stichting
mathematisch
centrum

MC

AFDELING MATHEMATISCHE BESLISKUNDE BW 47/75 MAY

ARIE HORDIJK

CONVERGENT 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): 60J05, 90C40

CONVERGENT DYNAMIC PROGRAMMING

by

Arie Hordijk*

ABSTRACT

In this paper we investigate the dynamic programming problem for which the total absolute return is finite for each policy, we call this the convergent dynamic programming problem.

KEY WORDS & PHRASES: *Denumerable dynamic programming, nearly conserving policies, existence optimal policies.*

*) This paper was written while the author was visiting the department of Operations Research, Stanford University. It was supported by National Science Foundation Grant ENG 74-12378 and GK-18339 and Office of Naval Research Contract N 00014-67-A-0112-0050.

1. INTRODUCTION AND SUMMARY

In BLACKWELL [1] the positive dynamic programming problem (p.d.p.), i.e., all returns are nonnegative, was investigated. In STRAUCH [8] the negative dynamic programming problem (n.d.p.), i.e., all returns are nonpositive, was studied. In HORDIJK [6] we introduced the assumption that the total absolute return is finite for each policy. Let us call this the convergent dynamic programming problem (c.d.p.). It should be noted that, in case the supremum of the total expected returns is finite for a p.d.p problem it is also of the c.d.p. type. Moreover, it seems to us that with the exception of problems of sequential statistical decision type (see Section 9 of HORDIJK [6]) the c.d.p. case also covers the interesting n.d.p. problems.

The c.d.p. problem with a denumerable state space was investigated to some extent in Section 6 of HORDIJK [6]. The optimality equation was derived and criteria for a stationary policy to be optimal were given. A few other results for c.d.p. are given in this paper.

In section 2 it is proved that the supremum of the total expected return over the nearly conserving policies equals the supremum over all policies, i.e., equals the value function. In section 3 it is shown that the property: the supremum of the total expected return over the conserving equals the value function, provides a characterization for the existence of optimal policies. Moreover it is proved there that the existence of an optimal policy implies the existence of a stationary optimal policy. An analogous result for the p.d.p. was obtained in BLACKWELL [2] and in ORNSTEIN [7].

In section 4 the c.d.p. with a finite state space is studied. A simple counterexample is given, showing that the existence of an optimal policy is not always guaranteed. Moreover, some conditions implying the existence of an optimal stationary policy are provided.

In the remainder of this section we introduce the notations used in this paper.

We are concerned with a dynamic system which at times $t = 0,1,\ldots,$ 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 $\mathcal{P}(i)$. 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 $\mathcal{P}(i)$ consists of probability measures $p(i,.)$. (We allow that with positive probability the system "breaks down" or "disappears", so $p(i,j) \geq 0$, $i,j \in E$ and $p(i,E) := \sum_{j \in E} p(i,j) \leq 1$, $i \in E$.) Let $\mathcal{P}$ be the set of all stochastic matrices P with $p(i,\cdot) \in \mathcal{P}(i)$ for each $i \in E$. Hence $\mathcal{P}$ has the *product property:* with P_1 and P_2 the set $\mathcal{P}$ also contains all those P with for every $i \in E$ in the i^{th} row of P either the i^{th} row of P_1, or the i^{th} row of P_2.

A policy R for controlling the system is a sequence of decision rules for the times $t = 0,1,\ldots,,$ 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 $0,1,\ldots,$ $t-1$ 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 a sequence $P_0, P_1, \ldots, \in \mathcal{P}$, 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 [4], generalized in STRAUCH & VEINOTT [9] that we do not loose generality by restricting the class of policies to the Markov policies. Indeed, if the immediate return is a concave function in P and $\mathcal{P}$ contains all randomized decision rules then the supremum of the total expected return over all policies equals the supremum over the Markov policies (see Section 13 of HORDIJK [6]). In this paper we assume that the above suprema are equal and we shall only use Markov policies except in the proof of Theorem 2.1 where we use also nonmemoryless 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 $p(i,\cdot)$ is taken then an immediate return depending on i and $p(i,\cdot)$ is incurred. Let $r_p(i)$ be the immediate return when taking decision $p(i,\cdot)$(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 $p(i,\cdot) =$ $= q(i,\cdot)$ then $r_p(i) = r_Q(i)$.

The expectation of the cost at time n when starting in state i at time zero and using policy $R = (P_0, P_1, \ldots)$ 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

$$\mathbb{E}_R\ r(\underline{x}_n) = P_0\, P_1 \ldots P_{n-1}\, r_{P_n} .$$

In section 4 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. Finally, for vectors x,y we write $x \le y$ resp. $x < y$ if $x(i) \le y(i)$ resp. $x(i) < y(i)$ for all 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. NEARLY CONSERVING POLICIES

We assume in this paper that

$$(2.0.1) \qquad \sup_R \mathbb{E}_{i,R} \sum_{n=0}^{\infty} |r(x_n)| < \infty .$$

As pointed out in Section 13 of HORDIJK [6] relation (2.0.1) is equivalent to assuming that the total absolute return is finite for each policy including policies which are randomized, i.e., for which the decision rules are randomizations over the original class of decisions. This result is essential due to Derman, Strauch and Veinott (see the references given in the introduction or alternatively see [DERMAN] [3] Theorem 7.1).

Let v be the value function, i.e.,

$$v(i) = \sup_R v_R(i),$$

with

$$v_R(i) := \mathbb{E}_{i,R} \sum_{n=0}^{\infty} r(\underline{x}_n) \text{ for all } i \in E.$$

For constant ρ and vector w we introduce the class of (ρ,w)-nearly-conserving decision rules

$$\mathcal{P}_{\rho,w} = \{P : r_P + Pv \geq v - \rho w\}.$$

Let $\mathcal{R}_{\rho,w}$ denote the class of policies with decision rules in $\mathcal{P}_{\rho,w}$.

THEOREM 2.1: *Given any constant* $\rho > 0$ *and any vector* w *such that* $w \geq v$ *and* $w(i) > 0$ *for all* $i \in E$, *it holds that*

$$(2.1.1) \qquad \sup_{R \in \mathcal{R}_{\rho,w}} \mathbb{E}_{i,R} \sum_{n=0}^{\infty} r(\underline{x}_n) = v(i) \quad \text{for all } i \in E.$$

PROOF: Choose constant δ with $0 < \delta < \rho$ and $\delta < 1$.

Given any $i \in E$ there exists an $R^i = (P_0^i, P_1^i, \ldots)$ such that

$$(2.1.2) \qquad v_{R^i}(i) > v(i) - \delta^2 w(i).$$

Define Markov time $\underline{\tau}$ as follows

$$(2.1.3) \qquad \tau(i_0,\ldots,i_k,\ldots) = \min\{k \geq 0 : v_{R^{i_0}_k}(i_k) \leq v(i_k) - \delta w(i_k)\}$$

where $R_k^i = (P_k^i, P_{k+1}^i, \ldots)$ and write

$$(2.1.4) \qquad v(i,j,k) \quad \text{for } v_{R_k^i}(j).$$

Writing the total expected return under policy R^i as the sum of the return until Markov time $\underline{\tau}$ plus the total expected return thereafter, we find

$$(2.1.5) \qquad \mathbb{E}_{i,R^i} \sum_{n=0}^{\underline{\tau}-1} r(\underline{x}_n) + \sum_{j \in E} \sum_{k=1}^{\infty} \mathbb{P}_{R^i}[\underline{x}_{\underline{\tau}} = j, \underline{\tau} = k \mid \underline{x}_0 = i]\, v(i,j,k).$$

Denote R for the nonmemoryless policy with decision rule at the n^{th} decision point P_n^i when the starting state is i, so R is the "composition" of the R^i's. We write $v_{R,\underline{\tau}}$ for the vector with i^{th} component $\mathbb{E}_{i,R^i} \sum_{n=0}^{\tau-1} r(\underline{x}_n)$.

Combining the relations (2.1.2), (2.1.3) and (2.1.5) we obtain the following relation, written in vector notation,

$$(2.1.6) \qquad v_{R,\underline{\tau}} + \sum_j \mathbb{P}_R[\underline{x}_{\underline{\tau}} = j] \{v(j) - \delta w(j)\} > v - \delta^2 w.$$

It can be proved that (see HORDIJK [6] Theorem 2.21)

$$(2.1.7) \qquad v_{R,\underline{\tau}} + \sum_j \mathbb{P}_R[\underline{x}_{\underline{\tau}} = j]\, v(j) \le v.$$

An intuitive reasoning assuming the existing of optimal policies for this relation is: the total return when following a certain policy R until Markov time $\underline{\tau}$ and controlling the system in an optimal way thereafter does not exceed the total return when using an optimal policy from time zero on.

The relations (2.1.6) and (2.1.7) together yield

$$(2.1.8) \qquad \sum_j \mathbb{P}_R[\underline{x}_{\underline{\tau}} = j]\, w(j) < \delta w.$$

Since $w \ge v$ it follows from (2.1.6) and (2.1.8),

$$(2.1.9) \qquad v_{R,\underline{\tau}} > v - (\delta^2+\delta)w.$$

Let R^* be the "periodic" policy which follows policy R until time $\underline{\tau}$, note that always $\underline{\tau} \geq 1$, and starts with a new period at time $\underline{\tau}$. To be more formal the decision which R^* prescribes at time $T = j_0+j_1+\ldots+j_{k-1}+k+n$ is $p_n^{i_{k0}}(i_{kn},.)$ when the history is

$$i_{00},\ i_{01},\ldots,i_{0j_0},i_{10},\ldots,i_{1j_1},\ldots,i_{k0},\ldots,i_{kn}$$

and

$$\underline{\tau}(i_{\nu0},i_{\nu1},\ldots,i_{\nu j_\nu}) = j_\nu + 1 \quad \text{for } 0 \le \nu \le k-1$$

and

$$\underline{\tau}(i_{k0},i_{k1},\ldots,i_{kn}) > n.$$

Note that in case $\underline{\tau}$ is infinite sor some sample path $(i_0,i_1,\ldots)$ the policy R^* is not a periodic policy.

It follows from the definition of Markov time $\underline{\tau}$ that if R^* takes decision rule P at some time then, recall that v is not smaller than the total return for each policy and use (2.1.3)

$$(2.1.10) \qquad r_p + Pv > v-\delta w.$$

Now since $\delta < \rho$ we conclude that P is an element of $\mathcal{P}_{\rho,w}$.
Since $\mathbb{E}_{i,R^*} \sum_{n=0}^{\infty} |r(\underline{x}_n)| < \infty$ for all $i \in E$, it holds that

$$(2.1.11) \qquad v_{R^*} = \lim_{N\to\infty} [v_{R,\underline{\tau}} + P_{\underline{\tau}}\, v_{R,\underline{\tau}} + P^2_{\underline{\tau}}\, v_{R,\underline{\tau}} + P^3_{\underline{\tau}}\, v_{R,\underline{\tau}} + \ldots + P^N_{\underline{\tau}} v_{R,\underline{\tau}}]$$

where $P_{\underline{\tau}}$ is the matrix with (i,j)-entry $\mathbb{P}_R[\underline{x}_{\underline{\tau}} = j \,|\, \underline{x}_0 = i]$ and $P^n_{\underline{\tau}}$ is the n-fold matrix product of it.

From relation (2.1.6) we obtain

$$P_{\underline{\tau}} v > v - v_{R,\underline{\tau}} - \delta^2 w.$$

Hence

$$(2.1.12) \qquad P^k_{\underline{\tau}} v \geq P^{k-1}_{\underline{\tau}} v - P^{k-1}_{\underline{\tau}} v_{R,\underline{\tau}} - \delta^2\, P^{k-1}_{\underline{\tau}} w \quad \text{for } k = 1,2,\ldots$$

Using the relation (2.1.9) we obtain

$$v_N := \sum_{k=0}^{N} P^k_{\underline{\tau}}\, v_{R,\underline{\tau}} > \sum_{k=0}^{N-1} P^k_{\underline{\tau}}\, v_{R,\underline{\tau}} + P^N_{\underline{\tau}} v - (\delta^2+\delta)\, P^N_{\underline{\tau}} w.$$

Hence in view of (2.1.12) with $k = N$

$$v_N > \sum_{k=0}^{N-2} P^k_{\underline{\tau}}\, v_{R,\underline{\tau}} + P^{N-1}_{\underline{\tau}} v - \delta^2\, P^{N-1}_{\underline{\tau}} w - (\delta^2+\delta)\, P^N_{\underline{\tau}} w\ .$$

Now repeating this argument successively for k = N-1, N-2,...,1 we obtain

$$v_N > v - \delta^2 \sum_{k=0}^{N} P_{\underline{\tau}}^k w - \delta P_{\underline{\tau}}^N w.$$

Using relation (2.1.8) we find

$$v_{R^*} = \lim_{N\to\infty} v_N \geq v - \delta^2(1-\delta)^{-1} w.$$

Since δ can be chosen arbitrarily small we conclude that relation (2.1.1) is valid. □

If v is not bounded from above then also $w \geq v$ is not bounded from above. The question then raises whether there always exists a bounded vector w for which relation (2.1.1) is satisfied. In particular does it always hold that

$$\sup_{R\in \mathcal{R}_{1,e}} v_R = v,$$

where e is the unit vector, i.e., e(i) = 1 for all $i \in E$ and $v_R = \mathbb{E}_R \Sigma_{n=0}^{\infty} r(\underline{x}_n)$. The answer is no, as shown by the following counterexample.

COUNTEREXAMPLE: $E = \{(n,m) \mid n = 1,2,\ldots \text{ and } m = 0,1,\ldots,2^n\}$.

There are two decisions in the states (n,0), n =1,2,..., with zero return and

$$p_1((n,0),(m,k)) = \begin{cases} \frac{1}{2} & \text{for } m = n+1 \text{ and } k = 0 \\ 0 & \text{otherwise} \end{cases}$$

and

$$p_2((n,0),(n,1)) = 1.$$

In state (n,k), $k \geq 1$ there is only one decision with return $(1-\frac{1}{n})$ and

$p((n,k),(n,k+1)) = 1$ for $k \le 2^n - 1$ and $p((n,2^n),(m,k)) = 0$ for all (m,k).

If we start in state $(n,0)$ and take decision one until state $(m,0)$ is reached then the total return equals $(1-\frac{1}{m})\,2^n$. Hence for the value function v we have $v((n,0)) = 2^n$ and $v((n,k)) = (1-\frac{1}{n})\,2^n - k + 1$ for $k \ge 1$. Herewith it is easily checked that a decision rule in $\mathcal{P}_{1,e}$ must take decision 1 in all states $(n,0)$, $n = 1,2,\ldots$. So $\mathcal{R}_{1,e}$ contains only one policy with zero total expected return in the states $(n,0)$, $n = 1,2,\ldots$.

3. CHARACTERIZATION OF THE EXISTENCE OF OPTIMAL POLICIES

A decision rule P is conserving if $r_P = v - Pv$. A policy $R = (P_0,P_1,\ldots)$ is conserving if $i \in E_m$ implies $r_{P_m}(i) = v(i) - P_m v(i)$, where $E_m := \{j : \mathbb{P}_{\ell,R}[\underline{x}_m = j] > 0$ for some $\ell \in E\}$. A policy R is equalizing if $\lim_{n\to\infty} \mathbb{E}_R\, v(\underline{x}_n) = 0$ (the notions "conserving" and "equalizing" are adapted from DUBINS AND SAVAGE [5]. It is proved in Section 4 of HORDIJK [6] that a policy is optimal if and only if it is conserving and equalizing. Hence if there exists a policy R which is optimal then

$$\sup_{R\in\mathcal{R}_0} v_R = v, \qquad (3.0.1)$$

where $\mathcal{R}_0 = \{(P_0,P_1,\ldots) \mid P_t$ is conserving for all $t\}$.

The following converse is true.

THEOREM 3.1: *The relation* (3.0.1) *implies the existence of a stationary optimal policy*

PROOF: Define

$$\mathcal{P}_0 = \{P \in \mathcal{P} : r_P + Pv = v\},$$

i.e., P_0 is the class of conserving decision rules. We consider the p.d.p. problem with return structure r_P^+ (i.e.,$r_P^+(i) = \max(r_P(i),0)$) and P_0 as class of decision rules. Let v^+ be the value function of this p.d.p. problem. In view of a theorem of Ornstein (see ORNSTEIN [7] or HORDIJK [6] Theorem 13.7) there exists for any $0 < \epsilon < 1$ a Q such that

$$\sum_{n=0}^{\infty} Q^n r_Q^+ \geq (1-\epsilon)v^+.$$

Hence, since $v \leq v^+$

$$\limsup_{n\to\infty} Q^n v \leq (1-\epsilon)^{-1} \lim_{n\to\infty} Q^n \sum_{k=0}^{\infty} Q^k r_Q^+ = 0.$$

Also,

$$\liminf_{n\to\infty} Q^n v \geq \lim_{n\to\infty} Q^n \sum_{k=0}^{\infty} Q^k r_Q = 0.$$

Consequently

$$\lim_{n\to\infty} Q^n v = 0. \tag{3.1.1}$$

Hence the stationary policy Q^∞ is conserving and equalizing. This implies that Q^∞ is optimal.

Indeed, iterating the equation

$$r_Q + Qv = v,$$

N times we obtain

$$\sum_{n=0}^{N} Q^n r_Q + Q^{N+1} v = v.$$

Letting N tend to infinity we find with (3.1.1) that $\sum_{n=0}^{\infty} Q^n r_Q = v$. □

As a corollary of the above results we state:

THEOREM 3.2: *If there exists an optimal policy in* c.d.p. *then there exists also a stationary optimal policy.*

A direct consequence of Theorems 2.1 and 3.1 is

THEOREM 3.3: *There exists a stationary optimal policy when $\mathcal{P}$ has a finite number of elements.*

PROOF: Since $\mathcal{P}$ is finite there is a pair ρ,w for which (2.1.1) is true and moreover $r_P + Pv = v$ for each $P \in \mathcal{P}_{\rho,w}$. Hence relation (3.0.1) is fulfilled. □

As a corollary we conclude:

There is a stationary optimal policy when E *is finite and the number of decisions in each state is finite.*

4. FINITE STATE SPACE

In this section we assume that the state space E has a finite number of elements. It was shown already in Section 3 that the existence of a stationary optimal policy is guaranteed when in addition the number of decisions in each state is finite. For more general decision spaces a optimal policy may fail to exist as the following counterexample shows.

COUNTEREXAMPLE:

$$E = \{1,2,3\};\ p_{a,b}(1,i) = \begin{cases} a & \text{for } i = 2 \\ b & \text{for } i = 3 \end{cases}$$

with $a \geq 0$, $b \geq 0$, $a+b \leq 1$ and $b \leq a^{1/2}$; there is only one decision in the states 2 and 3 and $p(2,i) = p(3,i) = 0$ for $i = 1,2,3$; in state 1 the immediate return is zero for all decisions and in states 2 respectively 3 the immediate returns are 1 respectively 2. Then

$$v(1) = \sup\{\frac{a + 2b}{a + b}: a > 0,\ b > 0,\ a+b \leq 1,\ b \leq a^{\frac{1}{2}}\} = 2.$$

However, this supremum is not attained for any pair (a,b). We note that for this example the set P is compact with respect to the product topology and $r_P(i)$ does not depend on the decision rule P and so is certainly continuous in P, for all $i \in E$.

Since the existence of optimal policies is not always guaranteed we conclude this section with some sufficient conditions.

THEOREM 4.1: *If*

$$(4.1.1) \qquad w(i) := \sup_{R \in R_0} \mathbb{E}_{i,R} \sum_{n=0}^{\infty} |r(x_n)| > 0, \quad \text{for all } i \in E$$

then there exists a stationary optimal policy.

PROOF: We consider the c.d.p. problem with P_0 as class of decision rules and $|r_P|$ as immediate returns. Similar as in Theorem 3.1 we find a $Q \in P_0$ such that $\lim_{n\to\infty} Q^n w = 0$. Using now the fact that E is finite we have for a certain constant b that $|v(i)| \leq bw(i)$ for all i. Hence $\lim_{n\to\infty} Q^n v = 0$ and Q is conserving and equalizing. Consequently Q^∞ is optimal. □

As a corollary of the above theorem we state:

THEOREM 4.2: *Each of the following two conditions imply the existence of a stationary optimal policy:*

a. *$\mathcal{P}$ is compact,* $r_P(i)$ *is an upper semicontinuous function in* P *for all* $i \in E$ *and the immediate return is always nonzero, i.e.,* $r_P(i) \neq 0$ *for all* $i \in E$ *and* $P \in \mathcal{P}$.

b. *For each pair of states* (i,j) *there is a* $P \in \mathcal{P}_0$ *such that state* j *can be reached from state i with positive probability when using policy* P^∞ *and* v *is a nonconstant vector.*

PROOF: From the facts that $r_P(i) + Pv(i)$ is an upper semicontinuous function in P and $\mathcal{P}$ is compact it follows that $\mathcal{P}_0$ is a nonvoid set. Hence condition a implies relation (4.1.1). This proves the first part of the theorem.

Assume condition b is true. For arbitrary $j \in E$ it is possible to find a $P \in \mathcal{P}_0$ such that under P state j is reached with positive probability from each state $i \in E$. Since E is finite it follows that state j is reached with probability 1 from each state $i \in E$. Since P is conserving it follows that v(i) is equal to v(j) plus the expected cost under P^∞ until reaching state j. Now if $w(i) = 0$ (w is defined in (4.1.1)) then this expected cost is zero and hence $v(i) = v(j)$. Consequently $w(i) = 0$ for some $i \in E$ implies that v is a constant vector. From condition b it follows then $w(i) > 0$ for all $i \in E$ and we can apply theorem 4.1. □

REFERENCES

[1] BLACKWELL, D., *Positive Dynamic Programming*. Proc. Fifth Berkeley Sympos. Math. Stat. and Proob., Vol. 1, 415-418, 1967.

[2] BLACKWELL, D., *On stationary Policies*. J. Roy. Statist. Soc. Ser. A. 133, 33-38, 1970.

[3] DERMAN, C. *Finite State Markovian Decision Processes*. Academic Press, New York, 1970.

[4] DERMAN, C. & R. STRAUCH, *A note on Memoryless Rules for Controlling Sequential Control Processes*. Ann. Math. Statist. 37, 276-278, 1966.

[5] DUBINS, L.E. & L.J. SAVAGE, *How to gamble if you Must: Inequalities for Stochastic Processes*. McGraw-Hill, New York, 1965.

[6] HORDIJK, A. *Dynamic Programming and Markov Potential Theory*. Mathematical Centre Tract No. 51, Amsterdam, 1974.

[7] ORNSTEIN, D. *On the existence of Stationary Optimal Strategies*. Proc. Amer. Math. Soc. 20, 563-569, 1969.

[8] STRAUCH, R. *Negative Dynamic Programming*. Ann. Math. Statist. 37, 871-889, 1966.

[9] STRAUCH, R. & A.F. VEINOTT Jr. *A property of Sequential Control processes*. Rand McNally, Chicago, Illinois, 1966.