stichting
mathematisch
centrum

MC

AFDELING MATHEMATISCHE BESLISKUNDE BW 48/75 SEPTEMBER

A. HORDIJK & K. SLADKÝ

SENSITIVE OPTIMALITY CRITERIA IN COUNTABLE STATE DYNAMIC PROGRAMMING

Prepublication

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

Sensitive optimality criteria in countable state dynamic programming *)

by

A. Hordijk & K. Sladký

ABSTRACT

Discrete time Markov decision processes with a countable state space are investigated. Under a condition of Liapunov function type the Laurent expansion of the total discounted expected return for the various policies is derived. Moreover, the equivalence of the sensitive optimality criteria as introduced by Veinott, is shown.

KEY WORDS & PHRASES: *Markov decision processes, discrete time, countable state space, equivalence sensitive optimality criteria, Lapunov function criterion.*

*) This paper is not for review; it is meant for publication in a journal.

1. INTRODUCTION AND SUMMARY

This paper investigates discrete time Markov decision processes with a countable state space and arbitrary decision sets. Under a condition of Liapunov function type introduced in section 2, we derive in section 3 the Laurent expansion of the total discounted expected return for the various policies. This extends the wellknown results of MILLER & VEINOTT [6] to the denumerable state case. In section 4 we give for each policy the asymptotic expansion of the m-fold summation of the infinite stream of expected returns. Using the results of sections 3 and 4 we prove in section 5 that a policy is n-discount optimal if and only if it is n-average optimal. This shows the equivalence of the sensitive optimality criteria as introduced by VEINOTT [10], [11], [12]. Section 5 extends results of LIPPMAN [4], MANDL [5], SLADKY [7] and VEINOTT [10], [11], [12]. Moreover, the results of section 5 guarantee the existence of stationary n-discount (n-average) optimal policies.

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 $P \in \mathcal{P}$. A memoryless or Markov policy R is sequence $P_1, P_2, \ldots \in \mathcal{P}$, 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 [2], generalized in STRAUCH & VEINOTT [8] that we do not loose generality by restricting the class of policies to the Markov policies, (see also section 13 of HORDIJK [3]. In this paper we shall only use Markov policies.

A memoryless policy which takes at all times the same decision rule, i.e., $P^\infty := (P,P,\ldots)$, $P \in \mathcal{P}$ 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 $P, Q \in \mathcal{P}$ with $p(i,.) = q(i,.)$ then $r_P(i) = r_Q(i)$.

The expectation of the cost 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

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

We shall use the notation P_R^{t-1} for the matrix $P_1 \cdots P_{t-1}$, where P_R^{t-1} is the unit matrix for 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 is a metric space. We assume that $\mathcal{P}$ is compact and r_P is continuous in P i.e. for each $i \in E$ the limit of $r_{P_m}(i)$ is $r_P(i)$ as P_n converges to P. Note that these assumptions are automatically fulfilled if $\mathcal{P}(i)$ is finite for all $i \in E$. For vectors x,y with i-th components x(i), y(i) we write $x \leq y$ resp. $x < y$ if $x(i) < y(i)$ for all $i \in E$ resp. $x(i) \leq y(i)$ for all i and $x(i) \neq y(i)$ for some i; for vectors x, x_n, n = 1,2,..., 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 $i \in E$.

For the vector with i-th component the sum of the expected discounted (to time zero) returns up to time T when starting in state i, using policy $R = (P_1,P_2,\ldots)$ and for discountfactor α, we write $v_T^{(\alpha)}(R)$, where α is the

interest rate i.e. $\rho = (1-\alpha)\alpha^{-1}$ or $\alpha = (1+\rho)^{-1}$. Hence

$$v_T^{\rho}(R) = \sum_{t=1}^{T} \alpha^t E_R\, r(\underset{\sim}{x}_t) = \sum_{t=1}^{T} \alpha^t P_R^{t-1}\, r_{P_t}.$$

Let $v^{\rho}(R)$ denote $\lim_{T\to\infty} v_T^{\rho}(R)$. Under the assumptions of section 2 all expectations, sums and limits which we use, exist and converge (cf. [3A] section 3). Following VEINOTT [11] we say that policy R^* is n-discount optimal with $n = -1,0,1,2,\ldots$, if

$$\liminf_{\rho\downarrow 0} \rho^{-n}[v^{\rho}(R^*) - v^{\rho}(R)] \geq 0,$$

for each policy R.
Let $v_T^1(R)$ denote the vector of expected returns under policy R up to time T i.e.

$$v_T^1(R) = \sum_{t=1}^{T} E_R\, r(\underset{\sim}{x}_t) = \sum_{t=1}^{T} P_R^{t-1}\, r_{P_t},$$

and define recursively for $n \geq 1$

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

Again following Veinott we call policy R^* n-average optimal with $n = -1,0,1,2,\ldots$, if

$$\liminf_{T\to\infty} \frac{1}{T}[v_T^{n+2}(R^*) - v_T^{n+2}(R)] \geq 0, \quad \text{for each policy R.}$$

2. ASSUMPTIONS AND PRELIMINARY RESULTS.

Throughout this paper, 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 $y_0(i) \geq \max_P |r_P(i)|$ and $y_0(i) \geq 1$ for all $i \in E$ and for $m = 0,1,\ldots$

$$(2.0.1) \qquad y_m + {}_0Py_{m+1} \leq y_{m+1}$$

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

(2.0.2) Py_m is continuous in P,

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. In HORDIJK [3] section 5 where this type of condition was introduced it is shown that for queuing models with y_0 a polynomial of degree 1 in state i then y_1, is a polynomial of degree 2. Similar y_m is a polynomial of degree m + 1.
Further in the case that $r_P(i)$ is bounded and the simultaneous Doeblin condition is satisfied then as is shown in HORDIJK [3] section 12.6 all y_m's are bounded vectors.

We conclude that the above assumptions are satisfied in an interesting class of countable state Markov decision processes, such as stationary inventory models with backlogging and waiting line models (see also HORDIJK [3A] sections 2.1 and 2.2).

2.1. THEOREM. *There are a sequence of vectors* $g,u_0,u_1,\ldots$, *with* g *a constant vector* $|u_m| \leq k_m y_m$ *for some constant* k_m *and a monotone decreasing sequence of nonempty compact subsets of* P *say* $P = P_{-1} \supset P_0 \supset P_1 \supset \ldots$ *such that for*

$$ (2.1.1) \qquad \overset{0}{\psi}_P := r_p - g + Pu_0 - u_0 $$

and

$$ (2.1.2) \qquad \psi^m_P := - u_{m-1} + Pu_m - u_m, \quad m = 1,2,\ldots $$

it holds that

(2.1.3) $\psi_P^m = 0$ *for* $P \in \mathcal{P}_m$

and

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

PROOF. The proof proceeds by induction on m. For the vectors g resp. u_0 we can take g identically equal to the g_0 of (5.4.3) in [3] and u_0 equal to the v of (5.4.5) in [3]. Suppose $g, u_0, u_1, \ldots, u_m$ and $\mathcal{P}_{-1} \supset \mathcal{P}_0 \supset \ldots \supset \mathcal{P}_m$ are found. The problem of finding u_{m+1} and $\mathcal{P}_{m+1}$ is again the problem studied in section 5.6 of [3]. For completeness we give here a slightly different proof of it.

For $R = (P_1, P_2, \ldots)$ an arbitrary policy we find by iterating the inequality

$$y_m + {}_0P y_{m+1} \leq y_{m+1}.$$

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

$$\sum_{t=1}^{T} {}_0P_1 \cdots {}_0P_{t-1}\, y_m + {}_0P^T\, y_{m+1} \leq y_{m+1}.$$

Since ${}_0P^T y_m \geq 0$ for all T it follows that

(2.1.5) $$\sum_{t=1}^{\infty} {}_0P_1 \cdots {}_0P_{t-1} y_m \leq y_{m+1}$$

The inequality $|u_m| \leq k_m y_m$ gives

(2.1.6) $$\sup_R \sum_{t=1}^{\infty} {}_0P_1 \cdots {}_0P_{t-1}\, |u_m| \leq k_m\, y_{m+1}.$$

Let e denote the unit vector i.e. all components equal to 1. Then $e \leq y_0 \leq y_m$ and consequently also

$$\sum_{t=1}^{\infty} {}_0P_1 \cdots {}_0P_{t-1}\, e \leq y_{m+1} < \infty,$$

for each policy $R = (P_1, P_2, \ldots)$.

Define constant

(2.1.7) $$g_m := \operatorname*{Sup}_{R \in \mathcal{R}_m} \frac{\sum_{t=1}^{\infty} {}_0P_1 \cdots {}_0P_{t-1}(-u_m)(0)}{\sum_{t=1}^{\infty} {}_0P_1 \cdots {}_0P_{t-1}\, e(0)}$$

with $R = (P_1,P_2,\ldots) \in \mathcal{R}_m$ if $P_k \in \mathcal{P}_m$ for $k = 1,2,\ldots$ and where ${}_0P_1\cdots{}_0P_{t-1}x(0)$ is the zero component of the vector ${}_0P_1\cdots{}_0P_{t-1}\,x$.

Define vector

$$(2.1.8) \qquad v := \sup_{R\in\mathcal{R}_m} \sum_{t=1}^{\infty} {}_0P_1\cdots{}_0P_{t-1}(-u_m - g_m e).$$

Then as a direct consequence of (2.1.7) we have that $v(0) = 0$.
Furthermore it is wellknown that v satisfies Bellman's optimality equation.

$$(2.1.9) \qquad v = \sup_{P\in\mathcal{P}_m} (-u_m - g_m e + {}_0Pv).$$

Now, since $v(0) = 0$ we can take the matrix P instead of ${}_0P$ in the right hand side of the above equality. Further if we take as new vector u_m the old one minus g_m times the unit vector then since $\Sigma_j\, p(i,j) = 1$ the relations (2.1.3) and (2.1.4) remain true for m. However, for the new vector u_m relation (2.1.9) reads with u_{m+1} for v

$$u_{m+1} = \sup_{P\in\mathcal{P}_m} (-u_m + Pu_{m+1}),$$

Hence

$$\psi_P^{m+1} \leq 0 \text{ for all } P \in \mathcal{P}_m.$$

Moreover, since $\mathcal{P}_m$ is compact and the right hand is continuous in P there are P's for which the right-hand side is maximal.

Define

$$\mathcal{P}_{m+1} = \{P \in \mathcal{P}_m : -u_m + Pu_{m+1} - u_{m+1} = 0\}.$$

Then $\mathcal{P}_{m+1}$ is nonempty and closed, and as a closed subset of a compact set in a metric space again compact.

Finally since $|u_m| \leq k_m y_m$ we have that the first vector u_{m+1} satisfies

$$|u_{m+1}| < (k_m+g_m)\, y_{m+1}. \quad \square$$

Using the inequality $|u_m| \leq k_m y_m$ we can derive the following inequality which we need in the sequel.

2.2. LEMMA. *Each policy* $R = (P_1, P_2, \ldots)$ *satisfies*

$$(2.2.1) \qquad \sum_{t=1}^{\infty} \alpha^t P_R^{t-1} |u_m| \leq \rho^{-1} k_m y_{m+1}(0)$$

and

$$(2.2.2) \qquad \lim_{T\to\infty} T^{-1} P_R^{T-1} |u_m| = 0.$$

PROOF. Using the last exit decomposition of state 0 before time t+1 (cf. CHUNG [1] p.46) it follows

$$(2.2.3) \qquad P_R^{t-1} y_m(i) = \sum_{k=1}^{t} P_1 \ldots P_{k-1}(i,0)\ {}_0P_k \cdots {}_0P_{t-1}\ y_m(0)$$

$$\leq \sum_{k=1}^{t} {}_0P_k \cdots {}_0P_{t-1}\ y_m(0) \leq y_{m+1}(0).$$

Hence,

$$\sum_{t=1}^{\infty} \alpha^t P_R^{t-1} |u_m| \leq k_m \sum_{t=1}^{\infty} \alpha^t y_{m+1}(0) = \rho^{-1} k_m y_{m+1}(0).$$

Relation (2.2.2) is obvious with (2.2.3). □

2.3. REMARK. For the specialized case that $\mathcal{P}$ consists of one element P i.e. $\mathcal{P} = \{P\}$ it is clear from theorem 2.1 that for the sequence $g(P)$, $u_0(P)$, $u_1(P), \ldots$, now depending on P, holds that $\psi_P^m = -u_{m-1}(P) + Pu_m(P) - u_m(P) \equiv 0$ for $m = 1,2,\ldots$.

3. LAURENT EXPANSION OF THE DISCOUNTED EXPECTED RETURN

In this section we focus on the discounted expected return for discount-factors α near 1 or small interest rates. Under the assumptions of section 2 we can expand the discounted expected return for the various interest rates as a Laurent series in powers of ρ in a neigbourhood of $\rho = 0$.

3.1. THEOREM. *For each policy* $R = (P_1, P_2, \ldots)$, *all* $M = 1,2,\ldots$ *and all* $T = 1,2,\ldots$ *it holds that*

$$(3.1.1)\quad v_T^{\rho}(R) = (1-\alpha^T)\rho^{-1}g + \sum_{m=0}^{M-1} \rho^m [u_m - \alpha^{T+m+1} P_R^{T+m} u_m + \sum_{t=1}^{T+m} \alpha^t P_R^{t-1} \psi_{P_t}^m] + \\ - \rho^M \sum_{t=1}^{T+M} \alpha^t P_R^{t-1} u_{M-1}.$$

PROOF.

$$(3.1.2)\quad v_T^{\rho}(R) = \sum_{t=1}^{T} \alpha^t P_R^{t-1} r_{P_t}$$

$$= \sum_{t=1}^{T} \alpha^t P_R^{t-1} [r_{P_t} - g + P_t u_0 - u_0] +$$

$$+ (1-\alpha^T)\rho^{-1}g - \sum_{t=1}^{T} \alpha^t (P_R^t u_0 - P_R^{t-1} u_0)$$

$$= (1-\alpha^T)\rho^{-1}g + \sum_{t=1}^{T} \alpha^t P_R^{t-1} \psi_{P_t}^0 - \rho \sum_{t=1}^{T+1} \alpha^t P_R^{t-1} u_0 +$$

$$+ u_0 - \alpha^{T+1} P_R^T u_0$$

$$= (1-\alpha^T)\rho^{-1}g + \rho^0 [u_0 - \alpha^{T+1} P_R^T u_0 + \sum_{t=1}^{T} \alpha^t P_R^{t-1} \psi_{P_t}^0] +$$

$$- \rho \sum_{t=1}^{T+1} \alpha^t P_R^{t-1} u_0.$$

Similarly, for $m = 1,2,\ldots$

$$(3.1.3)\quad \sum_{t=1}^{T+m} \alpha^t P_R^{t-1} u_{m-1} = - \sum_{t=1}^{T+m} \alpha^t P_R^{t-1} (-u_{m-1} + P_t u_m - u_m) +$$

$$\sum_{t=1}^{T+m} \alpha^t (P_R^t u_m - P_R^{t-1} u_m)$$

$$= - \sum_{t=1}^{T+m} \alpha^t P_R^{t-1} \psi_{P_t}^m + \rho \sum_{t=1}^{T+m+1} \alpha^t P_R^{t-1} u_m$$

$$- u_m + \alpha^{T+m+1} P_R^{T+m} u_m$$

$$= - \{[u_m - \alpha^{T+m+1} P_R^{T+m} u_m + \sum_{t=1}^{T+m} \alpha^t P_R^{t-1} \psi_{P_t}^m] +$$

$$- \rho \sum_{t=1}^{T+m+1} \alpha^t P_R^{t-1} u_m\}.$$

Substituting (3.1.3) for m = 1 in (3.1.2) and then substituting (3.1.3) for m = 2 in the result etc. gives after (M-1) substitutions the expression (3.1.1). □

Theorem 3.1 *together with lemma* 2.2 *gives*

$$(3.1.4) \qquad v^\rho(R) = \rho^{-1} g + \sum_{m=0}^{M-1} \rho^m [u_m + \sum_{t=1}^{\infty} \alpha^t P_R^{t-1} \psi_{P_t}]$$

$$+ \rho^{M-2} r(\rho, R),$$

where $\lim_{\rho \downarrow 0} r(\rho, R) = 0$, *uniformly in all policies* R.

This relation (3.1.4) *together with remark* 2.3 *yield for stationary policy* P^∞

$$(3.1.5) \qquad v^\rho(P^\infty) = \rho^{-1} g(P) + \sum_{m=0}^{M-1} \rho^m u_m(P) + O(\rho^{M-1}).$$

Relation (3.1.5) is a partial Laurent series in powers of ρ. The question raises then whether also a complete Laurent series is true i.e. $M = \infty$. Indeed, without pursuing this result here we state that under the Doeblincondition for P with bounded return vector r_P, M can be taken equal to infinity (see sections 11 and 12 of [3]).

4. ASYMPTOTIC BEHAVIOUR OF THE TOTAL EXPECTED RETURN.

We derive in this section an asymptotic expansion of $v_T^m(R)$ as $T \to \infty$.

4.1. THEOREM. *For each policy* $R = (P_1, P_2, \ldots)$, *all* $M = 1,2,\ldots$ *and all* $T = 1,2,\ldots$ *the following equality is satisfied*

$$(4.1.1) \qquad v_T^M(R) = \binom{T+M-1}{M} g + \sum_{\ell=0}^{M-1} \sum_{k=\ell}^{M-1} \binom{k}{\ell} \binom{T+M-k-2}{M-k-1} u_\ell + \\ - P_R^T \sum_{\ell=0}^{M-1} \binom{M-1}{\ell} u_\ell + \phi_T^M(R).$$

where

$$(4.1.2) \qquad \phi_T^1(R) = \sum_{t=1}^{T} P_R^{t-1} \psi_{P_t}^0$$

and

$$(4.1.3) \qquad \phi_T^{m+1}(R) = \sum_{t=1}^{T} \left(\phi_t^m(R) + \sum_{\ell=0}^{m-1} \binom{m-1}{\ell} P_R^{t-1} \psi_{P_t}^{\ell+1} \right).$$

PROOF. It is easily proved by induction on k that for all $T = 1,2,\ldots$, and all $k = 1,2,\ldots$

$$(4.1.4) \qquad \sum_{t=1}^{T} \binom{t+k-1}{k} = \binom{T+k}{k+1}$$

From relation (3.1.3) for $\rho = 0$ or $\alpha = 1$ we find for $\ell = 0,1,\ldots$

$$(4.1.5) \qquad - \sum_{t=1}^{T} P_R^{t-1} u_\ell = u_{\ell+1} - P_R^T u_{\ell+1} + \sum_{t=1}^{T} P_R^{t-1} \psi_{P_t}^{\ell+1}.$$

Hence,

$$(4.1.6) \qquad - \sum_{t=1}^{T} P_R^t u_\ell = u_\ell + u_{\ell+1} - P_R^T u_\ell - P_R^T u_{\ell+1} + \sum_{t=1}^{T} P_R^{t-1} \psi_{P_t}^{\ell+1}.$$

The proof of (4.1.1) proceeds by induction on M. Relation (3.1.2) for $\rho = 0$ yields

$$(4.1.7) \qquad v_T^1(R) = Tg + u_0 - P_R^T u_0 + \sum_{t=1}^{T} P_R^{t-1} \psi_{P_t}^0.$$

Substituting (4.1.2) for the last term on the right hand side we find that (4.1.7) is (4.1.1) for M = 1 (note that we use the convention $\binom{k}{o} = 1$ for k = 0,1,...).

By induction hypothesis assume that (4.1.1) is true for M = m and all T = 1,2,... . Hence using relation (4.1.4) we find

$$(4.1.8)\qquad v_T^{m+1}(R) = \sum_{t=1}^{T} v_t^m(R)$$

$$= \binom{T+m+1-1}{m+1} g + \sum_{\ell=0}^{m-1} \sum_{k=\ell}^{m-1} \binom{k}{\ell} \binom{T+m+1-k-2}{m+1-k-1} u_\ell +$$

$$- \sum_{t=1}^{T} P_R^t \sum_{\ell=0}^{m-1} \binom{m-1}{\ell} u_\ell + \sum_{t=1}^{T} \phi_t^m(R).$$

The last two terms give with relation (4.1.6)

$$(4.1.9)\qquad \sum_{\ell=0}^{m-1} \binom{m-1}{\ell} \left(u_\ell + u_{\ell+1} - P_R^T u_\ell - P_R^T u_{\ell+1} \right) +$$

$$+ \sum_{t=1}^{T} \left[\phi_t^m(R) + \sum_{\ell=0}^{m-1} \binom{m-1}{\ell} P_R^{t-1} \psi_{P_t}^{\ell+1} \right].$$

Using (4.1.3) and $\binom{m-1}{\ell} + \binom{m-1}{\ell-1} = \binom{m}{\ell}$ we find that (4.1.9) equals

$$(4.1.10)\qquad \sum_{\ell=0}^{m} \binom{m}{\ell} u_\ell - P_R^T \sum_{\ell=0}^{m+1-1} \binom{m+1-1}{\ell} u_\ell + \phi_T^{m+1}(R).$$

Substituting (4.1.10) in (4.1.8) gives relation (4.1.1) for M = m + 1. □

4.2. REMARK. For stationary policy P we find with g(P), $u_0(P), u_1(P), \ldots$ defined in remark 2.3 that for all M = 1,2,... .

$$(4.2.1)\qquad v_T^M(P^\infty) = \binom{T+M-1}{M} g(P) + \sum_{\ell=0}^{M-1} \sum_{k=\ell}^{M-1} \binom{k}{\ell} \binom{T+M-k-2}{M-k-1} u_\ell(P) - P^T \sum_{\ell=0}^{M-1} \binom{M-1}{\ell} u_\ell(P).$$

5. EQUIVALENCE OF SENSITIVE OPTIMALITY CRITERIA.

In this section we prove the equivalence of the sensitive optimality criteria as introduced by Veinott. Actually we shall prove that policy R is n-discount optimal if and only if it is n-average optimal. For stationary policy R the assumptions of section 2 are sufficient for nonstationary policy R we need an extra condition (relation (5.1.4) which is always satisfied if the decision sets $P(i)$ are finite.

5.1. LEMMA. *If policy* $R = (P_1, P_2, \ldots)$ *is such that for certain* $i \in E$

$$(5.1.1) \qquad P_R^{t-1}\, \psi_{P_t}^{m}(i) = 0,$$

for all $m = 0,1,\ldots,m_0-1$ *and all* $t = 1,2,\ldots$
then each of the following two conditions imply the property:
for each $\varepsilon > 0$ *there exists a finite negative integer* h_ε *such that for non-negative constants* $\varepsilon_1, \varepsilon_2, \ldots$ *with* $\sum_{t=1}^{T} \varepsilon_t \le T.\varepsilon$,

$$(5.1.2) \qquad P_R^{t-1}\, \psi_{P_t}^{m_0+1}(i) \le h_\varepsilon\, P_R^{t-1}\, \psi_{P_t}^{m_0}(i) + \varepsilon_t.$$

The conditions are

$$(5.1.3) \qquad a. \quad R \text{ is a stationary policy i.e. } P_1 = P_2 = \ldots$$

$$(5.1.4) \qquad b. \quad \limsup_{t\to\infty} \left| \frac{\psi_{P_t}^{m_0+1}(i)}{\psi_{P_t}^{m_0}(i)} \right| 1_{\{\psi_{P_t}^{m_0}(i) \neq 0\}} < \infty.$$

REMARK. Condition b is essentially a condition on the derivatives of $\psi_P^{m_0}$ and $\psi_P^{m_0+1}$ with respect to P. It is always satisfied if the set of decisions in i is finite.

PROOF. For any P and any j it follows from theorem 2.1 that

$$(5.1.5) \qquad \left\{ \begin{array}{l} \text{the first nonzero element of} \\ \{\psi_P^{0}(j), \psi_P^{1}(j), \ldots\} \\ \text{is negative} \end{array} \right.$$

Assume that for some t and some j with $P_R^{t-1}(i,j)$ positive we have that

$$\psi_{P_t}^{m_0}(j) > 0. \tag{5.1.6}$$

Then define

$$m = \min \{\exists \ell \text{ with } P_R^{t-1}(i,\ell) > 0 \text{ and } \psi_{P_t}^{m}(\ell) \neq 0\} \tag{5.1.7}$$

then it follows from (5.1.5) and (5.1.6) that $m < m_0$. Moreover, for ℓ with $P_R^{t-1}(i,\ell) > 0$ we have by (5.1.7) that $\psi_{P_t}^{n}(\ell) = 0$ for $n < m$. Hence from (5.1.5) if $P_R^{t-1}(i,\ell) > 0$ then

$$\psi_{P_t}^{m}(\ell) \leq 0$$

Consequently

$$P_R^{t-1}\, \psi_{P_t}^{m}(i) < 0,$$

which is in contradiction with assumption (5.1.1). Conclusion

$$\psi_{P_t}^{m_0}(j) \leq 0 \quad \text{for all } j \text{ with } P_R^{t-1}(i,j) > 0. \tag{5.1.8}$$

Also, from the above arguments if for j with $P_R^{t-1}(i,j) > 0$

$$\psi_{P_t}^{m_0}(j) = 0 \text{ then } \psi_{P_t}^{m_0+1}(j) \leq 0 \tag{5.1.9}$$

Relations (5.1.8) and (5.1.9) imply for each j with $P_R^{t-1}(i,j) > 0$ the existence of a negative integer h(j) such that

$$\psi_{P_t}^{m_0+1}(j) \leq h(j)\, \psi_{P_t}^{m_0}(j) \tag{5.1.10}$$

Now if the decision $p_t(j,.)$ is the same for all t, which is the case if R is a stationary policy, or when relation (5.1.4) is satisfied then we have for all t = 1,2,...

$$\psi_{P_t}^{m_0+1}(j) < h(j)\, \psi_{P_t}^{m_0}(j) \text{ if } P_R^{t-1}(i,j) > 0 \tag{5.1.11}$$

Let constant c be such that

$|\psi_P^{m_0+1}| \le c\, y_{m_0+1}$ for all P. That such a constant c exists follows from theorem 2.1. In fact, for c we can take $k_{m_0} + 2k_{m_0+1} + y_{m_0+1}(0)$.

Now choose $\varepsilon > 0$ arbitrarily, then there exists a finite set A_ε such that for all policies $R = (P_1, P_2, \ldots)$

$$(5.1.12) \qquad \sum_{j \notin A_\varepsilon} \sum_{t=1}^{\infty} {}_0P_1 \cdots {}_0P_{t-1}(i,j)\, y_{m_0+1}(j) < c^{-1}.\varepsilon.$$

The proof of the existence of A_ε is not short. However, it is easy to state the facts implying the above result.

They are:

a. The set $\mathcal{R}$ of all policies R is a compact set, where $R_n = (P_{n1}, P_{n2}, \ldots)$ converges to $R_\infty = (P_{\infty 1}, P_{\infty 2}, \ldots)$ if and only if $\lim_{n\to\infty} P_{nk} = P_{\infty k}$ for all $k = 1,2,\ldots$.

The proof of the compactness of $\mathcal{R}$ is a direct application of the well-known diagonal procedure and the compactness of $\mathcal{P}$.

b.

$$(5.1.13) \qquad \xi(R) := \sum_{t=1}^{\infty} {}_0P_1 \cdots {}_0P_{t-1}\, y_{m_0+1}$$

is a continuous function of policy $R = (P_1, P_2, \ldots)$. The proof of this is direct from the fact that (cf. (2.1.5))

$$\sum_{t=T+1}^{\infty} {}_0P_1 \cdots {}_0P_{t-1}\, y_{m_0+1} \le {}_0P_1 \cdots {}_0P_T\, y_{m_0+2},$$

and the fact that the right-hand side of this inequality tends to zero uniformly in R (for a proof see the first part of the proof of lemma 5.7 of [3] or lemma 3.7 in [3A]).

Now we sketch the proof of relation (5.1.12). Take sequence of finite subsets $A_1 \subset A_2 \subset \ldots$, such that $\bigcup_{n=1}^{\infty} A_n = E$ and suppose for each n there is a policy R_n such that for R_n relation (5.1.12) is not satisfied with A_n for A_ε. Let R_∞ be a limit of some subsequence R_{n_k}. Then it follows from the continuity of $\xi(R)$ that (cf. [3] lemmas 4.11 and 4.12)

$$\sum_{j \notin \bigcup_{n=1}^{\infty} A_n} \sum_{t=1}^{\infty} {}_0P_{\infty 1} \cdots {}_0P_{\infty t-1}(i,j)\, y_{m_0+1}(j) \geq c^{-1} \cdot \varepsilon,$$

which is clearly not true. By contradiction we find that finite set A_ε does exist.

Using the last exit decomposition of state 0(cf. lemma 2.2), we find for policy $R = (P_1, P_2, \ldots)$, $t = 1,2,\ldots$

$$\sum_{j \notin A_\varepsilon} P_R^{t-1}(i,j)\, |\psi_{P_t}^{m_0+1}(j)| \leq$$

$$\leq c \sum_{j \notin A_\varepsilon} \sum_{k=1}^{t} P_1 \cdots P_{k-1}(i,0)\, {}_0P_k \cdots {}_0P_{t-1}(0,j)\, y_{m_0+1}(j)$$

$$\leq c \sum_{j \notin A_\varepsilon} \sum_{k=1}^{t} {}_0P_k \cdots {}_0P_{t-1}(0,j)\, y_{m_0+1}(j).$$

Hence

$$\frac{1}{T} \sum_{t=1}^{T} \sum_{j \notin A_\varepsilon} P_R^{t-1}(i,j)\, |\psi_{P_t}^{m_0+1}(j)| \leq$$

$$\leq \frac{c}{T} \sum_{k=1}^{T} \sum_{j \notin A_\varepsilon} \sum_{t=k}^{T} {}_0P_k \cdots {}_0P_{t-1}(0,j) y_{m_0+1}(j)$$

$$\leq \varepsilon$$

Consequently for

$$(5.1.14) \qquad \varepsilon_t := \sum_{j \notin A_\varepsilon} P_R^{t-1}(i,j)\, |\psi_{P_t}^{m_0+1}(j)|$$

we have that

$$\sum_{t=1}^{T} \varepsilon_t \leq T.\varepsilon.$$

Define

$$h_\varepsilon = \min_{i \in A_\varepsilon} h(i)$$

then $h_\varepsilon > -\infty$

and from (5.1.8), (5.1.11) and (5.1.14) for t=1,2,...

$$P_R^{t-1}\,\psi_{P_t}^{m_0+1}(i) \le \sum_{j\in A_\varepsilon} P_R^{t-1}(i,j)\,\psi_{P_t}^{m_0+1}(j) + \varepsilon_t$$

$$\le h_\varepsilon \sum_{j\in A_\varepsilon} P_R^{t-1}(i,j)\,\psi_{P_t}^{m_0}(j) + \varepsilon_t$$

$$\le h_\varepsilon\, P_R^{t-1}\,\psi_{P_t}^{m_0}(i) + \varepsilon_t. \quad \square$$

With the preliminary results of lemma 5.1 we are in a position to prove the first main result of this section.

5.2. THEOREM. *Policy* $R = (P_1,P_2,\ldots)$ *is* n*-discount optimal if and only if*

(5.2.1) $\quad P_R^{t-1}\,\psi_{P_t}^{m} = 0$ *for* $m = 0,1,\ldots,n$ *and* $t = 1,2,\ldots$

and

(5.2.2) $\quad \lim_{\rho\downarrow 0} \rho \sum_{t=1}^{\infty} \alpha^t\, P_R^{t-1}\,\psi_{P_t}^{n+1} = 0$

PROOF. We first prove that condition (5.2.1) is necessary by showing that if (5.2.1) is not satisfied then R is not n-discount optimal.

So let us assume that for some i

(5.2.3) $\quad P_R^{t_0-1}\,\psi_{P_t}^{m_0}(i) \neq 0$ for some t_0 and some $m_0 \le n$.

and

(5.2.4) $\quad P_R^{t-1}\,\psi_{P_t}^{m} = 0$ for $m = 0,1,\ldots,m_{0-1}$ and $t = 1,2,\ldots$.

Then as shown in lemma 5.1, $P_R^{t-1}\,\psi_{P_t}^{m_0} \le 0$ for all t and hence

(5.2.5) $\quad P_R^{t_0-1}\,\psi_{P_{t_0}}^{m_0}(i) < 0.$

Let P be an element of P_{n+1}. Then $\psi_p^1 = \ldots = \psi_p^{n+1} = 0$

Hence from (3.1.4)

$$(5.2.6)\qquad v^{\rho}(R) - v^{\rho}(P^{\infty}) = \rho^{m_0} \sum_{t=1}^{\infty} \alpha^t \left[P_R^{t-1} \psi_{P_t}^{m_0} + \rho\, P_R^{t-1} \psi_{P_t}^{m_0+1} \right] + \rho^{m_0} r(\rho,R,P^{\infty}),$$

with $\lim_{\rho \downarrow 0} r(\rho,R,P^{\infty}) = 0$.

Using (5.1.2) we find that the right hand side multiplied by ρ^{-m_0} is smaller than or equal to

$$\sum_{t=1}^{\infty} \alpha^t P_R^{t-1} \psi_{P_t}^{m_0}(i)(1+\rho h_{\varepsilon}) + \rho \sum_{t=1}^{\infty} \alpha^t \varepsilon_t + r(\rho,R,P^{\infty})(i)$$

For the second term we have the inequality

$$\rho \sum_{t=1}^{\infty} \alpha^t \varepsilon_t = \rho^2 \sum_{t=1}^{\infty} \alpha^t \sum_{t=1}^{\infty} \alpha^t \varepsilon_t$$

$$= \rho^2 \alpha \sum_{t=1}^{\infty} \alpha^t \sum_{k=1}^{t} \varepsilon_k$$

$$\leq \rho^2 \varepsilon \sum_{t=1}^{\infty} \alpha^t t \leq \varepsilon \alpha^{-1}$$

Hence for $\varepsilon < |P_R^{t_0-1} \psi_{P_{t_0}}^{m_0}(i)|$ we obtain

$$\liminf_{\rho \downarrow 0} \rho^{-m_0}[v^{\rho}(R) - v^{\rho}(P^{\infty})] \leq P_R^{t_0-1} \psi_{P_{t_0}}^{m_0}(i) + \varepsilon < 0.$$

Consequently R is not m_0-discount optimal and a fortiori also not n-discount optimal. The conclusion is that condition (5.2.1) is necessary. Assuming that (5.2.1) is satisfied then we have (cf. lemma 5.1)

$$(5.2.7)\qquad P_R^{t-1} \psi_{P_t}^{n+1} \leq 0 \quad \text{for all } t.$$

Relation (5.2.6) for $m_0 = n$ gives

$$\liminf_{\rho \downarrow 0} \rho^{-n} |v^{\rho}(R) - v^{\rho}(P^{\infty})| = \liminf_{\rho \downarrow 0} \rho \sum_{t=1}^{\infty} \alpha^t P_R^{t-1} \psi_{P_t}^{n_0+1}$$

With (5.2.7) we conclude that for R to be n-discount optimal also condition (5.2.2) must be true. In that case the limit of $\rho^{-n}[v^{\rho}(R) - v^{\rho}(P^{\infty})|$

as $\rho \downarrow 0$ does exist.

Using exactly the same arguments it is straightforward to prove that for $P \in P_{n+1}$ and arbitrary R

$$(5.2.8) \qquad \liminf_{\rho\downarrow 0} \rho^{-n} [v^{\rho}(P^{\infty}) - v^{\rho}(R)] \geq 0,$$

i.e. P^{∞} is n-discount optimal.

The proof that conditions (5.2.1) and (5.2.2) are sufficient is now simple. Indeed, for arbitrary policy R^* we have from (5.2.8)

$$\liminf_{\rho\downarrow 0} \rho^{-n}[v^{\rho}(P^{\infty}) - v^{\rho}(R^*)] \geq 0$$

with

$$\lim_{\rho\downarrow 0} \rho^{-n}[v^{\rho}(R) - v^{\rho}(P^{\infty})] \geq 0$$

we obtain

$$\liminf_{\rho\downarrow 0} \rho^{-n}[v^{\rho}(R) - v^{\rho}(R^*)] \geq 0. \quad \square$$

A very similar theorem for n-average optimality shall be proved now.

5.3. THEOREM. *Policy* $R = (P_1, P_2, \ldots)$ *is* n-*average optimal if and only if*

$$(5.3.1) \qquad P_R^{t-1} \psi_{P_t}^m = 0 \quad \textit{for } m = 0,1,\ldots,n \textit{ and } t = 1,2,\ldots$$

and

$$(5.3.2) \qquad \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} P_R^{t-1} \psi_{P_t}^{n+1} = 0.$$

PROOF. The proof proceeds very similar to that of theorem 5.2.

For the necessity of (5.3.1) assume relations (5.2.3) and (5.2.4). From relations (4.1.2) and (4.1.3) it follows that

$$\phi_T^m(R) = 0 \quad \text{for } m = 0,1,\ldots m_0 \quad \text{and } T = 1,2,\ldots$$

$$\phi_T^{m_0+1}(R) = \sum_{t=1}^{T} P_R^{t-1} \psi_{P_t}^{m_0}$$

$$(5.3.3)\qquad \phi_T^{m_0+2}(R) = \sum_{t=1}^{T} \left\{ \left(\sum_{k=1}^{t} P_R^{k-1}\, \psi_{P_k}^{m_0} + m_0\, P_R^{t-1}\, \psi_{P_t}^{m_0} \right) + P_R^{t-1}\, \psi_{P_t}^{m_0+1} \right\}$$

$$= \sum_{t=1}^{T} P_R^{t-1} \left[(T+m_0+1-t)\, \psi_{P_t}^{m_0} + \psi_{P_t}^{m_0+1} \right].$$

For $P \in \mathcal{P}_{n+1}$ then $\phi_t^m(P^\infty) = 0$ for $m = 0,1,\ldots, n+2$ and $t = 1,2,\ldots$.
Hence from relation (4.1.1) we find

$$(5.3.4)\qquad v_T^{m_0+2}(R) - v_T^{m_0+2}(P^\infty) = (P^T - P_R^T) \sum_{\ell=0}^{m_0+1} \binom{m_0+1}{\ell} u_\ell$$

$$+ \sum_{t=1}^{T} P_R^{t-1} \left[(T+m_0+1-t)\, \psi_{P_t}^{m_0} + \psi_{P_t}^{m_0+1} \right].$$

For $\varepsilon < P_R^{t_0-1} \left| \psi_{P_{t_0}}^{m_0}(i) \right|$ let T_0 be such that $T_0 + m_0 + 1 > - h_\varepsilon$. Since in Cesaro limits a finite number of terms can be omitted we find with (2.2.2), for the i-th component of

$$\liminf_{T\to\infty} \frac{1}{T} \left[v_T^{m_0+2}(R) - v_T^{m_0+2}(P^\infty) \right]$$

$$\leq \liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T-T_0} \left[(T+m_0+1-t) + h_\varepsilon \right] P_R^{t-1}\, \psi_{P_t}^{m_0}(i) + \frac{1}{T} \sum_{t=1}^{T} \varepsilon_t$$

$$\leq P_R^{t_0}\, \psi_{P_t}^{m_0}(i) + \varepsilon < 0.$$

Consequently R is not m_0-average optimal and a fortiori not n-average optimal.

Assuming that (5.3.1) is satisfied we have again relation (5.2.7). Relation (5.3.4) for $m_0 = n$ gives

$$(5.3.5)\qquad \liminf_{T\to\infty} \frac{1}{T} \left[v_T^{n+2}(R) - v_T^{n+2}(P^\infty) \right] = \liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} P_R^{t-1}\, \psi_{P_t}^{n+1}.$$

With (5.2.7) we conclude that for R to be n-average optimal also condition (5.3.2) must be true. In that case the limit instead of limes inferior in expression (5.3.5) can be taken. The rest of the proof is strictly similar to that of theorem 5.2. □

5.4. COROLLARY. *Policy* $R = (P_1, P_2, \ldots)$ *is* n-*discount optimal if and only if it is* n-*average optimal.*

PROOF. From theorems 5.2 and 5.3 it follows that the only thing we have to prove is that under the condition (5.2.1) conditions (5.2.2) and (5.3.2) are equivalent. Indeed, since $P_R^{t-1}\,\psi_{P_t}^{n+1} \leq 0$ for all t it follows from a wellknown Abel and Tauber theorem that (cf. TITCHMARSH [9] p. 224-229)

$$\lim_{\rho \downarrow 0} \rho \sum_{t=1}^{\infty} \alpha^t P_R^{t-1}\,\psi_{P_t}^{n+1} = 0$$

if and only if

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} P_R^{t-1}\,\psi_{P_t}^{n+1} = 0. \quad \square$$

5.5. COROLLARY. *Since for each* n *the subset* P_n *of* P *is not empty, it follows from theorem* 5.2. *and* 5.3 *that for each* n *there exists a stationary policy which is* n-*discount optimal and also* n-*average optimal.*

REFERENCES.

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

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

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

[3A] HORDIJK, A; (1975). *Regenerative Markov decision models*. Mathematical Centre Report BW 49, Amsterdam.

[4] LIPPMAN, S.A. (1969). *Criterion equivalence in discrete dynamic programming*. Operations Res. 17, 920-923.

[5] MANDL, P. (1971). *On the variance in controlled Markov chains*. Kybernetika 7, 1-12.

[6] MILLER, B.L. & A.F. VEINOTT, Jr. (1969). *Discrete dynamic programming with a small interest rate*. Ann. Math. Statist. 40, 366-370.

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

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

[9] TITCHMARSH, E.C. (1939). *The theory of functions*, (second edition). Oxford University Press.

[10] VEINOTT, A.F., Jr. (1966). *On finding optimal policies in discrete dynamic programming with no discounting*. Ann. Math. Statist. 37, 1284-1294.

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

[12] VEINOTT, A.F., Jr. *Dynamic programming and stochastic control*. Unpublished class notes.