Markov games with unbounded rewards

EINDHOVEN UNIVERSITY OF TECHNOLOGY

Department of Mathematics

PROBABILITY THEORY, STATISTICS AND OPERATIONS RESEARCH GROUP

Memorandum COSOR 76-05

Markov games with unbounded rewards

by

J. Wessels

Eindhoven, March 1976

The Netherlands

Markov games with unbounded rewards

by

J. Wessels

Summary. 2-person zero-sum Markov games with the total expected reward criterion are considered. The one period rewards are not supposed to be bounded. However, it is assumed that the values of the one period games in each state constitute a vector in a Banach space in which the transition probabilities are contracting. This game is proved to possess a value vector and optimal stationary strategies (in a weakened sense). Furthermore, it is exhibited how the value vector and optimal strategies may be computed using successive approximations.

1. Introduction. In this paper we will consider a dynamic system with a countable state space $S = \{1,2,\ldots\}$, which is observed at discrete points in time $t = 0,1,\ldots$. The dynamic behaviour of the system may be influenced by two players (P_1 and P_2). If the system is in state i at time t, player P_1 may select an action k from a finite set K_i (nonempty) and P_2 may select an action ℓ from a finite set L_i (nonempty). As a result of such choices the state j will be observed at time $t+1$ with probability $p(j;i,k,\ell)$, moreover P_1 obtains a (possibly negative) reward $r(i,k,\ell)$ from P_2.

When analyzing games of this type with the total expected reward criterion, one usually requires discounting of future rewards and boundedness of the reward function (see e.g. Shapley [6], van der Wal [7], Himmelberg e.a. [3], Parthasarathy [5]). Another condition which has been used is positiveness of the reward function (see e.g. [3], [5]).

For discounted Markov games with bounded rewards one usually applies the theory of contracting operators in the space of all bounded functions on S (real valued functions on S will be called vectors henceforth). As in the theory of Markov decision processes the conditions might be weakened by using other spaces (compare [8]). In this paper such an approach will be combined with another weakening, viz. with respect to r only conditions will be given for the values of the matrix games with entries $r(i,k,\ell)$ for any i.

This latter weakening of the conditions has as a consequence, that the expected total rewards may not be defined properly for all pairs of decision rules. This difficulty is met by adapting the definition of the criterion and by adapting the definition of a saddle point.

In section 2 we will give definitions, notations and assumptions together with some properties of basic tools. In section 3 it will be proved that the Markov game pocesses a value and optimal decision rules (stationary Markov strategies). In section 4 it will be shown how one might approximate the value and find ε-optimal decision rules. Section 5 is devoted to some remarks and extensions.

2. Preliminaries. A *mixed action for player* P_1 in state i is a probability distribution on K_i (notations: f(i,k) for the probabilities, f(i,·) for the distribution, F_i for the set of mixed actions; analogously for P_2 with notations g(i,ℓ), g(i,·), G_i).
A *policy for player* P_1 determines a mixed action for P_1 in any state (notations: f for the policy, f(i,·) for the mixed action in state i, $\mathbf{F}$ for the set of all policies; analogously for P_2 with notations g, g(i,·), $\mathbf{G}$).
A *strategy for player* P_1 is a sequence $\pi = (\pi_0,\pi_1,\ldots)$, such that π_n maps H_n into $\mathbf{F}$ for n = 1,2,... and $\pi_0 \in \mathbf{F}$. H_n is the set of all paths until time n $(i_0,k_0,\ell_0,\ldots,i_{n-1},k_{n-1},\ell_{n-1})$ with $i_m \in S$, $k_m \in K_{i_m}$, $\ell_m \in L_{i_m}$. Γ is the set of all strategies for P_1. Analogously $\rho = (\rho_0,\rho_1,\ldots)$ is an element of G, the set of all strategies for P_2.
$\pi \in \Gamma$ is said to be a *Markov strategy for* P_1 if π_n maps all paths of H_n on the same element in $\mathbf{F}$. A Markov strategy for P_1 may be characterized by a sequence of policies $(f_0,f_1,\ldots) \in \mathbf{F}^\infty$. Analogously $\mathbf{G}^\infty$ denotes the set of Markov strategies for P_2.
$\pi \in \Gamma$ is said to be *stationary* if π is a Markov strategy and π_n does not depend on n. A stationary strategy for P_1 may be characterized by a policy $f \in \mathbf{F}$. The terms stationary strategy and policy will be used deliberately.

With respect to the transition probabilities we suppose

$$p(j;i,k,\ell) \geq 0 \quad \text{for } i,j \in S,\ k \in K_i,\ \ell \in L_i,$$

$$\sum_{j\in S} p(j;i,k,\ell) \leq 1 \quad \text{for } i \in S,\ k \in K_i,\ \ell \in L_i.$$

The transition distributions might be completed by introducing an extra absorbing state 0 with

$$p(0;i,k,\ell) = 1 - \sum_{j \in \$} p(j;i,k,\ell) \quad \text{for } i \in \$.$$

With this artificial extension in mind we can define a probability measure on the set of all infinite paths for each starting state $i \in \$$ and each pair of strategies π,ρ using the transition probabilities $p(j;i,k,\ell)$ at time t and π_t, ρ_t for determining the actual k and ℓ. This measure is denoted by $\mathbf{P}_{i,\pi,\rho}$. Expectations with respect to this measure are denoted by $\mathbf{E}_{i,\pi,\rho}$, furthermore $\mathbf{E}_{\pi,\rho}$ denotes a vector of expectations. (The set of vectors $v = (v(1),v(2),\ldots)$ with $v(i) \in \bar{\mathbb{R}}$ is denoted by $\bar{\mathbb{R}}^{\infty}$.)
The state at time t, the actions chosen by player P_1 and P_2 at time t, the path until time t are denoted by the random variables S_t, K_t, L_t, H_t.
$P^t(\pi,\rho)$ denotes the matrix with (i,j)-entry $\mathbf{P}_{i,\pi,\rho}(S_t = j)$, so
$P^1(f,g) =: P(f,g)$ gets the (i,j)-entry

$$\sum_{k,\ell} f(i,k)g(i,\ell)p(j;i,k,\ell) .$$

$r(f,g)$ is the vector with i-th entry

$$\sum_{k,\ell} f(i,k)g(i,\ell)r(i,k,\ell) .$$

Definition 2.1. Let V be a vector valued function defined on a subset $\mathcal{D}$ of $F \times G$. $V(\pi,\rho) \in \bar{\mathbb{R}}^{\infty}$ for $(\pi,\rho) \in \mathcal{D}$.
π^*,ρ^* is said to be a *saddle point of* V iff

1) $(\pi,\rho^*),(\pi^*,\rho) \in \mathcal{D}$ for all $(\pi,\rho) \in F \times G$.
2) $V(\pi,\rho^*) \leq V(\pi^*,\rho^*) \leq V(\pi^*,\rho)$ for all $(\pi,\rho) \in F \times G$, where inequalities should be read componentwise.

$V(\pi^*,\rho^*)$ is called the *value vector of the game with respect to the criterion* V; π^*,ρ^* are called *optimal strategies for* P_1 *and* P_2 *for the game with criterion* V.

Note that a game can possess at most one value vector with respect to V. We will use the following criterion vector V in this paper.

Definition 2.2. $\mathcal{D} \subset F \times G$ is defined as the set of pairs (π,ρ) which satisfy at least one of the following conditions

a) $$\mathbf{E}_{i,\pi,\rho} \sum_{n=0}^{\infty} r^{+}(\pi_n(H_n),\rho_n(H_n))(S_n) < \infty \quad \text{for all } i \in S .$$

b) $$\mathbf{E}_{i,\pi,\rho} \sum_{n=0}^{\infty} r^{-}(\pi_n(H_n),\rho_n(H_n))(S_n) < \infty \quad \text{for all } i \in S .$$

With

$$r^{+}(f,g)(i) := \max\{0, r(f,g)(i)\}$$

$$r^{-}(f,g)(i) := \max\{0, -r(f,g)(i)\} .$$

We define V as a vector valued function on $\mathcal{D}$ by

$$V(\pi,\rho) := \mathbf{E}_{\pi,\rho} \sum_{n=0}^{\infty} r(\pi_n(H_n),\rho_n(H_n))(S_n)$$

$$= \sum_{n=0}^{\infty} \mathbf{E}_{\pi,\rho}\, r(\pi_n(H_n),\rho_n(H_n))(S_n) .$$

$V(\pi,\rho)$ may contain some entries equal to ∞ or some entries $-\infty$.

Note that $V(\pi,\rho)$ is not defined as the expectation of the sum of terms like $r(S_t,K_t,L_t)$. This would require stronger convergence conditions.

Now we will give our assumptions on rewards and transition probabilities.

Assumption 2.1.

a) μ is a given vector with positive components.
b is the vector with entries $\mu^{-1}(i)$.

b) $$\exists_{\beta \in (0,1)} \sum_{j} p(j;i,k,\ell) b(j) \le \beta b(i) \quad (\text{for all } i \in S,\ k \in \mathbf{K}_i,\ \ell \in \mathbf{L}_i),$$

the minimal allowed β-value will be denoted by β henceforth.

Assumption 2.1.b may be written in another way if we use μ for the construction of a space of vectors $V \subset \mathbb{R}^{\infty}$

$$V := \{v \in \mathbb{R}^{\infty} \mid \sup_i |v(i)|\mu(i) < \infty\}$$

V is a Banach space when we introduce the following norm

$$\|v\| := \sup_i |v(i)|\mu(i) .$$

Now we may introduce operator norms for matrices or linear operators. Then assumption 2.1.b is equivalent to

$$\exists_{\beta\in(0,1)} \|P(f,g)\| \le \beta \text{ for all (degenerated) } f,g .$$

Assumption 2.1 implies for $t = 0,1,\ldots$ and $\pi \in F$, $\rho \in G$

$$\|P^t(\pi,\rho)\| \le \beta^t .$$

We denote by $\bar{r}$ the vector with i-th component equal to the value of the matrix game with entries $r(i,k,\ell)$. $\bar{f} \in \mathbb{F}$ and $\bar{g} \in \mathbb{G}$ denote optimal decision rules for this sequence of matrix games.

Assumption 2.2. $\bar{r} \in V$.

By V^+ we denote the set of vectors $w \in \bar{\mathbb{R}}^\infty$ such that $w \ge v$ for some $v \in V$.
V^- contains those vectors $w \in \bar{\mathbb{R}}^\infty$ with $w \le v$ for some $v \in V$.
$V^\pm$ is the union of V^+ and V^-.

$P(f,g)$ is properly defined as a linear operator on V, V^+, V^-, $V^\pm$;
$P(f,g)$ maps each of these sets into itself. "Properly defined" means that $P(f,g)v(i)$ is independent of the order of summation.
$P(f,g)$ is monotone on V, V^+, V^-, $V^\pm$.
$P(f,g)$ is contracting on V with contraction radius $\|P(f,g)\| \le \beta < 1$.

$\mathcal{D}$ contains at least $(\bar{f},\bar{g})$. In general, if f and g are such that $r(f,g) \in V^\pm$, then $(f,g) \in \mathcal{D}$ and $V(f,g)$ is defined. $\|V(\bar{f},\bar{g})\| \le (1-\beta)^{-1}\|\bar{r}\|$.
In general, if $r(f,g) \in V$, V^+, V^- then $V(f,g) \in V$, V^+, V^-. Note that $r(f,\bar{g}) \in V^-$, $r(\bar{f},g) \in V^+$ for all $f \in \mathbb{F}$, $g \in \mathbb{G}$.

Definition 2.3. Let $f \in \mathbb{F}$, $g \in \mathbb{G}$, then $L(f,g)$ is defined as a mapping of $V^\pm$ into $\bar{\mathbb{R}}^\infty$ by

$$L(f,g)v := r(f,g) + P(f,g)v \quad \text{for } v \in V^\pm .$$

Since $r(f,g)$ is not necessarily in $V^\pm$ the same holds for $L(f,g)v$.

Lemma 2.1. Let $f \in \mathbb{F}$, $g \in \mathbb{G}$.

a) if $r(f,g) \in V$, $L(f,g)$ maps V into V and $L(f,g)$ is contracting on V with contraction radius $\| P(f,g) \| \leq \beta < 1$. The fixed point of $L(f,g)$ in V is $V(f,g)$.

b) if $r(f,g) \in V^-$, then $L(f,g)$ maps V^- into V^-.

c) if $r(f,g) \in V^+$, then $L(f,g)$ maps V^+ into V^+.

d) $L(f,g)$ is monotone on $V^{\pm}$.

e) if $v \in V$, $r(f,g) \in V^{\pm}$, then $L^n(f,g)v \to V(f,g)$ (for $n \to \infty$), componentwise.

Proof. Straightforward. As an example we indicate the proof of e.
$r(f,g) = r + w$ with $w \geq 0$, $r \in V$ if $r(f,g) \in V^+$. Then

$$L^n(f,g)v = \sum_{k=0}^{n-1} P^k(f,g)[r + w] + P^n(f,g)v ,$$

which converges to

$$\sum_{k=0}^{\infty} P^k(f,g)[r + w] \quad \text{for } n \to \infty . \qquad \square$$

Definition 2.4. Let $v \in V$. We define $Uv \in \mathbb{R}^{\infty}$ by

$$Uv := \max_{f \in \mathbb{F}} \min_{g \in \mathbb{G}} L(f,g)v ,$$

where max min is defined componentwise.

$(Uv)(i)$ is the value of the finite matrix game with entries

$$r(i,k,\ell) + \sum_j p(j;i,k,\ell)v(j) .$$

Lemma 2.2.

a) U maps V into V (monotonously).

b) Let $W := \{v \in V \mid \|v\| \leq (1-\beta)^{-1}\|\bar{r}\|\}$. $UW \subset W$.

c) U is contracting on V with contraction radius $\gamma \leq \beta$.

Proof.

a) The monotonicity is trivial.

Let $v \in V$. It will be proved that $Uv \in V$.

$$Uv \leq \max_f\{\min_g r(f,g) + \max_g P(f,g)v\}$$

$$\leq \bar{r} + \max_f \max_g P(f,g)v$$

$$\leq \bar{r} + \beta \| v \|_b \ .$$

Analogously

$$Uv \geq \max_f \{\min_g r(f,g) + \min_g P(f,g)v\}$$

$$\geq \bar{r} + \min_f \min_g P(f,g)v$$

$$\geq \bar{r} - \beta \| v \|_b \ .$$

Hence

$$\| Uv \| \leq \| \bar{r} \| + \beta \| v \| \ .$$

b) If $\| v \| \leq (1-\beta)^{-1} \| \bar{r} \|$, then

$$\| Uv \| \leq \| \bar{r} \| + \beta (1-\beta)^{-1} \| \bar{r} \| = (1-\beta)^{-1} \| \bar{r} \| \ .$$

c) Let $v,w \in V$.

$$Uv \leq U(w + \| v - w \|_b) =$$

$$\leq \max_f \min_g [r(f,g) + P(f,g)w + \| v - w \| P(f,g)b]$$

$$\leq Uw + \beta \| v - w \|_b \ .$$

So $Uv - Uw \leq \beta \| v - w \|_b$. By interchanging v and w and combining the two results we obtain $\| Uw - Uw \| \leq \beta \| v - w \|$.

Let $v \in V$, then $Uv = L(f,g)v$ for certain $f \in \mathbb{F}$, $g \in \mathbb{G}$. Hence

$$r(f,g) + P(f,g)v \in V \ ,$$

which implies $r(f,g) \in V$.

So by successive application of U one finds a sequence of policies f_n, g_n with $r(f_n,g_n) \in V$:

Choose $v_0 \in V$, define $v_n := Uv_{n-1}$ for $n = 1,2,\ldots$ and f_n, g_n such that

$$Uv_{n-1} = L(f_n, g_n)v_{n-1} .$$

Now we have $v_n \in V$, $v_n \to v^*$ (in norm) for $n \to \infty$, with v^* the unique solution of $Uv = v$, $v \in V$. $r(f_n, g_n) \in V$, $(f_n, g_n) \in \mathcal{D}$.
Denote by f^*, g^* policies which satisfy

$$L(f^*, g^*)v^* = Uv^* = v^* ,$$

then

$$V(f^*, g^*) = v^*, \ (f^*, g^*) \in \mathcal{D} .$$

The vector v^* is a natural candidate for being the value vector of the game and the stationary strategies f^*, g^* for being optimal strategies.

3. The value vector and optimal strategies

Lemma 3.1. Let $f \in \mathbb{F}$, $g \in \mathbb{G}$, then (f, g^*), $(f^*, g) \in \mathcal{D}$ and $V(f, g^*) \in \bar{V}$, $V(f^*, g) \in \underline{V}$. Furthermore

$$V(f, g^*) \leq V(f^*, g^*) \leq V(f^*, g) .$$

Proof.

$$L(f, g^*)v^* \leq L(f^*, g^*)v^* = v^* .$$

So

$$r(f, g^*) + P(f, g^*)v^* \leq v^* ,$$

hence $r(f, g^*) \in \bar{V}$.
Lemma 2.1(d,e) now implies

$$V(f, g^*) = \lim_{n \to \infty} L^n(f, g^*)v^* \leq v^* .$$

Lemma 3.2. Let $\pi := (f_0, f_1, \ldots) \in \mathbb{F}^\infty$, $\rho := (g_0, g_1, \ldots) \in \mathbb{G}^\infty$, then $V(\pi, g^*)$, $V(f^*, \rho)$ are defined and elements of $\bar{V}$, $\underline{V}$ respectively.
Furthermore

$$V(\pi, g^*) \leq V(f^*, g^*) \leq V(f^*, g) .$$

Proof.

$$r(f_t,g^*) \le v^* - P(f_t,g^*)v^* \le (1+\beta)\|v^*\|b \ .$$

So

$$r^+(f_t,g^*) \le (1+\beta)\|v^*\|b \ ,$$

$$E_{\pi,g^*} \sum_{t=0}^{\infty} r^+(f_t,g^*)(S_t) \le (1+\beta)\|v^*\| \sum_{t=0}^{\infty} P^t(\pi,g^*)b \le$$

$$\le (1+\beta)(1-\beta)^{-1}\|v^*\|b < \infty \ .$$

Hence

$$V(\pi,g^*) \in V^- \ .$$

With lemma 2.1(d,e) we obtain for any N

$$L(f_0,g^*) \ldots L(f_N,g^*)v^* \le v^* \ .$$

Hence

$$\sum_{t=0}^{N} P^t(\pi,g^*)r(f_t,g^*) + P^{N+1}(\pi,g^*)v^* \le v^* \ .$$

The second term in the left hand side of this inequality converges to zero for $N \to \infty$ (componentwise and in norm) whereas the first part converges to $V(\pi,g^*)$. Hence

$$V(\pi,g^*) \le v^* = V(f^*,g^*) \ .$$

Theorem 3.1. v^* - the unique solution of $Uv = v$, $v \in V$ - is the value vector for the game with criterion V. Any pair of stationary strategies f^*,g^* satisfying

$$L(f,g^*)v^* \le v^* \le L(f^*,g)v^*$$

is optimal for the game with criterion V, i.e.

$$(\pi,g^*),\ (f^*,\rho) \in \mathcal{D} \text{ if } \pi \in F,\ \rho \in G$$

and

$$V(\pi,g^*) \le V(f^*,g^*) = v^* \le V(f^*,\rho) \quad \text{for } \pi \in F,\ \rho \in G \ .$$

Proof. It has to be proved that $V(\pi,g^*)$, $V(f^*,\rho)$ are defined and are elements of V^-, V^+ respectively. Furthermore the saddle point property has to be proved.

Let $\pi = (\pi_0,\pi_1,\ldots) \in F$, then (as in the proof of lemma 3.2)

$$\mathbf{E}_{\pi,g^*}\, r^+(\pi_n(H_n),g^*)(S_n) \le \mathbf{E}_{\pi,g^*}(1+\beta)\|v^*\|b(S_n)$$

$$\le (1+\beta)\|v^*\|\beta^n b \ .$$

So

$$V(\pi,g^*) \le (1+\beta)(1-\beta)^{-1}\|v^*\|b \ .$$

Now it suffices to prove

$$\sup_{\pi\in F} V(\pi,g^*) = v^* \ .$$

We know already

$$\sup_{\pi\in \mathbb{F}^\infty} V(\pi,g^*) = v^* \ .$$

Consider the Markov decision process with state space $\mathbb{S}$, strategies $\pi \in F$ based on the action sets $\mathbb{F}_i$. $P(f,g^*)$ is the matrix of transition probabilities if the policy f is applied. $r(f,g^*)$ is the reward vector for policy f. For this Markov decision process we have

$$\sup_{\pi\in\mathbb{F}^\infty} \mathbf{E}_{\pi,g^*} \sum_{n=0}^{\infty} r^+(\pi_n(H_n),g^*)(S_n) \le (1+\beta)(1-\beta)^{-1}\|v^*\|b \ .$$

Note that for this Markov decision process $\mathbb{F}^\infty$ is the set of all nonrandomized Markov strategies and F is the set of all nonrandomized strategies. We now may use the theorem of van Hee [1], which states that for the Markov decision process with total expected reward vector $V(\pi,g^*)$ for the strategy π

$$\sup_{\pi\in F} V(\pi,g^*) = \sup_{\pi\in\mathbb{F}^\infty} V(\pi,g^*) \ .$$

Note that all suprema should to be taken componentwise. □

4. Successive approximations

Since the value vector v^* is the unique fixed point of the operator U on V, it is trivial to give successive approximations for v^*.
Choose $v_0 \in V$, then $v_n := Uv_{n-1}$ for $n = 1,2,\ldots$ constitute a sequence of vectors in V which converge (in norm) to v^*

$$v_n - \beta(1-\beta)^{-1}\| v_n - v_{n-1} \|b \le v^* \le v_n + \beta(1-\beta)^{-1}\| v_n - v_{n-1} \|b \ .$$

As in discounted Markov games (compare van der Wal [7]) better estimates may be found for v^* in the n-th iteration. In a similar way estimates may be found for the quality of the policies found in the n-th iteration.
For these estimates we use similar notations as van der Wal [7]

$$\lambda_n := \inf_i \ (v_n(i) - v_{n-1}(i))\mu(i) \ ,$$

$$\nu_n := \sup_i \ (v_n(i) - v_{n-1}(i))\mu(i) \ ,$$

$$a_n := \begin{cases} \sup\limits_{i,\ell} \mu(i) \sum\limits_j p(j;i,f_n(i),\ell)b(j) & \text{if } \lambda_n < 0 \ , \\ \inf\limits_{i,\ell} \mu(i) \sum\limits_j p(j;i,f_n(i),\ell)b(j) & \text{if } \lambda_n \ge 0 \ . \end{cases}$$

$$b_n := \begin{cases} \inf\limits_{i,k} \mu(i) \sum\limits_j p(j;i,k,g_n(i))b(j) & \text{if } \nu_n < 0 \ , \\ \sup\limits_{i,k} \mu(i) \sum\limits_j p(j;i,k,g_n(i))b(j) & \text{if } \nu_n \ge 0 \ , \end{cases}$$

here f_n, g_n denote policies which satisfy

$$L(f,g_n)v_{n-1} \le L(f_n,g_n)v_{n-1} \le L(f_n,g)v_{n-1} \ .$$

Now we have the following bounds for v^*, $V(f_n,g_n)$, $V(f_n,\rho)$, $V(\pi,g_n)$ (the latter two are defined and in V^+, V^-)

a) $$v_n + \frac{a_n\lambda_n}{1-a_n} b \le v^* \le v_n + \frac{b_n\nu_n}{1-b_n} b \ ,$$

b) $$V(f_n,\rho) \ge v_n + \frac{a_n\lambda_n}{1-a_n} b \quad \text{for all } \rho \in G \ ,$$

c) $$V(\pi,g_n) \le v_n + \frac{b_n \nu_n}{1 - b_n} b \quad \text{for all } \pi \in F ,$$

d) $$v_n + \frac{a_n \lambda_n}{1 - a_n} b \le V(f_n,g_n) \le v_n + \frac{b_n \nu_n}{1 - b_n} b .$$

Proof. d) is a direct consequence of b) and c).
a) is also a consequence of b) and c), viz. $V(f_n,g^*) \le V(f^*,g^*) = v^*$ and $V(f^*,g_n) \ge V(f^*,g^*) = v^*$.
We will prove c).
That $(\pi,g_n) \in \mathcal{D}$ is seen as follows:

$$L(f,g_n)v_{n-1} \le L(f_n,g_n)v_{n-1} = v_n \quad \text{for all } f \in \mathbb{F} .$$

Hence

$$r(f,g_n) \le v_n - P(f,g_n)v_{n-1} \le [\| v_n \| + \beta \| v_{n-1} \|]b \quad \text{for all } f \in \mathbb{F} .$$

So for any $f \in \mathbb{F}$ we have

$$r^+(f,g_n) \le [\| v_n \| + \beta \| v_{n-1} \|]b .$$

Hence

$$\sum_{t=0}^{\infty} \mathbf{E}_{\pi,g_n} r^+(\pi_t(H_t),g_n)(S_t) \le (1 - \beta)^{-1}[\| v_n \| + \beta \| v_{n-1} \|]b .$$

Using the same theorem of van Hee [1] as in the proof of theorem 3.1 (with g_n instead of g^*) we obtain

$$\sup_{\pi \in F} V(\pi,g_n) = \sup_{\pi \in \mathbb{F}^\infty} V(\pi,g_n) .$$

So it remains to prove

$$V(\pi,g_n) \le v_n + \frac{b_n \nu_n}{1 - b_n} b \quad \text{for } \pi \in \mathbb{F}^\infty .$$

$$L(f,g_n)v_{n-1} \le v_n \le v_{n-1} + \nu_n b \quad \text{(by definition)} .$$

Hence for all $f_0^0, f_1^0 \in \mathbb{F}$

$$L(f_0^0,g_n)L(f_1^0,g_n)v_{n-1} \le L(f_0^0,g_n)[v_{n-1} + \nu_n b]$$

$$= L(f_0^0,g_n)v_{n-1} + \nu_n P(f_0^0,g_n)b$$

$$\le v_n + \nu_n b_n b \quad .$$

In this way we obtain for all $f_0^0,\dots,f_N^0 \in \mathbb{F}$

$$L(f_0^0,g_n)L(f_1^0,g_n) \dots L(f_N^0,g_n)v_{n-1} \le v_n + \nu_n(b_n + \dots + b_n^N)b \quad .$$

This implies for $\pi = (f_0^0,f_1^0,\dots) \in \mathbb{F}^\infty$

$$V(\pi,g_n) \le v_n + \nu_n b_n (1 - b_n)^{-1} b \quad . \qquad \square$$

5. Remarks and extensions

a) If $|r(i,k,\ell)| \le Mb(i)$ for all i,k,ℓ and certain $M > 0$, then $\mathcal{D} = F \times G$ and

$$V(\pi,\rho) = \mathbb{E}_{\pi,\rho} \sum_{n=0}^{\infty} r(S_n,K_n,L_n) \quad ,$$

which converges absolutely. In this situation we have $V(\pi,\rho) \in \mathcal{V}$.
So, in this way we have already a generalization of the usual situation.
A comparison with discounted Markov games and discounted semi-Markov games may be made by incorporating the discountfactor in the transition probabilities.

b) The finite action sets $\mathbb{K}_i$ and $\mathbb{L}_i$ may be replaced by compact subsets of some Euclidean space if we add the condition that $r(i,k,\ell)$ and $p(j;i,k,\ell)$ are continuous in (k,ℓ).

c) If in each state i the action set $\mathbb{L}_i$ contains only one element we obtain a Markov decision process with the following requirements (we delete all parameters relating to the second player P_2)

$$\|P(f)\| \le \beta < 1$$

$$\max_{k \in \mathbb{K}_i} r(i,k) =: \bar{r}(i) \quad \text{with } \bar{r} \in \mathcal{V} \quad .$$

Now we have for any f that $r(f) \le \bar{r}$, hence $r(f) \in \mathcal{V}^-$. These properties imply that $V(\pi)$ exists for all π.
The finiteness of $\mathbb{K}_i$ is not essential if we replace "max" by "sup". The successive appproximations v_n may be replaced by approximations of Uv_{n-1}

in the following way ($\delta > 0$).

Choose f_n such that

$$v_n := L(f_n)v_{n-1} \geq \max\{v_{n-1}, Uv_{n-1} - \delta b\} ;$$

this is possible if v_0 is chosen such that

$$v_0(i) < (Uv_0)(i) \quad \text{for all } i \in S .$$

This sequence $\{v_n\}$ generates the same sequence of estimates for v^* and $V(f_n)$ as given in [8] for the condition

$$|r(i,k)|\mu(i) \leq M \quad \text{for all } i \in \mathbb{S},\ k \in K_i .$$

d) Stopping times may be used to generate a class of successive approximation methods as has been shown by van der Wal [7]. This can also be done in our situation.

e) An interesting situation, which may be treated within the frame work of this paper, is the following.
Suppose at each time $t = 0,1,\ldots$ a honest matrix game is played (the entries may depend on the history but the actual game is always honest). Then for both players strategies which are optimal in each single matrix game (i.e. prudent strategies with value 0) are overall optimal. No extra restrictions on matrix entries are necessary.

f) For characterizations of the situations in which assumption 2.1.b $\|P(f,g)\| \leq \beta < 1$ is satisfied we refer to [2], since in this respect there is no essential difference between Markov decision processes and Markov games.

References

[1] K.M. van Hee, Markov strategies in dynamic programming. Memorandum COSOR 75-20 (October 1975), Dept. of Mathematics, Eindhoven University of Technology.

[2] K.M. van Hee, J. Wessels, Markov decision processes and strongly excessive functions. Memorandum COSOR 75-22 (November 1975), Dept. of Mathematics, Eindhoven University of Technology.

[3] C.H. Himmelberg, T. Parthasarathy, T.E.S. Raghavan, F.S. Van Fleck, Existence of p-equilibrium and optimal stationary strategies in stochastic games. Proc. Amer. Math. Soc. (to appear).

[4] J. MacQueen, A modified dynamic programming method for Markovian decision problems. J. Math. Anal. Appl. 14 (1966), 38-43.

[5] T. Parthasarathy, Discounted, positive, and noncooperative stochastic games. Intern. J. Game Th. 2 (1973), 25-37.

[6] L.S. Shapley, Stochastic games. Proc. Nat. Acad. Sci. USA 39 (1953), 1095-1100.

[7] J. van der Wal, The method of successive approximations for the discounted Markov game. Memorandum COSOR 75-02 (March 1975), Dept. of Mathematics, Eindhoven University of Technology.

[8] J. Wessels, Markov programming by successive approximations with respect to weighted supremum norms. J. Math. Anal. Appl. (to appear).