stichting
mathematisch
centrum

MC

AFDELING MATHEMATISCHE BESLISKUNDE
(DEPARTMENT OF OPERATIONS RESEARCH)

BW 73/77 MAART

A. FEDERGRUEN

ON THE FUNCTIONAL EQUATIONS IN UNDISCOUNTED AND SENSITIVE DISCOUNTED STOCHASTIC GAMES

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).

AMS(MOS) subject classification scheme (1970): 90D05, 90D15, 93E05

On the functional equations in undiscounted and sensitive discounted stochastic games

by

A. Federgruen

ABSTRACT

This paper considers two person zero-sum sequential games with finite state and action spaces. We consider the pair of functional equations *(f.e.)* that arises in the undiscounted infinite stage model, and show that a certain class of successive approximation schemes is guaranteed to converge to a solution pair whenever a stationary equilibrium policy with respect to the average return per unit time criterion *(AEP)* exists. Existence of the latter thus implies the existence of a solution to this pair of *f.e.* whereas the converse implication is shown only to hold under special circumstances.

In addition to this pair of *f.e.*, a complete sequence of *f.e.* has to be considered when analyzing more sensitive optimality criteria that make further selections within the class of *AEPs*. A number of characterizations and interdependences between the existence of solutions to the *f.e.* and existence of stationary sensitive optimal equilibrium policies are obtained.

KEY WORDS & PHRASES: *stochastic games; functional equations; average return per unit time criterion; sensitive optimality criteria; equilibrium policies*

0. INTRODUCTION AND SUMMARY

This paper considers two-person zero-sum Stochastic Renewal Games *(SRG's)* with finite state space $\Omega = \{1,\ldots,N\}$ and in each state $i \in \Omega$, two finite sets K(i) and L(i) of actions available to player 1 and 2 resp. We speak of a state as being observed for only an instant. The moment state i is observed, the two players choose an action, or a randomization of actions out of K(i) and L(i) resp. When the actions $k \in K(i)$, and $\ell \in L(i)$ are chosen in state i, then

(1) the probability that state j is the next state to be observed, is given by $P_{ij}^{k,\ell} \geq 0$ $(\sum_{j=1}^{N} P_{ij}^{k,\ell} = 1)$

(2) the period of time until the next observation of time, is a random variable t, with conditional probability distribution function $F_{ij}^{k,\ell}(.)$ given that j is the next state of the system

(3) for each $x \geq 0$, $R_{ij}^{k,\ell}(x)$ denotes the expected income earned by player 1 from player 2, during the first x units of time, given that state j is the next state of the system and $t \geq x$.

The discrete time case, where each transition takes exactly *one* unit of time, is known as the stochastic games-model (cf. e.g. [14],[18]) and will be denoted as the *SDG-case*. When one of the two players has only one action in each state of the system, the SRG and SDG model reduce to a Markov Renewal Program, *(MRP)* and pure Markov Decision Problem *(MDP)* resp. If the payoffs are discounted at the interest rate $r > 0$, the SRG-game is called the *r-discount game*. Let V(r) denote the vector the s-th component of which indicates the value of the r-discount game with initial state $s \in \Omega$. The existence of a value for the r-discount game goes back to SHAPLEY [18].

In a recent paper, BEWLEY and KOHLBERG [2] gave a description of the asymptotic behaviour of V(r), as the interest rate r decreases to zero, by deriving a series expansion of V(r), for all r sufficiently small. When there is no reason to discount future rewards, or whenever the infinite stage game model serves as an approximation to the model where the planning horizon is finite though large, the *average return per unit time*

criterion, in one of its possible specifications (cf. BEWLEY and KOHLBERG [3]) is the first criterion to be considered.

It is known from GILLETTE [8] that one or both players may fail to have equilibrium policies with respect to the average return per unit time criterion *(AEPs)*. Recurrency conditions with respect to the transition probability matrices *(tpms)* associated with the stationary policies have been found under which the existence of an AEP is guaranteed for each possible combination of rewards. (cf. HOFFMANN and KARP [9], SOBEL [20], ROGERS [15], STERN [21] and FEDERGRUEN [6].

In this paper we show that in undiscounted SRG's, a pair of functional vector equations arises which is the natural analogue of the corresponding ones in Markov Decision Theory (cf. [5],[11],[17]). We show that, in complete analogy to the structure of MRP's, the existence of a solution to this pair of functional equations is a necessary condition for the existence of a stationary AEP.

We give a constructive proof, showing that a specific class of successive approximation schemes converges to a solution of this pair of functional equations *(f.e.)*

For the case where the optimal average return per unit time is independent of the initial state of the system, these successive approximation schemes provide an algorithm to locate AEPs whenever existing, as is pointed out in FEDERGRUEN [7]. Conversely and in contrast with what is known to be the case in ordinary MRP's, it is shown that the existence of a fixed point of the pair of functional equations only needs to be *sufficient* for the existence of an AEP when the asymptotic average value (cf. section 2) is independent of the initial state of the system.

This is explained by showing that a pair of policies which satisfies the two optimality equations for some solution pair, is merely guaranteed to meet some *partial* optimality result (cf. prop. 2.4).

The above results are obtained in section 2, after giving the notation in section 1.

In section 3, we give some properties of the optimality equations, both for the general multichain and for the unichain-case. Since only the tails of the streams of rewards matter when considering the average

return per unit time criterion, more sensitive optimality criteria are needed to make further selections within the class of AEPs. As a consequence, we next consider the extension to the SRG-model of the sensitive discount and cumulative average optimality criteria that have been formulated and studied in the literature on MDPs (cf. MILLER and VEINOTT [13], VEINOTT [22], SLADKY [19], DENARDO [4] e.a.) In section 4, we show that in addition to the above mentioned pair of f.e. an entire sequence of coupled f.e. arises when considering these sensitive optimality criteria. We prove that this sequence has a solution when all of the tpm's associated with the pure stationary policies are unichained. Moreover, we extend the results obtained for the average return per unit time criterion to the entire set of sensitive optimality criteria.

1. NOTATION AND PRELIMINARIES

For each finite set A, let $\|A\|$ denote the number of elements it contains. E^n denotes the n-dimensional Euclidean space. If $A = [A_{ij}]$ is a matrix, let $|A| = \max_{i,j} |A_{ij}|$, and let val A indicate the value of the corresponding matrix game. Note that for any pair of matrices A,B of equal dimension

$$(1.1) \qquad |\text{val } A - \text{val } B| \leq |A-B|$$

(Let (x^A,y^A) and (x^B,y^B) be equilibrium pairs of actions in the matrix games A and B; then $\min_{i,j}(A_{ij}-B_{ij}) \leq x^B (A-B)y^A \leq x^B Ay^A - x^B By^A \leq \text{val } A - \text{val } B \leq x^A Ay^B - x^A By^B \leq x^A (A-B)y^B \leq \max_{i,j}(A_{ij}-B_{ij})$).

For each state $i \in \Omega$, let $K(i) = \{x \in E^{\|K(i)\|} \mid x \geq 0, \sum_{k=1}^{\|K(i)\|} x_k = 1\}$ denote the set of all randomized actions available to player 1 in state i. Similarly, $L(i) = \{y \in E^{\|L(i)\|} \mid y \geq 0, \sum_{\ell=1}^{\|L(i)\|} y_\ell = 1\}$ indicates the set of all randomized actions available to player 2 in state $i \in \Omega$.

For every $i \in \Omega$, any tableau of numbers $[c_i^{k,\ell}]$, $k = 1,\ldots,\|K(i)\|$; $\ell = 1,\ldots,\|L(i)\|$ and for each pair of closed convex subsets $\tilde{K}(i) \subseteq K(i)$ and $\tilde{L}(i) \subseteq L(i)$, we denote by

(1.2) $\quad [\tilde{K}(i),\tilde{L}(i)]^{[c_i^{k,\ell}]}$

the two-person zero-sum game which has $\tilde{K}(i)$ and $\tilde{L}(i)$ as the action sets for player 1 and 2 resp. and where the payoff to player 1, is given by

$$\sum_{k=1}^{\|K(i)\|} \sum_{\ell=1}^{\|L(i)\|} x_k \, c_i^{k,\ell} \, y_\ell,$$

when the players choose action $x \in \tilde{K}(i)$ and $y \in \tilde{L}(i)$ resp. The minimax value of this game is indicated by $\mathrm{val}_{[\tilde{K}(i),\tilde{L}(i)]}[c_i^{k,\ell}]$. When $\tilde{K}(i) = K(i)$ and $\tilde{L}(i) = L(i)$ we use the abbreviated notation $[c_i^{k,\ell}]$ to indicate the game in (1.1). The following lemma is immediate from KARLIN ([12], pp. 63):

LEMMA 1.1. *Fix* $i \in \Omega$. *Let* $\tilde{K}(i)$ *and* $\tilde{L}(i)$ *be closed convex polyhedral subsets of* $K(i)$ *and* $L(i)$. *Then the sets of optimal actions in any one of the two-person zero-sum games in* (1.1) *are again closed convex polyhedral subsets of* $\tilde{K}(i)$ *and* $\tilde{L}(i)$ *and thus of* $K(i)$ *and* $L(i)$.

A player's policy is a rule which prescribes for each stage $t = 1,2,\ldots$ which (randomized) action to choose in dependence on the current state and the entire history of the game up to that stage. A policy is said to be *stationary*, if it prescribes actions which depend merely upon the current state of the system, regardless of the stage of the game, and its history up to this stage. Note that a stationary strategy f (h) for player 1 (2) is characterized by a tableau $[f_{ik}]$ ($[h_{i\ell}]$) satisfying $f_{ik} \geq 0$ and $\sum_{k\in K(i)} f_{ik} = 1$ ($h_{ik} \geq 0$ and $\sum_{\ell\in L(i)} h_{i\ell} = 1$), where $f_{ik}(h_{i\ell})$ is the probability that the k-th (ℓ-th) alternative is chosen when entering state $i \in \Omega$. We let $\Phi(\Psi)$ denote the set of all stationary policies for player 1 (2).

When a positive interest rate r is introduced into the model, income earned at time t is discounted by the factor e^{-rt}. When in state i, the players choose action $k \in K(i)$ and $\ell \in L(i)$, the one-step expected r-discounted reward for player 1 is given by:

$$(1.3) \qquad \rho_i^{k,\ell}(r) = \sum_j P_{ij}^{k,\ell} \int_0^\infty \int_0^x e^{-rt} \, d\, R_{ij}^{k,\ell}(t) \, d\, F_{ij}^{k,\ell}(x)$$

and let $q_i^{k,\ell} = \rho_i^{k,\ell}(0)$ denote the one-step expected *(undiscounted)* reward. Let

$$\tau_{ij}^{k,\ell} = \int_0^\infty x \, d \, F_{ij}^{k,\ell}(x) < \infty$$

denote the expected *conditional* holding time in state i, when the players choose actions $k \in K(i)$, $\ell \in L(i)$ and given that the next state observed is state j. Likewise, let

$$T_i^{k,\ell} = \sum_j H_{ij}^{k,\ell} = \sum_j P_{ij}^{k,\ell} \, \tau_{ij}^{k,\ell}$$

denote the expected *unconditional* holding time in state i, when $k \in K(i)$, and $\ell \in L(i)$ are the actions chosen, and assume $T_i^{k,\ell} > 0$ $(i \in \Omega,\ k \in K(i),\ \ell \in L(i))$.

Associated with each pair $(f,h) \in \Phi \times \Psi$ are N-component reward vector q(f,h), holding time vector T(f,h), and the matrices P(f,h) and H(f,h):

$$q(f,h)_i = \sum_{k \in K(i)} \sum_{\ell \in L(i)} f_{ik} \cdot q_i^{k,\ell} \cdot h_{i\ell} \qquad ; i \in \Omega$$

$$T(f,h)_i = \sum_{k \in K(i)} \sum_{\ell \in L(i)} f_{ik} \cdot T_i^{k,\ell} \cdot h_{i\ell} \qquad ; i \in \Omega$$

$$P(f,h)_{ij} = \sum_k \sum_\ell f_{ik} \cdot P_{ij}^{k,\ell} \cdot h_{i\ell} \qquad ; i, j \in \Omega$$

$$H(f,h)_{ij} = \sum_k \sum_\ell f_{ik} \cdot H_{ij}^{k,\ell} \cdot h_{i\ell} \qquad ; i, j \in \Omega.$$

For any pair $(f,h) \in \Phi \times \Psi$ define the stochastic matrix $\Pi(f,h)$ as the Cesaro limit of the sequence $\{P^n(f,h)\}_{n=1}^\infty$. Denote by n(f,h) the number of subchains (closed, irreducible sets of states) for P(f,h) and let $R(f,h) = \{i \in \Omega \mid \Pi(f,h)_{ii} > 0\}$ i.e. R(f,h) is the set of recurrent states for P(f,h).

We then have:

$$(1.5) \qquad \Pi(f,h)_{ij} = \sum_{m=1}^{n(f,h)} \Phi_i^m(f,h).\ \pi_j^m(f,h)$$

where $\pi^m(f,h)$ is the unique equilibrium distribution of $P(f,h)$ on the m^{th} subchain $C^m(f,h)$ and $\Phi_i^m(f,h)$ is the probability of absorption in $C^m(f,h)$, starting from state $i \in \Omega$. Observe that $\sum_{i=1}^{N} \pi_i^m(f,h) = 1$ and $\pi^m(f,h)\ P(f,h) = \pi^m(f,h)$. For each $(f,h) \in \Phi \times \Psi$, we define the gain rate vector $g(f,h)$ such that $g(f,h)_i$ represents the long run average expected return per unit time when the initial state is i, and policies f,h are used. We thus have

$$(1.6) \qquad g(f,h)_i = \begin{cases} g^m(f,h) = \langle \pi^m(f,h) q(f,h) \rangle / \langle \pi^m(f,h) T(f,h) \rangle, & i \in C(f,h) \\ \sum_{m=1}^{n(f,h)} \Phi_i^m(f,h)\ g^m(f,h) & , \ i \in \Omega \backslash R(f,h) \end{cases}$$

Next, we recall that $V(r)$, the value vector of the r-discount game, satisfies the equation:

$$(1.7) \qquad V(r)_i = \mathrm{val}[\rho_i^{k,\ell}(r) + \sum_j m_{ij}^{k,\ell}(r)\ V(r)_j], \quad i \in \Omega,\ r > 0$$

where $$m_{ij}^{k,\ell}(r) = P_{ij}^{k,\ell} \int_0^\infty e^{-rt}\ d\, F_{ij}^{k,\ell}(t) \geq 0.$$

BEWLEY and KOHLBERG [2] recently showed for the discrete time case (SDG's) that $V(r)$ may be expressed as a real fractional power of *Puiseux* series in r, for all interest rates r that are sufficiently close to 0. More specifically, there exists an integer $M \geq 1$ such that:

$$(1.8) \qquad V(r) = \sum_{k=-\infty}^{M} a^{(k)}\ r^{-k/M}$$

This result carries easily over to the general SRG-case. We henceforth assume that $\rho_i^{k,\ell}(r)$ and $m_{ij}^{k,\ell}(r)$ have a Taylor series expansion ($i \in \Omega$; $k \in K(i)$; $\ell \in L(i)$):

LEMMA 1.2. $V(r)$ *has a Puiseux series expansion as in* (1.8).

PROOF: The proof goes along lines with section 11 in [2]. Note that $\sum_j m_{ij}^{k,\ell}(r) < 1$ for all $r > 0$ and all $i \in \Omega$; $k \in K(i)$ and $\ell \in L(i)$. Observe next, from a standard contraction mapping argument that

(1.9) the equation $x_i = \text{val}[\rho_i^{k,\ell}(r) + \sum_j m_{ij}^{k,\ell}(r)\, x_j]$, $i \in \Omega$ has a (unique) solution for all values of the parameters $\rho_i^{k,\ell}(r)$ and $m_{ij}^{k,\ell}(r)$ such that $m_{ij}^{k,\ell}(r) > 0$ and $\sum_j m_{ij}^{k,\ell}(r) < 1$.

Since (1.9) is a sentence in elementary algebra (cf. cor. 9.2 in [2]) it follows from Tarski's principle (cf. section 11 in [2]) that (1.9) is true over any real closed field, if it is true over the reals. Finally, the set of all real Puiseux series was shown to be a closed ordered field (cf. section 10 in [2]) which completes the proof that $V(r)$ has an expansion of the type $\sum_{k=-\infty}^{K} a^{(k)} r^{-k/M}$. The fact that $a^{(k)} = 0$, for all $k > M$ finally follows from the proof of th. 7.2 in [2] and the observation that

$$(1.10) \qquad \sum_j m_{ij}^{k,\ell}(r) \leq 1 - \tfrac{1}{2} r\, T_{\min}, \quad i \in \Omega,\ k \in K(i),\ \ell \in L(i)$$

where $T_{\min} = \min_{i,k,\ell} T_i^{k,\ell} > 0$. To verify (1.10) note that $e^{-rt} \leq 1 - \frac{1}{2}rt$ for all r sufficiently close to zero, and use the definition of $m_{ij}^{k,\ell}(r)$. □

We recall from the example in section 14 of [2] that in general $V(r)$ cannot be expressed as a rational function or *Laurent* series in r as is known to be the case in ordinary MRP's (cf. [4],[13]). The vector $a^{(M)}$ in (1.8) is called the *asymptotic average value* vector. Finally it was shown in [2] that in the *discrete-time case* (of SDG's), $a^{(M)} = \lim_{n\to\infty} v(n)/n$ where $v(n)$ is the vector, whose i-th component denotes the value of the n-step game with initial state i.

2. THE AVERAGE RETURN CRITERION; A PAIR OF FUNCTIONAL EQUATIONS

In this section we are concerned with the average return per unit time criterion, i.e. we evaluate any pair of (possibly non-stationary)

policies for players 1 and 2, by considering for each initial state $i \in \Omega$:

$$(2.1) \qquad g(\varphi,\psi)_i = \liminf_{N\to\infty}(E_{\varphi,\psi} \sum_{n=1}^{N} \rho_n)/(E_{\varphi,\psi} \sum_{n=1}^{N} \tau_n); \quad i \in \Omega.$$

where $\rho_n(\tau_n)$ denotes the payoff to player 1 (the length of the period) in between of the n-1-st and the n-th observation of state. $E_{\varphi,\psi}$ indicates the expectation, given the players' policies φ and ψ. A number of equivalent criteria have been formulated in [3].

A pair of policies (φ^*,ψ^*) is called an AEP, if and only if for every policy $\varphi \in \Phi$, and $\psi \in \Psi$:

$$(2.2) \qquad g(\varphi,\psi^*)_i \le g(\varphi^*,\psi^*)_i \le g(\varphi^*,\psi)_i, \quad \text{for all } i \in \Omega.$$

It is known from GILLETTE [8] that one or both players may fail to have gain-optimal policies. For the discrete-time case, the existence of a stationary AEP was shown to be guaranteed for every possible combination of one-step expected rewards $q_i^{k,\ell}$ if the matrix function $\Pi(f,h)$ is continuous on $\Phi \times \Psi$ (cf. [6], section 4).

Continuity of $\Pi(f,h)$ on its turn is guaranteed to hold when the number of subchains $n(f,h)$ is continuous, i.e. constant on $\Phi \times \Psi$, and a (finitely verifiable) sufficient condition for the latter was obtained in [6], with respect to the chain structure of the set of *pure* policies. Lemma 2-1 shows that these results carry over to the general SRG-case:

LEMMA 2.1: *Let* $n(f,h)$ *be constant on* $\Phi \times \psi$. *Then there exists a* AEP.

PROOF. Consider the SDG which has Ω as state space, K(i) and L(i) as the action spaces in $i \in \Omega$, and the following transition probabilities and one-step expected rewards:

$$(2.3) \qquad \begin{cases} \tilde{P}_{ij}^{k,\ell} = \tau/T_i^{k,\ell}[P_{ij}^{k,\ell} - \delta_{ij}] + \delta_{ij}; \; i,j \in \Omega, \; k \in K(i), \; \ell \in L(i). \\ \tilde{q}_i^{k,\ell} = q_i^{k,\ell}/T_i^{k,\ell} \qquad ; \quad i \in \Omega, \; k \in K(i), \; \ell \in L(i) \end{cases}$$

where $\delta_{ij} = \begin{cases} 0 \text{ if } i \neq j \\ 1 \text{ otherwise} \end{cases}$ an where τ has to be chosen such that $0 < \tau \leq \min_{i,k,\ell} T_i^{k,\ell}/(1-P_{ii}^{k,\ell})$.

This data-transformation which was first introduced in [16], has the property of leaving for every pair of policies $(f,h) \in \Phi \times \Psi$, both the chain structure (e.g. the number of subchains) and the gain-rate vector unaltered. Hence, $n(f,h)$ is constant on $\Phi \times \Psi$, in the transformed SDG as well, which implies (cf. [6], th. 3) the existence of a policy pair (f^*,h^*) for which $g(f,h^*) \leq g(f^*,h^*) \leq g(f^*,h)$ for all $f \in \Phi$, $h \in \Psi$. Finally one easily concludes that (f^*,h^*) is an AEP, i.e. an equilibrium pair of policies within the largest possible class of policies (cf. e.g. [6]). □

As a special case for the condition in lemma 2-1, we have that stationary AEP exist, if

(U): every pair of pure stationary policies is unichained.

In this section, we show that the following pair of optimality equations arises when analyzing the average return criterion:

$$(2.4) \qquad g_i = \text{val}[\textstyle\sum_{j=1}^{N} P_{ij}^{k,\ell} g_j], \qquad i \in \Omega$$

$$(2.5) \qquad v_i = \text{val}_{[K(i,g),L(i,g)]}[q_i^{k,\ell} - \textstyle\sum_j H_{ij}^{k,\ell} g_j + \sum_j P_{ij}^{k,\ell} v_j], \quad i \in \Omega.$$

where for each $i \in \Omega$, and each solution g^* to (2.4), $K(i,g^*)$ and $L(i,g^*)$ are the sets of optimal actions in the matrix game in (2.4) with $g = g^*$. Note from lemma 1-1 that the sets $K(i,g)$ and $L(i,g)$ are in fact the convex hulls of a finite number of extreme points such that the games to the right of (2.5) may be interpreted as simple matrix games. Observe finally that (2.4) and (2.5) are the natural extension of the well-known optimality equations in undiscounted MRPs (cf. HOWARD [11], DENARDO and FOX [5]).

We say that a pair of policies (f^*,h^*) satisfies the optimality equations (2.4) and (2.5), if for some solution (g^*,v^*), $(f^*(i), h^*(i))$ is an equilibrium pair of actions in the matrix games to the right of (2.4) and (2.5). First we show that a solution pair to the equations (2.4) and (2.5) exists whenever a stationary AEP exists. Our proof is a

constructive one; in fact we show that a certain class of successive approximation schemes converges to a solution pair of (2.4) and (2.5) whenever a stationary AEP exists. These schemes are the natural analogue of a value iteration method which was introduced and studied by HORDIJK and TIJMS [10] for undiscounted MDPs, and a special case of which was first used by BATHER [1]. First of all, observe that $a^{(M)}$, the asymptotic average value vector (cf. (2.1)) is a solution to (2.4):

$$(2.6) \qquad a_i^{(M)} = \mathrm{val}[\textstyle\sum_j P_{ij}^{k,\ell}\, a_j^{(M)}], \quad i \in \Omega$$

as is easily verified by inserting (1.10) into both sides of (1.9), multiplying the resulting equality by $r > 0$, letting r tend to zero, and by interchanging the limit and value-operation (cf. (1.1)).
We next consider a related SRG, with Ω as state space. For each $i \in \Omega$, $k \in K(i)$, $\ell \in L(i)$ let,

$$(2.7) \qquad \tilde{q}_i^{k,\ell} = q_i^{k,\ell} - \textstyle\sum_j H_{ij}^{k,\ell}\, a_j^{(M)};$$

be the associated one-step expected rewards. Both the transition probabilities and the transition time distribution remain unaltered. Moreover we restrict in each state $i \in \Omega$ the set of (randomized) actions available for player 1 to $K(i,a^{(M)})$ and the set of actions for player 2 to $L(i,a^{(M)})$. $\tilde{V}(r)$, $\tilde{a}^{(k)}$, $k = -\infty,\ldots,\tilde{M}$ and for each $f \in X_{i=1}^{N} K(i,a^{(M)})$, $h \in X_i L(i,a^{(M)})$ the quantities $\tilde{q}(f,h)$ and $\tilde{g}(f,h)$ are defined in complete analogy to $V(r)$, $a^{(k)}$, $k = -\infty,\ldots,M$; $q(f,h)$ and $g(f,h)$. Before introducing the successive approximation schemes we first need the following theorem:

THEOREM 2.2: *Assume there exists a stationary* AEP $(f^*,h^*) \in \Phi \times \Psi$ *in the original stochastic renewal game* (SRG). *Then*

(a) $a^{(M)} = g(f^*,h^*)$

(b) $a^{(k)} = 0$, $k = 1,\ldots,M-1$

(c) *every policy* $f \in \Phi$ ($h \in \Psi$) *which is gain-optimal for player 1 (or 2) in the original SRG, is gain-optimal in the transformed SRG*

(d) *there exists a constant* $B > 0$, *and an integer* $\tilde{M} \geq 1$, *such that for all*

$r, s > 0$ *sufficiently small:*

$$(2.8) \qquad |\tilde{V}(r) - \tilde{V}(s)| \leq B\,|r^{1/\tilde{M}} - s^{1/\tilde{M}}|$$

PROOF.

(a),(b): go along lines with lemma 7.11 in [3]:
Let $\sum_{k=-\infty}^{1} A^{(k)} r^{-k}$ [$\sum_{k=-\infty}^{1} B^{(k)} r^{-k}$] be the Laurent series expansion of $W^1(r)$ [$W^2(r)$], the total discounted return to player 1 [2] in the MRP that results when player 2 [1] ties himself down to policy h^*[f^*]. Since f^*[h^*] is gain-optimal in this MRP conclude that $A^{(1)} = B^{(1)} = g(f^*,h^*)$. Finally, parts (a) and (b) follow from the inequalities $W^2(r) \leq V(r) \leq W^1(r)$.

(c) Fix a stationary AEP (f^*,h^*) and $i \in \Omega$; recall (e.g. from th. 1 in [5]) that for any $f \in \Phi$ and $h \in \Psi$:
$P(f,h^*)\, g(f^*,h^*) \leq P(f^*,h^*)\, g(f^*,h^*) = g(f^*,h^*) \leq P(f^*,h)\, g(f^*,h^*)$
such that:
$P(f,h^*)\, a_i^{(M)} \leq P(f^*,h^*)\, a_i^{(M)} = a_i^{(M)} \leq P(f^*,h)\, a_i^{(M)}$, thus proving that $f^*(i) \in K(i,a^{(M)})$ and $h^*(i) \in L(i,a^{(M)})$ for all $i \in \Omega$, or in other words the feasibility of f^* and h^* in the transformed game. We next show, that (f^*,h^*) is an AEP in the transformed SRG, with $\tilde{g}(f^*,h^*) = 0$, by proving:

$$(2.9) \qquad \text{(a)}\ \tilde{g}(f,h^*) = g(f,h^*) - a^{(M)}, \text{ for all } f \in X_i\, K(i,a^{(M)})$$
$$\text{(b)}\ \tilde{g}(f^*,h) = g(f^*,h) - a^{(M)}, \text{ for all } h \in X_i\, L(i,a^{(M)})$$

Confining ourselves to (2.9) (a) (the proof of (b) being analogous) fix $f \in X_i\, K(i,a^{(M)})$, and observe by iterating the equality $a^{(M)} = P(f,h^*)\, a^{(M)}$, that:

$$a^{(M)} = \begin{cases} c^{(m)} & \text{for all } i \in C^m(f,h^*);\ m = 1,\ldots,n(f,h^*) \\ \sum_{m=1}^{n(f,h)} \Phi_i^m(f,h^*)\, c^{(m)}, & \text{for all } i \in \Omega \backslash R(f,h^*) \end{cases}$$

Then,

$$\tilde{g}^{(m)}(f,h^*) = \frac{\langle \pi^m(f,h^*), q(f,h^*) - H(f,h^*)a^{(M)} \rangle}{\langle \pi^m(f,h^*), T(f,h^*) \rangle}$$

$$= g^{(m)}(f,h^*) - c^{(m)}(\textstyle\sum_i \pi^m(f,h^*)_i \sum_j H(f,h^*)_{ij}) / \langle \pi^m(f,h^*), T(f,h^*) \rangle$$

$$= g^{(m)}(f,h^*) - c^{(m)}; \; m = 1,\ldots,n(f,h^*)$$

and conclude that $\tilde{g}(f,h^*)_i = \sum_{m=1}^{n(f,h^*)} \Phi_i^m(f,h^*)\, \tilde{g}^{(m)}(f,h^*) = g(f,h^*)_i - a_i^{(M)}$, for all $i \in \Omega$.

(d) The proof of part (b) shows that $\tilde{a}^{(\tilde{M})} = 0$ as well as the existence of a stationary AEP in the transformed model, and the latter implies by applying part (a) to the transformed game, that $\tilde{a}^{(k)} = 0$ for $k = 1,\ldots,\tilde{M}-1$ as well i.e. for all r sufficiently small:
$|\tilde{V}(r) - \tilde{a}^{(0)}| = r^{1/\tilde{M}} B(r)$ where $B(r) = |\sum_{k=0}^{\infty} a^{(-k-1)}\, r^{k/\tilde{M}}|$ is a function which is continuous in $r = 0$. Hence there exists a scalar $B > 0$ and a number $r^* > 0$ such that for all $r < r^*$:
$|\tilde{V}(r) - \tilde{a}^{(0)}| \le Br^{1/\tilde{M}}$, i.e. for all $r, s < r^*$:
$\tilde{V}(r) - \tilde{V}(s) \le \tilde{V}(r) - \tilde{a}^{(0)} + \tilde{V}(s) - \tilde{a}^{(0)} \le Br^{1/\tilde{M}} - Bs^{1/\tilde{M}} \le B|r^{1/\tilde{M}} - s^{1/\tilde{M}}|$ □

We next introduce the following successive approximation scheme:

$$y(n)_i = \mathrm{val}_{[K(i,a^{(M)}),\, L(i,a^{(M)})]}[\tilde{q}_i^{k,\ell} + \textstyle\sum_j m_{ij}^{k,\ell}(r_n) y(n-1)_j] \tag{2.10}$$

where $\{r_n\}_{n=1}^{\infty}$ is a sequence of interest rates, with $\lim_{n\to\infty} r_n = 0$. Under the assumption that a stationary AEP exists, the following theorem exhibits the existence of a solution pair to the optimality equations (1.4) and (1.5) by showing in analogy to th. 1 in HORDIJK and TIJMS [10] that the sequence $\{y(n)\}_{n=1}^{\infty}$ converges under specific conditions on $\{r_n\}_{n=1}^{\infty}$:

THEOREM 2.3. *Assume the original* SRG *has a stationary* AEP. *Then:*
(a) $(a^{(M)}, \tilde{a}^{(0)})$ *is a solution pair to the* f.e. (2.4) *and* (2.5)
(b) *Let* $\{r_n\}_{n=1}^{\infty}$ *satisfy the conditions:*
(1) $(1-r_1') \ldots (1-r_n') \to 0$, as $n \to \infty$
(2) $\sum_{j=2}^{n} (1-r_n') \ldots (1-r_j')\, |r_j'^{1/\tilde{M}} - r_{j-1}'^{1/\tilde{M}}| \to 0$, as $n \to \infty$
where $r_j' = \frac{1}{2} r_j \min_{i,k,\ell} T_i^{k,\ell}$.

Then $\lim_{n\to\infty} y(n) = \tilde{a}^{(0)}$.

PROOF.

(a) part (a) follows immediately from part (b) by letting n tend to infinity on both sides of (2.10) and by observing that the value of a matrix game depends continuously upon its entries (cf. (1.1).

(b) Note that

$$(2.11)\qquad \tilde{V}(r_n)_i = \mathrm{val}_{[K(i,a^{(M)}),L(i,a^{(M)})]}[\tilde{q}_i^{k,\ell} + \textstyle\sum_j m_{ij}^{k,\ell}(r_n)\,\tilde{V}(r_n)_j],\ i \in \Omega$$

and conclude from (1.1) that $|y(n+1) - \tilde{V}(r_n)_i| \le$
$(\max_{i,k,\ell} \sum_j m_{ij}^{k,\ell}(r_n)).\ |y(n) - \tilde{V}(r_n)| \le (1-\frac{1}{2}r_n.T_{min})\ |y(n) - \tilde{V}(r_n)|$
$= (1-r'_n)\ |y(n) - \tilde{V}(r_n)|$, for all $i \in \Omega$,
where the second inequality was shown in (1.12) to hold for all r_n sufficiently close to zero, i.e. for all $n \ge n_0$ (say). As a consequence of theorem 2.2, part (d), we may fix an integer $n_1 \ge n_0$ such that for all $m \ge n_1$:
$|\tilde{V}(r_{m+1}) - \tilde{V}(r_m)| \le B|r_{m+1}^{1/\tilde{M}} - r_m^{1/\tilde{M}}| = B'|r'^{1/\tilde{M}}_{m+1} - r'^{1/\tilde{M}}_m|$, where $B' = (2/T_{min})^{1/\tilde{M}}B$.
We conclude that for all m = 1,2, ... :

$$|y(n_1+m) - \tilde{V}(r_{n_1+m})| \le (1-r'_{n_1+m})\ |y(n_1+m-1) - \tilde{V}(r'_{n_1+m-1})|$$
$$+ B'(1-r'_{n_1+m})\ |r'^{1/\tilde{M}}_{n_1+m} - r'^{1/\tilde{M}}_{n_1+m-1}|$$

and by iterating this inequality, we finally obtain:

$$|y(n_1+m) - \tilde{V}(r_{n_1+m})| \le (1-r'_{n_1+m}) \ldots (1-r'_{n_1+1})\ |y(n_1) - \tilde{V}(r_{n_1})|$$
$$+ B' \sum_{j=n_1+1}^{n_1+m} (1-r'_{n_1+m}) \ldots (1-r'_j)\ |r'^{1/\tilde{M}}_j - r'^{1/\tilde{M}}_{j-1}|.$$

It follows from part (d) of th. 2.2 that $\lim_{n\to\infty} \tilde{V}(r_n) = \tilde{a}^{(0)}$ which in view of the above inequality enables us to conclude that $\lim_{n\to\infty} y(n) - \tilde{a}^{(0)} = \lim_{n\to\infty} y(n) - \tilde{V}(r_n) = 0$. □

We note that with $\tilde{q}_i^{k,\ell}$ redefined by $\tilde{q}_i^{k,\ell} = q_i^{k,\ell} - a_i^{(M)}\ T_i^{k,\ell}$ the analysis of th. 2.2 and th. 2.3 leads just as well to the existence of a solution to the following pair of f.e., whenever a stationary AEP exists:

$$(2.12) \qquad g_i = \mathrm{val}[\textstyle\sum_j P_{ij}^{k,\ell}\, g_j], \quad i \in \Omega$$

$$(2.13) \qquad v_i = \mathrm{val}_{[K(i,g),L(i,g)]}[q_i^{k,\ell} - g_i\, T_i^{k,\ell} + \textstyle\sum_j P_{ij}^{k,\ell}\, v_j], \quad i \in \Omega.$$

We next observe that conditions (1) and (2) of theorem 2.3, part (b) are satisfied for any choice:

$$r_n = n^{-b}, \quad \text{with } 0 < b \leq 1$$

as has been verified in FEDERGRUEN [7].

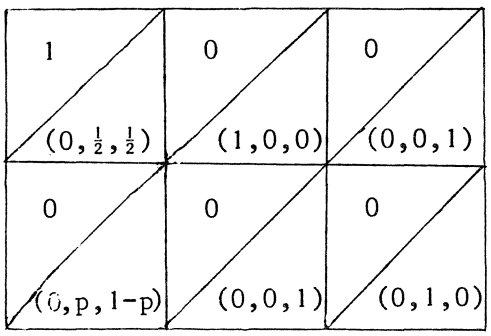
In addition, we note that when the asymptotic average value is independent of the initial state of the system, i.e. when $a_i^{(M)} = \langle a^M \rangle$ for all $i \in \Omega$, the f.e. (2.4) and (2.5), as well as (2.12) and (2.13) reduce to the single (vector)-equation:

$$(2.14) \qquad v_i^* = \mathrm{val}[q_i^{k,\ell} - \langle g^* \rangle\, T_i^{k,\ell} + \textstyle\sum_j P_{ij}^{k,\ell}\, v_j^*], \quad i \in \Omega$$

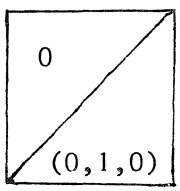
the discrete-time version of which has been considered in HOFFMAN and KARP [9]. In this case, the convergence result of part (b) of the previous theorem leads to a method for approximating the asymptotic average value by lower and upper bounds as well as for finding for both players and any $\varepsilon > 0$ stationary policies which are ε-optimal with respect to the average return criterion (cf. FEDERGRUEN [7]).

EXAMPLE 1 *shows that whereas the existence of a solution pair to the* f.e. (2.4) *and* (2.5) *is a necessary condition for the existence of a stationary AEP, it may fail to be sufficient:*

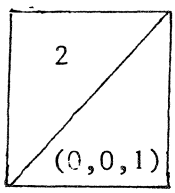
EXAMPLE 1: (all $\tau_{ij}^{k,\ell} = 1$; $i,j \in \Omega$; $k \in K(i)$, $\ell \in L(i)$.



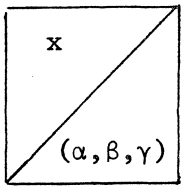
state 1



state 2



state 3

The notation 

means that if the players choose the row and column corresponding to this box, then player 2 pays player 1 the amount x and the next state is 1, with probability α, 2 with probability β and 3 with probability $\gamma = 1-\alpha-\beta$. Take $p = \frac{3}{4}$ and verify first that $a^{(M)} = [1,0,2]$. Note that $a_2^{(M)} = 0$ and $a_3^{(M)} = 2$, and verify that $x = 1$ is the *unique* solution to the equation:

$$(2.14) \qquad x = \text{val}\begin{bmatrix} 1 & x & 2 \\ 0.5 & 2 & 0 \end{bmatrix}$$

by distinguishing between the cases $x > 1$ and $x < 1$. Finally recall from (2.6) that $a_1^{(M)}$ is a solution to (2.14). We next verify that $a^{(M)}$ in combination with any vector $v^* \in E^3$ satisfying $v_1^* = \frac{1}{2}v_2^* + \frac{1}{2}v_3^* - \frac{1}{2}$, is a solution pair to (2.4) and (2.5).
Note that $K(1,a^{(M)}) = \{[1,0]\}$ and $L(1,a^{(M)}) = \{[y_1,y_2,y_3] | y_3 = 0, y_1+y_2=1, y_1 \geq \frac{2}{3}\}$, such that in this example (2.5) becomes:

$$v_1^* = \min\{\tfrac{1}{2}v_2^* + \tfrac{1}{2}v_3^*;\ -\tfrac{1}{3} + \tfrac{1}{3}v_1^* + \tfrac{1}{3}v_2^* + \tfrac{1}{3}v_3^*\};\ v_2^* = v_2^*;\ v_3^* = v_3^*$$

We next verify that there is no stationary AEP in this game. Note that a stationary policy for player 1 is completely specified by the probability x with which action 1 in state 1 is chosen. Likewise, a stationary policy for player 2, is specified by the probability vector (y_1,y_2,y_3) with which the available actions in state 1 are randomized.

If $x = 1$ and $y_2 = 1$: $g(f,h)_1 = 0$

if $x = 1$ and $y_2 < 1$: $g(f,h)_1 = (y_1+2y_3)/(y_1+y_3)$

if $x < 1$: $g(f,h)_1 = (\frac{1}{2}+\frac{1}{2}x)y_1+2y_2+2xy_3$

which shows that no pair of stationary policies is an AEP.

We conclude that in general, and in contrast to what is known to be the case for ordinary MRPs, a policy pair (f^*,h^*) *which satisfies the optimality equations* (2.4) *and* (2.5) *for some solution pair* (g^*,v^*), *does not need to be an AEP.*

Example 1, with $p = \frac{1}{2}$ shows that this may even be the case when stationary AEPs do exist:
Note that with $p = \frac{1}{2}$, $g^* = [1,0,2]$ satisfies (2.4), by verifying:

$$1 = \text{val}\begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 0 \end{bmatrix},$$

and conclude that $K(1,g^*) = \{[x,1-x] \mid \frac{1}{2} \leq x \leq 1\}$ and $L(1,g^*) = \{[1,0,0]\}$.
The optimality equation (2.5) thus becomes:

$$v_1^* = \max\{\tfrac{1}{2}v_1^* + \tfrac{1}{2}v_2^*;\ \tfrac{1}{2}v_1^* + \tfrac{1}{2}v_2^* - \tfrac{1}{2}\};\ v_2^* = v_2^*;\ v_3^* = v_3^*$$

such that g^* in combination with any vector $v^* \in E^3$ satisfying $v_1^* = \frac{1}{2}v_2^* + \frac{1}{2}v_3^*$, is a solution pair to (2.4) and (2.5). Verify that any pair of policies (f^*,h^*) with $\frac{1}{2} \leq f_{11}^* < 1$ and $h_{11}^* = 1$ is a stationary AEP, whereas the only pair of policies which satisfies the optimality equations (2.4) and (2.5) for (g^*,v^*) has $f_{11}^* = 1$ and $h_{11}^* = 1$ and is *not* an AEP.
Observe finally (by considering the gain rate of one of the AEPs) that $g^* = a^{(M)}$.

We conclude that even when stationary AEPs do exist, such policy pairs do not necessarily need to be found within the class of (pairs of) policies that satisfy the optimality equations for some solution pair (the existence of which follows from th. 2.3).

Whereas the above examples illustrate that no *full* optimality results may be obtained for policy pairs that satisfy the optimality equations (2.4) and (2.5) for some solution pair (g^*,v^*), in proposition 2.4 below a *restricted optimality result* is derived.

PROPOSITION 2.4: *Let* (f^*,h^*) *be a policy pair which satisfies the optimality equations* (2.4) *and* (2.5) *for some solution pair* (g^*,v^*).
Then $g(f^*,h^*) \leq g(f^*,h)$ *for all policies* h, *for which for all* $i \in \Omega$.

(2.15) $\quad g_i^* = P(f^*,h)g_i^* \Rightarrow h(i)$ is an *optimal* action in the matrix game in (2.4);

with the same *restricted* optimality result holding for player 1, when player 2 ties himself down to policy h^*.

PROOF. Fix a policy h which satisfies (2.15). Recall from the proof of part (a) of th. 2.2 that g^* is constant on each of the subchains of $P(f^*,h)$ and conclude that:

(1) $g^*_i \leq P(f^*,h)g^*_i$, with

(2) $v_i \leq q(f^*,h)_i - g^*_i\, T(f^*,h)_i + P(f^*,h)v^*_i$, for all states $i \in R(f^*,h)$ for which (1) holds with strict equality. Apply the proof of lemma 4, part (a) in [5] to verify that $g^* \leq g(f^*,h)$, with strict equality holding for $h = h^*$. □

Let in example 1, $f^k(h^k)$, $k = 1,2$ be the pure policy for player 1 (2) which has $1 = f^k_{1k} = (h^k_{1k})$. Note that both for $p = \frac{1}{2}$ and $p = \frac{3}{4}$, (f^1,h^1) satisfies the restricted optimality result of prop. 2.4, but fails to be an AEP, since $0 = g(f^1,h^2)_1 < g(f^1,h^1)_1 = 1$. Observe that h^2 satisfies $g^* = P(f^1,h^2)g^*$ but $h^2(1)$ is *not* an *optimal* action in the matrix game in (2.4).

Finally note that, whereas a stationary AEP doesnot need to satisfy *both* optimality equations (2.4) and (2.5) for any solution pair (g^*,v^*) (cf. example 1 with $p = \frac{1}{2}$), it will certainly have to satisfy the first f.e. for $g^* = a^{(M)}$.

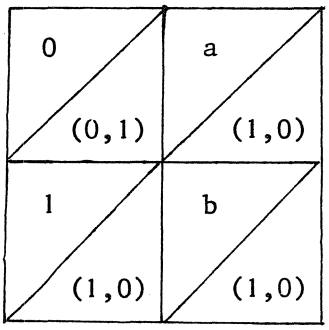
REMARK. In ordinary MRP's, a policy f, in order to be maximal gain, needs to satisfy the *second* optimality equation (2.5) *only* in its *recurrent* states (cf. th. 3.1 part (e) of [17]). In the general SDG or SRG model however, we couldnot weaken the prerequisite in proposition 2.4, to the assumption:

(a) $(f^*(i), h^*(i)$ is an equilibrium pair of actions in the matrix game in (2.4) for *every* $i \in \Omega$

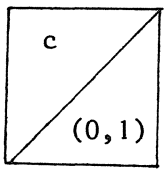
(b) $(f^*(i), h^*(i))$ is an equilibrium pair of actions in the matrix game in (2.5) for *every* $i \in R(f^*,h^*)$

even when confining ourselves to the *restricted* optimality result in prop. 2.4, as is illustrated by example 2:

EXAMPLE 2: Consider the *SDG*-model:



state 1



state 2

Let $f^k(h^k)$; k = 1,2 be defined as above. Take a = 0, b = 2, c = 1. Note that (f^2,h^1) is an AEP such that $a^{(M)} = [1,1]$ and verify that in this example $(a^{(M)},[0,1])$ is a solution pair to (2.4) and (2.5). Note that (f^1,h^1) satisfies (2.4) in every $i \in \Omega$, and (2.5) in every $i \in R(f^1,h^1) = \{2\}$; however $0 = g(f^1,h^2)_1 < g(f^1,h^1)_1 = 1$ in spite of h^2 satisfying condition (2.15) in proposition (2.4).

Prop. 2.4 makes clear that a policy pair which satisfies (2.4) and (2.5) may fail to be an AEP, only when one of the sets $K(i,g^*)$ or $L(i,g^*)$ $(i\in\Omega)$ is a *strict* subset of $K(i)$ or $L(i)$. As a consequence no problems arise when the asymptotic average value is independent of the initial state of the system:

COROLLARY 2.5. *Assume* $a_i^{(M)} = \langle a^{(M)} \rangle$ *for all* $i \in \Omega$. *Then the following statements are equivalent:*

(I) $a^{(k)} = 0$, *for* k = 1,...,M-1

(II) *there exists a stationary* AEP

(III) *the functional equations* (2.4) *and* (2.5) *have a solution pair* $(a^{(M)},v^*)$.

In addition, under either one of (I), (II) or (III), any policy pair which satisfies the funct. eq. (2.4) and (2.5) for some solution pair $(a^{(M)},v^*)$ in an AEP.

PROOF.

(I) ⇒ (II): Consider the SRG, in which merely the one-step expected rewards are tranformed as in (2.7), i.e. $\tilde{q}_i^{k,\ell} = q_i^{k,\ell} - T_i^{k,\ell} a^{(M)}$. Note that various transformations on $\tilde{\rho}_i^{k,\ell}(r)$ imply this transformation on the $\tilde{q}_i^{k,\ell} = \tilde{\rho}_i^{k,\ell}(0)$. Choose e.g. $\tilde{\rho}_i^{k,\ell}(r) = \rho_i^{k,\ell}(r) - \frac{a^{(M)}}{r}(\sum_j m_{ij}^{k,\ell}(r)-1)$ (to verify the latter, consider the first two terms in the Taylor series expansion of $m_{ij}^{k,\ell}(r)$). Finally, subtract $a^{(M)}/r$ from both sides of (2.7) to obtain: $V(r)_i - a^{(M)}/r = \text{val}[\tilde{\rho}_i^{k,\ell}(r) + \sum_j m_{ij}^{k,\ell}(r)[V(r) - \frac{a^{(M)}}{r}]]$, $i \in \Omega$ and conclude that $\tilde{V}(r) = V(r) - a^{(M)}/r$; next apply th. 2.3

(II) ⇒ (I): cf. th. 2.2 part (b)

(III) ⇒ (II): follows from prop. 2.4, by taking any pair of policies which satisfies the funct. eq. (2.4) and (2.5) for $(a^{(M)}, v^*)$, thus proving the last assertion of the corollary, at one blow. □

REMARK. The implication (III) ⇒ (II) even holds for a denumerable state space (cf. e.g. [6], th. 2). Observe that when the asymptotic average value does depend upon the initial state of the system, (I) and (II) do not need to be equivalent, i.e. (I) may fail to imply (II); as an example take the Big Match (cf. [8]) which has even a Laurent series expansion for V(r), i.e. which has M = 1 (cf. [3], section 8).

3. SOME PROPERTIES OF THE SOLUTION SPACE OF THE OPTIMALITY EQUATIONS

In this section, we discuss a number of properties of the functional equations (2.4) and (2.5) which we will need in the following section. We first observe that in general (2.4) and (2.5) may fail to have a solution pair, just like there may fail to be (stationary) AEPs. As an example, take ex.2 with a = 1, b = c = 0, which appeared first in STERN [21] and was used in BEWLEY and KOHLBERG ([3], sect. 11). Note from [3], that this example has $\underline{0}$ as its asymptotic average value vector, but has no stationary AEP and apply cor. 2.5 (or alternatively note that in this example M = 2, and $a^{(1)} = [1,0]$; and apply cor. 2.5).

Next, whenever a solution pair (g^*,v^*) exists to the optimality equations (2.4) and (2.5), the v^*-part of the solution is obviously not uniquely determined (note e.g. that if (g^*,v^*) is a solution pair, then so is (g^*,v^*+c1) for any scalar c; cf. also [17], where a complete characterization of the solution space was given, for the case of ordinary MRPs). In addition, since a pair of policies (f^*,h^*) which satisfies the optimality equations, does not need to be an AEP, it is still unclear to us whether the g^*-part is always uniquely determined. All of the above difficulties arise, in view of the chain structure being discontinuous on $\phi \times \Psi$ in the general multichain case. Theorem 3.1 below gives a number of characterizations with respect to the optimality equations, under condition (U). Since in the following section, optimality equations of a slightly more general structure will appear, we formulate and derive our results with respect to the f.e.:

$$(3.1) \qquad x_i = \mathrm{val}_{[\hat{K}(i),\hat{L}(i)]}[a_i^{k,\ell} + \textstyle\sum_j P_{ij}^{k,\ell}\, x_j], \quad i \in \Omega$$

$$(3.2) \qquad y_i = \mathrm{val}_{[\tilde{K}(i,x^*),\tilde{L}(i,x^*)]}[c_i^{k,\ell} - \textstyle\sum_j H_{ij}^{k,\ell}\, x_j + \sum_j P_{ij}^{k,\ell}\, y_j], \; i \in \Omega$$

where for each $i \in \Omega$, $\hat{K}(i)$ and $\hat{L}(i)$ are closed convex polyhedral subsets of $K(i)$ and $L(i)$, and where for each solution x^* to (3.1), $\tilde{K}(i,x^*)$ and $\tilde{L}(i,x^*)$ are the sets of optimal actions in the matrix games in (3.1) with $x = x^*$; $a_i^{k,\ell}$ and $c_i^{k,\ell}$ are given quantities ($i \in \Omega$, $k \in K(i)$, $\ell \in L(i)$).

THEOREM 3.1.

(a) (3.1) *has a solution* x^*, *if and only if*

(3.3) *the* SDG *with* $\tilde{q}_i^{k,\ell} = a_i^{k,\ell}$ *has a stationary* AEP, *and* $\underline{0}$ *as its asymptotic average value vector*

(b) *Assume condition* (U) *to be satisfied. Then if* (3.3) *holds:*

(1) *The solution* x^* *to* (3.1) *is unique up to a multiple of* $\underline{1}$, *such that the sets* $\tilde{K}(i) = K(i,x^*)$ *and* $\tilde{L}(i) = L(i,x^*)$, $i \in \Omega$, *are uniquely determined.*

(2) *A solution* (x^*,y^*) *to* (3.1) *exists, where* x^* *is uniquely determined by:*

$$(3.4)\qquad x_i^* = x_i^o + \max_{f\in X_i\tilde{K}(i)} \min_{h\in X_i\tilde{L}(i)} \frac{\langle \pi^1(f,h), c(f,h)-H(f,h)r^o\rangle}{\langle \pi^2(f,h), T(f,h)\rangle}$$

$$= x_i^0 + \min_{h\in X_i\tilde{L}(i)} \max_{f\in X_i\tilde{K}(i)} \frac{\langle \pi^1(f,h), c(f,h)-H(f,h)x^o\rangle}{\langle \pi^1(f,h), T(f,h)\rangle}$$

where x^o *denotes some solution to* (3.1). *Moreover,* y^* *is unique up to a multiple of* $\underline{1}$.

PROOF.

(a) immediate form corollary 2.5

(b) (1): Let x^o, x^1 be two solutions to (3.1) and let (f^o,h^o) and (f^1,h^1) be two pairs of policies which satisfy (3.1) for x^o and x^1 resp. Note that: $x^o \le a(f^o,h^1) + P(f^o,h^1)\, x^o$ and $x^1 \ge a(f^o,h^1) + P(f^o,h^1)\, x^1$ and subtract the second inequality from the first one, in order to obtain: $x^o - x^1 \le P(f^o,h^1)\,[x^o-x^1]$, and by iterating the latter:

$$(3.5)\qquad [x^o-x^1]_i \le c_1 = \langle \pi^{(1)}(f^o,h^1),\, x^o - x^1\rangle, \quad i \in \Omega.$$

Similarly, we obtain

$$\langle \pi^{(1)}(f^1,h^o),\, x^o - x^1\rangle = c_2 \le [x^o-x^1]_i, \quad i \in \Omega.$$

We finally show $c_1 = c_2$, which proves part (a). Multiply both sides of (3.5) by $\pi^{(1)}(f^o,h^1)$ in order to conclude that $x_i^o - x_i^1 = c_1$, for all $i \in R(f^o,h^1)$. Similarly we obtain $x_i^o - x_i^1 = c_2$ for all $i \in R(f^1,h^o)$ which implies $c_1=c_2=c$, as a consequence of $R(f^1,h^o) \cap R(f^o,h^1) \neq \emptyset$, in view of assumption (U).

(2): Fix a solution x^o to (3.1), and consider the SRG, which has Ω as its state space, $\tilde{K}(i)$ and $\tilde{L}(i)$ as the sets of (randomized) alternatives available to player 1 and 2 and with one-step expected rewards $\tilde{q}_i^{k,\ell} = c_i^{k,\ell} - \sum_j H_{ij}^{k,\ell}\, x_j^o$ and unaltered transition probabilities and transition time distributions. Note from lemma 1.1 that each of the sets $\tilde{K}(i)$ and $\tilde{L}(i)$ may be considered as the set of randomizations of a finite number of (pure) alternatives. This, in combination with condition (U), implies as a result of lemma 2.1, and cor. 2.5 the existence of a solution to the f.e.:

$$y_i^* = \mathrm{val}_{[\tilde{K}(i),\tilde{L}(i)]}[c_i^{k,\ell} - \sum_j H_{ij}^{k,\ell} x_j^o - g_i^o T_i^{k,\ell} + \sum_j P_{ij}^{k,\ell} y_j^*], \quad i \in \Omega$$

where g^o is the asymptotic average value vector in this stochastic game with $g_i^o = \langle g^o \rangle$, $i \in \Omega$ in view of our unichainedness-assumption. This implies that $x^* = x^o + g^o$ is a solution to (3.2), thus showing the existence of a solution pair to (3.1) and (3.2). We next show that the x^*-part is uniquely determined, and derive its explicit expression. The fact that the y^*-part is unique up to a multiple of $\underline{1}$ follows as in part (b) (1). Let (x^*,y^*) be a solution to (3.1) and (3.2), and let (f^*,h^*) be a policy pair which satisfies (3.1) and (3.2) for (x^*,y^*). Let $g^o = x^* - x^o$, Then for each $h \in X_i \tilde{L}(i)$:

$$y^* \geq c(f^*,h) - H(f^*,h)\, x^o - \langle g^o \rangle\, T(f^*,h) + P(f^*,h)y^*.$$

Multiply this inequality by $\Pi(f^*,h)$ and conclude that:

$$(3.6) \qquad g^o \geq \frac{\langle \pi^1(f^*,h), c(f^*,h) - H(f^*,h)x^o \rangle}{\langle \pi^1(f^*,h), T(f^*,h) \rangle}$$

with strict equality holding for $h = h^*$. Likewise, one can show that:

$$g^o \leq \langle \pi^1(f,h^*),\, c(f,h^*) - H(f,h^*)x^o \rangle / \langle \pi^1(f,h^*)\, T(f,h^*) \rangle$$

for all $f \in X_i \tilde{K}(i)$, with strict equality for $f = f^*$, thus completing the proof of part (b). □

4. SENSITIVE DISCOUNT AND CUMULATIVE AVERAGE OPTIMALITY.

In this section, we consider a sequence of increasingly selective optimality criteria, which appears as the natural extension to the SRG-model of the sensitive discount (or cumulative average) optimality criteria, as formulated and studied in Markov Decision Theory (cf. e.g. [4],[13],[22]).

We call a policy pair (φ^*,ψ^*) a n-discount equilibrium pair of policies *(n-EP)* ($n = -1,0,\ldots$), if:

$$(4.1) \qquad \limsup_{r \downarrow 0} r^{-n}[V(\varphi^*,\psi^*)(r) - V(\varphi^*,\psi)(r)] \leq 0 \leq$$

$$\liminf_{r \downarrow 0} r^{-n}[V(\varphi^*,\psi^*)(r) - V(\varphi,\psi^*)(r)]$$

where $V(\varphi,\psi)(r)$ denotes the total discounted return to player 1, when the players use policies φ,ψ and when the rewards are discounted at rate r.

We restrict our analysis to the sensitive discount criteria for the discrete-time case of SDGs, in order to avoid too burdensome a notation. The extension of our results to the general SRG-case, is immediate and the analysis of the cumulative average optimality criteria is analogous to the one given below, with the same sequence of f.e. associated (note that for $n = -1$, equivalence of the two criteria was proven in BEWLEY and KOHLBERG [3]). E.g. whereas in the general SRG-model the expressions in the various f.e., to be considered below, become more complicated functions of the terms in the expansions of $\rho_i^{k,\ell}(r)$ and $m_{ij}^{k,\ell}(r)$ (cf. DENARDO [4]), the structure of each consecutive pair of f.e. is exactly identical to the one of (3.1) and (3.2). Consider the following sequence of optimality equations:

$$(4.2) \qquad g_i = \mathrm{val}[\textstyle\sum_j P_{ij}^{k,\ell}\, g_j], \quad i \in \Omega$$

$$(4.3) \qquad x(0)_i = \mathrm{val}_{[K(i,g),L(i,g)]}[q_i^{k,\ell} - g_i + \textstyle\sum_j P_{ij}^{k,\ell}\, x(0)_j], \quad i \in \Omega$$

$$(4.4) \qquad x(m)_i = \mathrm{val}_{[K^{(m)}(i,X(m-1)),L^{(m)}(i,X(m-1))]}[-x(m-1)_i + \textstyle\sum_j P_{ij}^{k,\ell}\, x(m)_j],$$
$$m = 1,2,\ldots,\ i \in \Omega$$

where $X(m)$ denotes the $m+2$-tuple of vectors $(g,x(0),\ldots,x(m))$, $m = 0,1,\ldots$. In addition for all $m = 1,2,\ldots$ and $i \in \Omega$ and any solution $X(m-1)$ to the first $m+1$ f.e. in (4.2)-(4.4), $K^{(m)}(i,X(m-1))$ and $L^{(m)}(i,(m-1))$ denote the sets of optimal actions in the $m+1$-st f.e.

For each stationary pair of policies (f,h) let:

$$(4.5) \qquad V(f,h)(r) = g(f,h)/_r + \textstyle\sum_{k=0}^{\infty} x^{(k)}(f,h) r^k$$

represent the Laurent series expansion of the total discounted return associated with (f,h). Finally, if x is a vector, we say x is *lexicographically non-negative* written $x \gtrdot 0$, if the first nonvanishing element of x is positive. Similarly, x is called *lexicographically positive*, written $x \gtrdot 0$ if $x \gtrdot 0$ and $x \neq 0$. We write $x \gtrdot (>) y$ or $y \lessdot (<) x$ if $x-y \gtrdot (>) 0$

THEOREM 4.1.

(a) *Let* (f^*,h^*) *be a stationary* n-EP (n = -1,0,...). *Then*

(1) *There exists a* n+3*-tuple* $(g^*,v,\ldots,x^{(n+1)})$ *which satisfies* (4.2), (4.3) *and the first* n+1 f.e. *of* (4.4).

(2) *In the Puiseux series expansion of* V(r), *we have:*

$$(4.6)\qquad a^{(-\ell M)} = \begin{cases} g(f^*,h^*), & \text{for } \ell = -1 \\ x^{(\ell)}(f^*,h^*), & \text{for } \ell = 0,\ldots,n \end{cases}$$

$$a^{(-\ell M-p)} = 0, \text{ for } \ell = -1,\ldots,n; \qquad p = 1,\ldots,M-1$$

(b) *Let* (f^*,h^*) *be a stationary* N-EP. *Then*

(1) (f^*,h^*) *is a* n-EP *for all* $n \geq N$

(2) V(r) *has a Laurent series expansion.*

PROOF.

(a) (1) For n = -1 the th. holds as a consequence of th. 2.3; hence we assume $n \geq 0$. Note that $h^*(f^*)$ is a n-optimal policy in the MDP which results for player 2 (1) when player 1 (2) ties himself down to policy $f^*(h^*)$. Use th. 4 of [22] to conclude that for all $i \in \Omega$, $f \in \Phi$, $h \in \Psi$:

$$(4.7)\qquad [\textstyle\sum_j P(f,h^*)_{ij} g(f^*,h^*)_j ; q(f,h^*)_i - g(f^*,h^*)_i + \sum_j P(f,h^*)_{ij} x^{(0)}(f^*,h^*)_j$$

$$; \ldots ; - x^{(n-1)}(f^*,h^*)_i + \textstyle\sum_j P(f,h^*)_{ij}\, x^{(n)}(f^*,h^*)_j] \leqslant$$

$$[g(f^*,h^*)_i ;\ x^{(0)}(f^*,h^*)_i ; \ldots ; x^{(n)}(f^*,h^*)_i] \leqslant$$

$$[\textstyle\sum_j P(f^*,h)_{ij} g(f^*,h^*)_j ; q(f^*,h)_i - g(f^*,h^*)_i + \sum_j P(f^*,h)_{ij} x^{(0)}(f^*,h^*)_j$$

$$; \ldots ; - x^{(n-1)}(f^*,h^*)_i + \textstyle\sum_j P(f^*,h)_{ij}\, x^{(n)}(f^*,h^*)_j]$$

with strict equality holding for $h = h^*$, and $f = f^*$. One easily concludes that $X(n) = [g(f^*,h^*);\ldots;x^{(n)}(f^*,h^*)]$ satisfy (4.2), (4.3) and the first n f.e. in (4.4). To prove that there exists a solution to the n+1-st f.e. in (4.4) as well, note that (f^*,h^*) is an AEP in the stochastic game, which has Ω as its state space, $K^{(n+1)}(i,X(n))$ and $L^{(n+1)}(i,X(n))$ as the action spaces in state $i \in \Omega$, and with one-step

expected rewards $\tilde{q}_i^{k,\ell} = -x^{(n)}(f^*,h^*)_i$, and transition probabilities $\tilde{P}_{ij}^{k,\ell} = P_{ij}^{k,\ell}$. (cf. e.g. DENARDO [4], p. 491). Finally note that this stochastic game has $\underline{0}$ as its asymptotic average value vector, and apply th. 2.3.

(2) We prove part (a) (2) by complete induction with respect to ℓ. Note that for $\ell = -1$, the equalities $a^{(-\ell M)} = g(f^*,h^*)$ and $a^{(-\ell M-p)} = 0$ for $p = 1,\ldots,M-1$ follow from the fact that (f^*,h^*) is a stationary AEP, using lemma 2.2. Assume that (4.6) holds for all $\ell = -1,\ldots,\ell^* < n$. Let

$$\sum_{k=-\infty}^{1} A^{(k)} r^{-k} [\sum_{k=-\infty}^{1} B^{(k)} r^{-k}]$$

be the Laurent series expansion of $W^1(r)[W^{(2)}(r)]$, the total discounted return to player 1 [2] in the MDP that results when player 2 [1] ties himself down to policy $h^*[f^*]$ and note that

$$A^{(1)} = B^{(1)} = g(f^*,h^*) \text{ and } A^{(-k)} = B^{(-k)} = x^{(k)}(f^*,h^*) \quad \text{for all } k = 0,\ldots,\ell^*+1 \tag{4.8}$$

in view of $f^*[h^*]$ being ℓ^*+1- optimal in this MDP. Observe that, $W^2(r) \leq V(r) \leq W^1(r)$, and conclude from (4.8) and the induction assumption that the coefficients of the terms with power strictly less than ℓ^*+1, in $W^1(r)$, $V(r)$ and $W^2(r)$ coincide. Since $A^{(-\ell^*-1)} = B^{(-\ell^*-1)}$ we conclude that (4.6) holds for $\ell = \ell^*+1$ as well.

(b) (1): It follows from VEINOTT ([23], p. 1646) that $f^*[h^*]$, since being N-optimal, is n-optimal for all $n \geq N$ in the MDP that results when player 2 [1] ties himself down to policy $h^*[f^*]$.

(b) (2): Immediate from (a) (2) and (b) (1). □

We observe that part (b) of th. 4.1 may not be extended to the general SRG-model, since it does not even hold in the general MRP-case (cf. [4], p. 489). However, part (b) generalizes proportion 6.4 in [3], where it was shown that $V(r)$ has a Laurent series expansion, if there exists a uniformly *discount optimal* pair of policies, i.e. a pair (f^*,h^*) which is optimal in the r-discount game for all $r > 0$ sufficiently small. We next observe, that whereas the existence of a solution to the first n+1 f.e. in (4.1) is a

necessary condition for the existence of a stationary n-EP, it certainly may fail to be sufficient, as was pointed out for the case n = -1, in section 2.

In analogy to prop. 2.4, the following *partial* optimality result may be obtained for any policy pair which satisfies (4.2),(4.3) and the first n+1 f.e. in (4.4) for some solution $(g^*,x^*(0),\ldots,x^*(n+1))$:

PROPOSITION 4.2. *Fix* n = -1,0,... . *Let* $(g^*,x^*(0),\ldots,x^*(n+1))$ *be a solution to* (4.2),(4.3) *and the first* n+1 f.e. *of* (4.3), *and let* $(f^*,h^*) \in \Phi \times \Psi$ *be a policy pair which satisfies these optimality equations for this solution. Then*

$$[g(f^*,h^*)_i; x^{(0)}(f^*,h^*)_i;\ldots;x^{(n)}(f^*,h^*)_i] \lesssim [g(f^*,h)_i;\ldots;x^{(n)}(f^*,h)_i]$$

holds in every $i \in \Omega$, *for those policies* h *for which:*

$$(4.9)\qquad \sum_j P(f^*,h)_{ij}\, g_j^* = g_i^* \Rightarrow h(i) \in L(i,g^*)$$

$$\{h(i) \in L(i,g^*) \text{ and } q(f^*,h)_i - g_i^* + \sum_j P(f^*,h)_{ij}\, x_j^{*(0)}\} \Rightarrow h(i) \in L^{(1)}(i,X^*(0))$$

$$\{h(i) \in L^{(m)}(i,X^*(m-1)) \text{ and } -x_i^{*(m-1)} + \sum_j P(f^*,h)_{ij}\, x_j^{*(m)}\} \Rightarrow h(i) \in L^{(m+1)}(i,X^*(m)),$$
$$1 \le m \le n+1$$

with the same restricted optimality result holding for policy f^*. □

We finally turn to the case where condition (U) is satisfied:

THEOREM 4.3. *Assume condition* (U) *holds. Then*

(a) *there exists a solution to the entire sequence of* f.e. (4.2),(4.3) *and* (4.4).

(b) *Fix* n = 0,1,... . *In the solution* $(g^*,x^*(0),\ldots x^*(n))$ *to* (4.2),(4.3) *and the first* n f.e. *of* (4.4), *we have* $(g^*,x^*(0),\ldots,x^*(n-1))$ *uniquely determined (explicit expressions of which may be obtained by a repeated application of* th. 3.1*), whereas* $x^*(n)$ *is unique up to a multiple of* $\underline{1}$.

PROOF. Part (a) follows from part (b), and part (b) is proven by complete induction with respect to n. Note that for n = 0, the assertion follows as

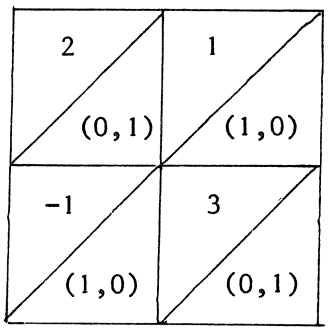
a special case of th. 3.1. Assume, it holds for some n = 0,1,... , We then have in particular that $x^*(n-1)$ (or g^* when n = 0) is uniquely determined and that the f.e.:

$$(4.10) \qquad x(n)_i = \mathrm{val}_{[K^{(n)}(i,X(n-1));L^{(n)}(i,X(n-1))]}[-\dot{x}(n-1)_i + \sum_j P_{ij}^{k,\ell} x(n)_j], \qquad i \in \Omega$$

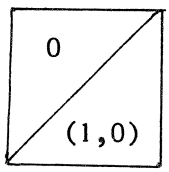
(or (4.3) in case n = 0) has a solution. Apply th. 3.1 to the combination of (4.10) (or (4.3) in case n = 0) and the n+1-st f.e. in (4.4), to verify that the assertion holds for n+1 as well. □

In SOBEL ([20], th. 2), it was asserted that a stationary 1-EP always exists under condition (U). In [6] we pointed out that the proof of this theorem is incorrect, and the next example shows that the asserted result itself *may fail to hold*:

EXAMPLE 3:



state 1



state 2

Note that $g(f,h) = (5xy-2x-4y+3)/(2+2xy-x-y)$ where $x = f_{11}$ and $y = h_{11}$. Conclude that $\{f^* | f^*_{11} = 1\}$ and $\{h^* | h^*_{11} \geq \frac{1}{3}\}$ are the sets of optimal (stationary) polices for players 1 and 2, with respect to the average return per unit time criterion. Note however that none of these policy pairs is 1-EP since $x^{(0)}(f^*,h) = 0$ when $h_{11} = 0$, whereas $x^{(0)}(f^*,h^*)_1 \geq \frac{1}{4}$ for all policies h^* which are gain optimal for player 2.

REFERENCES

[1] BATHER, J., *Optimal decision procedures for finite Markov Chains*, Part II, Adv. Appl. Prob. 5 (1973), 521-540.

[2] BEWLEY, T. and E. KOHLBERG, *The asymptotic theory of Stochastic Games*, to appear in Math. of O.R. (1976).

[3] BEWLEY, T. and E. KOHLBERG, *On Stochastic Games with stationary optimal strategies*, Tech. Report no. 23 Harvard Institute of Economic Research, Harvard University (1976).

[4] DENARDO, E., *Markov Renwal Programs with small interest rates*, Ann. of Math. Stat. 42 (1971), 477-496.

[5] DENARDO, E. and B. FOX, *Multichain Markov Renewal Programs*, SIAM, J. Appl. 16 (1968), 468-487.

[6] FEDERGRUEN, A., *On* N*-person Stochastic Games with denumerable state space*, Math. Center Report BW 67/76, (1976).

[7] FEDERGRUEN, A., *Successive approximation methods in undiscounted sequential games*, forthcoming (1977).

[8] GILLETTE, D., *Stochastic Games with zero stop probabilities in* M. Dresher et al. (eds.) *Contributions to the theory of games*, Vol. III (Princeton Univ. Press), Princeton, New Jersey (1957), 179-188.

[9] HOFFMAN, A. and R. KARP, *On non-terminating Stochastic Games*, Man. Sci. 12 (1966), 359-370.

[10] HORDIJK, A. and H. TIJMS, *A modified form of the iterative method of Dynamic Programming*, Ann. of Stat. 3 (1975), 203-208.

[11] HOWARD, R., *Dynamic Programming and Markov Processes*, Technology Press and Wiley, New York (1960).

[12] KARLIN, S., *Mathematical Methods and the Theory of Games*, Vol. I, Addison Wesley, London (1959).

[13] MILLER, B. & A. VEINOTT, Jr., *Discrete Dynamic Programming with a small interest rate*, Ann. Math. Stat. 40, 366-370.

[14] PARTHASARATHY, T. & M. STERN, *Markov Games a survey*, University of Illinois, Chicago (1976).

[15] ROGERS, P., *Nonzero-sum Stochastic Games*, Report ORC 69-8, Op. Res. Center, Univ. of California, Berkeley (1969).

[16] SCHWEITZER, P., *Iterative Solution of the Functional Equations of undiscounted Markov Renewal Programming*, J.M.A.A. 34 (1971), 495-501.

[17] SCHWEITZER, P. and A. FEDERGRUEN, *Functional Equations of Undiscounted Markov Renewal Programming*, Math. Center Report BW 70/77, (1976) (to appear in Math. O.R.)

[18] SHAPLEY, L., *Stochastic Games*, Proc. Nat. Acad. Sci. U.S.A. 39 (1953), 1095-1100.

[19] SLADKY, K., *On the set of optimal controls for Markov Chains with rewards*, Kybernetika 10, 350-367.

[20] SOBEL, M., *Noncooperative Stochastic Games*, Ann. of Math. Stat. 42 (1971), 1930-1935.

[21] STERN, M., *On Stochastic Games with limiting average payoff*, Ph.D. dissertion, Dept. of Math., Univ. of Illinois, Chicago Circle Campus (1975).

[22] VEINOTT, A. Jr., *Discrete Dynamic Programming with sensitive discount optimality criteria*, Ann. Stat. 40 (1969), 1635-1660.