stichting
mathematisch
centrum

MC

AFDELING MATHEMATISCHE BESLISKUNDE
(DEPARTMENT OF OPERATIONS RESEARCH)

BW 75/77 MEI

A. FEDERGRUEN

SUCCESSIVE APPROXIMATION METHODS IN UNDISCOUNTED STOCHASTIC GAMES

Preprint

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): 93E05

Successive approximation methods in undiscounted stochastic games *)

by

A. Federgruen

ABSTRACT

This paper considers undiscounted two person zero-sum sequential games with finite state and action spaces. Under conditions that guarantee the existence of stationary optimal strategies, we present two successive approximation methods for finding the optimal gain rate, a solution to the optimality equation, and for any $\varepsilon > 0$, ε-optimal policies for both players.

KEY WORDS & PHRASES: *stochastic games, successive approximation, average return per unit time criterion, equilibrium policies.*

---

*) This report will be submitted for publication elsewhere

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: $P_{ij}^{k,\ell} \geq 0$ denotes the probability that state j is the next state to be observed ($\Sigma_{j=1}^{N} P_{ij}^{k,\ell} = 1$); $q_i^{k,\ell}$ is the one-step expected reward earned by player 1 from player 2 and $T_i^{k,\ell}$ denotes the expected holding time in state i. Throughout this paper we assume that $T_i^{k,\ell} > 0$ ($i \in \Omega$; $k \in K(i)$; $\ell \in L(i)$).

The discrete time case, where each transition takes exactly one unit of time, is known as the stochastic games-model (cf. e.g. [9], [18]) and will be denoted as the *SDG-case*. If the payoffs are discounted at the interest rate $r > 0$, the SRG-game is called the *r-discounted game*. The existence of a value and of stationary optimal policies in the r-discount game goes essentially back to SHAPLEY [18]; in addition it is easily verified that value-iteration converges to the value of the game, in view of the value-iteration operator being a contraction mapping on E^N, the N-dimensional Euclidean space.

In the *undiscounted* version of the game, i.e. when the long run average return per unit time is the criterion to be considered one or both players may fail to have optimal policies as follows from an example in GILLETTE [9]. Both for this model and for the case of more general state and action spaces, recurrency conditions with respect to the transition probability matrices (*tpm's*) associated with the stationary policies have been obtained under which the existence of a stationary pair of equilibrium policies (*AEP*) is guaranteed (cf. HOFFMANN & KARP [11], SOBEL [19], ROGERS [15], STERN [20] and FEDERGRUEN [5]).

So far, very little attention has been paid to the actual computation of both the asymptotic average value g^* and of a solution v^* to the average return optimality equation (cf. section 1), under conditions that guarantee the existence of a stationary AEP.

In view of the fact that the value of (both the discounted and undiscounted version) of the game does not necessarily lie within the same ordered field as the parameters of the problem (cf. BEWLEY & KOHLBERG [2]) we cannot expect to find a finite algorithm in the sense that it involves a finite number of field-operations.

Two algorithms were given by HOFFMANN & KARP [11] and POLLATSCHEK & AVI-ITZHAK [14]. It was shown that the first algorithm converges to a stationary AEP, if all tpm's of the pure stationary policies are unichained and have no transient states. Although the second algorithm seems to compare favorably with the first one, as far as net running time and the required number of iterations is concerned, it is still unknown under which conditions its convergence is guaranteed.

In this paper, we provide two successive approximation methods for locating optimal policies for both players. In both algorithms, we obtain in addition at each step of the iteration procedure, upper and lower bounds for the asymptotic average value which converge to the latter as the number of iteration steps tends to infinity.

The first algorithm is an adaptation of a "modified" value-iteration method as introduced by BATHER [1] and as generalized by HORDIJK & TIJMS [12]. Its convergence is guaranteed whenever condition H1 below is satisfied.

(H1): (a) a stationary AEP exists
(b) the asymptotic average value g^* is independent of the initial state of the system.

The second algorithm is based upon the more elementary value-iteration method, and may successfully be applied whenever condition (H2) below holds:

(H2): each of the tpm's of the pure stationary policy pairs is unichained

Note that (H2) $\Rightarrow$ (H1) (cf. e.g. [5], th.3). Under (H2) we obtain in addition lower and upper bounds for the fixed point v^* of the optimality equation which in this case is unique up to a multiple of $\underline{1}$, where $\underline{1}$ is the N-vector with all components unity.

At each step of the procedure, both methods merely require the solution of N relatively small Linear Programs (the size of which is determined by the number of actions in K(i) and L(i), $i \in \Omega$). Especially for large scale

systems, i.e. when N >> 1, this compares favorably with the techniques used in [11] and [14] which require at each step of the procedure the solution of a system of at least N equations.

In section 1 we give the notation and some preliminaries. In section 2 and 3 we present our two successive approximation methods, analyse their convergence and convergence rate, and derive the lower and upper bounds for g^* and v^*.

1. NOTATION AND PRELIMINARIES

For any finite set A, let $\|A\|$ denote the number of elements it contains. 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 $\min_{i,j}(A_{ij} - B_{ij}) \leq x^B(A - B)y^A = x^B A y^A - x^B B y^A \leq \text{val } A - \text{val } B \leq x^A A y^B - x^A B y^B = x^A(A - B)y^B \leq \max_{i,j}(A_{ij} - B_{ij})$). For all $i \in \Omega$, and any set of numbers $\{c_i^{k,\ell} | k \in K(i), \ell \in L(i)\}$, $[c_i^{k,\ell}]$ denotes the $\|K(i)\| \times \|L(i)\|$ matrix, the (k,ℓ)-th entry of which is $c_i^{k,\ell}$.

For all $r > 0$, let V(r) denote the vector, the i-th component of which denotes the value of the r-discounted game, with initial state $i \in \Omega$. BEWLEY & KOHLBERG [2] recently showed for the discrete time case (SDG's) that V(r) may be expressed as a real fractional power or Puiseux series in r, for all interest rates r, sufficiently close to 0. More specifically, there exists an integer $M \geq 1$ such that:

$$(1.2) \qquad V(r) = g^*/r + \sum_{k=-\infty}^{M-1} a(k)\, r^{-k/M}$$

We call g^* the *asymptotic average value vector*. This result carries easily over to the general SRG-case (cf. [6], lemma 1.2).

A player's policy is a rule which prescribes for each stage t = 1,2,...

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 player1(2) is characterized by a tableau $[f_{ik}]$ ($[h_{i\ell}]$) satisfying $f_{ik} \geq 0$ and $\Sigma_{k\in K(i)} f_{ik} = 1$ ($h_{ik} \geq 0$ and $\Sigma_{\ell\in L(i)} h_{i\ell} = 1$), where $f_{ik}(h_{i\ell})$ is the probability that the k-th. (ℓ-th) alternative in K(i)(L(i)) is chosen when entering state $i \in \Omega$. We let $\Phi(\Psi)$ denote the set of all stationary policies for player 1(2). We associate with each pair $(f,h) \in \Phi \times \Psi$ a N-component reward vector q(f,h), the holding rate vector T(f,h) and a stochastic matrix P(f,h):

$$(1.3)\qquad q(f,h)_i = \sum_{k\in K(i)} \sum_{\ell\in L(i)} f_{ik}\, q_i^{k,\ell}\, h_{i\ell}\,; \quad i \in \Omega$$

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

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

Finally we define for any pair $(f,h) \in \Phi \times \Psi$ the stochastic matrix $\Pi(f,h)$ as the Cesaro limit of the sequence $\{P^n(f,h)\}_{n=1}^{\infty}$. Since being concerned with the long run average return per unit time criterion, we evaluate any pair (ϕ,ψ) of (possibly non-stationary) policies for players 1 and 2, by considering the gain rate vector $g(\phi,\psi)$:

$$(1.4)\qquad g(\phi,\psi)_i = \liminf_{n\to\infty} (E_{\phi,\psi} \sum_{\ell=1}^{n} \rho_\ell)/(E_{\phi,\psi} \sum_{\ell=1}^{n} \tau_\ell); \quad i \in \Omega$$

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. $E_{\phi,\psi}$ indicates the expectation given the player's policies ϕ and ψ. A number of equivalent criteria have been formulated in [3A].

A pair of policies (ϕ^*, ψ^*) is called an AEP, if and only if for every policy pair (ϕ, ψ)

$$(1.5)\qquad g(\phi,\psi^*)_i \leq g(\phi^*,\psi^*)_i \leq g(\phi^*,\psi)_i, \qquad \text{for all } i \in \Omega$$

One easily verifies (cf. e.g. [3A] and [6]) that if (f^*,h^*) is a stationary AEP, $g(f^*,h^*) = g^*$.

In [6], we showed that a pair of optimality equations arises when considering the average return per unit time criterion and we investigated the interdependences between the existence of a stationary AEP and a solution to this pair of optimality equations.

In the case where $g_i^* = \langle g^* \rangle$, $i \in \Omega$, i.e., when the asymptotic average value is independent of the initial state (cf. condition (H1)), this pair of optimality equations reduces to the single equation:

$$(1.6) \qquad v_i + g = \mathrm{val}[q_i^{k,\ell} - gT_i^{k,\ell} + \sum_j P_{ij}^{k,\ell} v_j], \quad i \in \Omega$$

and in this case we obtained the following equivalences:

LEMMA 1. *(cf. cor. 2.5 in [6]):*
Assume that $g_i^* = \langle g^* \rangle$, $i \in \Omega$. *Then the following statements are equivalent:*
(I) $a^{(h)} = 0$, $k = 1,\ldots,M-1$ (cf.(1.2))
(II) *there exists a stationary AEP*
(III) (1.6) *has a solution pair* (g,v)

In addition, under either one of (I), (II) or (III), any solution pair (g,v) has $g = g^*$, and any policy pair $(f^*,h^*) \in \Phi \times \Psi$ which satisfies the equation (1.6) i.e. which attains the N equilibria in (1.6) simultaneously for some solution pair (g,v) is an AEP. □

Finally we say that two undiscounted SRG's are equivalent if they have the same state and action spaces, and if the gain rate vector of any stationary pair of policies is identical in both SRG's. Now, consider the related SDG which has Ω as its 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:

$$(1.7) \qquad \begin{cases} \tilde{P}_{ij}^{k,\ell} = \tau/T_i^{k,\ell} \, [P_{ij}^{k,\ell} - \delta_{ij}] + \delta_{ij}; & i,j \in \Omega,\ k \in K(i),\ \ell \in L(i) \\ \tilde{q}_i^{k,\ell} = q_i^{k,\ell}/T_i^{k,\ell} & ; \quad i \in \Omega,\ k \in K(i),\ \ell \in L(i) \end{cases}$$

where δ_{ij} is the Kronecker-delta and where τ has to be chosen such that $0 < \tau \leq \min_{i,k,\ell} T_i^{k,\ell}/(1-P_{ij}^{k,\ell})$

This data-transformation which was first introduced in [16] turns every SRG into an equivalent SDG (cf. ibid.). Moreover, let $V = \{v \in E^N \mid v \text{ satisfies } (1.6)\}$ and let $\tilde{V}$ be the corresponding set in the transformed SDG:

LEMMA 2.

$$\tilde{V} = \{v \in E^V \mid \tau v \in V\} \tag{1.8}$$

PROOF. We show that the set to the right of (1.8) is included within $\tilde{V}$, the reversed inclusion being analogous. Fix $v \in V$ and rewrite (1.6) in a homogeneous form:

$$0 = \text{val}[q_i^{k,\ell} - gT_i^{k,\ell} + \textstyle\sum_j (P_{ij}^{k,\ell} - \delta_{ij})v_j], \quad i \in \Omega.$$

Let (f^*,h^*) attain the N equilibria in (1.6) and note that for all $f \in \Phi$, $h \in \Psi$:

$$q(f,h^*)_i - gT(f,h^*)_i + P(f,h^*)v_i \leq 0 \leq q(f^*,h)_i - gT(f^*,h)_i + P(f^*,h)v_i, \tag{1.9}$$

$$i \in \Omega$$

with strict equality for $f = f^*$ and $h = h^*$. Next multiply each of the left-hand (right-hand) inequalities in (1.9) by $T(f,h^*)_i^{-1}$ $(T(f^*,h)_i^{-1}) > 0$, use (1.7) and conclude that:

$$0 = \text{val}[\tilde{q}_i^{k,\ell} - g + \textstyle\sum_j \tilde{P}_{ij}^{k,\ell} (\tau^{-1}v)_j], \qquad i \in \Omega$$

which proves that $\tau^{-1}v \in \tilde{V}$. □

We conclude from lemma 2 and the fact that the original SRG and the transformed SDG are equivalent that:

(1) any stationary AEP in the original SRG is a stationary AEP in the transformed SDG, and vice versa

(2) the asymptotic average value is identical in both the original and the transformed game

(3) if v is a solution to the optimality equation (1.6) in the original SRG, then so is $\tau^{-1}v$ with respect to the transformed SDG.

2. A MODIFIED VALUE-ITERATION TECHNIQUE.

Throughout this section, we assume condition (H1) to hold, which implies in view of lemma 1, the existence of a solution pair (g^*,v^*) to the optimality equation (1.6). Fix τ_0, such that

$$(2.0) \qquad 0 < \tau_0 \leq \min_{i,k,\ell} T_i^{k,\ell}/(1-P_{ii}^{k,\ell}) \quad \text{(cf.(1.7))}$$

and consider the transformed SDG with $\tau = \tau_0$. Let $\{r_n\}_{n=1}^{\infty}$ be a sequence of interest rates such that $\lim_{n\to\infty} r_n = 0$, and let $\tilde{V}(r)$ denote the value vector of the r-discounted version of the transformed SDG. In view of condition (H1) the equivalence of the original SRG and the transformed SDG, as well as lemma 1, we conclude that $\tilde{V}(r)$ has for some integer $\tilde{M} \geq 1$, a Puiseux series expansion of the special type:

$$(2.1) \qquad \tilde{V}(r) = g^*/r + \sum_{k=-\infty}^{0} \tilde{a}(k) r^{-k/\tilde{M}}, \quad \text{for all r sufficiently close to 0.}$$

Applying the proof of th. 2.3 in [6] to SDG's we conclude that any scheme

$$(2.2) \qquad y(n+1)_i = \text{val}\,[\, q_i^{k,\ell} - g^* + (1+r_n)^{-1} \textstyle\sum_j \tilde{P}_{ij}^{k,\ell}\, y(n)_j\,], \quad i \in \Omega$$

with y(0) a given N-vector, has $\lim_{n\to\infty} y(n) = \tilde{a}^{(0)}$, provided that the sequence $\{r_n\}_{n=1}^{\infty}$ satisfies the conditions:

$$(2.3) \qquad (1 - r_n) \ldots (1 - r_2) \to 0, \quad \text{as } n \to \infty$$

$$(2.4) \qquad \sum_{j=2}^{n} (1 - r_n) \ldots (1 - r_j) \left| r_j^{1/\tilde{M}} - r_{j-1}^{1/\tilde{M}} \right| \to 0, \quad \text{as } n \to \infty$$

where $\tilde{a}^{(0)}$ is a solution to the optimality equation associated with the transformed SDG.

LEMMA 3. *Conditions (2.3) and (2.4) are satisfied for any choice:*

$$r_n = n^{-b} \qquad \textit{with} \qquad 0 < b \leq 1$$

PROOF. Note using the mean value theorem that $n^b - (n-1)^b \leq 1$ for all $n = 1,2,\ldots$ and use this inequality in order to verify that:

$$(1-r_n) \ldots (1-r_2) = \frac{(n^b-1)}{n^b} \frac{((n-1)^b-1)}{(n-1)^b} \ldots \frac{(2^b-1)}{2^b} \leq \frac{2^b-1}{n^b}$$

which proves (2.3). Next we apply the mean value theorem to verify that

$$\sum_{j=2}^{n} (1-r_n) \ldots (1-r_j) \mid r_j^{\tilde{M}^{-1}} - r_{j-1}^{\tilde{M}^{-1}} \mid \leq$$

$$b\tilde{M}^{-1} \sum_{j=2}^{n} \frac{(n^b-1)}{n^b} \ldots \frac{(j^b-1)}{j^b} (j-1)^{-b\tilde{M}^{-1}-1} \leq b\tilde{M}^{-1} n^{-b} \sum_{j=2}^{n} j^{b(1-\tilde{M}^{-1})-1} \leq$$

$$b\tilde{M}^{-1} n^{-b} \int_{1}^{n} x^{b(1-\tilde{M}^{-1})-1} \, dx =$$

$$\begin{cases} bn^{-b} \ell n(n), & \text{if } \tilde{M} = 1 \\ (\tilde{M}-1)^{-1} n^{-b\tilde{M}^{-1}}, & \text{otherwise} \end{cases}$$

which proves (2.4). □

REMARK 1. For the MDP- i.e. one player - case, lemma 3 indicates a larger range of permitted values for b, than the one that was obtained in [12] (p.206, remark) using a different analysis.

Observe that the sequence $\{y(n)\}_{n=1}^{\infty}$ cannot be computed in view of g^* being unknown. We circumvent this numerical difficulty as in WHITE [21], i.e. we define the sequences $\{\hat{y}(n)\}_{n=1}^{\infty}$ and $\{G(n)\}_{n=1}^{\infty}$ by:

$$(2.5) \qquad \hat{y}(n+1)_i = y(n+1)_i - y(n+1)_N = \text{val}[\tilde{q}_i^{k,\ell} + (1+r_n)^{-1} \textstyle\sum_j \tilde{P}_{ij}^{k,\ell} \hat{y}(n)_j] - G(n+1); \quad i \in \Omega; \quad n = 0,1,2,\ldots$$

$$(2.6)\qquad G(n+1) = \mathrm{val}[\tilde{q}^{k,\ell}_N + (1+r_n)^{-1} \textstyle\sum_j \tilde{P}^{k,\ell}_{Nj}\, \hat{y}(n)_j];$$

$$i \in \Omega;\quad n = 0,1,2,\ldots$$

where $\hat{y}(0)_i = y(0)_i - y(0)_N$; $\quad i \in \Omega$.

THEOREM 1. *For all* $n = 1,2,\ldots$ *let:*

$$(2.7)\qquad L(n+1) = \min_i\{\mathrm{val}[\tilde{q}^{k,\ell}_i + (1+r_n)^{-1} \textstyle\sum_j \tilde{P}^{k,\ell}_{ij}\, \hat{y}(n)_j] - (1+r_n)^{-1}\hat{y}(n)_i\}$$

$$U(n+1) = \max_i\{\mathrm{val}[\tilde{q}^{k,\ell}_i + (1+r_n)^{-1} \textstyle\sum_j \tilde{P}^{k,\ell}_{ij}\, \hat{y}(n)_j] - (1+r_n)^{-1}\hat{y}(n)_i\}$$

(a) *Let* (f^*,h^*) *be a stationary AEP and for any* $n = 1,2,\ldots$ *let* $(f_n,h_n) \in \Phi \times \Psi$ *be any pair of policies which attain the N equilibria to the right of (2.5) simultaneously. Then*

(1) $L(n) \le G(n) \le U(n) \qquad n = 1,2,\ldots$

(2) $L(n+1) \le g(f_n,h^*)_i \le g^* \le g(f^*,h_n)_i \le U(n+1); \quad i \in \Omega$

(b) *If* $\{r_n\}_{n=1}^{\infty}$ *satisfies (2.3) and (2.4), then:*

$$\lim_{n\to\infty} L(n) = \lim_{n\to\infty} G(n) = \lim_{n\to\infty} U(n) = g^*$$

$$\lim_{n\to\infty} \hat{y}(n) = \tilde{a}^{(0)} - \langle \tilde{a}^{(0)}_N \rangle \underline{1} \in \tilde{V}$$

where $\underline{1}$ *is the vector, the components of which are unity.*

PROOF.

(a) (1) Note from (2.5) that $\hat{y}(n)_N = 0$ for all $n = 0,1,2,\ldots$, hence

$$L(n) \le \mathrm{val}[\tilde{q}^{k,\ell}_N + (1+r_n)^{-1} \textstyle\sum_j \tilde{P}^{k,\ell}_{Nj}\, \hat{y}(n)_j] = G(n) \le U(n)$$

(2) The inner equalities are immediate from the fact that (f^*,h^*) is a stationary AEP. We next prove the most left inequality $L(n+1) \le g(f_n,h^*)$, the proof of $g(f^*,h_n) \le U(n+1)$ being analogous. Note that for all $i \in \Omega$:

$$L(n+1) + (1+r_n)^{-1}\hat{y}(n)_i \le \mathrm{val}[\tilde{q}^{k,\ell}_i + (1+r_n)^{-1}\textstyle\sum_j \tilde{P}^{k,\ell}_{ij}\hat{y}(n)_j] \le$$

$$\le \tilde{q}(f_n,h^*)_i + (1+r_n)^{-1}\tilde{P}(f_n,h^*)\hat{y}(n)_i\ ,$$

and multiply both sides of this inequality by $\Pi(f_n, h^*) \geq 0$.

(b) Recall that $\lim_{n\to\infty} y(n) = \tilde{a}^{(0)} \in \tilde{V}$. Next we observe that if $v \in \tilde{V}$ then so is $v + c\underline{1}$ for all scalars c. Hence,

$$\lim_{n\to\infty} \hat{y}(n) = \lim_{n\to\infty} y(n) - \langle y(n)_N \rangle \underline{1} = \tilde{a}^{(0)} - \langle \tilde{a}^{(0)}_N \rangle \underline{1} \in \tilde{V}$$

This in combination with the fact that the "val"-operator is (Lipschitz) continuous (cf. (1.1) and (1.6)) imply:

$$\lim_{n\to\infty} L(n) = \min_i \{ val[\tilde{q}_i^{k,\ell} + \sum_j \tilde{P}_{ij}^{k,\ell} \tilde{a}_j^{(0)}] - \tilde{a}_i^{(0)} \} = \min_i g^* =$$

$$g^* = \max_i \{ val[\tilde{q}_i^{k,\ell} + \sum_j \tilde{P}_{ij}^{k,\ell} \tilde{a}_j^{(0)}] - \tilde{a}_i^{(0)} \} = \lim_{n\to\infty} U(n)$$

which together with part (a)(1) completes the proof of (b). □

REMARK 2. When taking $r_n = n^{-b}$ for some b, with $0 < b \leq 1$, the approach to the limits in part (b) of the above theorem, exhibits a convergence rate which is of the order

$$\begin{cases} O(n^{-b} \ell n\, n), & \text{if } \tilde{M} = 1 \\ O(n^{-b\tilde{M}^{-1}}), & \text{otherwise} \end{cases}$$

as follows from the proof of lemma 3 and th.2.3 in [6]. We note that the bounds for g^* in part (a)(2) generalize the bounds ODONI [10] and HASTINGS [13] obtained for the MDP-case.

We summarize this section by specifying an algorithm which approximates g^*, as well as a solution $v \in V$, and which finds for any $\varepsilon > 0$, ε-optimal policies for both players:

ALGORITHM 1.

Step 0: Fix τ_0 satisfying (2.0) and transform the SRG with $(q_i^{k,\ell}; P_{ij}^{k,\ell}; T_i^{k,\ell})$ into an equivalent SDG with $(\tilde{q}_i^{k,\ell}; \tilde{P}_{ij}^{k,\ell})$ using the transformation formulae (1.7). Fix a sequence $\{r_n\}_{n=1}^{\infty}$ satisfying (2.3) and (2.4); e.g. take $r_n = n^{-b}$, with $0 < b \leq 1$. Set $n = 0$; fix $y(0) \in E^N$ and $\varepsilon > 0$.

Step 1: Calculate $\hat{y}(n+1)$, $L(n+1)$, $G(n+1)$ and $U(n+1)$ from $\hat{y}(n)$ using (2.5), (2.6) and (2.7)

Step 2: If $U(n+1) - L(n+1) < \varepsilon$, determine a stationary policy pair (f_n, h_n) which attains the N equilibria to the right of (2.5) simultaneously, use $f_n(h_n)$ as an ε-optimal policy for player 1(2); $G(n+1)$ as an ε-approximation for g^* and $\tau_0 \hat{y}(n+1)$ as an approximation for a solution $v \in V$. Otherwise, increment n by one and return to step 1.

3. VALUE-ITERATION. A SUFFICIENT CONDITION FOR CONVERGENCE

Once again, we fix τ_0 satisfying (2.0) and consider the transformed SDG. In this section we discuss the asymptotic behaviour of the sequence:

$$(3.1) \qquad v(n+1)_i = Qv(n)_i, \qquad i \in \Omega$$

where $Qx_i = \text{val}[\tilde{q}_i^{k,\ell} + \sum_j \tilde{P}_{ij}^{k,\ell} x_j]$, $i \in \Omega$ and $v(0) \in E^N$ is a given N-vector. Note that the Q operator is monotonous, satisfying the basic properties:

$$(3.2) \qquad Q(x+c\underline{1}) = Qx + c\underline{1} \quad \text{for all scalars } c; \quad x \in E^N$$

$$(3.3) \qquad (x-y)_{\min} \leq (Qx - Qy)_{\min} \leq (Qx - Qy)_{\max} \leq (x-y)_{\max}; \quad x,y \in E^N$$

where (3.3) is easily verified by applying the Q-operator to both sides of the inequalities $y + (x-y)_{\min}\underline{1} \leq x$ and $x \leq y + (x-y)_{\max}\underline{1}$, using its monotonicity as well as (3.2).

Note that $v(n)_i$ may be interpreted as the value of the n-stage game in the transformed SDG when starting in state i and given some final amount $v(0)_j$ is earned by player 1 from player 2, when ending up in state j.

Whereas we still have $\lim_{n\to\infty} \frac{v(n)}{n} = g^*$ (cf. BEWLEY & KOHLBERG [2], th. 3.2) the difference $\{v(n) - ng^*\}_{n=1}^{\infty}$ does not need to be bounded, as is known to be the case in the one player - model (cf. BROWN [4], th. 4.3).

In fact, BEWLEY & KOHLBERG [3] proved the existence of a number $B > 0$ and a Puiseux series in n,

$$W(n) = ng^* + \sum_{k=-\infty}^{\hat{M}-1} b^{(k)} n^{k/\hat{M}}$$

such that $|v(n) - W(n)| < B \log (n+1), \qquad n = 1,2,\ldots$

LEMMA 4. $\{v(n) - ng^*\}_{n=1}^{\infty}$ *is bounded under condition (H1)*

PROOF. Note from lemma 1, that (H1) implies the existence of a solution $v \in \tilde{V}$. Next, use (1.1) in order to conclude that:

$$|v(n) - ng^* - v| \leq |val[\tilde{q}_i^{k,\ell} + \sum_j \tilde{P}_{ij}^{k,\ell} v(n-1)_j] -$$

$$- val[\tilde{q}_i^{k,\ell} + \sum_j \tilde{P}_{ij}^{k,\ell} (v+(n-1)g^*)]| \leq |v(n-1) - (n-1)g^* - v|$$

□

It is known from Markov Decision Theory that even in case $\{v(n) - ng^*\}_{n=1}^{\infty}$ is bounded, the sequence may fail to converge if some of the tpm's of the pure stationary policy pairs happen to be periodic (in [17], SCHWEITZER & FEDERGRUEN obtained for the MDP-case the necessary and sufficient condition for $\{v(n) - ng^*\}_{n=1}^{\infty}$ to converge for all $v(0) \in E^N$).

In this section we analyse the behaviour of (3.1) under condition (H2) which is a stronger version of (H1) (cf. section 0).

First however we need the following notation: Note that in view of (3.2), it is possible to restrict the analysis of the Q-operator on a N-1 dimensional subspace like $\tilde{E}^N = \{x \in E^N \mid x_N = 0\}$ by considering the following *reduction* $\hat{Q}$ of the Q-operator:

$$\hat{Q} : \tilde{E}^N \to \tilde{E}^N : x \to Qx - \langle Qx_N \rangle \underline{1}$$

Accordingly, define $\hat{v}(n)_i = v(n)_i - v(n)_N = \hat{Q}\,\hat{v}(n-1)$, $i \in \Omega$. (Note the similarity with the reduction in WHITE [21] and of $\{y(n)\}_{n=1}^{\infty}$ to $\{\hat{y}(n)\}_{n=1}^{\infty}$ in (2.5)).

We call a function L(x) on a vector space X, a *Lyapunov* function with *origin* $x^* \in X$, if:

(3.4) (1) L(x) is continuous on X

(2) $L(x) \geq 0$ and $L(x) = 0 \Leftrightarrow x = x^*$.

We have not been able to obtain a straighforward analysis of the behaviour of $\{v(n)\}_{n=1}^{\infty}$ or $\{\hat{v}(n)\}_{n=1}^{\infty}$. However, the study of difference equations

of the type (3.1) may be greatly facilitated with the help of Lyapunov functions, as is shown by the following lemma.

LEMMA 5. *Let* $L(n)$ *be a Lyapunov function on a vector space* X, *with origin* x^*. *For* $n = 2,3,\ldots$ *let* A^n *denote the n-fold application of an operator* $A:X \to X$ *i.e.* $A^{n+1}x = A(A^n x)$. *Then,*

$$\lim_{n\to\infty} A^n x = x^*, \quad \text{for all } x \in X, \quad \text{if}$$

(3.5)
(1) $L(Ax) \leq L(x)$, for all $x \in X$
(2) there exists an integer $J \geq 1$ such that $L(A^J x) < L(x)$, for all $x \neq x^*$

□

Lemma 5 is an immediate adaptation of th. 10.4 in ZANGWILL [22]. In the context of Markov Decision Theory, the use of Lyapunov functions, and in-particular of lemma 5, was first pointed out in [7].

Now, under (H2), the solution to the optimality equation (1.6) is unique up to multiple of $\underline{1}$, as was shown in [6], th. 3.1, i.e. on $\tilde{E}^N$ there exists a *unique* solution $v^* \in \tilde{V}$.

We next observe that both $L_1(x)$ and $L_2(x)$ are Lyapunov function on $\tilde{E}^N$ with v^* as origin, where

$$L_1(x) = \| x - v^* \|_d$$

$$L_2(x) = \| \hat{Q}\, x - x \|_d = \| Q\, x - x \|_d,$$

with $\| x \|_d = \max_i x_i - \min_i x_i$ (cf. BATHER [1]). $L_1(x)$ obviously satisfies both conditions in (3.3); $L_2(x) \geq 0$ is immediate as well, its continuity on $\tilde{E}^N$ follows from the continuity of the "val"-operator (cf. (1.1)) and $\| Qx - x \|_d = 0 \Leftrightarrow$ there exists a scalar g, such that $Qx - x = \langle g \rangle 1 \Leftrightarrow x \in \tilde{V} \cap \tilde{E}^N \Leftrightarrow$ $\Leftrightarrow x = v^*$).

Note that $L_2(x)$ has the advantage of being computable in each point $x \in \tilde{E}^N$.

We next recall that under (H2), the tpm's of all stationary policy pairs are unichained, and in addition have all diagonal entries strictly positive

when choosing τ_0 strictly less than the upperbound on τ in (2.0). In [8], th.4 FEDERGRUEN, SCHWEITZER & TIJMS showed that this implies the following "scrambling-type" condition for all pairs of N-tuples of pure policy pairs $\{(f_1,h_1);\ldots;(f_N,h_N)\}$ and $\{(f_1',h_1');\ldots;(f_N',h_N')\}$:

$$(3.6)\qquad \sum_{j=1}^{N} \min[P(f_N,h_N)\ldots P(f_1,h_1)_{i_1 j}; P(f_N',h_N')\ldots P(f_1',h_1')_{i_2 j}] > 0$$

$$\text{for all } i_1 \neq i_2.$$

Observe that for all $i_1,i_2 \in \Omega$ the expression to the left of the above inequality is a continuous function on $[X_{\ell=1}^{N} \Phi \times \Psi]^2$ which can be embedded as a compact subset of a Euclidean space. Hence there exists a uniform scrambling coefficient $\alpha > 0$, such that

$$(3.7)\qquad \sum_{j=1}^{N} \min[P(f_N,h_N)\ldots P(f_1,h_1)_{i_1 j}; P(f_N',h_N')\ldots P(f_1',h_1')_{i_2 j}] > \alpha$$

$$\text{for all } i_1 \neq i_2;\ (f_\ell,h_\ell) \text{ and } (f_\ell',h_\ell') \in \Phi \times \Psi \ (\ell = 1,\ldots,N)$$

This enables us to prove the convergence of $\{\hat{v}(n)\}_{n=1}^{\infty}$ under (H2). Let $\ell(n+1) = [Q\,\hat{v}(n) - \hat{v}(n)]_{min}$ and $u(n+1) = [Q\,\hat{v}(n) - \hat{v}(n)]_{max}$ for all $n = 0,1,\ldots$ Define $g(n+1) = [Q\,\hat{v}(n)]_N$

THEOREM 2.

(a) both $L_1(x)$ *and* $L_2(x)$ *satisfy (3.4) with* $J = N$*; hence* $\lim_{n\to\infty} \hat{v}(n) = v^*$ for all $v(0) \in E^N$

(b) $\ell(n) \leq \ell(n+1) \leq g(n+1) \leq u(n+1) \leq w(n)$ *for all* $n = 1,2,\ldots$ $\lim_{n\to\infty} \ell(n) = \lim_{n\to\infty} g(n) = \lim_{n\to\infty} w(n) = g^*$

(c) Let $(f^*,h^*) \in \Phi \times \Psi$ *be an AEP, and for all* $n = 1,2,\ldots$ *let* $(f_n,h_n) \in \Phi \times \Psi$ *be any pair of policies which attain the N equilibria to the right of (3.1) simultaneously. Then* $\ell(n+1) \leq g(f_n,h^*) \leq g^* \leq g(f^*,h_n) \leq w(n+1)$

PROOF.

(a) We merely show that $L_1(x)$ satisfies (3.4), the proof for $L_2(x)$ being analogous. Use (3.3) to verify that $L_1(\hat{Q}x) = \|\hat{Q}x - v^*\|_d = \|Qx - Qv^*\|_d \leq$

$\leq \|x - v\|_d = L_1(x)$. Next we obtain part (2) of condition (3.4) by showing that:

$$(3.8) \qquad L_1(\hat{Q}^N x) = L_1(Q^N x) < (1-\alpha)L_1(x), \quad \text{for all } x \in X$$

where the proof of (3.8) goes along lines with the proof of th.5 in [8], using (3.7).

(b) The proof of $\ell(n+1) \leq m(n+1) \leq w(n+1)$ is analogous to the proof of part (a)(1) in th.1; next note that $\ell(n+1) = [Q\hat{v}(n) - \hat{v}(n)]_{min} = [Q(Q\hat{v}(n-1)) - Q\hat{v}(n-1)]_{min} \leq [Q\hat{v}(n-1) - \hat{v}(n-1)]_{min} = \ell(n)$, where the inequality part follows from (3.3). The monotonicity of $\{w(n)\}_{n=1}^{\infty}$ is shown in complete analogy.

(c) cf. proof of th. 1 part (a)(2). □

Observe that (3.8) is stronger than condition (3.4)(2), since the latter does not require the existence of some integer $J \geq 1$, for which

$$\sup_{x \in X} L(A^J x)/L(x) < 1.$$

In fact (3.8) shows that the approach to all of the limits in parts (a) and (b) of the above theorem exhibits a *geometric* rate of convergence, which is considerably better than the rates we obtained in section 2, for algorithm 1 (cf. Remark 2). In this particular case, it is even possible to show (along lines with the proof of th. 5 in [8]) that $\hat{Q}$ is a N-step contraction mapping on $\tilde{E}^N$, i.e. $\|Q^N x - Q^N y\|_d \leq (1-\alpha)\|x - y\|_d$, for all $x,y \in E^N$ and the latter leads to the following bounds on v^*:

$$(3.9) \qquad \hat{v}(nN+r)_i - \alpha^{-1}(1-\alpha)^N \|v(N) - v(0)\|_d \leq v_i^* \leq$$

$$\hat{v}(nN+r)_i + \alpha^{-1}(1-\alpha)^N \|v(N) - v(0)\|_d; \quad i \in \Omega;\ n = 1,2\ldots;\ r = 0,\ldots N-1.$$

(for a proof cf. [8], th. 6 part(a)).

Finally we conclude that merely replacing $\{\hat{y}(n)\}_{n=1}^{\infty}$, L(n), G(n), U(n) by $\{\hat{v}(n)\}_{n=1}^{\infty}$, $\ell(n)$, g(n) and u(n) resp. and taking τ_0 *strictly less* than the upperbound on τ in (2.0), we obtain under (H2) a *second* algorithm for approximating v^*, g^* and for locating ε-optimal policies for both players

(note that in this case (3.9) may be used as a stopping criterion for getting ε-approximations for v^*).

REFERENCES

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

[2] BEWLEY, T. & E. KOHLBERG, *The asymptotic theory of Stochastic Games*, Math. of O.R. 1 (1976), pp. 197-208.

[3] _____ & _____, *The asymptotic solution of a recursive equation arising in Stochastic Games*, (to appear in Math of O.R.) (1976).

[3A] _____ & _____, *On Stochastic Games with stationary optimal strategies*, Tech. Report no. 23, Harvard Institute of Economic Research, Harvard University (1976).

[4] BROWN, B., *On the iterative method of Dynamic Programming, on a finite state space discrete time Markov Process*, Ann. of Math. Statistics 36, (1965), pp. 1279-1285.

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

[6] _____, *On the functional equations in undiscounted and sensitive discounted stochastic games*, Math. Center Report BW 73/77 (1977).

[7] _____ & P.J. SCHWEITZER, *On the use of Lyapunov functions in Markov Decision Theory*, (forthcoming).

[8] _____ , _____ & H.C. TIJMS, *Contraction mappings underlying undiscounted Markov Decision Problems*, Math. Center Report BW 72/77 (1977)(to appear in J.M.A.A.).

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

[10] HASTINGS, N., *Bounds on the gain of a Markov Decision Process*, Op. Res. 19 (1971), pp. 240-244.

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

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

[13] ODONI, A., *On finding the maximal gain for Markov Decision Processes*, O.R. 17 (1969), pp. 857-860.

[14] POLLATSCHEK, M. & B. AVI-ITZHAK, *Algorithms for Stochastic Games with geometrical interpretation*, Man. Sci. 15 (1969), pp. 399-415.

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

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

[17] SCHWEITZER, P.J. & A. FEDERGRUEN, *The asymptotic behaviour of undiscounted value iteration in Markov Decision Problems*, Math. Center Report BW 44/76 (to appear in Math. of O.R.)(1976).

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

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

[20] STERN, M., *On stochastic Games with limiting average payoff*, Ph.D. dissertation, Dept. of Math., University of Illinois, Chicago Circle Campus (1975).

[21] WHITE, D., *Dynamic Programming, Markov Chains and the method of successive approximations*, J.M.A.A. 6 (1963), pp. 373-376.

[22] ZANGWILL, W., *Nonlinear Programming, a unified approach*, Prentice Hall, inc; Englewood Cliffs, N.J. (1969).