95) we have

$$(2.2) \qquad R(t) = O(t) \quad \text{as } t \to \infty$$

and using a standard Tanberian theorem (c.f. FELLER [6], p.421) gives

$$(2.3) \qquad r(s) = O(1/s). \quad \text{as } s \downarrow 0$$

The following theorem summarizes a useful result concerning the series expansion of $q(s)$ (c.f. FELLER [6] for the case of a distribution function).

2.7. <u>THEOREM</u>. *If* Q *has a finite* m^{th} *moment* $Q_m \overset{\text{def}}{=} \int_{[0,\infty)} x^m\, Q(dx)$ *then the following series expansion is valid in a neighborhood of* $s = 0$

$$q(s) = \sum_{i=0}^{m} Q_i \frac{(-1)^i}{i!} s^i + o(s^m)$$

Let $P \overset{\text{def}}{=} Q(\infty)$. Note that $P = Q_0$ in theorem 2.7.P is called the matrix of transition probabilities of the embedded Markov chain of the Markov renewal process. Let P^* be the (C,1) limit of P. P^* satisfies $P^*P = PP^* = P^*P^* = P^*$ and $P^*1 = 1$ (c.f. DOOB [5], p.175). P can have several simple ergodic sets E_m, $m = 1,\ldots,n$, say and a possibly empty set of transient states T. The states of a simple ergodic set have identical row vectors in the matrix P^*. The elements of these row vectors satisfy $P^*_{ij} > 0$ if i and j are in the same simple ergodic set and $P^*_{ij} = 0$ otherwise. If $\pi(m)$ denotes the common row vector of P^* of the m^{th} ergodic set E_m then a row of P^* corresponding to a transient state i satisfies $P^*_i = \sum_{m=1}^{n} t_{im}\pi(m)$ where t_{im} is the probabiltiy of absorption in the set E_m.

The matrix $[I - P + P^*]$ is invertible (c.f. KEMENY and SNELL [10]) and its inverse is usually called the fundamental matrix and is denoted by Z. In the sequel we will use the matrix $H \overset{\text{def}}{=} Z - P^*$ rather than Z itself.

2.8. LEMMA. (c.f. DENARDO [4], p.482). *Let vector* $a \in \mathbb{R}^n$ *satisfy* $P^*a = 0$ *and let vector* $b \in \mathbb{R}^n$ *be arbitrary then*

(i) *A vector* $x \in \mathbb{R}^n$ *satisfies* $[I - P]x = a$ *if and only if* $x = Ha + y$ *for a vector* $y \in \mathbb{R}^n$ *satisfying* $y = P^*y$.

(ii) *If* $[I - P]x = a$ *and* $P^*Q_1x = P^*b$ *then* $x = Ha + y$ *with* y_i *for* $i \in E_m$, $m = 1,\dots,n$, *being the quotient of scalar products*

$$y_i = \frac{\langle\pi(m),\ [c - Q_1Ha]\rangle}{\langle\pi(m),\ Q_1 1\rangle}$$

and, denoting the common value of the y_i, $i \in E_m$ *by* $y(m)$,

$$y_i = \sum_{m=1}^{n} t_{im}\, y(m) \qquad \text{for } i \in T$$

(iii) $r(s)a = o(1/s)$

2.9. DEFINITION. Let L be a normed N-dimensional vector space with norm $\|u\| = \max_{j\in J} u_j$, $u \in L$. Let M be the collection of all functions $U: (-\infty,\infty)\to L$ with the following properties

(1) $U(t) = 0$ for $t \in (-\infty,0)$

(2) U_j is Borel measurable for $j \in J$

(3) $\|U(t)\|$ is bounded on finite intervals

2.10. THEOREM. (ÇINLAR [2] p.137). *The integral equation*

$$V(t) = G(t) + \int_{y\in[0,t)} Q(dy)V(t - y)$$

has a unique solution $V(t) \in M$ *for any vector* $G(t) \in M$ *given by*

$$V(t) = \int_{y\in[0,t)} R(dy)G(t-y)$$

If we define the Laplace transforms of $V(t)$ and $G(t)$ by $v(s)$ and $g(s)$ and transform the integral equation of theorem 2.10 then we obtain using (2.1)

(2.4) $$v(s) = r(s)g(s) = [I - q(s)]^{-1}g(s).$$

2.11. DEFINITION. An action $k \in K(i)$, $i \in J$ is called an ε-time action if for $\varepsilon \geq 0$

$$Q_{ij}^{k}(t) = \begin{cases} Q_{ij}^{k}(\infty) = P_{ij}^{k} & \text{for } t \geq \varepsilon \\ 0 & \text{for } t < \varepsilon \end{cases}$$

and a function $\varepsilon \to G_i^k(\varepsilon)$ is specified, representing the direct income earned at time ε and being analytic at the origin $\varepsilon = 0$.

As a consequence of this definition $G_i^k(\varepsilon)$ can be expanded into a Taylor series in a neighborhood of $\varepsilon = 0$, given by

(2.5) $$G_i^k(\varepsilon) = \sum_{j=0}^{\infty} \frac{\varepsilon^j}{j!} G_i^k(0)^{(j)}$$

where $G_i^k(0)^{(j)}$ denotes the j^{th} derivative of $G_i^k(\varepsilon)$ at $\varepsilon = 0$.

For a usual action the expected income earned in a time period of length min $(\underline{t}_{n+1} - \underline{t}_n, t - \underline{t}_n)$ is denoted by $G_i^k(t)$. It is assumed that $G_i^k(t) \in M$ with $N = 1$ for $t \in [0,\infty)$ (see definition 2.9) and that $G_i^k(t)$ is a directly Riemann integrable function of t (c.f. FELLER [6] p.348 for the concept of direct Riemann integrability). Let $g_i^k(s)$ denote the Laplace transform of $G_i^k(t)$ and let $G_i^{k(j)}$ be the j^{th} moment of $G_i^k(t)$. If $G_i^k(t)$ has finite m^{th} moment then the following series expansion of $g_i^k(s)$ is valid

(2.6) $$g_i^k(s) = \sum_{j=0}^{m} \frac{(-s)^j}{j!} G_i^{k(j)} + o(s^m)$$

3. A PARTIAL LAURENT EXPANSION FOR $v(s,\varepsilon)$

In a Markov renewal decision problem, with s being the interest rate, v(s) represents also the expected discounted income vector for a fixed policy. If there are ε-time actions involved we consider ε as a second variable in v(s), q(s) and g(s) thus rewriting (2.4) as

(3.1) $$[I - q(s,\varepsilon)]\, v(s,\varepsilon) = g(s,\varepsilon)$$

In this section we derive a partial Laurent expansion of $v(s,\varepsilon)$ in the variables s and ε for a fixed policy $f \in F$. Let A be the subset of states in which ε-time actions are applied by a fixed policy f. Again we drop the dependence of f in the notation throughout this section. The $|\bar{A}|$-dimensional subvector $g(s,\varepsilon)_{\bar{A}}$ depends only on s. If $G_i^{(\ell)}$ exists for all $i \in \bar{A}$ then

$$g(s,\varepsilon)_{\bar{A}} = \sum_{\ell=0}^{m} \frac{(-s)^{\ell}}{\ell\,!} G_{\bar{A}}^{\ell} + o(s^m)$$

where $G_{\bar{A}}^{(\ell)}$ is the $|\bar{A}|$-vector with elements $G_i^{(\ell)}$. Let $G_{\bar{A}}^{(\ell)-} \overset{\text{def}}{=} (-1)^{\ell} G_{\bar{A}}^{(\ell)}/\ell!$ then $g(s,\varepsilon)$ becomes

(3.1) $$g(s,\varepsilon)_{\bar{A}} = \sum_{\ell=0}^{m} s^{\ell} G_{\bar{A}}^{(\ell)-} + o(s^m)$$

For $g(s,\varepsilon)_A$ we have by definition 2.11 and in virtue of (2.6)

$$g(s,\varepsilon)_A = e^{-s\varepsilon} G(\varepsilon)_A$$

$$= e^{-s\varepsilon} \sum_{j=0}^{\infty} \frac{\varepsilon^j}{j!} G_A^{(j)}$$

where $G_A^{(j)}$ abbreviates $G(0)_A^{(j)}$. Let $G_A^{(j)+} \overset{\text{def}}{=} G_A^{(j)}/j!$ and substituting the Taylor expansion of $e^{-s\varepsilon}$ yields

$$(3.2)\qquad g(s,\varepsilon)_A = \sum_{\ell=0}^{\infty} \frac{(-s\varepsilon)^{\ell}}{\ell!} \sum_{j=0}^{\infty} \varepsilon^j G_A^{(j)+}$$

$$= \sum_{\ell=0}^{\infty} \frac{(-s)^{\ell}}{\ell!} \sum_{j=\ell}^{\infty} \varepsilon^j G_A^{(j-\ell)+}$$

For the submatrix $q(s,\varepsilon)_{AJ}$ we have

$$q(s,\varepsilon)_{AJ} = e^{-s\varepsilon}(Q_0)_{AJ}$$

Substituting the Taylor expansion of $e^{-s\varepsilon}$ yields

$$q(s,\varepsilon)_{AJ} = \sum_{\ell=0}^{\infty} \frac{(-s\varepsilon)^{\ell}}{\ell!} (Q_0)_{AJ}$$

Defining $(Q_{\ell})_{\bar{A}J} \overset{\text{def}}{=} \frac{(-1)^{\ell}}{\ell!} (Q_0)_{AJ}$ we have

$$(3.3)\qquad q(s,\varepsilon)_{AJ} = \sum_{\ell=0}^{\infty} (s\varepsilon)^{\ell} (Q_{\ell})_{\bar{A}J}$$

If the m^{th} moment $(Q_m)_{ij}$ exists for $i \in \bar{A}$, $j \in J$ and a fixed policy $f \in F$, then we have by theorem 2.7 defining $(Q_{\ell})_{\bar{A}J} \overset{\text{def}}{=} (-1)^{\ell}(Q_{\ell})_{\bar{A}J}/\ell!$ for $\ell = 0,1,2,\ldots$

$$(3.4)\qquad q(s,\varepsilon)_{\bar{A}J} = \sum_{\ell=0}^{m} s^{\ell}(Q_{\ell})_{\bar{A}J} + o(s^m)$$

3.1. <u>THEOREM</u>. *Consider a fixed policy* $f \in F$. *Let* A *be the set of states with* ε*-time actions. Suppose that* $(Q_{n+2})_{\bar{A}J}$ *and* $G_{\bar{A}}^{(n+1)}$ *are finite. Then we have the following partial Laurent expansion for* $v(s,\varepsilon)$

$$(3.5)\qquad v(s,\varepsilon) = \sum_{i=-1}^{n} s^i \sum_{j=0}^{\infty} \varepsilon^j V(i,j) + (s^n)$$

where $V(i,j)$ *is an* N*-vector for each fixed* $i \in \{-1,0,1,\ldots,n\}$ *and* $j \in \{0,1,2,\ldots\}$ *which is the unique solution of the equations*

$$(3.6)\qquad \begin{cases} [I-P]V(i,j) = a(i,j) \\ P^*Q_1^*\ V(i,j) = P^*b(i,j) \end{cases} \text{with } Q_1^* \stackrel{\text{def}}{=} \begin{bmatrix} 0 \\ (Q_1)_{\bar{A}J} \end{bmatrix}$$

with $a(i,j) \in \mathbb{R}^N$ *given for* $i \geq 0$ *and* $j \geq 0$ *by*

$$a(i,j) = \begin{bmatrix} \sum_{k=1}^{i+1} \frac{(-1)^k}{k!}(Q_0)_{AJ}V(i-k,j-k) + \frac{(-1)^i}{i!}\frac{G_A(j-i)}{(j-i)!} \\ \\ \sum_{k=1}^{i+1} \frac{(-1)^k}{k!}(Q_k)_{\bar{A}J}V(i-k,j) + \begin{cases} \frac{(-1)^i}{i!} & G_{\bar{A}}^{(i)} \quad \text{for } j = 0 \\ 0 & \text{for } j \neq 0 \end{cases} \end{bmatrix}$$

and $a(i,j) = 0$ *otherwise and with* $b(i,j) \in \mathbb{R}^N$ *given for* $i \geq -1$ *and* $j \geq 0$ *by*

$$b(i,j) = \begin{bmatrix} \sum_{k=1}^{i+2} \frac{(-1)^k}{k!}(Q_0)_{AJ}V(i+1-k,j-k) + \frac{(-1)^{i+1}}{(i+1)!}\frac{G_A^{(j-i-1)}}{(j-i-1)!} \\ \\ \sum_{k=2}^{i+2} \frac{(-1)^k}{k!}(Q_k)_{\bar{A}J}V(i+1-k,j) + \begin{cases} \frac{(-1)^{i+1}}{(i+1)!} & G_{\bar{A}}^{(i+1)} \quad \text{for } j = 0 \\ 0 & \text{for } j \neq 0 \end{cases} \end{bmatrix}$$

and $b(i,j) = 0$ *otherwise.*

PROOF. Primarily we prove $v(s,\varepsilon)$ to be of the form

$$(3.7)\qquad v(s,\varepsilon) = \sum_{i=-1}^{n} s^i \sum_{j=0} \varepsilon^j\ V(i,j) + f_n(s,\varepsilon)$$

with the $V(i,j)$ being the unique solution of (3.6) c.f. lemma 2.8(ii) and sub-sequently prove that $f_n(s,\varepsilon) = o(s^n)$. If we substitute (3.1)...(3.4) and (3.7) in (3.1) we obtain

$$(3.8)\qquad [I - q(s,\varepsilon)]f_n(s,\varepsilon) + \sum_{i=-1}^{n} s^i \sum_{j=0}^{\infty} \varepsilon^j\ V(i,j) +$$

$$-\left[\begin{array}{l}\displaystyle\sum_{\ell=0}^{n+2}(s\varepsilon)^{\ell}(Q_{\ell})^{-}_{AJ}\\ \\ \displaystyle\sum_{\ell=0}^{n+2} s^{\ell}(Q_{\ell})^{-}_{\bar{A}J}\end{array}\right]\sum_{i=-1}^{n} s^{i}\sum_{j=0}^{\infty}\varepsilon^{j}\,V(i,j)+o(s^{n+1})=$$

$$=\left[\begin{array}{l}\displaystyle\sum_{k=0}^{n+1}\frac{(-1)!}{k!}\,s^{k}\sum_{\ell=k}^{\infty}\varepsilon^{\ell}G_{A}^{(\ell-k)+}\\ \\ \displaystyle\sum_{k=0}^{n+1} s^{k}\,G_{\bar{A}}^{(k)-}\end{array}\right]+o(s^{n+1})$$

Noting that

$$\sum_{k=0}^{n+2} s^{k}(Q_{k})^{-}_{\bar{A}J}\sum_{i=-1}^{n} s^{i}\sum_{j=0}^{\infty}\varepsilon^{j}\,V(i,j)=$$

$$\sum_{i=-1}^{n} s^{i}\sum_{k=0}^{i+1}\sum_{j=0}^{\infty}\varepsilon^{j}(Q_{k})^{-}_{\bar{A}J}\,V(i-k,j)+$$

$$+\,s^{n+1}\sum_{j=0}^{\infty}\varepsilon^{j}\sum_{k=1}^{n+2}(Q_{k})^{-}_{\bar{A}J}\,V(n+1-k,j)+o(s^{n+1})$$

and

$$\sum_{k=0}^{n+2}(\varepsilon s)^{k}(Q_{k})^{-}_{AJ}\sum_{i=-1}^{n} s^{i}\sum_{j=0}^{\infty}\varepsilon^{j}\,V(i,j)=$$

$$\sum_{i=-1}^{n} s^{i}\sum_{j=0}^{\infty}\varepsilon^{j}\sum_{k=0}^{i+1}(Q_{k})^{-}_{AJ}\,V(i-k,j-k)+$$

$$+\,s^{n+1}\sum_{j=0}^{\infty}\varepsilon^{j}\sum_{k=1}^{n+2}(Q_{k})^{-}_{AJ}\,V(n+1-k,j-k)+o(s^{n+1}).$$

(3.8) is equivalent to

$$(3.9)\qquad [I-q(s,\varepsilon)]\, f_n(s,\varepsilon) + \sum_{i=-1}^{n} s^i \sum_{j=0}^{\infty} \varepsilon^j\, V(i,j) +$$

$$- \begin{bmatrix} \sum_{i=-1}^{n} s^i \sum_{j=0}^{\infty} \varepsilon^j \sum_{k=0}^{i+1} (Q_k)_{AJ}^{-}\, V(i-k,j-k) \\ \\ \sum_{i=-1}^{n} s^i \sum_{j=0}^{\infty} \varepsilon^j \sum_{k=0}^{i+1} (Q_k)_{\bar{A}J}^{-}\, V(i-k,j) \end{bmatrix} +$$

$$- s^{n+1} \begin{bmatrix} \sum_{j=0}^{\infty} \varepsilon^j \sum_{k=1}^{n+2} (Q_k)_{AJ}^{-}\, V(n+1-k,j-k) \\ \\ \sum_{j=0}^{\infty} \varepsilon^j \sum_{k=1}^{n+2} (Q_k)_{\bar{A}J}^{-}\, V(n+1-k,j) \end{bmatrix} + o(s^{n+1}) =$$

$$\begin{bmatrix} \sum_{k=0}^{n} \frac{(-1)^k}{k!} s^k \sum_{\ell=k}^{\infty} \varepsilon^{\ell} G_A^{(\ell-k)+} \\ \\ \sum_{k=0}^{n} s^k\, G_{\bar{A}}^{(k)-} \end{bmatrix} + s^{n+1} \begin{bmatrix} \frac{(-1)^{n+1}}{(n+1)!} \sum_{\ell=n+1}^{\infty} \varepsilon^{\ell} G_A^{(\ell-n-1)+} \\ \\ G_{\bar{A}}^{(n+1)-} \end{bmatrix}$$

This equation holds for sufficiently small s and ε if and only if for $i = -1,0,1,\dots,n$, $j = 0,1,\dots$

$$(3.10)\qquad V(i,j) - \begin{bmatrix} \sum_{k=0}^{i+1} (Q_k)_{AJ}^{-}\, V(i-k,j-k) \\ \\ \sum_{k=0}^{i+1} (Q_k)_{\bar{A}J}^{-}\, V(i-k,j) \end{bmatrix} = \begin{bmatrix} \frac{(-1)^i}{i!} G_A^{(j-i)+} \\ \\ \begin{cases} G_{\bar{A}}^{(i)-} & j=0 \\ 0 & j \neq 0 \end{cases} \end{bmatrix}$$

and

$$(3.11)\qquad [I - q(s,\varepsilon)]\, f_n(s,\varepsilon) = o(s^{n+1}) +$$

$$+\, s^{n+1}\left[\begin{array}{l} \frac{(-1)^{n+1}}{(n+1)!} \sum_{\ell=n+1}^{\infty} \varepsilon^{\ell}\, G_A^{(\ell-n-1)} + \sum_{j=0}^{\infty} \varepsilon^j \sum_{k=1}^{n+2} (Q_k^-)_{AJ} V(n+1-k, j-k) \\ \\ G_{\bar{A}}^{(n+1)-} + \sum_{j=0}^{\infty} \varepsilon^j \sum_{k=1}^{n+2} (Q_k)_{\bar{A}J}^{-}\, V(n+1-k, j) \end{array}\right]$$

both are satisfied. It is easily verified that (3.10) is equivalent to

$$[I - P]V(i,j) = a(i,j).$$

Premultiplying $[I - P]V(i+1,j) = a(i+1,j)$ by P^* yields $P^*a(i+1,j) = 0$. Further by the definitions of $a(i,j)$ and $b(i,j)$

$$b(i,j) = a(i+1,j) + Q_1^* V(i,j)$$

which, premultiplied by P^*, yields the second equation of (3.6). That (3.6) has a unique solution in $V(i,j)$ follows from lemma 2.8(ii).

Solving for $f_n(s,\varepsilon)$ in (3.11) yields

$$f_n(s,\varepsilon) = [I - q(s,\varepsilon)]^{-1} \sum_{j=0}^{\infty} \varepsilon^j\, a(n+1,j) s^{n+1}$$

$$+ [I - q(s,\varepsilon)]^{-1}\, o(s^{n+1})$$

Because $[I - q(s,\varepsilon)]^{-1} = O(1/s)$ for $\varepsilon \geq 0$ (c.f. (2.3)) and $P^* a(n+1,j) = 0$ permits us to apply lemma 2.8(iii) we obtain

$$f_n(s,\varepsilon) = o(1/s)s^{n+1} + O(1/s)o(s^{n+1})$$

$$= o(s^n) \qquad \square$$

4. THE COMPUTATION OF n-DISCOUNT, m-TIME OPTIMAL POLICIES.

We review first the case that $\varepsilon = 0$ which is covered by DENARDO [4]. Using the terminology of BLACKWELL [1] a policy $f^* \in F$ is *s-optimal* if $v_{f^*}(s,0) \geq v_f(s,0)$ for all $f \in F$. A policy $f^* \in F$ is *optimal* if it is s-optimal for sufficiently small positive s. It is shown in [4] that an optimal policy exists if for each $f \in F$ $v_f(s,0)$ either has an isolated singularity or is analytic at the origin $s = 0$. We define the following sequence of sets recursively

$$(4.1) \qquad \begin{cases} F(-2,0) \overset{\text{def}}{=} F \\ F(n,0) \overset{\text{def}}{=} \{f^* \in F(n-1,0) \colon V_{f^*}(n,0) \geq V_f(n,0) \text{ for } f \in F(n-1,0)\} \end{cases}$$

provided that the required moments, $(Q_{n+2})_{\bar{A}J}$ and $G_{\bar{A}}^{(n+1)}$, exist for all $f \in F(n,0)$. We define the set $F(\infty,0)$ by

$$(4.2) \qquad F(\infty,0) \overset{\text{def}}{=} \begin{cases} F(n,0) \text{ if } F(n,0) \text{ contains one policy} \\ \lim_{n\to\infty} F(n,0) \text{ otherwise} \end{cases}$$

It is shown in [4] that $F(\infty,0)$ is exactly the set of policies optimal in the Blackwell sense if it either contains a single policy or several policies each of which has a Laurent expansion about the origin. Hence $F(\infty,0)$ is non-empty and each $f \in F(\infty,0)$ is optimal. A direct consequence of definition (4.1) is that

$$(4.3) \qquad F(n,0) \subseteq F(n-1,0) \qquad \text{for } n = -1,0,1,\ldots .$$

Because $F(\infty,0) \neq \emptyset$ also $F(n,0) \neq \emptyset$ for all n.

A policy $f \in F(-1,0)$ is the familiar gain-optimal policy. We reserve the name *Markov renewal program* here for a Markov renewal decision problem in which only a gain-optimal policy is required. Such a policy can be computed by means of policy iteration or linear programming (c.f. DENARDO [3]) and requires only the matrices P and Q_1^* and the vector $b(-1,0) = G^{(0)}$ for each policy. The Markov renewal program computing a policy $f \in F(-1,0)$ can be summarized by the 5-tuple

$$(4.4)\qquad (J,F,P(\cdot),Q_1^*(\cdot),b_{(\cdot)}(-1,0))$$

In the sequel a vector $u \in \mathbb{R}^N$ will be called a *gain vector* if and only if it satisfies $P^*u = u$ and a *value vector* if and only if it is a solution of $[I - P]u = a$ with $a \neq 0$.

4.1. LEMMA. *For each* $f \in F$ *and* $m = 0,1,\ldots$ *the vector* $V_f(-1,m)$ *is a gain vector*

PROOF. By theorem 3.1 $V_f(-1,m)$ satisfies

$$[I - P(f)]V_f(-1,m) = a_f(-1,m) = 0$$

By lemma 2.8(i) with $a = 0$ we have $V_f(-1,m) = P^*(f)\ V_f(-1,m)$ implying the assertion.

We define now the following sequence of sets recursively by

$$(4.5)\qquad \begin{cases} F(-1,-1) \overset{\text{def}}{=} F \\ F(-1,m) \overset{\text{def}}{=} \{f^* \in F(-1,m-1)\colon V_{f*}(-1,m) \geq V_f(-1,m) \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{for } f \in F(-1,m-1)\} \end{cases}$$

4.2. THEOREM. *A policy* $f \in F(-1,m)$ *can be computed for* $m = 0,1,2,\ldots$ *by applying policy iteration to the Markov renewal program*

$$(4.6)\qquad (J,F(-1,m-1),\ P(\cdot),\ Q_1^*(\cdot),\ b_{(\cdot)}(-1,m))$$

PROOF. By lemma 4.1 $V_f(-1,m)$ is a gain-vector for each $f \in F(-1,m-1)$. Hence a policy $f \in F(-1,m)$ is a gain optimal with respect to (4.6).

4.3. COROLLARY. *A policy* $f \in F(-1,m)$ *exists for each* $m \in \mathbb{N}$.

PROOF. The set $F(-1,0)$ is non empty under the conditions stated at the beginning of this section. Because a gain-optimal policy in a Markov renewal program (4.6) attains the N maxima simultaneously and $b_f(-1,m)$ exists for all $f \in F$ and $m \in \mathbb{N}$ we have the assertion.

To complete the definition of $F(n,m)$ we define for $n = 0,1,\ldots$ and $m = 0,1,2,\ldots$

$$(4.6) \qquad \begin{cases} F(n,-1) \overset{\text{def}}{=} F(n-1,\infty) \\ \\ F(n,m) \overset{\text{def}}{=} \{f^* \in F(n,m-1) : V_{f*}(n,m) \geq V_f(n,m) \text{ for } f \in F(n,m-1)\} \end{cases}$$

provided that the moments $(Q_{n+2})_{\bar{A}J}$ and $G_{\bar{A}}^{(n+1)}$ are finite for $f \in F$. Note that (4.6) redefines $F(n,0)$. A policy $f \in F(n,m)$ is called an *n-discount, m-time optimal policy.*

4.4. LEMMA. *If* $(Q_{n+2})_{\bar{A}J}$ *and* $G_{\bar{A}}^{(n+1)}$ *are finite for* $f \in F$ *and* u *is a value vector satisfying* $[I - P]u = a(n,m)$ *then* $V(n,m) - u$ *is a gain vector for* $n = -1,0,1,\ldots, m = 0,1,\ldots$

PROOF. By theorem 3.1 $V(n,m)$ satisfies $[I - P]V(n,m) = a(n,m)$. Subtracting the equation satisfied by u yields

$$[I - P][V(n,m) - u] = 0$$

which implies that $V(n,m) - u$ is a gain vector by lemma 2.8(i).

4.5. THEOREM. *If* $(Q_{n+2})_{\bar{A}J}$ *and* $G_{\bar{A}}^{(n+1)}$ *are finite for each* $f \in F(n,-1)$ *then an n-discount, m-time optimal policy can be computed by applying policy iteration to the Markov renewal program*

$$(4.7) \qquad (J, F(n,m-1), P(\cdot), Q_1^*(\cdot), b_{(\cdot)}(n-1,m) - Q_1^*(\cdot)x)$$

where x *is a value vector satisfying* $[I - P(f)]x = a_f(n,m)$ *for all* $f \in F(n-1,m)$.

PROOF. Note first that because of the definitions (4.1), (4.5) and (4.6) we have

$$F(n-1,m) \supseteq F(n-1,\infty) \stackrel{\text{def}}{=} F(n,-1) \supseteq F(n,m-1)$$

Hence the vector x is defined for and shared by all $f \in F(n,m-1)$. Let $V^*(n,m) = V_f(n,m)$ for $f \in F(n,m)$, $n = -1,0,1,\dots$ and $m = 0,1,\dots$ $(V^*(-1,m),x)$ is a solution, unique in $V^*(-1,m)$, of the system of equations for $f \in F(-1,m)$

$$\begin{cases} [I - P(f)]V^*(-1,m) = 0 \\ [I - P(f)]x = a_f(0,m) = b_f(-1,m) - Q_1^*(f)V^*(-1,m) \end{cases}$$

We note that the same vector x satisfies the second equation for all $f \in F(-1,m)$ as a consequence of the termination conditions of policy iteration applied to (4.6). For each $f \in F(-1,m)$ $(V_f(0,m) - x,\ u)$ is a solution, unique in $V_f(0,m) - x$, of the system of equations

$$\begin{cases} [I - P(f)][V_f(0,m) - x] = 0 \\ [I - P(f)]u = a_f(1,m) = b_f(0,m) - Q_1^*(f)V_f(0,m) \\ \qquad\qquad = b_f(0,m) - Q_1(f)x - Q_1^*(f)[V_f(0,m) - x] \end{cases}$$

Because the vector x is shared by each $f \in F(0,m-1)$ and $V_f(0,m) - x$ is a gain vector, maximizing the gain vector in the Markov renewal program (4.7) yields a policy $f \in F(0,m)$. Because the argument repeats for $n > 0$ we have the assertion.

From theorem 4.4 and 4.7 we conclude that an n-discount, m-time optimal policy can be computed by applying policy iteration or linear programming to a sequence of $(n+2)(m+1)$ Markov renewal programs. If $m < \infty$ we may better define $F(i,-1) \stackrel{\text{def}}{=} F(i-1,m)$ $i = 0,1,2,\dots$ instead of using definition (4.6). If (i,j) represents the Markov renewal program applied to compute a policy in F(i,j) and the computational order is denoted by an arrow, then the scheme is as follows

$$(-1,0) \to (-1,1) \to \dots \to (-1,m) \to$$
$$\to (0,0) \to (0,1) \to \dots \to (0,m) \to$$
$$\dots$$
$$\to (n,0) \to (n,1) \to \dots \to (n,m)$$

To illustrate the method developed in this section we apply it to the numerical example presented in section 1. Suppose we start the iteration with policy 1, computing a (-1)-discount, 0-time optimal policy.

Policy evaluation (policy 1)

Solve:

$$[I - P]V(-1,0) = a(-1,0) = 0$$
$$[I - P]x = a(0,0) = b(-1,0) - Q_1^* V(-1,0)$$

Here we have $b(-1,0) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $Q_1^* = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$.

Hence

$$x_1 - \tfrac{1}{2}x_1 - \tfrac{1}{2}x_2 = 1 - 0$$
$$x_2 - x_1 = 0 - V(-1,0)_1$$

To obtain a solution put $x_2 = 0$ yielding $x_1 = 2$ and $V(-1,0)_1 = V(-1,0)_2 = 2$.

Policy improvement

$$\max_{f \in F} \{b_f(-1,0) - Q_1^*(f)V(-1,0) + P(f)\, x\} =$$
$$= \max \left\{ \begin{bmatrix} 2 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 \\ 0 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} 2 \\ 2 \end{bmatrix} + \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} 2 \\ 0 \end{bmatrix} \right\}$$
$$= \max \left\{ \begin{bmatrix} 2 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 \\ 0 \end{bmatrix} - \begin{bmatrix} 0 \\ 2 \end{bmatrix} + \begin{bmatrix} 0 \\ 2 \end{bmatrix} \right\} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$$

Hence $F(-1,0) = \{$policy 1, policy 2$\}$ and $V^*(-1,0) = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$

Next we compute a policy $f \in F(-1,1)$.

Policy evaluation (policy 1)

Solve:

$$[I - P]\ V(-1,1) = a(-1,1) = 0$$
$$[I - P]x = a(0,1) = b(-1,1) - Q_1^* \ V(-1,1)$$

We have $b(-1,1) = \begin{bmatrix} -2 \\ 0 \end{bmatrix}$ and thus

$$x_1 - \tfrac{1}{2}x_1 - \tfrac{1}{2}x_2 = -2 - 0$$
$$x_2 - x_1 \qquad = \ 0 - V(-1,1)_1$$

Putting $x_2 = 0$ yields $x_1 = -4$ and $V(-1,1)_1 = V(-1,1)_2 = -4$

Policy improvement

$$\max_{f \in F(-1,0)} \{b_f(-1,1) - Q_1^*(f)\ V(-1,1) + P(f)x\} =$$

$$= \max_{f \in F(-1,0)} \left\{ \begin{bmatrix} -4 \\ 0 \end{bmatrix}, \begin{bmatrix} -2 \\ 0 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} -4 \\ -4 \end{bmatrix} + \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} -4 \\ 0 \end{bmatrix} \right\}$$

$$= \max_{f \in F(-1,0)} \left\{ \begin{bmatrix} -4 \\ 0 \end{bmatrix}, \begin{bmatrix} -2 \\ 0 \end{bmatrix} \right\} = \begin{bmatrix} -2 \\ 0 \end{bmatrix}$$

Hence $F(-1,1) = \{\text{policy } 2\}$ and also $F(\infty,\infty) = \{\text{policy } 2\}$. *

5. REFERENCES

[1] BLACKWELL, D., *Discrete dynamic programming*, Ann. Math. Statistics 33, p. 719-726 (1962).

[2] ÇINLAR, E. *Markov renewal theory*, Adv. Appl. Prob. 1, p. 123-187 (1969).

* Calculation of V(0,0) and V(1,0) shows that both policies are ∞-discount optimal in the sense of VEINOTT [14] and DENARDO [4].

[3] DENARDO, E.V. *Computing bias-optimal policies in discrete and continuous Markov decision problems*, Oper. Res. 18, P. 279-289 (1970).

[4] DENARDO, E.V., *Markov renewal programs with small interest rates*, Ann. Math. Statistics, 42, p. 477-496 (1971).

[5] DOOB, J., *Stochastic processes*, Wildy, New York (1953).

[6] FELLER, W., *An introduction to probability theory and its applications*, 2, Wiley, New York (1966).

[7] HORDYK, A. and SLADKÝ, K., *Sensitive optimality criteria in countable state dynamic programming*, Prepublication, Mathematical Centre Report BW48/75, September 1975.

[8] HOWARD, R.A., *Dynamic programming and Markov processes, Technology Press*, M.I.T., Cambridge (1960).

[9] JEWELL, W.S., *Markov renewal programming, I: formulation, finite return models, II: infinite return models, example*, Oper. Res. 11, p. 938-971 (1963).

[10] KEMENY, J.G. and SNELL, J.L., *Finite Markov chains*, van Nostrand, Princeton (1961).

[11] ROSS, S.M., *Applied probability models with optimization applications*, Holden-Day, San Francisco (1970).

[12] ROTHBLUM, U.G., *Normalized Markov decision chains I; sensitive discount optimality*, Oper. Res. 23, p. 785-795 (1975).

[13] SKADKÝ, K., *On the set of optimal controls for Markov chains with rewards*, Kybernetika 10, 350-367 (1974).

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

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