Zero-Sum Stochastic Games

A Survey

Koos Vrieze
University of Limburg
P.O. Box 616, 6200 MD Maastricht, The Netherlands

We present a survey of stochastic models of competitive behaviour between two players in a dynamic setting. After presenting the formal model of stochastic games, we consider discounted reward, average reward, total reward, and structured stochastic games. We conclude with some approximation algorithms for computing the value of discounted- and average reward stochastic Games.

1. Introduction

In this paper we present a survey on zero-sum Stochastic Games. These games are stochastic models of competitive behaviour between two players in a dynamic setting. They include as special cases static noncooperative games, repeated games with complete information and Markovian Decision Problems. These games are played at a series of discrete time points, called stages or decision moments. At each stage a zero-sum game has to be played, which is one out of a well-defined finite set of zero-sum games. Let S be this set of games. If at a certain stage a certain game, say game $s \in S$, has to be played then we say that our system is in state s. The dynamics of Stochastic Games are of the Markovian type in the sense that the state of the system at a certain decision moment only depends (in a stochastic way) on the state at the preceding decision moment and on the actions chosen by the players at that preceding decision moment.

If in the above setting one player were a dummy in each game $s \in S$, hence reducing these games to optimization problems, then the above model reduces to the classical Markovian Decision Problem. Thus Stochastic Games can be regarded as a fusion of two other types of decision problems, namely matrix games and Markovian Decision Problems. The zero-sum games to be played at each stage are matrix games. The dynamics of the Stochastic Game are embedded in a discrete time Markov Process.

The theory of Stochastic Games started with the fundamental paper of Shapley [26]. Strikingly the theory of Stochastic Games and the theory of Markovian Decision Problems, so close related, evolved for many years along separate lines. The techniques commonly used in the early approaches to Stochastic Games stem from the theory of functions and from fixed point theorems. Only in the late 1970's was the interrelationship between these two

fields of research fully recognized. From that time on many new results emerged by combining techniques from these two research areas. Different Dutch researchers contributed to this development. We mention Federgruen [9], Vrieze and Tijs [39], Tijs and Vrieze [32], Van der Wal [34], and Hordijk and Kallenberg [16]. Nowadays, new results in the theory of Stochastic Games usually grow out of a suitable injection of properties of static games into the framework of a Markovian Decision Problem.

Let us come back to the model. The basic component of zero-sum Stochastic Games is the matrix game (for an introduction to matrix games, see Owen [21]). A matrix game is a decision problem for two players, who simultaneously and independently have to choose an action out of a (finite) decision set. The action dependent payoffs to the players are conflicting in the sense that they add up to zero. Obviously, such a decision problem can be represented by a matrix M of numbers m_{ij}, where m_{ij} equals by convention the payoff to player 1, if player 1 chooses his i-th action and player 2 his j-th action. By the rules of the game, the payoff to player 2 in this case equals $-m_{ij}$. Obviously the goal of player 1 is to maximize the outcome of the game (by choosing in an appropriate way some row of M, while player 2 tries to minimize the outcome (by choosing in an appropriate way some column of M). Now a Stochastic Game is nothing else but repeatedly playing matrix games, at well-defined discrete decision moments, according to the following rules. There is fixed a finite number, say z, of matrix games. At each decision moment the players are informed about the matrix game at hand at that moment. Next, both players make a choice (simultaneously and independently) out of the action sets available in this matrix game. These choices result not only in a payoff (as in a matrix game), but also in an action dependent probability measure on the set of matrix games. Next, according to this probability measure, a chance experiment is carried out to determine the matrix game to be played at the following decision moment. So, in Stochastic Games, at each decision moment, the players have short term as well as long term interests: short term in the sense that the chosen action determines some immediate payoff at that moment; long term in the sense that the actions determine the dynamic behaviour of the system at that moment, giving rise to intentions of the players of steering the system to more favourable matrix games.

As already stated, Stochastic Games can be regarded as extensions of Markov decision problems (cf. Denardo [8]). A, say maximizing, Markov decision problem is defined analogously to Stochastic Games, with the restriction that each of the z matrix games consists of a single column. This reflects the fact that in Markov decision problems, we have to do with only one decision maker who has to choose a row out of that single column. The immediate rewards and the dynamics of the system are defined completely similarly to Stochastic Games. Obviously, a minimizing Markov decision problem can be represented as a Stochastic Game, where each of the matrix games consists of a single row. Thus a zero-sum Stochastic Game can be viewed as the extension of the multi-stage decision problem with one decision maker to the case of the multi-stage decision problem with two decision makers having strictly opposed

interests. The theory of Markov decision problems evolved in the late fifties and the early sixties (Bellman [1], Blackwell [4,5]). At that time, people working in the field of Markov decision problems seemed to be unaware of the pioneering work of Shapley. Some of the essential theorems were derived for the (simpler) case of Markov decision problems a decade later than Shapley did for Stochastic Games. In this sense, Shapley was ahead of his time.

This paper is organised as follows. In Section 2 the notion of Stochastic Game is formally introduced. Strategies are defined and three possible evaluation criteria are given. By an evaluation criterion we mean a rule which prescribes how to combine the immediate payoffs on the different decision moments into one overall criterion. We consider three criteria: (i) 'discounting', introduced by Shapley [26], (ii) 'average', introduced by Gillette [14] and (iii) 'total', introduced by Thuijsman and Vrieze [31].

In Section 3 discounted reward Stochastic Games are treated. It is shown how properties of the solution sets of matrix games projected into the Shapley optimality equations for discounted Stochastic Games, lead to the characterizing properties of the solution sets of discounted Stochastic Games. Most of the results are derived from Vrieze and Tijs [39] and Tijs and Vrieze [32].

In Section 4 we look at average reward Stochastic Games. Bewley and Kohlberg [2,3] introduced the idea of using the notion of Puiseux series in these type of games. The emphasis will lie on characterizations of games for which both players possess optimal stationary strategies. Most of the results in this section come from Vrieze [37] and Thuijsman [30].

In Section 5 total reward Stochastic Games are considered. In a certain sense, this criterion appears to have properties similar to the average criterion (Thuijsman and Vrieze [31]). The relations between the total reward criterion with the other two criteria are exposed in this section. Most of the results in this respect stem from Thuijsman [30].

In Section 6 we handle a number of so-called structured Stochastic Games, i.e., subclasses of games, determined by placing restrictions on the reward and transition data. Examples are Stochastic Games where the dynamics of the games can only be influenced by one player (cf. Parthasarathy and Raghavan [22] and Vrieze [37]) or games where the reward function can be written as a sum of a term only depending on the state and of a term only depending on the actions (cf. Sobel [29] and Parthasarathy et al. [23]).

Finally in Section 7 we give some approximation algorithms for computing the value and ϵ-optimal stationary strategies for the discounted reward as well as for the average reward Stochastic Games. Most of these algorithms are extensions of algorithms for Markovian Decision Problems to Stochastic Games. (cf. Van der Wal [34], Federgruen [9], Hordijk and Kallenberg [16]).

2. The stochastic game model

In this section we state the formal definition of a Stochastic Game and the way it is played.

A Stochastic Game consists of a sequence of matrix games $M_1, M_2, \ldots, M_t, \ldots$ to be played consecutively, where $M_t \in S$, $t = 1,2,\ldots$, with S a finite set of

matrix games. Notation: $S = \{1,2,...,z\}$. A general element of S will always be denoted by the symbol s. If $M_t = s$, then we say that our system is in state s at decision moment t.

Matrix game s has size $m_s \times n_s$. Hence player 1 (the row player) has available the set $A_s := \{1,2,...,m_s\}$ as his set of pure actions in state s, while player 2 (the column player) has available the set $B_s := \{1,2,...,n_s\}$ as his set of pure actions in state s.

The (i,j)-th entry of matrix games s is denoted by $r_s(i,j)$ being by definition the payoff to player 1 to be paid by player 2 if in state s player 1 chooses action i and player 2 chooses action j.

The dynamics of the Stochastic Game are represented by transition probabilities. Let $\mathbb{P}(S) := \{(x_1,x_2,...,x_z);\ x_s \geqslant 0,\ \Sigma_{s=1}^{z} x_s = 1\}$ be the set of probability measures on the set of matrix games S. Then to each pair of actions (i,j) of the players in a state s there is associated a probability measure $p_s(i,j) \in \mathbb{P}(S)$ with the following meaning: if in state s player 1 chooses i and player 2 chooses j then the probability that at the next decision moment matrix game $t \in S$ has to be played equals the t-th component of $p_s(i,j)$, notation $p(t \mid s,i,j)$.

According to the number of decision moments two types of Stochastic Games can be distinguished: finite horizon and infinite horizon. In this paper we will concentrate on the latter one, where the set of decision moments is supposed to be the set $\mathbb{N}$ of natural numbers. At the end of this section it will be made clear that Stochastic Games with a finite number of decision moments can be identified with matrix games.

A Stochastic Game is played as follows. A starting state $s_1 \in S$ is given to both players at decision moment 1. Both players simultaneously and independently choose an action out of their respective available action sets. Say, this results in action $i_1 \in A_{s_1}$ for player 1 and action $j_1 \in B_{s_1}$ for player 2. Then two things happen. First there is an immediate payoff $r_{s_1}(i_1,j_1)$ to player 1 from player 2 and second, the system moves to a next state according to the probability measure $p_{s_1}(i_1,j_1)$ where $p(t \mid s_1,i_1,j_1)$, for each $t \in S$ equals the probability that this next state will be state t. Then at decision moment 2 both players are informed about the new current state. Here the game proceeds as if it starts again, etc. We assume perfect recall and complete information, i.e. at each decision moment both players perfectly remember all past states and actions that have actually occurred and both players know each function r_s and all mappings p_s completely.

As usually in non-cooperative game theory we allow the players to select at each decision moment a (pure) action according to the specification of a mixed action. Since, at each decision moment, the players have full knowledge of the history of the game up to that moment, they may use this knowledge in specifying their mixed action. Furthermore, this mixed action may depend on the stage number. Formally, let h_n be the history of the game at decision moment n, i.e. $h_n := (s_1,i_1,j_1,s_2,i_2,j_2,...,s_{n-1},i_{n-1},j_{n-1})$, where at decision moment k, $s_k \in S$, $i_k \in A_{s_k}$ and $j_k \in B_{s_k}$ have occurred for $k = 1,2,...,n-1$. Let $\mathbb{P}(A_s)(\mathbb{P}(B_s))$

denote the set of mixed actions for player 1 (player 2) in state $s \in S$, i.e.

$$\mathbf{P}(A_s) := \{(x_1, x_2, \ldots, x_{m_s}); x_i \geqslant 0 \text{ and } \sum_{i=1}^{m_s} x_i = 1\}$$

and

$$\mathbf{P}(B_s) := \{(y_1, y_2, \ldots, y_{n_s}); y_j \geqslant 0 \text{ and } \sum_{j=1}^{n_s} y_j = 1\}.$$

Then a behaviour strategy for player 1, notation π_1, can be associated with a function π_1 on the set of triples (s,n,h_n) with $\pi_1(s,n,h_n) \in \mathbf{P}(A_s)$ for each s,n,h_n. Such a strategy is used as follows: if at decision moment n the state equals s and if history h_n has occurred, then player 1 chooses his pure action according to the mixed action $\pi_1(s,n,h_n)$. A behaviour strategy for player 2, notation π_2, is analogously defined.

Three special types of strategies are discerned.

- First, a pure strategy is a strategy where $\pi_k(s,n,h_n)$ specifies with probability one some pure action for each $s \in S$, history h_n and decision moment n.
- Second, a Markov strategy is a strategy where at each decision moment the mixed action only depends on the stage number and on the current state and not on the history of the game. Hence a Markov strategy for player k is a function π_k on the set of pairs (s,n), with the same interpretation as above for behaviour strategy.
- Third, a stationary strategy is a strategy where at each decision moment the mixed action only depends on the current state and not on the stage number or the history of the game. For stationary strategies we introduce an apart notation, ρ for player 1 and σ for player 2. Then $\rho = \{\rho(s); s \in S\}$ with $\rho(s) \in \mathbf{P}(A_s)$ and $\sigma = \{\sigma(s); s \in S\}$ with $\sigma(s) \in \mathbf{P}(B_s)$ and when player 1 decides to play a stationary strategy ρ, then each time the system is in state s he will choose his pure action according to $\rho(s)$; similarly for player 2.

If both players specify a strategy, say π_1 and π_2, then for a fixed starting state $s \in S$, this will determine a probability measure on the set of histories h_n up to decision moment n. We will denote these probabilities by $P_n(s, \pi_1, \pi_2, h_n)$. From these probabilities we can derive two things.

- First, since $P_{n-1}(s, \pi_1, \pi_2, .)$ can be interpreted as some marginal distribution of $P_n(s, \pi_1, \pi_2, .)$ it follows by the Kolmogorov extension theorem, that the sequence $(P_n(s, \pi_1, \pi_2, .), n = 1,2,\ldots)$ can be extended to a unique probability measure on the set of infinite sequences $(s_1, i_1, j_1, s_2, i_2, j_2, \ldots)$.
- Second, for each decision moment n, the marginal distribution of triples (s_n, i_n, j_n) occurring at decision moment n can be computed.

Let $P_{s\pi_1\pi_2}(s_n, i_n, j_n)$ denote the probability that the triple (s_n, i_n, j_n) occurs at decision moment n if player 1 plays π_1, player 2 plays π_2 and the starting state is s. Then we can compute the expected payoffs at the different decision moments. Let $R(n)$ be the stochastic variable denoting the payoff at decision moment n, then

$$E_{s\pi_1\pi_2}[R(n)] := \sum_{s_n,i_n,j_n} P_{s\pi_1\pi_2}(s_n,i_n,j_n) r_{s_n}(i_n,j_n). \tag{1}$$

Already Shapley [26] showed that for stationary strategies this expression can considerably be simplified.

Let $E_{\pi_1\pi_2}[R(n)] := (E_{1\pi_1\pi_2}[R(n)],...,E_{z\pi_1\pi_2}[R(n)])$. Then for a pair of stationary strategies ρ and σ it holds that

$$E_{\rho\sigma}[R(n)] = P^{n-1}(\rho,\sigma) r(\rho,\sigma) \tag{2}$$

where $P^k(\rho,\sigma)$ equals the k-fold product of the $z \times z$-matrix $P(\rho,\sigma)$ and where the (s,t)-th element of $P(\rho,\sigma)$, notation $p(t \mid s,\rho,\sigma)$, equals: $p(t \mid s,\rho,\sigma) := \Sigma_{i=1}^{m_s} \Sigma_{j=1}^{n_s} p(t \mid s,i,j) \rho_i(s) \sigma_j(s)$. Further $r(\rho,\sigma) = (r_1(\sigma), r_2(\rho,\sigma),..., r_z(\rho,\sigma))$ with $r_s(\rho,\sigma) := \Sigma_{i=1}^{m_s} \Sigma_{j=1}^{n_s} r_s(i,j) \rho_i(s) \sigma_j(s)$.

The interpretation is as follows: $p(t \mid s,\rho,\sigma)$ being the (s,t)-th element of $P(\rho,\sigma)$, equals the probability that the system moves in one step to state $t \in S$ if in state $s \in S$ player 1 plays $\rho(s)$ and player 2 plays $\sigma(s)$. It can easily be shown by induction, that the (s,t)-th element of $P^{n-1}(\rho,\sigma)$ equals the probability that at decision moment n the system is in state t if it starts at decision moment 1 in state s and if the players play the stationary strategies ρ and σ.

Obviously $r_s(\rho,\sigma)$ is the expected immediate reward in state s when player 1 plays $\rho(s)$ and player 2 plays $\sigma(s)$.

Now expression (2) is immediate. Observe that expression (2) specifies simultaneously the expected payoff at stage n for all z specific plays with starting state respectively 1,2,...,z.

Summarizing the above, we see that, associated with a pair of strategies (π_1,π_2) and a specific starting state s, there is a sequence of expected payoffs $(E_{s\pi_1\pi_2}[R(n)],\ n = 1,2,...)$. In order to compare the worth of strategies, an evaluation criterion is needed, i.e. a rule which uniquely associates a real number to such a sequence. In this paper we consider three evaluation rules.

- First, the discounted reward criterion, defined as

$$v^\beta(s,\pi_1,\pi_2) := (1-\beta) \sum_{n=1}^{\infty} \beta^{n-1} E_{s\pi_1\pi_2}[R(n)]. \tag{3}$$

  Here $\beta \in (0,1)$ is the discount factor, reflecting the interest rate. The factor $1-\beta$ is only meant for normalization purposes.

- Second, the average reward criterion, defined as

$$v(s,\pi_1,\pi_2) := \liminf_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} E_{s\pi_1\pi_2}[R(n)]. \tag{4}$$

  Since the limit of the right-hand side of (4) does not need to exist, a further specification is necessary. The choice of lim inf ('the worst case') is more or less arbitrary. However the results for average reward Stochastic Games do not change when lim inf is replaced by lim sup or any convex combination of them.

- Third, the total reward criterion, defined as

$$v^T(s,\pi_1,\pi_2) = \liminf_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} \sum_{n=1}^{k} E_{s\pi_1\pi_2}[R(n)]. \tag{5}$$

In general this expression may be $+\infty$ or $-\infty$. We will apply this criterion to a class of Stochastic Games where this expression makes sense. Observe that in case where $\Sigma_{n=1}^{\infty} E_{s\pi_1\pi_2}[R(n)]$ exists, we have $v^T(s,\pi_1,\pi_2) = \Sigma_{n=1}^{\infty} E_{s\pi_1\pi_2}[R(n)]$.

The expressions (3), (4) and (5) specify the expected payoffs to player 1. By definition the payoffs to player 2 are the negatives of these expressions.

As solution concept for zero-sum Stochastic Games the usual concept for zero-sum games in normal form is adopted. The players seek strategies which guarantee them a payoff as high as possible. Since the payoffs, for any criterion, are defined as the payoffs to player 1 (and as minus the payoffs to player 2), this concept leads to the following. Let $c(s,.,.)$ stand for any criterion. Then player 1 tries to find a strategy which guarantees him a payoff as close as possible to $\sup_{\pi_1} \inf_{\pi_2} c(s,\pi_1,\pi_2)$. Analogously, player 2 tries to find a strategy which guarantees him an expected payoff as close as possible to $\inf_{\pi_1} \sup_{\pi_2} c(s,\pi_1,\pi_2)$. Whenever

$$c(s) := \sup_{\pi_1} \inf_{\pi_2} c(s,\pi_1,\pi_2) = \inf_{\pi_2} \sup_{\pi_1} c(s,\pi_1,\pi_2), \tag{6}$$

we say that the game is strictly determined for starting state s. In that case, the number $c(s)$ is called the value of the game for starting state s. A Stochastic Game is said to have a value if for each starting state the game is strictly determined. A strategy which guarantees a player the value of the game up to ϵ, $\epsilon \geqslant 0$, is called ϵ-optimal, so for player 1, π_1^* is ϵ-optimal if

$$\inf_{\pi_2} c(s,\pi_1^*,\pi_2) \geqslant c(s) - \epsilon \tag{7}$$

and π_2^* is ϵ-optimal for player 2 if

$$\sup_{\pi_1} c(s,\pi_1,\pi_2^*) \leqslant c(s) + \epsilon. \tag{8}$$

A 0-optimal strategy is called optimal.

We conclude this section by a remark on Stochastic Games with a finite number of decision moments. For all three evaluation criteria these games can be reformulated (and therefore handled) as a matrix game. Observe that for a game with finite horizon both players have a finite number of pure strategies. When we display the Stochastic Game for a fixed starting state as a game in extensive form, then these sets of pure strategies for the players coincide with the set of pure strategies for the players in the extensive form game. By the results of Kuhn [19], it follows that these games can be regarded as matrix games, merely by numbering the pure strategies (being the sets of pure actions in the corresponding matrix game) and by relating to each pair of pure strategies the expected payoff as assigned by the evaluation rule to the finite stream of expected payoffs of these strategies. Hence (Von Neumann [35]) the

value of these finite horizon games exists for any criterion and both players possess optimal strategies, consisting of mixtures of pure strategies.

3. The discounted reward criterion

Though in fact Shapley considered 'stopping' Stochastic Games with the total reward criterion, his 1953 paper can be seen as the start of the theory on Stochastic games in general and of the discounted reward criterion in particular. The mathematical techniques used in discounted reward games are similar to those used in stopping total reward games. A stopping Stochastic Game is a game with the property $q(s,i,j) := 1-\Sigma_{t=1}^{z} p(t\,|\,s,i,j)>0$ for each s,i and j. $q(s,i,j)$ equals the stopping probability in state s when the players select action i and j respectively. A discounted reward Stochastic Game can be formulated as a total reward stopping Stochastic Game by adapting the transition probabilities. Take for the stopping game $1-\beta$ as the stopping probability and take $\beta p(t\,|\,s,i,j)$ as the probability of moving from s to t for actions i and j. It can be verified that for each pair of strategies the total reward in this stopping game equals the discounted reward in the original game.

We now state the main theorem, due to Shapley [26], of discounted reward Stochastic Games. For that purpose, define for $v = (v_1,v_2,\ldots,v_z)\in\mathbb{R}^z$, for each $s\in S$, the matrix game

$$M_s^{\beta}(v) := [(1-\beta)r_s(i,j)+\beta\sum_{t=1}^{z} p(t\,|\,s,i,j)v_t]_{i=1}^{m_s}{}_{j=1}^{n_s}. \tag{9}$$

Further $\mathrm{Val}(M_s^{\beta}(v))$ will denote the minmax value of this matrix game.

THEOREM 3.1.

(a) *Discounted reward Stochastic Games are strictly determined.*

(b) *The value, say* $v^{\beta} := (v_1^{\beta},v_2^{\beta},\ldots,v_z^{\beta})$ *equals the unique solution to the following set of functional equations:*

$$v_s = \mathrm{Val}(M_s^{\beta}(v)), \text{ for each } s\in S \tag{10}$$

(c) *A stationary strategy,* ρ, *for player 1 is optimal if and only if, for each* $s\in A(s)$, ρ_s *is an optimal action for player 1 in* $M_s^{\beta}(v^{\beta})$. *A similar result holds for player 2.*

The proof of this theorem is based on the fact that the right-hand side of (9) represents a contraction mapping with contraction factor β on the $\mathbb{R}^z$. Hence Banach's contraction mapping theorem yields a unique solution (fixed point) to the set of equations (10), which turns out to be the value of the game. An alternative proof of Theorem 3.1 is given in Vrieze [37]. There the set of equations (10) is formulated as a non-linear programming problem (linear object function subject to quadratic constraints). Application of the Kuhn-Tucker conditions to this NLPP gives a constructive proof of all parts of Theorem 3.1.

In an essential way Theorem 3.1 makes use of matrix game theory. Indeed several structural properties of discounted reward Stochastic Games can be found by suitable injection of matrix game properties. For instance, by (c) of

Theorem 3.1 and by the Bohnenblust, Karlin and Shapley [7] characterization of solution sets of matrix games, we derive (Vrieze and Tijs [39]):

THEOREM 3.2. *The set of optimal stationary strategies O_1^β, for player k, $k = 1,2$, in the discounted Stochastic Game is equal to the Cartesian product*

$$X_{s=1}^{z} O_k^\beta(s),$$

where $O_k^\beta(s)$ is the convex polyhedron of optimal, mixed actions of player k in the matrix game $M_s^\beta(v^\beta)$.

Also the Shapley-Snow [27] results concerning the extreme optimal actions for matrix games can be extended to Stochastic Games (Vrieze and Tijs [39]).

THEOREM 3.3. *Let ρ^E be an extreme point of O_1^β and σ^E be an extreme point of O_2^β. Then there exists a stochastic subgame from which ρ^E and σ^E can be computed in the Shapley-Snow manner. (Here a subgame arises when pure actions are deleted from one or several states for one or both players).*

Notice that this theorem gives a method, though not an efficient one, of computing the extreme optimal stationary strategies of $O_1^\beta \times O_2^\beta$ by looking at the finite number of stochastic subgames in which at each state both players have the same number of pure actions.

A next theme that lends itself to conveying properties of matrix games to Stochastic Games, is perturbation theory. In the first place, from Theorem 3.1, part (b) and the fact that the value of a matrix game is a continuous function of the entries of a matrix game, it follows that (Tijs and Vrieze [32]):

THEOREM 3.4. *The value of discounted reward Stochastic Games, considered as a function on the parameters (rewards, transitions, discount factor) is a continuous one.*

Also with respect to the sets of ϵ-optimal stationary strategies ($\epsilon \geqslant 0$) a continuity statement can be made (Tijs and Vrieze [32]):

THEOREM 3.5. *Let $O_k^\beta(\epsilon)$ be the set of ϵ-optimal stationary strategy to player k, $k \in \{1,2\}$. Then $O_k^\beta(\epsilon)$ is an upper semi-continuous multimap on the parameters of the Stochastic Game.*

A useful implication of Theorem 3.5 is the following observation. Take $\epsilon > 0$. Then for any two games S and $\tilde{S}$ which are 'close enough' to each other it holds that $O_k^\beta(\epsilon) \cap \tilde{O}_k^\beta(\epsilon) \neq \varnothing$. Moreover it can be shown that $O_k^\beta(\epsilon) \subset \tilde{O}_k^\beta(\epsilon + c\delta)$, where c is some number determined by the parameters of S and δ is the distance between S and $\tilde{S}$.

In practical situations small deviations in the exact values of the game parameters are inevitable. Theorems 3.4 and 3.5 show that small changes in the

game parameters induce only small changes in the solution of the game. This property increases the reliability and practicability of discounted reward Stochastic Games.

A last property for discounted reward Stochastic Games that we will mention is a topological one. Fix action spaces A_s, B_s, $s \in S$. Let SG be the class of Stochastic Games with these action spaces. Let USG be the subclass of SG for which both players have a unique optimal stationary strategy. Notice from Theorem 3.5 that this unique optimal stationary strategy varies continuously over USG. (If a player possesses in a discounted reward Stochastic Game a unique optimal stationary strategy, then this strategy is his only optimal strategy, stationary or not. For the class of matrix games of fixed size, the subclass of matrix games with unique optimal actions for both players is a dense and open subset with respect to this class (Bohnenblust, Karlin and Shapley [7]). For Stochastic Games an analogous result, proved in an analogous way, holds (Tijs and Vrieze [32]).

THEOREM 3.6. *The set USG is a dense and open subset with respect to SG.*

In fact Theorem 3.6 states that each open neighbourhood in SG of a fixed Stochastic Game has a non-void intersection with the set USG. Even more it can be shown that each such neighbourhood contains elements of USG that have the same value as the fixed Stochastic Game. Furthermore, since USG is open, it follows that a 'generic' Stochastic Game is one belonging to USG.

4. THE AVERAGE REWARD CRITERION

Average reward Stochastic Games are considerably more difficult to analyse then discounted reward Stochastic Games. The reason comes from the fact that the expected average reward is not continuous over the strategy space (in contrast to the discounted case).
For example:

(This game has two basic matrix games, 1 and 2. For matrix game 1, player 1 has one action (one row) while player 2 has two actions (two columns). A box $\boxed{r \diagdown t}$ means: payoff r and a transition to state t with probability 1. State 2 is degenerate for both players: they both have only one action available; furthermore, this state is absorbing, since once being there, the game will stay in state 2 for ever).

If player 2 plays the mixed action $(1-\epsilon,\epsilon)$ in state 1, then the average reward equals 0 as long as $\epsilon > 0$, while for $\epsilon = 0$ the average reward equals 1 for starting state 1.

Average reward Stochastic Games were introduced by Gillette [14]. He considered games with perfect information (in each state one of the players has only one action available) and irreducible Stochastic Games (games where for each pair of stationary pure strategies (ρ,σ) the associated stochastic matrix $P(\rho,\sigma)$ (cf. (2)) has a single ergodic class and no transient states). Blackwell and Ferguson [6] used a slightly modified version of an example of Gillette to show for average reward Stochastic Games that, in general, the players need not possess optimal strategies. Even more for this example, called the big match, one of the players has no ϵ-optimal strategy within the class of (semi)-Markov strategies for $\epsilon>0$ small enough.

For a long time it was an open question whether all average reward Stochastic Games have a value. Only about 1980 was this question answered in the affirmative by Mertens and Neyman [20]. Before that time, results for special cases of average reward Stochastic Games had been obtained by several authors. The emphasis was mainly on the existence of optimal stationary strategies for the players. Hoffman and Karp [15] treated irreducible Stochastic Games. Their approach is based on results of Markov decision theory. Kohlberg [18] analyzed so-called 'repeated games with absorbing states'. These are games where all but one of the states are absorbing and where the remaining state is transient or recurrent, depending on the strategies played. The big match belongs to this class of games. Kohlberg showed that these games have a value which can be found by considering the n-step game and letting n tend to infinity. Later it appeared that Kohlberg's approach indicated the way in which in the general case the existence of the value can be shown. Bewley and Kohlberg [2,3] demonstrated in an elegant way some of the relationships between the discounted game, the n-step game and the average reward game. Below we shall explain their use of the field of real Puiseux series.

Stationary strategies are the best manageable type of strategies. Therefore, it is natural to characterize the class of games for which both players possess (ϵ)-optimal stationary strategies. A characterization of Stochastic Games with optimal stationary strategies for both players was given in Vrieze [37]. Further, in Tijs and Vrieze [33] it was shown that for both players there are always states which are easy to them, i.e. when the game starts in such a state then the respective player can guarantee the value of the game by playing an appropriate stationary strategy.

In Filar et al. [13] another interesting question is settled, namely they give an algorithm which yields the (ϵ)-best stationary strategies among the stationary strategies with respect to the average reward criterion, even in the case when there are no (ϵ)-optimal stationary strategies.

Some of the above mentioned results will be worked out now. We start with the introduction of Puiseux series. Let for a positive integer M: $F_M := \{\Sigma_{k=0}^{\infty} c(k)(1-\beta)^{k/M};\quad c(k)\in\mathbb{R}$ and such that the series $\Sigma_{k=0}^{\infty} c(k)(1-\beta)^{k/M}$ converges for all β sufficiently close to $1\}$. Thus, the members of F_M are power series in $(1-\beta)^{1/M}$. Let $F := \bigcup_{M=1}^{\infty} F_M$, then it can be checked that F is an ordered field and F is called the field of real

Puiseux series. Bewley and Kohlberg [2] extended Shapley's equations (cf. (10)) for discounted reward Stochastic Games in the following way:

The set of equations (cf. (9) and (10)):

$$x_s = \mathrm{Val}(M_s^\beta(x)) \text{ for each } s \in S, \tag{11}$$

where $x_s \in F$ and $x = (x_1, x_2, \ldots, x_z) \in F^z$, and where β is interpreted as a variable, is called the limit discount equation. That for $x \in F^z$ the right-hand side of (11) belongs to F is a consequence of the ordered field preserving property of the value function (Weyl [43]). Notice that (11) is a set of equations in the function space F^z and that for a fixed $\beta \in \mathbb{R}$, (11) is equivalent to (10). Bewley and Kohlberg [2,3] proved the following:

THEOREM 4.1.

(a) *The set of equations (11) has a unique solution in F^z, say*

$$x_s^* = \sum_{k=0}^{\infty} c_s(k)(1-\beta)^{k/M}, \; s = 1,2,\ldots,z.$$

(b) $\sum_{k=0}^{\infty} c_s(k)(1-\beta)^{k/M}$, $s = 1,2,\ldots,z$, *is the value of the β-discounted reward Stochastic Game for all β sufficiently close to 1.*

(c) $c_s(0) = \lim_{\beta \to 1} \sum_{k=0}^{\infty} c_s(k)(1-\beta)^{k/M} = \lim_{n \to \infty} n^{-1} FV_s(n)$. *Here $FV_s(n)$ is the total reward value of the finite horizon Stochastic Game with n decision moments and starting state s.*

(d) *If player 1 has, for each $s \in S$, a real action $\rho(s)$ such that $\rho(s)$ guarantees player 1 $\mathrm{Val}(M_s^\beta(x^*)) + O(1-\beta)$ in the matrix game $M_s^\beta(x^*)$, then $\rho = (\rho_1, \rho_2, \ldots, \rho_z)$ is an optimal stationary strategy in the average reward game. An analogous statement holds for player 2.*

Translated in terms of discounted reward Stochastic Games, part (c) of Theorem 4.1 states that $\lim_{\beta \to 1} v_s^\beta$ exists for each $s \in S$ and that this limit equals the limit of the average reward values for finite horizon games with the same starting state. Later on, Mertens and Neyman [20] showed that these limits also equal the average reward value for the infinite horizon game. From part (d) one deduces immediately that stationary strategies which are uniformly discount optimal are also average reward optimal. (A strategy is uniform discount optimal if it is optimal for each discount factor β close enough to 1). The following theorem is due to Mertens and Neyman [20]:

THEOREM 4.2.

(a) *Average reward Stochastic Games have a value.*

(b) *ϵ-optimal strategies can be constructed from the solutions to the limit discount equation by computing for each stage a discount factor with the aid of rules depending on the history of the game up to that stage. Next the action at that stage with that history can be chosen as an optimal action in Shapley's equation for the computed discount factor.*

The proof of this theorem is based on the results of Bewley and Kohlberg, using a martingale property.

Certainly, history dependent strategies are terrible to handle. Application of stationary strategies is more easy, since players only have to look at the current state. Below we give a characterization of Stochastic Games for which both players have optimal stationary strategies (Vrieze [37]).

THEOREM 4.3. *Both players possess optimal stationary strategies if and only if the following set of equations has a solution* $g, v(1), v(2) \in \mathbb{R}^z$:

$$g_s = \mathrm{Val}_{A_s \times B_s}\left(\sum_{t=1}^{z} p(t \mid s,.,.) g_t\right), \text{ each } s \in S \tag{12}$$

$$v_s(1) = \mathrm{Val}_{E_s(1) \times B_s}\left(r_s(.,.) + \sum_{t=1}^{z} p(t \mid s,.,.) v_t(1)\right), \text{ each } s \in S \tag{13}$$

$$v_s(2) = \mathrm{Val}_{A_s \times E_s}(2)\left(r_s(.,.) + \sum_{t=1}^{z} p(t \mid s,.,.) v_t(2)\right), \text{ each } s \in S. \tag{14}$$

(Here $\mathrm{Val}_{C \times D}(f(.,.))$, *means the value of the matrix game on the polytope with extreme points the sets* C *and* D *for player 1 and player 2 respectively; for* $(c,d) \in C \times D$ *the payoff equals* $f(c,d)$; A_s *and* B_s *have the usual meaning and* $E_s(k)$, $k = 1,2$ *are the sets of extreme optimal actions for player* k *in the matrix game (12).)*

For each solution to (12)—(14), $g = (g_1, g_2, \ldots, g_z)$, is the same, equalling the average reward value of the Stochastic Game. Equation (12) can be interpreted as a conservative property in the sense that the players should take care that they remain in their 'good states' during the play. Equations (13) for player 1 and (14) for player 2 reflect the equalizing property in the sense that within their good states the average rewards have to approach the value. From a solution to (12)—(14) optimal stationary strategies can be constructed. Namely, let $\rho = (\rho_1, \rho_2, \ldots, \rho_z)$ be such that, for each $s \in S$, ρ_s is an optimal action for player 1 in matrix game (13). Then ρ is optimal. Likewise one can construct an optimal stationary strategy for player 2 from (14).

We already saw that optimal stationary strategies need not exist. One can wonder if for certain starting states one or both players can guarantee themselves the value for that starting state with the aid of stationary strategies. In Tijs and Vrieze [33] it is shown that both players, in every game, have at least one state for which this is the case. It is still an open problem to characterize for a player his whole set of such 'easy' states.

We finish this section with some remarks on games for which the value does not depend on the initial state. Bewley and Kohlberg [3] already showed that for games for which $\lim_{\beta \to 1} v_s^\beta$ is the same for each $s \in S$, the value of the game exists and that both players possess optimal Markov strategies. The following theorem can be found in Vrieze [37].

THEOREM 4.4.

(a) *For an average reward Stochastic Game the value is independent of the initial state if and only if, for some number* $g \in \mathbb{R}$, *for each* $\epsilon > 0$, *the following set of equations has a solution for* $v(\epsilon) = (v_1(\epsilon), v_2(\epsilon), \ldots, v_z(\epsilon))$:

$$|v_s(\epsilon) + g - \mathrm{Val}_{A_s \times B_s}(r_s(.,.) + \sum_{t=1}^{z} p(t \mid s,.,.)v_t(\epsilon))| \leqslant \epsilon, \tag{15}$$

for each $s \in S$.

(b) *Both players have optimal stationary strategies in an average reward Stochastic Game with value independent of the initial state if and only if equation (15) has a solution for* $\epsilon = 0$ *i.e. if and only if*

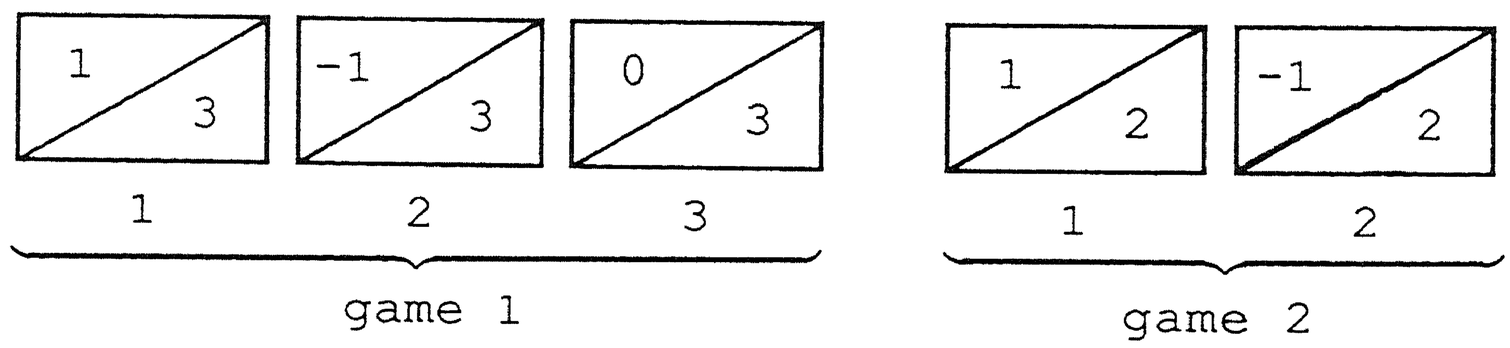
$$v_s + g = \mathrm{Val}_{A_s \times B_s}(r_s(.,.) + \sum_{t=1}^{z} p(t \mid s,.,.)v_t) \tag{16}$$

for some $g \in \mathbb{R}$ *and* $v \in \mathbb{R}^z$.

In part (a) as well as in part (b) of Theorem 4.4 the value of the game is g for each starting state. In part (a) ϵ-optimal stationary strategies can be constructed by taking optimal actions in the matrix games in (15). In part (b) optimal stationary strategies result by taking optimal actions in the matrix games in (16).

5. TOTAL REWARD STOCHASTIC GAMES

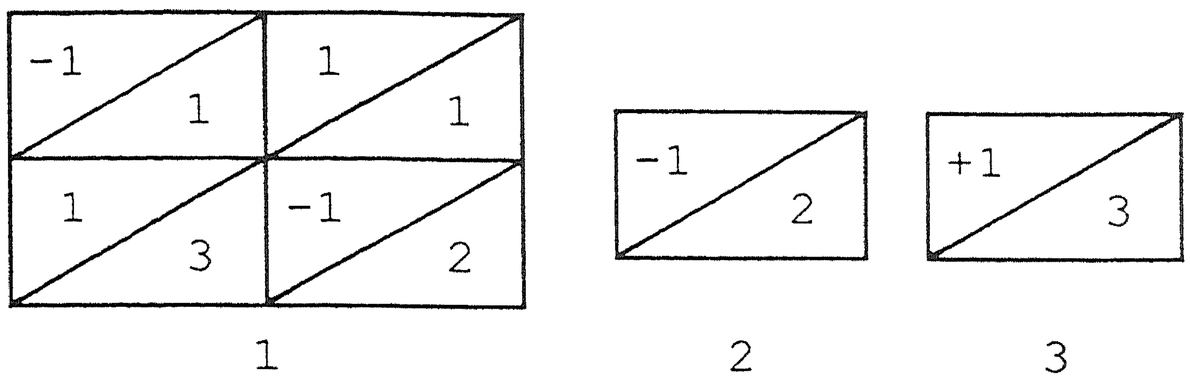
In Section 3 we already mentioned that in fact Shapley considered total reward Stochastic Games under the restriction of stopping transitions. In this section we apply the total reward criterion to Stochastic Games as defined in Section 2. The motivations for looking at the total reward criterion lies in the fact that this can be seen as a sensitive criterion in addition to the average reward criterion. For instance, consider the following examples:



For game 1, obviously the average reward value is (0,0,0). However player 1 would prefer to start in state 1 (getting total reward 1) and player 2 would prefer to start in state 2 (paying total reward -1). Likewise in game 2 the average reward value is (0,0) but player 1 likes to start in state 1, thus owning half of the time one unit and half of the time zero units. And player 2 likes to start in state 2, being due half of the time minus one unit and half of the time 0 units.

For both games the average reward criterion does not discriminate between

the states for the players, while the total reward criterion would do. In general, total reward Stochastic Games need not have a value, as can be seen from the following example:



It can easily be verified that for state 1:

$$\sup_{\pi_1}\inf_{\pi_2} v^T(1,\pi_1,\pi_2) = -\infty \neq 0 = \inf_{\pi_2}\sup_{\pi_1} v^T(1,\pi_1,\pi_2).$$

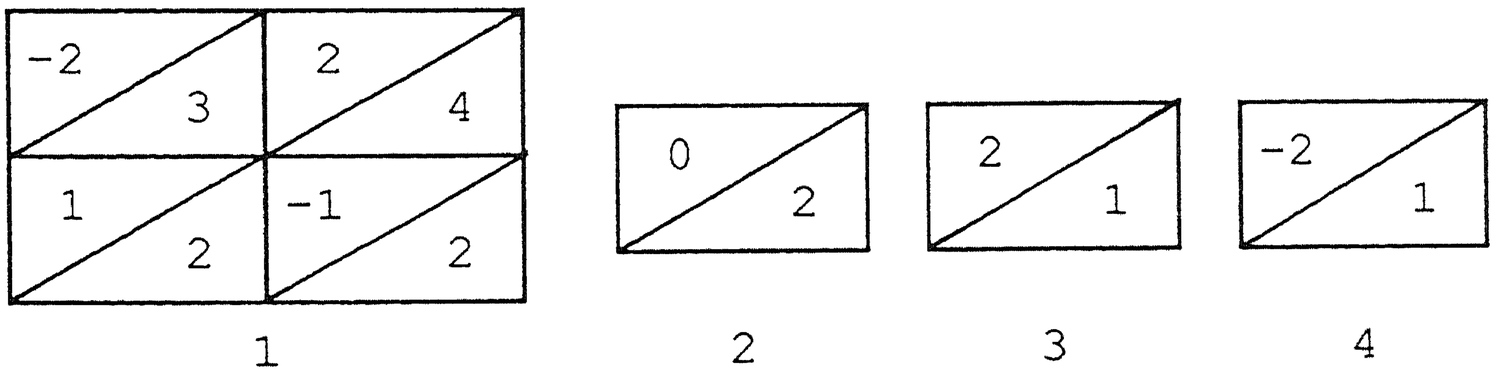
Researchers in Stochastic Games will have recognized the above game as being the big match of Blackwell and Ferguson [6].

It can be proved that in case one of the players has no optimal stationary strategy for an average reward Stochastic Game, then such a game has no total reward value. Therefore, we concentrate on games with the following property.

PROPERTY P1. *The Stochastic Game has average reward value (0,0,...,0) and both players possess optimal stationary strategies with respect to the average reward criterion.*

This class of games is introduced in Thuijsman and Vrieze [31]. It is still an open question whether for this class of games the total reward value always exists. It can easily be shown that property P1 implies that both $\sup_{\pi_1}\inf_{\pi_2} v^T(s,\pi_1,\pi_2)$ and $\sup_{\pi_2}\inf_{\pi_1} v^T(s,\pi_1,\pi_2)$ are finite.

The following example, called the bad match, elaborated in Thuijsman and Vrieze [31] shows that, in analyzing total reward games, similar problems as for average reward games are encountered.



For this game the total reward value equals $(0,0,2,-2)$. Player 1 has no total reward optimal strategy, in fact player 1 has no ϵ-optimal (semi-)Markov strategy for the game starting in state 1 (or state 3 or state 4).

Observe the similarity between this game (the bad match) with the big match. In Thuijsman and Vrieze [31] it is shown that the total reward ϵ-optimal history dependent strategies for player 1 in the bad match can be constructed along the same lines as average reward ones in the big match.

In spite of the similarity, the analysis of Mertens and Neyman [20] cannot straightforwardly be implemented to total reward Stochastic Games, which is mainly due to the fact that even under Property P1 streams of payoffs may occur for which the partial sums are not uniformly bounded.

Concerning the characterization of games with both players possessing total reward optimal stationary strategies, the following result can be mentioned (Vrieze and Thuijsman [38]):

THEOREM 5.1. *For a total reward Stochastic Game the value exists and both players have optimal stationary strategies if and only if the following set of functional equations have a solution: (variables:* $u = (u_1,u_2,...,u_z)$, $w(1) = (w_1(1), w_2(1),...,w_z(1)), w(2) = (w_1(2), w_2(2),...,w_z(2))$ *and* $\alpha \geqslant 0$*)*

$$u_s = \mathrm{Val}_{A_s \times B_s}(r_s(.,.) + \sum_{t=1}^{z} p(t \mid s,.,.)u_t), \text{ for each } s \in S \tag{17}$$

$$w_s(1) + u_s = \mathrm{Val}_{E_s(1) \times B_s}(\alpha r_s(.,.) + \sum_{t=1}^{z} p(t \mid s,.,.)w_t(1)),$$

$$\text{for each } s \in S \tag{18}$$

$$w_s(2) + u_s = \mathrm{Val}_{A_s \times E_s(2)}(\alpha r_s(.,.) + \sum_{t=1}^{z} p(t \mid s,.,.)w_t(2)),$$

$$\text{for each } s \in S \tag{19}$$

(Here $\mathrm{Val}_{C \times D}(f(...))$ *has the same meaning as in Theorem 4.3.)*

Observe the similarity of this theorem with Theorem 4.3. Analogously to the average reward case, the u part of any solution to (17) — (19) is the same, being the total reward value. Also here optimal stationary strategies can be constructed from optimal actions in the polyhedral games (18) and (19). Notice further that equation (17) is equivalent to Property P1 (cf. part (b) of Theorem 4.4). In case both players have total reward optimal stationary strategies it can be deduced from the limit discount equation that the total reward value equals $c(M) = (c_1(M),c_2(M),...,c_z(M))$, i.e. the coefficient of the factor $(1-\beta)$ in the Puiseux series solution to the limit discount equation. In this case $c(M)$ is also the leading term since under property P1 it holds that $c(0) = c(1) = ... = c(M-1) = 0$. In terms of the discount factor this property can

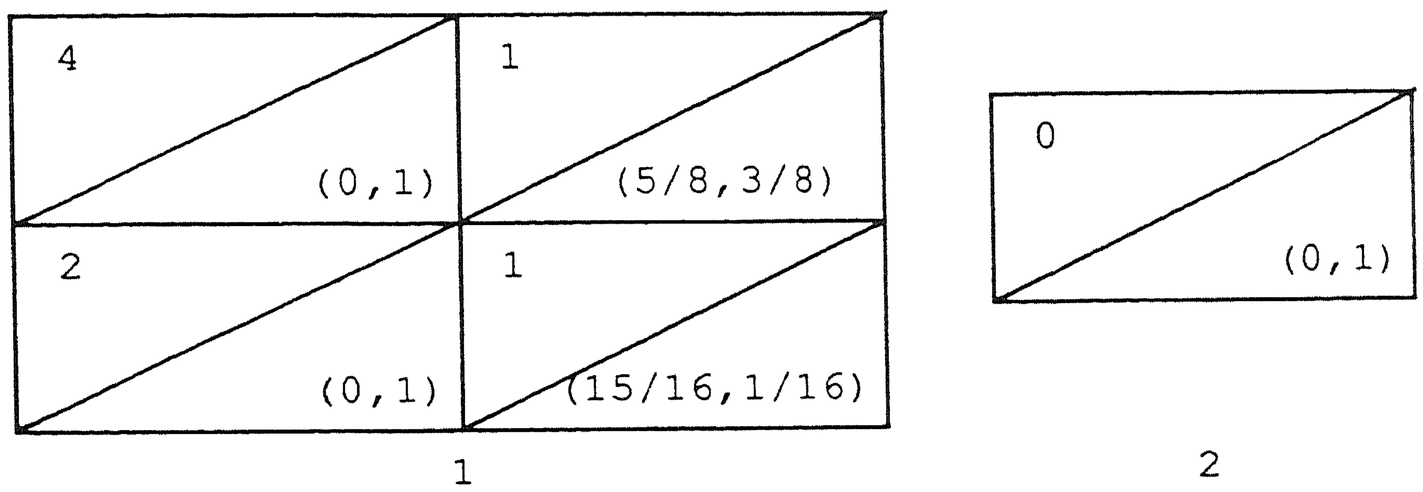
be stated as: total reward value $= \lim_{\beta\to 1}(1-\beta)^{-1}v^{\beta}$. In Vrieze and Thuijsman [38] it is conjectured that for games with the property P1 the total reward value always exists and equals $\lim_{\beta\to 1}(1-\beta)^{-1}v^{\beta}$.

6. Structured stochastic games

Since 1980 several subclasses of Stochastic Games have been studied. Here subclasses of Stochastic Games are classes of Stochastic Games determined by conditions and restrictions on the reward and/or transition functions.

There are two reasons for analyzing structured games. First, practical situations often lead to Stochastic Games with certain restrictive properties of the data of the game. Second, the value and optimal strategies can be computed more easily for structured games than in the general case.

In the general case neither the value nor optimal stationary strategies need to lie in the same ordered field as the data of the game. For instance the game with rational data:



has, for $\beta = 4/5$, discounted reward value $(\sqrt{8},0)$.

Stochastic games, for which the value and some pair of optimal stationary strategies, lie in the same ordered field as the data of the game, are said to have the orderfield property. For a game with the orderfield property one can expect to find a solution in a finite number of computation steps, resulting in an exact solution of the game. This observation gives further support to paying attention to structured games.

Successively we will mention the classes studied so far and give in short some characteristics. We do not state properties with respect to the total reward criterion, since this criterion has only recently been proposed in the literature. However, most of the properties concerning the average reward criterion will also hold for the total reward criterion, when, in addition to the structure on the rewards and transitions, Property P1 (cf. Section 5) is assumed to hold.

(a) One player controls transition.
In this class, only one of the players controls the transitions. Say player 2, then translation of such a property to the data gives: $p(t\,|\,s,i_1,j) = p(t\,|\,s,i_2,j)$ for each $i_1,i_2\in A_s$, each $j\in B_s$ and each $s,t\in S$. Hence we can

denote the transitions by $p(t|s,j)$. The orderfield property for this class of games for as well the discounted as the average reward criterion, was first shown by Parthasarathy and Raghavan [22]. For the discounted reward case they gave an LP algorithm. In Vrieze [36], and independently in Hordijk and Kallenberg [16], a constructive proof, using also an LP algorithm, for the average reward case can be found. Later on, this class of games was intensively studied by Filar [11,12] especially as an application to the travelling inspector model.

(b) Transitions with switching control.
For this class of games in each state only one of the players governs the transitions. However, unlike the 'one player controls transition' case, not in every state does this have to be the same player. This class of games was introduced by Filar [10]. For both the discounted and the average reward case he proved the orderfield property. A constructive proof for the discounted version can be found in Vrieze [37] and for the average version in Vrieze et al. [42]. In both cases the solution procedure consists of an iterative procedure of finite length, where at each iteration a suitable LP problem has to be solved.

(c) Separable reward and state independent transitions games are defined by the following structure: $r_s(i,j) = r_1(s)+r_2(i,j)$ and $p(t|s_1,i,j) = p(t|s_2,i,j)$ for each $s_1,s_2 \in S$. Hence we may write $p(t|i,j)$. As a consequence of the structure imposed the action sets of the players are the same for each state. These games are introduced in Parthasarathy et al. [23]. They showed that SER-SIT games have the orderfield property and that this class can be solved by solving a matrix game.
For the discounted version this matrix game is

$$[r_2(i,j)-\beta\sum_{t=1}^{z}p(t|i,j)r_1(t)]_{i=1}^{m}{}_{j=1}^{n}.$$

For the average version this matrix game is the limit of the β-discounted one's for β tending to one:

$$[r_2(i,j)-\sum_{t=1}^{z}p(t|i,j)r_2(t)]_{i=1}^{m}{}_{j=1}^{n}.$$

For both criteria both players have optimal myopic stationary strategies. Myopic means that the stationary strategy is even independent of the current state. A further result is that for the average case the value is independent of the initial state. SER-SIT games are also partially studied by Sobel [29].

(d) ARAT games.
Additive reward and additive transition games are introduced by Raghavan et al. [24]. ARAT games are defined by $r_s(i,j) = r_{1s}(i)+r_{2s}(j)$ and $p(t|s,i,j) = p_1(t|s,i)+p_2(t|s,j)$. Hence both the rewards and the transitions are additive with respect to both players. They proved the following results. For both the discounted and the average reward criterion both

players possess optimal stationary pure strategies. This property immediately implies the orderfield property. Furthermore, both players have uniformly discount optimal stationary pure strategies. These results follow straightforwardly from Shapley's equations (10), since the matrix game (9) can be decomposed into a term depending on i and a term depending on j for ARAT games.

(e) One player controls rewards for a game with two states. This class of games is introduced by Vrieze et al. [41]. It is defined by restriction to two states and $r_s(i,j) = \tilde{r}_s(i)$, i.e. the rewards only depend on the action of player 1. Also for this class of games the orderfield property turns out to hold for the discounted case. For the average reward criterion this is an open question.

We conclude this section with the remark that the time has come to characterize the subclass of games having the orderfield property. Two approaches look promising for the discounted reward criterion at least. One is established in the paper by Vrieze et al. [41]. To each set of stationary pure strategies of a player they add a set of stationary strategies of the other player in the following way:

Let Q be a set of stationary pure strategies for, say, player 1 and let $S(Q)$ be the stationary strategies of player 2 added to Q, then $\sigma \in S(Q)$ if and only if for each $\rho \in Q, \sigma$ is a best answer to ρ.

Vrieze et al. [41] showed that, in their case, $S(Q)$ is either void or a union of a finite number of disjoint polytopes with rational extremes (when the data is rational). In general this quality is sufficient for proving the orderfield property. And as such this idea can be used for characterizing the orderfield property in more generally settings.

A second approach can be found in Sinha [28]. He combined SER-SIT games and switching control games. For the discounted case he exploited a value iteration method based on Shapley's equations. In each step three connected LP problems have to be solved. In a finite number of steps the solution of the discounted reward game corresponds to an extreme point of a suitably chosen system of linear inequalities. In each iteration step this system returns together with some base (in LP terminology) to this system. Since each of these iterations approaches the value more closely and since there are a finite number of different bases, Sinha was able to prove that this procedure stops after a finite number of steps. There are indications that this method can be extended to a method that can be used in proving generally whether some subclass has the orderfield property (in the discounted case) or not.

7. Algorithms for stochastic games

In this final section we give a short review on algorithms for discounted and average reward Stochastic Games. The question is always, how to compute the value of the game together with a pair of stationary strategies (when existing). For solution methods for special subclasses we refer to Section 6.

7.1. Algorithms for discounted reward Stochastic Games

In the first place we mention the algorithm, which, in a natural way, arises from Shapley's proof of the existence of the value, namely (cf. (9) and (10)):

1. choose $v_0 = (v_0(1),...,v_0(z))$ arbitrarily,
2. let, for $\tau = 1,2,...,$: $v_\tau(s) := \mathrm{Val}(M_s^\beta(v_{\tau-1}))$, each $s \in S$.

This value iteration method approaches the value of the game exponentially fast, while at each iteration suboptimal stationary strategies can be deduced from the matrix $M_s^\beta(v_{\tau-1})$.

A second algorithm, proposed by Hoffman and Karp [15], can be called value oriented policy iteration. It runs as follows:

1. choose $v_0 = (v_0(1),v_0(2),...,v_0(z))$ arbitrarily,
2. let, for $\tau = 0,1,2,...,\sigma^\tau = (\sigma^\tau(1),\sigma^\tau(2),...,\sigma^\tau(z))$ be such that $\sigma^\tau(s)$ is an optimal action for player 2 in $M_s^\beta(v_\tau)$,
3. solve for player 1 the Markov decision problem which results when player 2 fixes σ^τ and let $v_{\tau+1}$ be the optimal value for this Markov decision problem; repeat from 2.

A third algorithm we will mention, is an extension of the Brown-Robinson scheme for matrix games to Stochastic Games (Vrieze and Tijs [40]). First they showed that the scheme can be applied to a converging sequence of matrix games. Next the contraction property of Shapley's value operator enables them to prove the convergence of the Brown-Robinson scheme when applied to discounted Stochastic Games.

The convergence rate is low (the same as in the case of matrix games), however at each iteration only simple calculations have to be done.

More about algorithms, especially viewed in a mathematical programming context (cf. also Vrieze [37]), can be found in Schultz [25].

Several facts about convergence rates for successive approximation schemes and value oriented policy iteration schemes can be found in Van der Wal [34].

7.2. Algorithms for average reward Stochastic Games

For average reward Stochastic Games there are still many open problems. The existing algorithms only solve special classes. Surely, by the result of Bewley and Kohlberg [2] (cf. Section 4), the average value, g, can be approached by computing the discounted value, v^β, and letting β tend to 1 ($g = \lim_{\beta \to 1} v^\beta$).

However there are no clear rules available for estimating the convergence rate. A further difficulty is that the players need not possess optimal stationary strategies.

We mention two algorithms.

The first one is the application of the Hoffman and Karp scheme to the class of irreducible Stochastic Games, i.e. games for which for each pair of stationary pure strategies the corresponding stochastic matrix (cf. (2)) has a single ergodic class and no transient states. They showed that their scheme converges to a solution of the following set of equations (in $g \in \mathbb{R}$ and $v = (v_1,...,v_z) \in \mathbb{R}^z$).

$$g + v_s = \mathrm{Val}_{A_s \times B_s}(r_s(.,.) + \sum_{t=1}^{z} p(t \mid s,.,.)v_t), \text{ for each } s \in S. \tag{20}$$

By Theorem 4.4, part (b) it follows at once that a solution to these equations is equivalent to a solution of the average reward Stochastic Game.

A second algorithm is due to Federgruen [9] and can be applied to games for which: (1) both players have optimal stationary strategies and (2a) the average value is independent of the initial states or (2b) the stochastic game is irreducible.

Federgruen's scheme is an extension of the modified value-iteration method of Hordijk and Tijms [17] for Markov decision problems to stochastic games. The idea is to choose a suitable sequence of discount factors tending to 1, obeying certain desired properties. Next at each step a (discounted) value iteration step is carried out, resulting in a scheme which converges to a solution of the set of equations (20). Related algorithms for the same classes can be found in Van der Wal [34].

Since playing stationarily is preferable to playing non-stationarily it is desirable to have an algorithm yielding $\sup_\rho \inf_{\pi_2} v(\rho,\pi_2)$ for player 1 and $\inf_\sigma \sup_{\pi_1} v(\pi_1,\sigma)$ for player 2.

Observe that these quantities are the bounds for the respective players of the average rewards that can be reached by playing stationarily. Only in the case both players possess ϵ-optimal stationary strategies are these bounds the same, equalling the average value of the game. Recently, in Filar et al. [13] a mathematical programming formulation is given yielding a solution to this problem: (variables $v^1=(v_1^1,\ldots,v_z^1)$, $v^2=(v_1^2,\ldots,v_z^2)$, $u^1=(u_1^1,\ldots,u_z^1)$, $u^2=(u_1^2,\ldots,u_z^2)$, $\rho=(\rho_1,\ldots,\rho_z)$ and $\sigma=(\sigma_1,\ldots,\sigma_z)$)

$$\inf \sum_{s=1}^{z} (v_s^1 + v_s^2)$$

subject to:

(a) $v_s^1 \geqslant \max_{i\in A_s} \{\Sigma_{t=1}^z p(t|s,i,\sigma_s) v_t^1\}$

(b) $v_s^1 + u_s^1 \geqslant \max_{i\in A_s} \{r_s(i,\sigma_s) + \Sigma_{t=1}^z p(t|s,i,\sigma_s) u_t^1\}$

(c) $v_s^2 \geqslant \max_{j\in B_s} \{\Sigma_{t=1}^z p(t|s,\rho_s,j) v_t^2\}$

(d) $v_s^2 + u_s^2 \geqslant \max_{j\in B_s} \{r_s(\rho_s,j) + \Sigma_{t=1}^z p(t|s,\rho_s,j) u_t^2\}$

All inequalities should hold for each $s \in S$.

References

1. R. Bellmann (1957). *Dynamic Programming,* Princeton University Press, Princeton.
2. T. Bewley, E. Kohlberg (1976). The asymptotic theory of Stochastic Game. *Math. of O.R. 1,* 197-208.
3. T. Bewley, E. Kohlberg (1978). On Stochastic Games with stationary optimal strategies. *Math. of O.R. 3,* 104-125.
4. D. Blackwell (1962). Discrete dynamic programming. *Ann. of Math. Stat. 36,* 719-726.
5. D. Blackwell (1965). Discounted dynamic programming. *Ann. of Math. Stat. 36,* 226-235.

6. D. BLACKWELL, T. FERGUSON (1968). The big match. *Ann. of Math. Stat. 39,* 159-163.
7. M.F. BOHNENBLUST, S. KARLIN, L.S. SHAPLEY (1950). Solutions of discrete two-person games. H.W. KUHN, A.W. TUCKER (eds.). Contributions to the theory of games, vol. I. *Ann. of Math. Studies 24,* 51-72, Princeton University Press, Princeton.
8. E.V. DENARDO (1982). *Dynamic Programming, Models and Applications,* Prentice Hall.
9. A. FEDERGRUEN (1984). *Markovian Control Problems; Functional Equations and Algorithms,* MC Tract 97, Centre for Mathematics and Computer Science, Amsterdam.
10. J.A. FILAR (1981). Ordered field property for Stochastic Games when the player who controls transitions changes from state to state. *JOTA 34,* 503-515.
11. J.A. FILAR (1984). Player aggregation in the travelling inspector model. *IEEE Transactions on Automatic Control, AC-30,* 723-729.
12. J.A. FILAR (1987). Quadratic programming and the single controller Stochastic Game. *Journal of Math. Analysis and applications,* to appear.
13. J.A. FILAR, T.A. SCHULTZ, F. THUIJSMAN, O.J. VRIEZE (1987). *Nonlinear Programming and Stationary Equilibria in Stochastic Games,* Techn., Rep. 87-18, UMBC.
14. D. GILLETTE (1957). Stochastic games with zero-stop probabilities. M. DRESHER, A.W. TUCKER, P. WOLFE (eds.). Contributions to the theory of games, vol. III. *Ann. of Math. Stud. 39,* 179-188, Princeton University Press, Princeton.
15. A. HOFFMAN, R. KARP (1966). On nonterminating Stochastic Games. *Man. Science 12,* 359-370.
16. A. HORDIJK, L. KALLENBERG (1981). Linear programming and Markov games II. O. MOESCHLIN, D. PALLASCHKE (eds.). *Game Theory and Math. Ec.,* 307-320, North-Holland, Amsterdam.
17. A. HORDIJK, H.C. TIJMS (1975). A modified form of the iterative method of dynamic programming. *Ann. of Stat. 3,* 203-208.
18. E. KOHLBERG (1974). Repeated games with absorbing states. *Ann. of Stat. 2,* 724-738.
19. H. KUHN (1953). Extensive games and the problem of information. H.W. KUHN, A.W. TUCKER (eds.). Contributions to the theory of games, vol. III. *Ann. of Math. Stud. 28,* 193-216, Princeton University Press, Princeton.
20. J.F. MERTENS, A. NEYMAN (1981). Stochastic games. *Int. J. of Game Theory 10,* 53-66.
21. G. OWEN (1968). *Game Theory,* Saunders Comp. Philadelphia.
22. T. PARTHASARATHY, T.E.S. RAGHAVAN (1981). An orderfield property for Stochastic Games when one player controls transition probabilities. *JOTA 33,* 375-392.
23. T. PARTHASARATHY, S.H. TIJS, O.J. VRIEZE (1984). Stochastic games with state independent transitions and separable rewards. G. HAMMER, D.

PALLASCHKE (eds.). *Selected Topics in Op. Res. and Math. Ec.*, 226-236, Springer-Verlag.
24. T.E.S. RAGHAVAN, S.H. TIJS, O.J. VRIEZE (1985). On Stochastic Games with additive rewards and transition structure. *JOTA 47,* 451-464.
25. T.A. SCHULTZ (1987). *Mathematical Programming and Stochastic Games,* Ph.D. Dissertation, Baltimore, Maryland.
26. L.S. SHAPLEY (1953). Stochastic games. *Proc. Nat. Acad. Sci. USA 39,* 1095-1100.
27. L.S. SHAPLEY, R.N. SNOW (1950). Basic solutions of discrete games. H.W. KUHN, A.W. TUCKER (eds.). Contributions to the theory of games, vol. I. *Ann. of Math. Stud. 24,* 27-35, Princeton University Press, Princeton.
28. S. SINHA (1986). *An Extension Theorem for the Class of Stochastic Games having Orderfield Property,* Report at the the Stat. Math. Division, Indian Statistical Institute, New Delhi.
29. M.J. SOBEL (1981). Myopic solutions of Markov decision processes and Stochastic Games. *Op. Res. 29,* 995-1009.
30. F. THUIJSMAN (1989). *Optimality and Equilibria in Stochastic Games,* Ph. D. Dissertation, Rijksuniversiteit Limburg, Maastricht.
31. F. THUIJSMAN, O.J. VRIEZE (1987). The bad match; a total reward Stochastic Game. *Op. Res. Spectrum 9,* 93-99.
32. S.H. TIJS, O.J. VRIEZE (1980). Perturbation theory for games in normal form and Stochastic Games. *JOTA 30,* 549-567.
33. S.H. TIJS, O.J. VRIEZE (1986). On the existence of easy initial states for undiscounted Stochastic Games. *Math. of O.R. 11,* 506-513.
34. J. VAN DER WAL (1981). *Stochastic Dynamic Programming,* MC Tract 139, Centre for Mathematics and Computer Science, Amsterdam.
35. J. VON NEUMANN (1928). Zur Theory der Gesellschaftsspiele. *Mathematiche Annalen 100,* 295-320.
36. O.J. VRIEZE (1981). Linear programming and undiscounted Stochastic Games in which one player controls transitions. *Op. Res. Spectrum 3,* 29-35.
37. O.J. VRIEZE (1987). *Stochastic Games with Finite State and Action Spaces,* CWI Tract 33, Centre for Mathematics and Computer Science, Amsterdam.
38. O.J. VRIEZE, F. THUIJSMAN (1986). Stochastic games and optimal stationary strategies, a survey. *Proceedings of the 11-th Symposium of Operations Research,* September 1986, Darmstadt.
39. O.J. VRIEZE, S.H. TIJS (1980). Relations between game parameters, value and optimal strategy spaces in Stochastic Games and construction of game with given solution. *JOTA 31,* 501-513.
40. O.J. VRIEZE, S.H. TIJS (1982). Fictitious play applied to sequences of games and discounted games. *Int. J. Game Theory 11,* 71-85.
41. O.J. VRIEZE, S.H. TIJS, T. PARTHASARATHY, C.A.J.M. DIRVEN (1986). A class of Stochastic Games with the ordered field property, submitted to *JOTA.*

42. O.J. VRIEZE, S.H. TIJS, T.E.S. RAGHAVAN, J.A. FILAR (1983). A finite algorithm for the switching control Stochastic Game. *Op. Res. Spectrum 5,* 15-24.
43. H. WEYL (1950). Elementary proof of a maximum theorem due to Von Neumann. H.W. KUHN, A.W. TUCKER (eds.). Contributions to the theory of games, vol. I. *Ann. of Math. Stud. 24,* Princeton University Press, Princeton.