stichting
mathematisch
centrum

MC

AFDELING MATHEMATISCHE BESLISKUNDE BW 49/75 JUNE

A. HORDIJK

REGENERATIVE MARKOV DECISION MODELS

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): 60J05, 90C40

Regenerative Markov decision models

by

A. Hordijk

ABSTRACT

Discrete time Markov decision processes with a countable state space are investigated. Under a condition guaranteeing the recurrence to a fixed state, the existence of stationary optimal policies with respect to discounted expected and average expected return is shown. Also sensitive discount optimal policies do exist and limit decision rules, as the discount-factor tends to one, of discounted optimal rules are bias or equivalently average-overtaken optimal. Finally, an iteration procedure to compute sensitive discount optimal policies is given.

KEY WORDS & PHRASES: *Discrete time dynamic programming; denumerable state space; compact decision sets; existence and computation of optimal policies.*

1. INTRODUCTION

It is clear that from practical viewpoint the class of discrete time Markov decision processes with a finite number of states and per state a finite number of decisions, is most important. And so it is evident that most of the research is devoted to this finite case.

However there are interesting models in which it is natural to have a denumerable state space. For example inventory models with backlogging for which the stochastic demand between two decision epochs is unbounded, or qeueuing models for which an unbounded stochastic number of customers may enter service facility between decision points. With these models in mind we intend to extend a number of important results from the finite to the denumerable case. The condition (see section 2) which we assume is in the finite case equivalent to the existence of some state, say state 0, such that state 0 is accessible under each policy from each state.
If the state space is denumerable then as is well known there is difference between accessiblity, with recurrence and positive recurrence. What we need is the strongest form of recurrence i.e. positive recurrence.

There are many models for which this assumption is not satisfied for all policies. For example take the waiting line model where at discrete times a decision can be taken on the service rate and where the decision to close the service facility is allowed. Well as is evident, when the latter is always chosen that no state is recurrent. The procedure which enables to obtain conclusions with the results of this paper is then to restrict the class of decision rules. To make this more explicit for the above model. Restrict the class of decision rules to those which always switch the service facility on when the number of customers waiting is larger than some large integer N. Sometimes one is not interested at all in decision rules not in this class since they are far from being part of optimal policies. If this is not the case there is still the freedom to adjust the integer N, allowing to obtain conclusions in the larger class of policies for which the switch-on level has a finite limes superior as time tends to infinity.

Each of the sections 3,4, and 5 start with summarizing the results contained in that section. In the remainder of this section we introduce notions

and notations used in this paper.

We are concerned with a dynamic system which at times $t = 1,2,\ldots$, is observed to be in one of a possible number of states. Let E denote the countable space of all possible states. If at time t the system is observed in state i then a decision must be chosen from a given set $P(i)$. The probability that the system moves to a new state j (the so-called transition probability) is a function only of the last observed state i and the subsequently taken decision. In order to avoid an over-burdened notation we shall identify the decision to be taken with the probability measure on E that is induced by it. Thus for each $i \in E$ the set $P(i)$ consists of probability measures $p(i,.)$.

Let P be the set of all stochastic matrices P with $p(i,.) \in P(i)$ for each $i \in E$. Hence P has the *product property:* with P_i, $i \in E$ the set P also contains that P with for every $i \in E$ the i^{th} row of P equal to the i^{th} row of P_i.

A policy R for controlling the system is a sequence of decision rules for the times $t = 1,2,\ldots$, where the decision rule for time t is the instruction at time t which prescribes the decision to be taken. This instruction may depend on the history, i.e., the states and decisions at times $1,\ldots, t-1$ and the state at time t. When the decision rule is independent of the past history except for the present state then it can be identified with a $P \in P$. A memoryless or Markov policy R is a sequence $P_1,P_2,\ldots, \in P$, where P_t denotes the decision rule at time t. P_t also gives the transition probabilities at time t. It follows from a theorem in DERMAN & STRAUCH [3], generalized in STRAUCH & VEINOTT [10] that we do not loose generality by restricting the class of policies to the Markov policies, (see also section 13 of HORDIJK [5]). In this paper we shall only use Markov policies.

A memoryless policy which takes at all times the same decision rule, i.e., $P^{\infty} := (P,P,\ldots)$, $P \in P$ is called a stationary policy.

When in state i decision $p(i,.)$ is taken then an immediate return depending on i and $p(i,.)$ is incurred. Let $r_p(i)$ be the immediate return when taking decision $p(i,.)$ the i^{th} row of matrix P in state i and write r_p for the vector with i^{th} component $r_p(i)$. Note that if $P, Q \in P$ with

$p(i,.) = q(i,.)$ then $r_P(i) = r_Q(i)$.

The expectation of the return at time n when starting in state i at time one and using policy $R = (P_1, P_2, \ldots)$ will be denoted by $\mathbb{E}_{i,R}\, r(\underline{x}_n)$, where $\underline{x}_n$ (random variables are underlined) is the state at time n. $\mathbb{E}_R\, r(\underline{x}_n)$ denotes the vector with i^{th} component $\mathbb{E}_{i,R}\, r(\underline{x}_n)$. It is easily seen that

$$\mathbb{E}_R\ r(\underline{x}_n) = P_1 P_2 \ldots P_{n-1}\ r_{P_n}.$$

We shall use the notation P_R^{t-1} for the matrix $P_1 \ldots P_{t-1}$.

We need a notion of convergence on $\mathcal{P}$. A sequence $P_n, n = 1,2,\ldots,$ is convergent to P if $\lim_{n\to\infty} p_n(i,j) = p(i,j)$ for all i and j. In this case we shall say that $\lim_{n\to\infty} P_n = P$. $\mathcal{P}$ with this product topology (see section 13 of [5]) is a metric space. We assume that $\mathcal{P}$ is compact and r_P is continuous in P i.e. for each $i \in E$ $\lim r_{P_n}(i) = r_P(i)$ as P_n converges to P. Note that these assumption are automatically fulfilled if $\mathcal{P}(i)$ is finite for all $i \in E$. For vectors x,y we write $x \le y$ resp. $x < y$ if $x(i) \le y(i)$ for all i resp. $x(i) \le y(i)$ for all i and $x(i) \neq y(i)$ for some i; for vectors x, x_n, $n = 1,2,\ldots$, we write $\lim_{n\to\infty} x_n = 0$ if $\lim_{n\to\infty} x_n(i) = 0$ for all $i \in E$ and $\lim_{n\to\infty} x_n = x$ if $\lim_{n\to\infty} x_n(i) = x(i)$ for all $i \in E$.

2. ASSUMPTIONS

We assume the existence of a state, say state 0, *and the existence of finite nonnegative vectors* $y_0, y_1, y_2, \ldots$.

Such that $y_0(i) \ge \sup_P |r_P(i)|$ *and* $y_0(i) \ge 1$ *for all* $i \in E$ *and for* $m = 0,1,\ldots$

$$y_m + {}_0P y_{m+1} \le y_{m+1}, \tag{2.0.1}$$

for all $P \in \mathcal{P}$ *and*

$$Py_m \text{ is continuous in } P, \tag{2.0.2}$$

where ${}_0P$ *is the matrix obtained from* P *by replacing the elements of the*

0-th *column by zeros i.e.*

$$ {}_0p(i,j) = \begin{cases} 0 & j = 0 \\ p(i,j) & j \neq 0. \end{cases} $$

For a finite state space the above assumption is equivalent to the condition that state 0 can be reached from each state under each stationary policy. For E denumerable we need that state 0 is positive recurrent under each stationary policy. More precisely (2.0.1) for m is equivalent to assuming that the supremum over all stationary policies of the total expected return, with immediate return in state i equal to $y_m(i)$, until reaching state 0 is finite. In fact, $y_{m+1}(i)$ can be taken as that supremum when starting state is i.

Assumptions of the above type were first introduced in HORDIJK [5]. As pointed out there (see sections 2.6, 2.7 and 5.12) for the special case that P consists of one element P and y_m is the unit vector e (i.e. e(i) = 1 for all $i \in E$) then condition (2.0.1) is closely related to a Foster-condition [4], and is equivalent to a condition called a Liapunov function criterion in KUSHNER [7].

The essential property what makes this condition work is the fact that y_{m+1} is a y_m-excessive function with respect to P (see chapter 2 of [5]).

Before we go on and use these conditions we want to point out that they are fulfilled in the following models (Without problems the reader can skip the subsections 2.1 and 2.2).

2.1. STATIONARY INVENTORY MODEL WITH BACKLOGGING

Let $\underline{y}_t$ denote the level of inventory at time t and let $\underline{\Delta}_t$ be the amount ordered after observing $\underline{y}_t$. Assume that delivery of the ordered units is instantaneous. Thus after the moment of ordering, the inventory level is $\underline{y}_t + \underline{\Delta}_t$. Suppose the sequence of demands $\underline{d}_t$, t = 1,2,..., for the product during each of the periods is a sequence of independent and identically distributed random variables with

$$\text{Prob. } [\underline{d}_t = j] = p_j \qquad \text{for } j = 0,1,\ldots \qquad \text{with } \sum_{j=0}^{\infty} p_j = 1.$$

We allow negative inventory, i.e. backlogging of demand, and consequently have a denumerable state space.

The decision which has to be made at times $t = 1,2,\ldots$ is the amount to be ordered. Now let $p_k(i,j)$ denote the transition probability to inventory level j when i units are available and k units are ordered. Then

$$p_k(i,j) = \text{Prob. [demand equals } i+k-j] = \begin{cases} p_{i+k-j} & \text{for } i + k \geq j \\ 0 & \text{otherwise.} \end{cases}$$

In all practical cases there will be a finite storage capacity. Also an infinitely large backlogging will not be convenient and so it seems that the following condition is natural. The set K(i) of available ordering decisions in state i satisfies.

(2.1.1) $\quad K(i) = \{k : a \leq i + k \leq b\}$ for all $i \in E$ for some integers a,b.

If moreover $p_j > 0$, $j = 0,1,\ldots$, then each stationary policy has no disjoint closed sets and state a is always accessible.

Now if we take state a as the special state of assumption (2.0.1) then it is straightforward to check that given y_m we can choose y_{m+1} as follows

$$y_{m+1} = y_m + \frac{1-q}{q} h_m ,$$

where $\quad q := \min_{0 \leq i \leq b-a} p_i$

and

$$h_m := \max\{ \sum_{j \neq i-a} p_{i-j}\, y_m(j) \Big(\sum_{j \neq i-a} p_{i-j} \Big)^{-1} : a \leq i \leq b\}.$$

Moreover,

$$h_m \le h_0\left(\frac{1+q}{q}\right)^{m-1} .$$

Note that since K(i) is finite for all $i \in E$ assumption (2.0.2) is trivially fulfilled.

2.2. WAITING LINE MODEL WITH CONTROLLABLE INPUT

Assume that the arrival process is a Poisson process with expected number of arrivals per unit time λ_p where p denotes the service price. Thus the input process can be controlled by the service price. It seems reasonable to assume that λ_p decreases as p increases. Let us assume further that the price p lies between the bounds p_1 and p_2, i.e. $p_1 \le p \le p_2$. Let F be the distribution of the service time $\underline{s}$. The times at which a decision on the price has to be taken are the times a person completes service. The state at that time is the number of people the departing customer leaves behind. We assume that the service time is independent of p.

The transition probabilities corresponding to price p equal

$$(2.2.1) \qquad p(i,j) = \begin{cases} 0 & \text{for } j < i-1, \\ k_{j-i+1}(p) & \text{for } j \ge i-1, \end{cases}$$

where $k_r(p)$ denotes the probability of r people arriving during service period, i.e.

$$(2.2.2) \qquad k_r(p) = \int_0^\infty e^{-\lambda_p s} (\lambda_p s)^r (r!)^{-1} \, dF(s).$$

For future reference we state that (2.2.2) implies

$$(2.2.3) \qquad \sum_{r=k}^{\infty} r(r-1)\ldots(r-k+1)\, k_r(p) = \lambda_p^k \, \mathbb{E}\underline{s}^k,$$

where it is assumed that $\mathbb{E}\underline{s}^k$ exists. Since $k_r(p)$, $r = 0,1,\ldots$, is a continuous function of λ_p it follows directly that $\mathcal{P}$ is compact if λ_p is a continuous function of p.

The following assumptions are made:

(2.2.4) $\rho^{-1} := 1 - \lambda_{p_1} \mathbb{E}\, \underline{s} > 0,$

(2.2.5) λ_p is a continuous function of p for $p_1 \le p \le p_2$,

(2.2.6) $r_p(i)$ is a continuous function of P for all $i \in E$.

We denote $i^{(n)}$ for the factorial product $\prod_{k=0}^{n-1} (i-k)$.
Similar to the Binomium of Newton we have for integers x,y

$$(x+y)^{(n)} = \sum_{k=0}^{n} \binom{n}{k} x^{(k)} y^{(n-k)} .$$

For matrix P corresponding to price p and $i \ge 1$ (note that $p(0,j) = p(1,j)$ for all j)

(2.2.7)
$$\begin{aligned} \sum_{j=1}^{\infty} p(i,j) j^{(n)} &= \sum_{j=i-1}^{\infty} k_{j-i+1}(p)\, j^{(n)} \\ &= \sum_{r=0}^{\infty} k_r(p) (r+i-1)^{(n)} \\ &= \sum_{k=0}^{n} \sum_{r=k}^{\infty} k_r(p) \binom{n}{k} r^{(k)} (i-1)^{(n-k)} \\ &= \sum_{k=0}^{n} \binom{n}{k} \mu_k(p) (i-1)^{(n-k)}, \end{aligned}$$

where from (2.2.3) $\mu_k(p) = \lambda_p^k \mathbb{E}\, \underline{s}^k$

With

$$i^{(k)} - (i-1)^{(k)} = k(i-1)^{(k-1)}$$

we find that

(2.2.8) $$\sum_{k=0}^{n} \binom{n}{k} \mu_k(p) (i-1)^{(n-k)} = \sum_{k=0}^{n} \binom{n}{k} \mu_k^*(p) i^{(n-k)} ,$$

where

$$\mu_0^*(p) = \mu_0(p) = 1 \text{ and } \mu_k^*(p) = \mu_k - k\mu_{k-1}^*(p).$$

Assume that for some $a > \rho$

$$\mu_{k+2}^*(P) \le a^k\, k!$$

Then

$$\mu_{k+2}^*(p) \le (an)^{(k)} \quad \text{for } n \ge k.$$

With induction we prove for $k = 1,2,\ldots$

$$(2.2.9) \qquad i^{(k)} + \sum_{j=1}^{\infty} p(i,j)\, f_k(j) \le f_k(i),$$

for some function f_k with

$$f_k(i) \le c_{k+1}\cdot((k+1)ai)^{(k+1)}$$

and for the c_k's some nondecreasing sequence of constants.

The proof of (2.2.9) for k=1 is similar to the one given below and will be omitted. Now assume (2.2.9) is true for $k = 1,2,\ldots, n-2$.
Then from (2.2.7) and (2.2.8),

$$(2.2.10) \qquad -n\mu_1^*(p)\, i^{(n-1)} - \sum_{k=2}^{n} \binom{n}{k} \mu_k^*(p)\, i^{(n-2)} + \sum_{j=1}^{\infty} p(i,j) j^{(n)} = i^{(n)}.$$

From (2.2.9) and $\mu_{k+2}^*(p) \le (an)^{(k)}$

$$(2.2.11) \qquad \sum_{k=2}^{n} \binom{n}{k} \mu_k^*(p) i^{(n-k)} + \sum_{j=1}^{\infty} p(i,j) f(j) \le f(j),$$

for some function f with $f(i) \le c_n\Big[\sum_{k=2}^{n} \binom{n}{k} (an)^{(k-2)}\ ((n+1-k)ai)^{(n+1-k)}\Big]$

Summing the inequalities (2.2.10) and (2.2.11)(i.e. left side plus left side and right side plus right side) and multiplying by ρn^{-1}, note that from assumption (2.2.4) $\rho n^{-1}(-n\mu_1^*(p)) \ge 1$, gives

$$i^{(n-1)} + \sum_{j=1}^{\infty} p(i,j)f^*(j) \leq f^*(i),$$

where

$$f^*(i) = \rho n^{-1}[i^{(n)} + f(i)].$$

Hence

$$f^*(i) \leq \rho n^{-1} c_n [i^{(n)} + n \sum_{k=2}^{n} \binom{n-1}{k-2} (an)^{(k-2)} ((n-1)ai)^{(n-1-(k-2))}]$$

$$\leq \rho n^{-1} c_n [i^{(n)} + n((n-1)ai+an)^{(n-1)}] \qquad (2.2.12)$$

$$\leq c_n (nai)^{(n)} \quad \text{for } i \geq n \geq 2.$$

For c_{n+1} we can take the maximum of (2.2.12) for $i = 1,\ldots, n-1$. To conclude it is straightforward from relation (2.2.9) that the assumptions (2.0.1) and (2.0.2) are satisfied when $\sup_P |r_P(i)|$ as function of i is bounded by some polynomial in i.

3. DISCOUNTED EXPECTED AND AVERAGE EXPECTED RETURN

In this section we focus attention on discounted and average expected return. We do not need the assumptions of section 2 in all strength in this section. It is sufficient to assume that relations (2.0.1) and (2.0.2) are true for $m = 0$ and $m = 1$.

For the vector with i-th *component the sum of all expected discounted returns when starting in state* i *and using policy* $R = (P_1, P_2, \ldots)$ *we write*

$$v_R^{\alpha} = \sum_{t=1}^{\infty} \alpha^{t-1} P_1 \ldots P_{t-1} r_{P_t},$$

where $0 < \alpha < 1$ *is the discount factor.*

We shall prove below that the above sum is absolutely convergent and so v_R^{α} is properly defined.

3.1. LEMMA. *From*

$$(3.1.1) \qquad x + {}_0P y \leq y,$$

for $x \geq 0$ *and for all* $P \in \mathcal{P}$ *it follows for arbitrary policy* $R = (P_1, P_2, \ldots)$ *that the total expected return (with immediate return vector* x*) until state* 0 *is reached is bounded by* y *i.e.*

$$(3.1.2) \qquad \sum_{t=1}^{\infty} {}_0P_1 \cdots {}_0P_{t-1} x \leq y$$

PROOF. Iterating the inequality

$$(3.1.3) \qquad x + {}_0P y \leq y$$

successively for $P_T, P_{T-1}, \ldots, P_1$ we obtain

$$\sum_{t=1}^{T} {}_0P_1 \cdots P_{t-1} x + {}_0P_1 \cdots {}_0P_T y \leq y.$$

Since $y \geq 0$ and hence ${}_0P_1 \cdots {}_0P_T y \geq 0$ for all T, we find as $T \to \infty$ relation (3.1.2). □

Using the last exit decomposition of state 0 (see CHUNG [2] p.46) the above stated absolute convergence follows now easily. Indeed,

3.2. LEMMA. *For any policy* $R = (P_1, P_2, \ldots)$ *and any* $i \in E$,

$$(3.2.1) \qquad \sum_{t=1}^{\infty} \alpha^{t-1} P_1 \cdots P_{t-1} |r_{P_t}|(i) \leq (1-\alpha)^{-1} y_1(0)$$

PROOF. From relation (2.0.1) with $m = 0$ and lemma 3.1 we conclude that for any policy $R = (P_1, P_2, \ldots)$

$$\sum_{t=k}^{\infty} {}_0P_k \cdots {}_0P_{t-1} |r_{P_t}|(0) \leq y_1(0).$$

Hence for any $i \in E$, using the last exit decomposition of state 0

$$\sum_{t=1}^{\infty} \alpha^{t-1} P_1 \ldots P_{t-1} |r_P|(i) =$$

$$\sum_{t=1}^{\infty} \sum_{k=1}^{t} \alpha^{t-1} P_1 \ldots P_{k-1}(i,0)_0 P_k \ldots {}_0P_{t-1} |r_{P_t}|(0)$$

$$\sum_{k=1}^{\infty} \sum_{t=k}^{\infty} \ldots \leq (1-\alpha)^{-1} y_1(0). \quad \square$$

For policy $R = (P_1, P_2, \ldots)$ *the vector of average expected return is defined as*

$$(3.2.3) \qquad g_R = \limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} P_1 \ldots P_{t-1} r_{P_t}.$$

Again the cesaro limit g_R is properly defined, since,

3.3. LEMMA. *For any policy* $R = (P_1, P_2, \ldots)$ *and any* $i \in E$,

$$(3.3.1) \qquad P_1 \ldots P_{T-1} |r_{P_T}|(i) \leq T y_1(0)$$

The proof uses similar arguments as in lemma 3.2.

Since v_R^α is properly defined we can introduce the components wise supremum ,

$$(3.3.2) \qquad v^\alpha = \sup_R v_R^\alpha.$$

From (3.2.1) we have

$$(3.3.3) \qquad |v^\alpha(i)| \leq (1-\alpha)^{-1} y_1(0) < \infty \quad \text{for all } i.$$

Under more general assumptions it can be shown that v^α satisfies the *optimality equation* (for a proof see [5])

$$(3.3.4) \qquad v^\alpha = \sup_{P\in\mathcal{P}} [r_P + \alpha P v^\alpha]$$

We assumed in the introduction that r_P is continuous. From (3.3.3) and

$Pe \leq Py_0$ and Py_0 is continuous in P, we conclude that also Pv^α is continuous. Since a continuous function has a maximal value on a compact set we obtain for certain $Q \in P$

$$(3.3.5) \qquad r_Q + \alpha Q v^\alpha = v^\alpha,$$

Such a Q is called v^α-*conserving*.

3.4. THEOREM. *For v^α-conserving matrix Q it holds that Q^∞ is α-discounted optimal.*

PROOF. For any T, iterating the equality (3.3.5) gives

$$\sum_{t=1}^{T} \alpha^{t-1} Q^{t-1} r_Q + \alpha^T Q^T v^\alpha = v^\alpha.$$

With (3.3.3) we can similar to relation (3.3.1) deduce that

$$(Q^T v^\alpha(i) \leq T(1-\alpha)^{-1} y_2(0).$$

Hence

$$\lim_{T\to\infty} \alpha^T Q^T v^\alpha = 0$$

and

$$\sum_{t=1}^{\infty} \alpha^{t-1} Q^{t-1} r_Q = v^\alpha$$

which completes the proof. □

Since v^α-conserving matrices do exist we conclude that the existence of a stationary policy which is optimal with respect to the α-discounted expected return within the class of all policies is guaranteed under our assumptions. Moreover, such a policy can be obtained from a v^α-conserving decision rule.

Next, we want to establish a solution of the optimality equation for

the average expected return. The technique we shall use is originally due to TAYLOR [11] and further developed by ROSS [8]. First we need a technical lemma.

3.5. LEMMA. *For all* $0 < \alpha < 1$ *the following inequalities are true*

$$(3.5.1) \qquad (1-\alpha)|v^{\alpha}(0)| \leq y_1(0)$$

$$(3.5.2) \qquad |v^{\alpha}(i) - v^{\alpha}(0)| \leq (1+y_1(0))y_1(i)$$

PROOF. Inequality (3.5.1) is immediate from (3.2.1).

Let $\underline{\tau}$ be the entry time of state 0 after time 1 i.e.

$$(3.5.3) \qquad \underline{\tau} = \min_{k\geq 2} \{\underline{x}_k = 0\}.$$

Then $\underline{\tau}$ is a Markov time. For Q a v^{α}-conserving policy we have that v^{α} is a potential with respect to substochastic matrix αQ. Hence from a well known theorem in Markov potential theory (see chapter 2 of [5]),

$$(3.5.4) \qquad v^{\alpha} = \mathbb{E}_{Q^{\infty}}\left[\sum_{t=1}^{\underline{\tau}-1} r(\underline{x}_t) + \mathbb{E}_{Q^{\infty}} v^{\alpha}(\underline{x}_{\underline{\tau}})\right].$$

Or in a different notation

$$(3.5.5) \qquad v^{\alpha} = \sum_{t=1}^{\infty} \alpha^{t-1}\, {}_0Q^{t-1}\, r_Q + \mathbb{E}_{Q^{\infty}} \alpha^{\underline{\tau}-1}\, v^{\alpha}(0).$$

Hence from (3.1.2) with y_0 for x and y_1 for y, for any $i \in E$

$$(3.5.6) \qquad |v^{\alpha}(i) - v^{\alpha}(0)| \leq y_1(i) + (1-\mathbb{E}_{Q^{\infty}} \alpha^{\underline{\tau}-1})|v^{\alpha}(0)|.$$

Further by the well known inequality $(1-\alpha^T) \leq (1-\alpha)\cdot T$

$$(3.5.7) \qquad (1-\mathbb{E}_{Q^{\infty}} \alpha^{\underline{\tau}-1})|v^{\alpha}(0)| \leq (1-\alpha)|v^{\alpha}(0)|\mathbb{E}_{Q^{\infty}}(\underline{\tau}-1).$$

The expected time until entering state 0 can be seen as a total expected return until entering state 0 with as immediate return vector the unit vector (see [5] 2.7).

Hence from (3.1.1) with e for x we find form the assumption in section 2 that

(3.5.8) $\quad \mathbb{E}_Q \underline{\tau} \leq y_1.$

Combination of the inequalities (3.5.1), (3.5.6), (3.5.7) and (3.5.8) yields (3.5.2). □

3.6. OPTIMALITY EQUATION FOR AVERAGE EXPECTED RETURN

Equation (3.3.4) specified for the i-th component gives

$$(3.6.1) \qquad v^\alpha(i) = \sup_{P\in\mathcal{P}} [r_P(i) + \alpha \sum_j p(i,j)v^\alpha(j)].$$

Consequently by subtracting $v^\alpha(0)$ from both sides

$$v^\alpha(i) - v^\alpha(0) = \sup_{P\in\mathcal{P}} [r_P(i) - (1-\alpha)v^\alpha(0) + \alpha \sum_j p(i,j)(v^\alpha(j) - v^\alpha(0))].$$

Since lemma (3.5) implies the boundedness of $|v^\alpha(i) - v^\alpha(0)|$ for $i \in E$ and $(1-\alpha)v^\alpha(0)$ as function of α, the diagonal procedure provides a sequence $\{\alpha_n\}$ with $0 < \alpha_n < 1$, $\alpha_n \to 1$ as $n \to \infty$ and a constant vector i.e. all components are equal, g together with a function v such that

$$\lim_{n\to\infty} (1-\alpha_n)v^{\alpha_n}(0) = g(0) \text{ and } \lim_{n\to\infty} v^{\alpha_n}(i) - v^{\alpha_n}(0) = v(i).$$

Moreover, from (3.5.2)

$$(3.6.3) \qquad |v| \leq (1+y_1(0))y_1$$

and hence with the dominated convergence theorem

$$\lim_{n\to\infty} \sum_j p(i,j)(v^{\alpha_n}(j) - v^{\alpha_n}(0)) = \sum_j p(i,j)v(j).$$

With (3.6.1) we conclude

$$v \geq \sup_{P \in \mathcal{P}} [r_P - g + Pv].$$

Moreover, since Py_1 is a continuous function of P it follows from a generalized dominiated convergence theorem (see ROYDEN [9] proposition 18 p.231) that for Q such that $Q = \lim_{n\to\infty} Q_{\alpha_n}$ and Q_α is v^α-conserving hence

$$v^{\alpha_n}(i) - v^{\alpha_n}(0) = r_{Q_{\alpha_n}}(i) - (1-\alpha_n)v^{\alpha_n}(0) + \alpha_n \sum_j q_{\alpha_n}(i,j)(v^{\alpha_n}(j)-v^{\alpha_n}(0)),$$

the following equation holds

$$v = r_Q - g + Qv.$$

Such a Q we call (g,v)-*conserving*.

By starting from the beginning with a suitable subsequence of discount factors tending to 1 we conclude:
Each limitpoint as α *tends to one of* v^α*-conserving decision rules is* (g,v)*-conserving* (cf. lemma 3.10).

Before we can prove that a (g,v)-conserving decision rule provides a stationary average optimal policy, we need two technical lemmas:

3.7. LEMMA. *Let* $x_0 := y_1$
and

$$x_{n+1} = \max_{P \in \mathcal{P}} {}_0Px_n$$

then x_n *is a decreasing sequence and*

$$\lim_{n\to\infty} x_n = 0.$$

PROOF. Since from assumption (2.0.!) $y_0 + {}_0Py_1 \leq y_1$ for all $P \in \mathcal{P}$ and $y_0 > 0$ we have that

$$x_1 = \max_{P \in \mathcal{P}} {}_0Py_1 \leq y_1 = x_0.$$

Now suppose $x_n \leq x_{n-1}$ then ${}_0P x_n \leq {}_0P x_{n-1}$ for all $P \in \mathcal{P}$ and hence

$$x_{n+1} = \max_{P\in\mathcal{P}} {}_0P x_n \leq \max_{P\in\mathcal{P}} {}_0P x_{n-1} = x_n.$$

Thus by induction x_n, $n = 0,1,\ldots$, is a decreasing sequence. Consequently

$$x := \lim_{n\to\infty} x_n$$

exists. Let P_n be such that $x_{n+1} = {}_0P_n x_n$ and subsequence n_k such that P_{n_k} has a limit say P. Then again using a generalized bounded convergence theorem we conclude that

$$x = {}_0Px.$$

Hence for all T

$$x = {}_0P^T x.$$

However

$$\sum_{t=1}^{\infty} {}_0P^{t-1} x \leq \sum_{t=1}^{\infty} {}_0P^{t-1} y_1 \leq y_2.$$

Hence

$$x = \lim_{T\to\infty} P^T x = 0. \quad \square$$

3.8. LEMMA. *For any policy* $R = (P_1, P_2, \ldots)$ *it holds*

$$\lim_{T\to\infty} P_1 \ldots P_T \, y_1/T = 0 \tag{3.8.1}$$

PROOF. Using again the last exit decomposition of state 0 we find with lemma 3.7

$$(3.8.2)\qquad P_1 \ldots P_T\, y_1(i) = \sum_{k=1}^{T+1} P_1 \ldots P_{k-1}(i,0)\; {}_0P_k \ldots {}_0P_T\, y_1(0)$$

$$\leq \sum_{k=1}^{T+1} P_1 \ldots P_{k-1}(i,0) x_{T-k}(0)$$

Using a lemma on Nörlund-means (cf. [2] p.22) we find

$$(3.8.3)\qquad \lim_{T\to\infty} \frac{\sum_{k=1}^{T+1} P_1 \ldots P_{k-1}(i,0) x_{T-k}(0)}{\sum_{k=1}^{T+1} P_1 \ldots P_{k-1}(i,0)} = \lim_{T\to\infty} x_T(0) = 0.$$

Combination of (3.8.2), (3.8.3) with the inequality

$$\sum_{k=1}^{T+1} P_1 \ldots P_{k-1}(i,0) \leq T + 1$$

implies relation (3.8.1). □

3.9. THEOREM. *For* (g,v)*-conserving decision rule* Q *it holds that* Q^∞ *has maximal average expected return. Moreover,*

$$g_{Q^\infty} = g$$

PROOF. For Q the equality $r_Q - g + Qv = v$ holds. Iterating this equality T times gives

$$\sum_{t=1}^{T} Q^{t-1} r_Q - g.T + Q^{T+1}v = v.$$

With lemma 3.8 and (3.6.3) we find

$$\lim_{T\to\infty} T^{-1} \sum_{t=1}^{T} Q^{t-1} r_Q = g.$$

Further for arbitrary policy $R = (P_1, P_2, \ldots)$ we find by interating the inequality

$$r_P - g + Pv \leq v$$

successively for P_T, $P_{T-1},\ldots,P_1$ that

$$\sum_{t=1}^{T} P_1 \ldots P_{t-1} r_{P_t} - g.T + P_1 \ldots P_T v \leq v.$$

Again with lemma 3.8 and (3.6.3) we obtain

$$\limsup_{T\to\infty} T^{-1} \sum_{t=1}^{T} P_1 \ldots P_{t-1}\, r_{P_t} \leq g. \quad \square$$

Since each limitpoint as α tends to one of v^α-conserving decision rules is (g,v)-conserving we conclude with the above theorem.

Each limitpoint as α tends to 1 of discounted-optimal policies is average-optimal.

As we shall show in the next section it is under general conditions even bias-optimal (or equivalently 0-discount optimal).

In subsection 3.6 we obtained v(i) as the limit of $v^\alpha(i) - v^\alpha(0)$ as α tends to one through certain sequence α_n. Actually the limit does exist.

3.10. LEMMA.

$$v(i) = \lim_{\alpha\uparrow 1} [v^\alpha(i) - v^\alpha(0)].$$

PROOF. Suppose w is another limit vector. Then from 3.6 again

$$|w| \leq (1+y_1(0))y_1$$

and with 3.6 and theorem 3.9

$$(3.10.1) \qquad w = \max_{P\in\mathcal{P}} (r_P - g + Pw).$$

Since v(0) = w(0) = 0 we have for Q a (g,v)-conserving rule

$$w \geq r_Q - g + {}_0Qw$$

and

$$v = r_Q - g + {}_0Qv.$$

Hence

$$v - w \leq {}_0Q(v-w).$$

Similar for Q^* a (g,w)-conserving rule

$$w - v \leq {}_0Q^*(w-v).$$

Consequently

$$|v - w| \leq \max_{P\in\mathcal{P}} {}_0P|v - w|.$$

Now from $|v - w| \leq 2(1+y_1(0))y_1$ and lemma 3.7 we conclude that $v - w = 0$. So all limit vectors are equal and the proof is complete. □

4. LAURENT EXPANSION AND SENSITIVE OPTIMALITY CRITERIA

In this section we focus on the discounted expected return for discountfactors α near 1 or equivalently small interest rates ρ ($\rho=(1-\alpha)\alpha^{-1}$ or $\alpha = (1+\rho)^{-1}$). Under the assumptions of section 2 we can expand the discounted expected return for stationary policies as a Laurent series in powers of ρ. The existence of n-discount optimal policies will be shown. Moreover, limit points as α-tends to 1 of discounted optimal decision rules are 0-discount optimal, at least in the class of stationary policies.

We first start with a technical lemma, which with a different proof can also be found in [6].

4.1. LEMMA. *There is a sequence of constant vectors* $g_0,g_1,\ldots$, *with* $|g_m(0)| \leq \Pi_{k=1}^{m}(1+y_k(0))y_{m+1}(0)$ *a sequence of vectors* $v_0,v_1,\ldots$, *with* $|v_m| \leq \Pi_{k=1}^{m+1}(1+y_k(0))y_{m+1}$, *and a monotone decreasing sequence of nonempty compact subsets of* P *say* $P = P_{-1}, P_0, P_1,\ldots$ *such that for*

$$(4.1.1) \qquad \psi_p^0 := r_p - g_0 + Pv_0 - v_0$$

and

$$(4.1.2) \qquad \psi_P^m := -v_{m-1} - g_m + Pv_m - v_m, \qquad m = 1,2,\ldots$$

it holds that

$$(4.1.3) \qquad \psi_P^m = 0 \qquad \text{for } P \in \mathcal{P}_m$$

and

$$(4.1.4) \qquad \max_{P\in\mathcal{P}_{m-1}} \psi_P^m = 0.$$

PROOF. The proof proceeds by induction on m. The g, v we found in section 3 suffice as g_0, v_0. Define

$$\mathcal{P}_0 = \{P \in \mathcal{P}: \; r_P - g_0 + Pv_0 = v_0\}$$

Since (g,v)-conserving policies do exist, we have that $\mathcal{P}_0$ is not empty. It is easily seen that $\mathcal{P}_0$ is closed and hence as a closed subset of a compact metric space again compact.

Assume g_0, v_0, g_1, $v_1,\ldots,g_{m-1}$, v_{m-1} and $\mathcal{P}_1$, $\mathcal{P}_0,\ldots,\mathcal{P}_{m-1}$ are found. The way of constructing g_m and v_m is strictly similar to that of g,v in section 3. In short we will repeat the various steps of finding g_m and v_m.

Introduce

$$v_m^\alpha = \sup_{P\in\mathcal{P}_{m-1}} -\sum_{t=1}^{\infty} \alpha^{t-1}P^{t-1}v_{m-1}.$$

Since for all $i \in E$

$$(4.1.5) \qquad |v_{m-1}(i)| \le \prod_{k=1}^{m-1} (1+y_k(0))y_m(i)$$

we find similar to (3.2.1) that

$$(4.1.6) \qquad |v_m^\alpha(i)| \le (1-\alpha)^{-1} \prod_{k=1}^{m} (1+y_k(0))y_{m+1}(i).$$

Moreover, similar to (3.5.2)

$$(4.1.7) \qquad |v_m^\alpha(i) - v_m^\alpha(0)| \le \prod_{k=1}^{m+1} (1+y_k(0))y_{m+1}(i).$$

Since v_m^α is a bounded vector, Pv_m^α is continuous in P and hence the optimality equation reads

$$v_m^\alpha = \max_{P\in\mathcal{P}_{m-1}} [-v_{m-1} + \alpha P v_m^\alpha].$$

Rewriting this as in 3.6 gives

$$v_m^\alpha(i) - v_m^\alpha(0) = \max_{P\in\mathcal{P}_{m-1}} [-v_{m-1}(i) - (1-\alpha)v_m^\alpha(0) + \sum_j p(i,j)(v_m^\alpha(j)-v_m^\alpha(0))].$$

Now choose sequence α_n of discountfactors tending to 1 (possible, from (4.1.6) and (4.1.7) such that

$$g_m(i) := \lim_{n\to\infty} (1-\alpha_n)v_m^\alpha(0)$$

and

$$v_m(i) := \lim_{n\to\infty} (v_m^\alpha(i)-v_m^\alpha(0)).$$

Then

$$v_m = \max_{P\in\mathcal{P}_{m-1}} [-v_{m-1} - g_m + Pv_m]$$

which is relation (4.1.4) for m.

Define

$$\mathcal{P}_m = (P\in\mathcal{P}_{m-1}: -v_{m-1}-g_m+Pv_m-v_m=0)$$

then $\mathcal{P}_m$ is a nonvoid closed subset of the compact set $\mathcal{P}_{m-1}$. □

4.2. In HORDIJK & SLADKY [6] for M any integer the following partial Laurent expansion is derived for the total discounted expected return

$$(4.2.1)\qquad \alpha v_R^{\alpha} = \rho^{-1} g_0 + \sum_{m=0}^{M-1} \rho^m [u_m + \sum_{t=1}^{\infty} \alpha^t P_R^{t-1} \psi_{P_t}^m] + O(\rho^M),$$

where $R = (P_1, P_2, \ldots)$, $\rho = \alpha^{-1}(1-\alpha)$ and $u_{m-1} = v_{m-1} + g_m$.

Following VEINOTT [12] we say that policy R^* is n-*discount optimal* with $n = -1, 0, 1, 2, \ldots$, if

$$(4.2.2)\qquad \liminf_{\alpha \uparrow 1} (1-\alpha)^{-n} [v_{R^*}^{\alpha} - v_R^{\alpha}] \geq 0,$$

for each policy R.

Let $v_T^1(R)$ denote the vector of expected returns under policy $R = (P_1, P_2, \ldots)$ up to time T i.e.

$$(4.2.3)\qquad v_T^1(R) = \sum_{t=1}^{T} P_R^{t-1} r_{P_t},$$

and define recursively for $n \geq 1$

$$(4.2.4)\qquad v_T^{n+1}(R) = \sum_{t=1}^{T} v_t^n(R).$$

Again following Veinott we call policy R^* n-*average optimal* if

$$(4.2.5)\qquad \liminf_{T \to \infty} \frac{1}{T} [v_T^{n+2}(R^*) - v_T^{n+2}(R)] \geq 0$$

for each policy R.

Using Laurent expansion (4.2.1) we proved in [6] when the number of actions is finite in each state then policy R^* is n-discount optimal if and only if it is n-average optimal. Moreover, for decision-rule $P \in P_{n+1}$ it holds that stationary policy P^{∞} is n-discount optimal. Note that (-1)-average optimality is optimality with respect to the average expected return. In section 3 we proved that limits of discounted optimal decision rules, as the discountfactor tends to one, are average optimal. It is easily checked that those limits are elements of P_0 and hence we could have used the above cited results by proving the average optimality of these limits. However, via the approach of section 3 we get rid of the assumption $P(i)$ finite for all $i \in E$.

In the sequel of this section we shall prove that those limits are 0-discount optimal in the class of stationary policies. From the results of [6] we know then in many cases they are in fact 0-discount optimal in the class of all policies. In the literature 0-discount optimal is also called bias optimal or 1-optimal the equivalent criterion 0-average optimal is also called average overtaken optimal.

We first need some technical lemmas.

4.3. LEMMA. *The vectors* g_n, v_n, $n = 0,1,2,\ldots$ *are unique. If for some* n *and vectors* h *and* w *with* h *constant vector,* $w(0) = 0$ *and* $|w| \leq cy_k$ *for some constant* c *and integer* k,

$$(4.3.1) \qquad w = \max_{P \in \mathcal{P}_n} [-v_n - h + Pw]$$

then

$$h = g_{n+1}$$

and

$$w = v_{n+1} .$$

PROOF. Similar to theorem 3.9 it holds that

$$g_{n+1} = h = \sup_{P \in \mathcal{P}_n} \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} P^{t-1}(-v_n).$$

The proof that $v_{n+1} = w$ proceeds similar to the proof that the solution w of (3.10.1) is unique in lemma 3.10. □

Also similar to 3.10 we have for any n

$$v_n(i) = \lim_{\alpha \uparrow 1} [v_n^{\alpha}(i) - v_n^{\alpha}(0)].$$

Since we know now that g_n, v_n are essentially unique we can express them also in a different way.

4.4. LEMMA. *Define constant vector* h *as*

$$(4.4.1) \qquad h(i) := \sup_{R \in \mathcal{R}_n} \frac{\sum_{t=1}^{\infty} {}_0P_R^{t-1}(-v_n)(0)}{\sum_{t=1}^{\infty} {}_0P_R^{t-1}e(0)} \quad ,$$

with $R = (P_1, P_2, \ldots) \in \mathcal{R}_n$ *if* $P_k \in \mathcal{P}_n$ *for* $k = 1,2,\ldots$ *and define vector* w *as*

$$w := \sup_{R \in \mathcal{R}_n} \sum_{t=1}^{\infty} {}_0P_R^{t-1}(-v_n - h).$$

Then $h = g_{n+1}$ *and* $w = v_{n+1}$.

PROOF. By straightforward verification it can be shown that (h,v) satisfies the conditions of lemma 4.3. □

A well known mean ergodic theorem says that the average expected return is equal to the expected return until reaching state 0 divided by the expected time until reaching state 0. Hence that the right hand of (4.4.1) is equal to the maximal average expected return is not surprising.

4.5. LEMMA. *For any* $P \in \mathcal{P}$ *there exist a sequence of vectors*

$$g_0(P),\ u_0(P),\ g_1(P),\ u_0(P),\ldots$$

such that all g's *are constant vectors,*

$$(4.5.1) \qquad r_P - g_0(P) + Pu_0(P) = u_0(P)$$

and

$$(4.5.2) \qquad -u_n(P) + Pu_{n+1}(P) = u_{n+1}(P) \qquad \text{for } n = 0,1,\ldots$$

Moreover, $g_n(P)$, $u_n(P)$, $n = 0,1,\ldots$ *are continuous as functions of* P.

PROOF. Apply lemma 4.1 for the specialized case that $\mathcal{P}$ consists of one element P i.e. $\mathcal{P} = \{P\}$. It is clear then that for the sequence $g_0(P)$, $v_0(P)$, $g_1(P)$, $v_1(P),\ldots$, now depending on P, holds that

$$-v_m(P) - g_{m+1}(P) + Pv_{m+1}(P) = v_{m+1}(P) \quad \text{for } m = 1,2,\ldots .$$

Now defining $u_m(P) = v_m(P) + g_{m+1}(P)$ we obtain (4.5.2). From lemma 4.4 we know that

$$g_{m+1}(P)(0) = \frac{\sum_{t=1}^{\infty} {}_0P^{t-1}(-v_m(P))(0)}{\sum_{t=1}^{\infty} {}_0P^{t-1}e(0)} .$$

Since from lemma 4.1

$$|v_m(P)| \le \prod_{k=1}^{m+1} (1+y_k(0))y_{m+1}$$

and since

$$\sum_{t=T+1}^{\infty} {}_0P^{t-1}y_{m+1} \le {}_0P^{T}y_{m+2}$$

it follows from lemma 3.7 with y_{m+2} for y_1 that for any $\varepsilon > 0$ there is an integer $T(\varepsilon)$ such that

$$\left| g_{m+1}(P)(0) - \frac{\sum_{t=1}^{T} {}_0P^{t-1}(-v_m(P))(0)}{\sum_{t=1}^{T} {}_0P^{t-1}e(0)} \right| \le \varepsilon, \quad \text{for all } P \in \mathcal{P}.$$

Hence $g_{m+1}(P)$ is continuous as function of P. With similar arguments it is straightforward to show that also $v_m(P)$ is continuous in P and so is $u_m(P)$. □

4.6. THEOREM. *For any matrix* $P \in \mathcal{P}$, *all integers* M

$$(4.6.1) \qquad \alpha v^{\alpha}_{P^{\infty}} = \sum_{t=1}^{\infty} \alpha^t P^{t-1} r_{P_t} = \rho^{-1} g_0(P) + \sum_{m=0}^{M-1} \rho^m u_m(P) + O(\rho^{M-1}).$$

PROOF. With lemma 4.5, and $\rho = \alpha^{-1}(1-\alpha)$,

$$(4.6.2) \qquad \sum_{t=1}^{\infty} \alpha^t P^{t-1} r_P = \sum_{t=1}^{\infty} \alpha^t P^{t-1}[r_P - g_0(P) + Pu_0(P) - u_0(P)] +$$

$$+ \rho^{-1} g_0(P) + u_0(P) - \rho \sum_{t=1}^{\infty} \alpha^t P^{t-1} u_0(p) ,$$

$$= \rho^{-1} g_0(P) + u_0(P) - \rho \sum_{t=1}^{\infty} \alpha^t P^{t-1} u_0(P) ,$$

and similar for any m = 0,1,...

$$(4.6.3) \qquad - \sum_{t=1}^{\infty} \alpha^t P^{t-1} u_m(P) = u_{m+1}(P) - \rho \sum_{t=1}^{\infty} \alpha^t P^{t-1} u_{m+1}(P).$$

Substitution of (4.6.3) for m = 0 in (4.6.2) then substitution of (4.6.3) for m = 1 in the result etc. gives (4.6.1) with restterm

$$(4.6.4) \qquad - \rho^M \sum_{t=1}^{\infty} \alpha^t P^{t-1} u_{M-1}(P).$$

Since

$$|u_{M-1}(P)| \le |v_{M-1}(P)| + |g_M(P)|$$

we find with lemma 4.1 that

$$\rho^M \sum_{t=1}^{\infty} \alpha^t P^{t-1} |u_{M-1}(P)| \le \rho^M (1-\alpha)^{-1} \prod_{k=1}^{M+1} (1+y_k(0)) y_{M+1}.$$

Hence the restterm is $O(\rho^{M-1})$ uniformly in $P \in \mathcal{P}$. □

It is said that vector $(x_1,\ldots,x_n)$ is *lexicographic larger than or equal* to vector $(y_1,\ldots,y_n)$ if the first nonzero element of $(x_1-y_1, x_2-y_2, \ldots, x_n-y_n)$ is positive.

From Laurent expansion (4.6.1) it is easily seen that P^∞ is n-discount optimal in the class of stationary policies if and only if for all $i \in E$ $(g_0(P)(i), u_0(P)(i), \ldots, u_n(P)(i))$ is lexicographic maximal.

From results of [6] we know that $P \in \mathcal{P}_{n+1}$ is n-discount optimal. From lemmas 4.1, 4.3 and 4.5 it follows for $P \in \mathcal{P}_{n+1}$ that $g_k(P) = g_k$, $k = 1,\ldots,n+1$ and $v_k(P) = v_k$, $k = 1,\ldots,n+1$ and hence $u_k(P) = u_k$, $k = 1,\ldots,n$. Combining these results we find that

$$(g_0, u_0, \ldots, u_n) = \max_{P \in \mathcal{P}} (g_0(P), u_0(P), \ldots, u_n(P)),$$

where the maximum is componentswise and lexicographic.

We can now give the final result of this section.

4.7. THEOREM. *If* P_n^∞ *is* α_n*-discounted optimal,* $\lim_{n\to\infty} \alpha_n = 1$ *and* $\lim_{n\to\infty} P_n = P_\infty$ *then*

$$(g_0(P_\infty), u_0(P_\infty)) = (g_0, u_0).$$

Hence P_∞ *is 0-discount or equivalently bias-optimal.*

PROOF. In section 3 we showed that $g(P_\infty) \geq g(P)$ for all $P \in \mathcal{P}$. Now for Q such that $g(Q) = g(P_\infty)$ we find with theorem (4.6) and the fact that P_n^∞ is α_n-discount optimal

$$\rho_n^{-1} g_0(P_n) + u_0(P_n) + \rho_n(u_1(P_n) + 0) \geq$$

$$\rho_n^{-1} g_0(Q) + u_0(Q) + \rho_n(u_1(Q) + 0),$$

where

$$\rho_n = \alpha_n^{-1}(1-\alpha_n).$$

With $g_0(P_n) \leq g_0(P_\infty) = g_0(Q)$ it follows then

$$u_0(P_n) - u_0(Q) \geq \rho_n(u_1(P_n) - u_1(Q) + 0).$$

Since $u_0(P)$ and $u_1(P)$ and continuous in P we find as n tends to infinity that $u_0(P_\infty) \geq u_0(Q)$. □

5. ITERATION PROCEDURE

In section 4 we discussed the existence of n-discount optimal policies. In BREIMAN [1] a device for computing average optimal policies in binary decision problems is given. In this section we give a similar iteration procedure for computing n-discount optimal policies.

Given v_P, $P \in \mathcal{P}^*$ with $|v_P| \leq cy_k$ for some integer k and some constant

c and compact product set $\mathcal{P}^* \subset \mathcal{P}$ compute for constant vector g

$$x_0(g) := \max_{P\in\mathcal{P}^*}(v_P - g)$$

and

$$x_{n+1}(g) := \max_{P\in\mathcal{P}^*}(v_P - g + {}_0P x_n(g))$$

then

$$x(g) := \lim_{n\to\infty} x_n(g)$$

exists. The limit x(g) depends continuously on g and there is exactly one g^* such that the zero component is zero i.e. $x(g^*)(0) = 0$. If for some g we find $x(g)(0) > 0$ then $g^* > g$ if $x(g)(0) < 0$ then $g^* < g$.

The pair $(g^*, x(g^*))$ is the unique solution (g,w) with g constant vector and $|w| \le c^* y_{k+1}$ for some constant c^* of the equation

$$w = \max_{P\in\mathcal{P}^*}(v_P - g + {}_0Pw). \tag{5.0.1}$$

Hence if $v_P = r_P$, $P \in \mathcal{P}^*$ and $\mathcal{P}^* = \mathcal{P}$ then as in section 3 and in lemma 4.3 $g^* = g_0$ and $x(g) = v_0$ and similar if $v_P = v_n$ and $\mathcal{P}^* = \mathcal{P}_n$ for some n then $g^* = g_{n+1}$ and $s(g^*) = v_{n+1}$.

So this scheme provides an iteration procedure to compute g_0, v_0 and also g_{n+1}, v_{n+1} and $\mathcal{P}_{n+1}$ when g_n, v_n and $\mathcal{P}_n$ are known.

To prove that $\lim_{n\to\infty} x_n(g)$ exists, let w be the solution of (5.0.1) i.e. (cf. lemma 4.4)

$$w = \sup_{R\in\mathcal{R}^*} \sum_{t=1}^{\infty} {}_0P_R^{t-1}(v_{P_t} - g)$$

then as in the proof of lemma 3.10

$$|x_{n+1}(g) - w| \le \max_{P\in\mathcal{P}^*} {}_0P|x_n(g) - w|.$$

Similar to the proof of lemma 3.7 we find then

$$\lim_{n\to\infty} x_n(g) = w.$$

As in lemmas 4.3 and 4.4

$$g^*(0) = \sup_{R\in R^*} \frac{\sum_{t=1}^{\infty} {}_0P_R^{t-1} v_{P_t}(0)}{\sum_{t=1}^{\infty} {}_0P_R^{t-1} e(0)}$$

Hence

$$w(0) = \sup_{R\in R^*} \sum_{t=1}^{\infty} {}_0P_R^{t-1}(v_{P_t} - g)(0) < 0$$

if $g > g^*$.

That x(g) is continuous in g follows again from the fact that x(g) can be approximated uniformly for all g's in any bounded interval by $x_T(g)$ i.e. for any $\varepsilon > 0$ and any state i and all g's such that $-\infty < g_1 \leq g(0) \leq g_2 < +\infty$ for any g_1, g_2 there is an integer T such that (cf. lemma 4.5)

$$|x(g)(i) - x_T(g)(i)| < \varepsilon.$$

REFERENCES

[1] BREIMAN, L. (1964), *"Stopping-rule problems"* in Applied Combinatorial Mathematics, E.F. Beckenbach (ed.), Wiley, New York.

[2] CHUNG, K.L. (1967), *Markov chains with stationary transition probabilities*, second edition, Springer, Berlin.

[3] DERMAN, C. & R. STRAUCH, (1966), *A note on memoryless rules for controlling sequential control processes*, Ann. Math. Statist. 37, 276-278.

[4] FOSTER, F.G. (1953), *On stochastic matrices associated with certain queueing processes*, Ann. Math. Statist. 24, 355-360.

[5] HORDIJK, A. (1974), *Dynamic programming and Markov potential theory*, Mathematical Centre Tract no. 51, Amsterdam.

[6] HORDIJK, A. & K. SLADKY, (1975), *Sensitive optimality criteria in countable state dynamic programming*, Mathematical Centre Report BW 48.

[7] KUSHNER, H. (1971), *Introduction to stochastic control*, Holt, Rinehart and Winston, New York.

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

[9] ROYDEN, H.L. (1968), *Real Analysis*, second edition, The Macmillan Company, London.

[10] STRAUCH, R. & A.F. VEINOTT, Jr. (1966), *A property of sequential control processes*, Rand McNally, Chicago, Illinois.

[11] TAYLOR, H.M. (1965), *Markovian sequential replacement processes*, Ann. Math. Statist. 36, 1677-1694.

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