JOURNAL OF OPTIMIZATION THEORY AND APPLICATIONS: Vol. 98, No. 1, pp. 175-196, JULY 1998

Total Reward Stochastic Games and Sensitive Average Reward Strategies

F. THUIJSMAN[1] AND O. J. VRIEZE[2]

Communicated by G. P. Papavassilopoulos

Abstract. In this paper, total reward stochastic games are surveyed. Total reward games are motivated as a refinement of average reward games. The total reward is defined as the limiting average of the partial sums of the stream of payoffs. It is shown that total reward games with finite state space are strategically equivalent to a class of average reward games with an infinite countable state space. The role of stationary strategies in total reward games is investigated in detail. Further, it is outlined that, for total reward games with average reward value 0 and where additionally both players possess average reward optimal stationary strategies, it holds that the total reward value exists.

Key Words. Stochastic games, total reward, average reward, value existence.

1. Introduction

In this paper, we consider two-person, zero-sum stochastic games. A stochastic game is a dynamical system that proceeds along an infinite countable number of decision times. In the two-player case, both players can influence the course of play by making choices out of well-defined action sets. Unless mentioned otherwise, we will assume throughout this paper that the system can only be in finitely many different states. The actions available to a player depend on the state of the system. When at a certain decision time the players, independently and simultaneously, have both made a choice, then two things happen:

(i) player II pays player I a state and action dependent amount;

[1] Associate Professor, Department of Mathematics, Maastricht University, Maastricht, Netherlands.

[2] Professor, Department of Mathematics, Maastricht University, Maastricht, Netherlands.

0022-3239/98/0700-0175$15.00/0 © 1998 Plenum Publishing Corporation

(ii) the system moves to the next decision time, and the state at that new decision time is determined by a chance experiment according to a probability measure determined by the present state and by the actions chosen by the players.

Thus, a stochastic game Γ is defined by $\langle S, A_1, A_2, r, p\rangle$, where:

(i) $S=\{1, 2, \dots, z\}$ is the state space;
(ii) $A_1=\{A_1(s)|s\in S\}$, with $A_1(s)=\{1, 2, \dots, m_1(s)\}$ the action set of player I in state s;
(iii) $A_2=\{A_2(s)|s\in S\}$, with $A_2(s)=\{1, 2, \dots, m_2(s)\}$ the action set of player II in state s;
(iv) r is a real-valued function on

$$\{(s, a_1, a_2)|s\in S, a_1\in A_1(s), a_2\in A_2(s)\};$$

(v) p is a probability vector-valued map on

$$\{(s, a_1, a_2)|s\in S, a_1\in A_1(s), a_2\in A_2(s)\},$$

i.e., $p(s, a_1, a_2)=(p(1|s, a_1, a_2), \dots, p(z|s, a_1, a_2))\in\mathbb{R}^z$, where $p(t|s, a_1, a_2)$ is the probability that the next state is t, when at state s the players choose a_1 and a_2, respectively.

The players are assumed to have complete information (i.e., they know S, A_1, A_2, r, p) as well as perfect recall (i.e., at any stage they know the history of play). They can use this information when playing the game. Plans of how to play the game, strategies, will be formally defined in Section 2. The specification of an initial state and a pair of strategies results in a stochastic process on the states and the actions and thus leads to an infinite countable stream of expected payoffs. In comparing the worth of strategies, such an infinite stream should be translated to one single number. Several evaluation rules have been studied in the literature.

In the initiating paper on stochastic games, Shapley (Ref. 1) introduced the discounted payoff criterion. Let π_i denote an arbitrary strategy for player i, $i=1, 2$, and let $r_\tau(s, \pi_1, \pi_2)$ denote the expected payoff to player I at decision time τ when the play starts in state s and when the players use π_1 and π_2. Now, the discounted payoff is defined as

$$\phi_\beta(s, \pi_1, \pi_2):=(1-\beta)\sum_{\tau=0}^{\infty}\beta^\tau r_\tau(s, \pi_1, \pi_2), \tag{1}$$

where $\beta\in(0, 1)$ is the discount factor and $1-\beta$ is a normalization factor. Since $r_\tau(s, \pi_1, \pi_2)$ is uniformly bounded, it easily follows that $\phi_\beta(s, \pi_1, \pi_2)$ always exists.

A second commonly used criterion is the average reward criterion, as introduced by Gillette (Ref. 2). This is defined as

$$\phi_a(s, \pi_1, \pi_2) = \liminf_{T\to\infty} (T+1)^{-1} \sum_{\tau=0}^{T} r_\tau(s, \pi_1, \pi_2). \tag{2}$$

Since the limit of the right-hand side of (2) does not need to exist, this criterion is usually introduced from the worst case viewpoint of player I. Observe from (2) that the average reward is the limit of the partial averages and thus can be considered as a Cesaro average payoff.

The third criterion which we like to mention and which will be studied extensively in this paper is the total reward criterion. This evaluation rule is formally defined as

$$\phi_t(s, \pi_1, \pi_2) := \liminf_{T\to\infty} (T+1)^{-1} \sum_{\tau=0}^{T} \sum_{n=0}^{\tau} r_n(s, \pi_1, \pi_2). \tag{3}$$

This criterion was introduced by Thuijsman and Vrieze (Ref. 3). Observe that the total reward can be interpreted as the Cesaro average of the partial sums of the stream of expected payoffs. Again, this limit does not need to exist and the worst case viewpoint of player I has been taken.

In Section 2, we discuss the main results in the theory of stochastic games. In Section 3, we motivate the total reward evaluation rule as a refinement with respect to the average reward criterion. In Section 4, we show that total reward games can be represented as average reward games at the expense of an infinite state space. In Section 5, for stationary strategies, we give several equivalent expressions with respect to the total reward as well as a complete characterization of games for which both players have optimal stationary strategies. Finally in Section 6, we show that, for games with average reward value 0 as well as with average reward optimal stationary strategies for both players, the total reward value exists. Since in general for guaranteeing nearly the total reward value the players need behavioral strategies, this result implies that, for games where the players have average reward optimal stationary strategies, they can play more sensitively by using behavioral strategies.

2. Peliminaries

In this section, we mention the most important results in the theory of stochastic games. First, we introduce the notions of strategies, solution of a game, and ϵ-optimal strategies. The most general type of strategy is a behavioral strategy. In stochastic games, it is assumed that, at every decision time, the players know not only the present state of the system but also the whole sequence of states and actions that have actually occurred in the past. Now, the randomized choice at decision time τ may depend on this known history

$$h_t := (s_0, a_{10}, a_{20}, s_1, a_{11}, a_{21}, \ldots, s_{\tau-1}, a_{1\tau-1}, a_{2\tau-2}, s_\tau).$$

Then, a behavioral strategy for player i, $i=1, 2$, can be defined as

$$\pi_i = \{\pi_i(0), \pi_i(1), \ldots\},$$

with

$$\pi_i(\tau): H_t \to \mathscr{P}(A_i(s_\tau)),$$

where H_τ is the set of possible histories up to decision time τ and $\mathscr{P}(A_i(s_\tau))$ is the set of randomized actions based on the pure action set $A_i(s_\tau)$, i.e.,

$$\mathscr{P}(A_i(s_\tau)) := \left\{ q \in \mathbb{R}^{m_i(s_\tau)} \,|\, q \geq 0, \sum_{j=1}^{m_i(s_\tau)} q_j = 1 \right\}.$$

A Markov strategy is a strategy that, with respect to the history of the game, only takes the current decision time into account. Formally, π_i is a Markov strategy if

$$\pi_i = (\pi_i(0), \pi_i(1), \ldots),$$

with

$$\pi_i(\tau) = (\pi_i(\tau, 1), \pi_i(\tau, 2), \ldots, \pi_i(\tau, z)),$$

$$\pi_i(\tau, s) \in \mathscr{P}(A_i(s)), \qquad \forall s \in S.$$

The most simple form of strategy is a stationary strategy, where the history up to the present state is neglected by the players. In this paper, we will denote a stationary strategy by f_i for player i and formally

$$f_i = (f_i(1), f_i(2), \ldots, f_i(z)),$$

with

$$f_i(s) \in \mathscr{P}(A_i(s)), \qquad \forall s \in S;$$

i.e., whenever the system is in state s, player i plays the randomized action $f_i(s)$ independent of the history of the game and independent of the decision time.

Now, we define a solution of the game. Let ϕ be either ϕ_β, or ϕ_a, or ϕ_t. The stochastic game is said to have a value

$$\phi^* = (\phi^*(1), \phi^*(2), \ldots, \phi^*(z)),$$

when, for all starting states $s \in S$;

$$\inf_{\pi_2} \sup_{\pi_1} \phi(s, \pi_1, \pi_2) = \phi^*(s) = \sup_{\pi_1} \inf_{\pi_2} \phi(s, \pi_1, \pi_2). \tag{4}$$

Observe that the left-hand side of (4) is the highest amount that player II would have to pay (by playing clever), while the right-hand side is the highest amount that player I can guarantee. For the evaluation rule ϕ, the strategy π_1^ϵ [π_2^ϵ] is called ϵ-optimal, with $\epsilon \geq 0$, for player I [II] if

$$\inf_{\pi_2} \phi(s, \pi_1^\epsilon, \pi_2) \geq \phi^*(s) - \epsilon, \qquad \text{for all } s \in S,$$

$$[\sup_{\pi_1} \phi(s, \pi_1, \pi_2^\epsilon) \leq \phi^*(s) + \epsilon, \qquad \text{for all } s \in S].$$

A 0-optimal strategy is called optimal. For discounted stochastic games, Shapley (Ref. 1) showed the existence of the value as well as the existence of optimal stationary strategies for both players. The discounted value ϕ_β^* is the unique solution to the following set of equations:

$$\phi_\beta^*(s) = \mathrm{Val}_{A_1(s) \times A_2(s)} \left[(1-\beta) r(s, \cdot, \cdot) + \beta \sum_{t=0}^{z} p(t|s, \cdot, \cdot) \phi_\beta^*(t) \right], \qquad s \in S. \tag{5}$$

In (5), the right-hand side denotes the value of the matrix game defined on the action sets $A_1(s)$ and $A_2(s)$ with payoffs

$$(1-\beta) r(s, a_1, a_2) + \beta \sum_{t=0}^{z} p(t|s, a_1, a_2) \phi_\beta^*(t), \qquad \text{for } (a_1, a_2) \in A_1(s) \times A_2(s).$$

For the discounted stochastic game, optimal stationary strategies $f_i^* = (f_i^*(1), \ldots, f_i^*(z))$ can be found by taking $f_i^*(s)$ optimal in (5). Bewley and Kohlberg (Ref. 4) extended Shapley's result in a very useful direction by showing that, for all β close to 1, ϕ_β^* can be expressed as a Puiseux series in $1-\beta$; i.e., there exist $M \in \{1, 2, \ldots\}$ and $c_0, c_1, c_2, \ldots \in \mathbb{R}^z$ such that

$$\phi_\beta^* = \sum_{k=0}^{\infty} c_k (1-\beta)^{k/M}. \tag{6}$$

For average reward stochastic games, which are also called undiscounted stochastic games, the existence proof of the value turned out to be more difficult. This is mainly due to the fact that, unlike the discounted reward, the average reward is not a continuous function of the strategies of the players. Mertens and Neyman (Ref. 5) showed the existence of the value of average reward stochastic games by providing a construction for ϵ-optimal behavioral strategies by choosing at every decision time τ an optimal action in a $\beta(h_\tau)$-discounted game, i.e., an action optimal in (5) for $\beta = \beta(h_\tau)$. In their procedure, as the notation $\beta(h_\tau)$ already indicates, the discount factor is being updated every decision time in dependence on the actual history.

In their proof, Mertens and Neyman used the result of Bewley and Kohlberg (Ref. 4) and showed that the vector c_0 in (6) is the average reward value ϕ_a^*. That in general the players do not possess ϵ-optimal stationary strategies for the average reward criterion was already known from a famous example by Blackwell and Ferguson (Ref. 6), called the big match; cf. Example 3.2 below.

For total reward stochastic games, not much is known. Thuijsman and Vrieze (Ref. 3) have shown that, in total reward stochastic games, one encounters similar problems as in average reward stochastic games. This aspect is briefly recalled in Example 3.4 in the next section.

3. Total Reward Stochastic Games and Sensitive Average Reward Strategies

Total reward stochastic games can be considered as refinements of average reward stochastic games. For an infinite stream of payoffs, the average is determined by the asymptotic behavior of this stream and ignores differences between streams of payoffs, whenever the averages are the same.

Example 3.1. See Fig. 1. This example shows the motivation for a refinement of the average reward criterion.

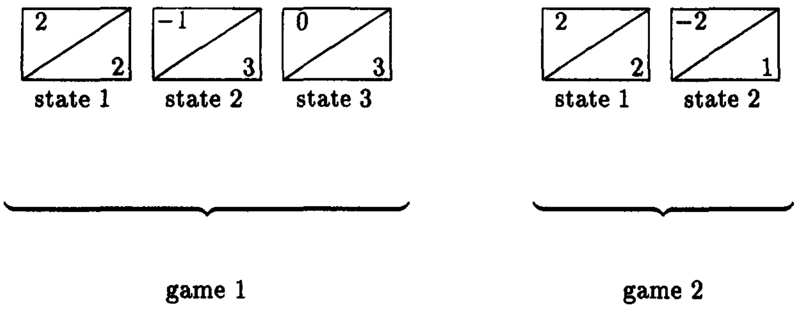
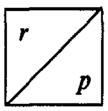


Fig. 1. Example 3.1: Two games to illustrate the definition of total rewards.

In the game representation, player I is always the row player and player II the column player. A box



denotes the immediate outcome of an action combination, i.e., payoff r to player I and payoff $-r$ to player II, and transitions to the next decision time according to p. When p is deterministic, i.e., the system moves to a certain state with probability 1, then usually this next state number is given in the right lower part of the box. When p is probabilistic, then this probability vector is given.

For game 1, the average reward value vector equals (0, 0, 0). However, player I would prefer to start in state 1 (getting total reward 1), while player II would prefer to start in state 2 (paying total reward -1, or equivalently, getting 1). Likewise for game 2, the average reward value vector equals (0, 0) and also in this game player I would like to start in state 1 (owning half of the time 2 and half of the time 0), while player 2 would like to start in state 2 (being indebted half of the time -2 and half of the time 0).

Example 3.1 shows that the total reward criterion can be interpreted as a refinement with respect to the average reward criterion, applied to games where, for every state, the average reward value is 0. But what about starting states with average reward value unequal to 0? Evidently, the total reward value for such a starting state exists, since playing an ϵ-optimal strategy with respect to the average reward assures as total reward $+\infty$ or $-\infty$, depending on the average reward value being positive or negative.

Example 3.2. See Fig. 2. This example, called the big match [cf. Blackwell and Ferguson (Ref. 6) for an average reward analysis], shows that, for states with average reward value 0, the total reward value may not exist if for other states the average reward value is not equal to 0. This game has average reward value vector $(0, 1, -1)$, while for the total rewards

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

Hence, the total reward value does not exist for state 1.

Example 3.2 suggests that, for the total reward criterion, it makes sense to restrict to games where the average reward value is 0 for every state. However, we need a further restriction.

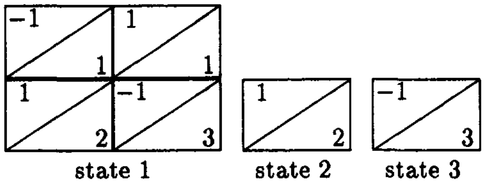


Fig. 2. Example 3.2: The big match.

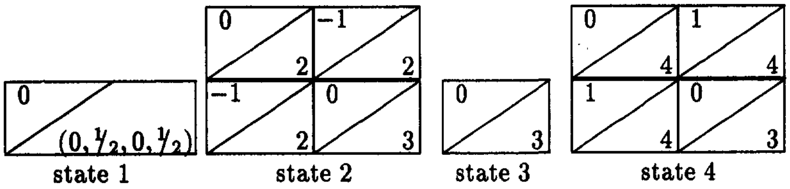


Fig. 3. Example 3.3: Although the average reward value is 0 for all states, the total reward value does not exist for state 1.

Example 3.3. See Fig. 3. In this example, the average reward value vector is (0, 0, 0, 0). However, for the total rewards,

$$-\infty = \sup_{\pi_1} \inf_{\pi_2} \phi_t(1, \pi_1, \pi_2) \neq \inf_{\pi_2} \sup_{\pi_1} \phi_t(1, \pi_1, \pi_2) = +\infty.$$

This can be seen as follows. Player I can play average reward optimal for initial states 3 and 4, but only ϵ-optimal for initial state 2. Thus, for any strategy of player I, an average reward δ-best reply by player II, $\delta > 0$, will yield an average reward of at most $-\epsilon + \delta$ for state 2 and at most δ for state 4. Hence, for initial state 1, the average reward is at most $-\epsilon/4$ for δ sufficiently small and therefore

$$\sup_{\pi_1} \inf_{\pi_2} \phi_t(1, \pi_1, \pi_2) = -\infty.$$

In view of these examples, we study the class of stochastic games characterized by property P1 below.

Property P1. The average reward value equals 0 for every initial state and both players possess optimal stationary strategies with respect to the average reward criterion.

Bewley and Kohlberg (Ref. 7) showed that property P1 implies property P2 below, and in Vrieze (Ref. 8) it can be found that P2 is equivalent to P1.

Property P2. The Puiseux series expansion of ϕ_β^* can be written as

$$\phi_\beta^* = \sum_{k=M}^{\infty} c_k (1-\beta)^{k/M}.$$

In the analysis below, property P2 shall also be used. However, since we motivated the total reward criterion as a refinement of the average reward criterion, our starting point will be property P1. Speaking of total rewards, we would like to evaluate a stream $r_0(s, \pi_1, \pi_2), r_1(s, \pi_1, \pi_2), \ldots$ by $\sum_{r=0}^{\infty} r_\tau(s, \pi_1, \pi_2)$. But, even if it is bounded, this sum may not exist; cf. Example 3.1, game 2. The next evaluation that one can think of is the Cesaro-limit of the row of partial sums, i.e.,

$$\lim_{T\to\infty} (T+1)^{-1} \sum_{t=0}^{T} \sum_{n=0}^{t} r_n(s, \pi_1, \pi_2).$$

For instance, it sounds fair that, for game 2 of Example 3.1, starting in state 1, the stream of payoffs, with partial sums 2, 0, 2, 0, ..., is evaluated as 1, since 1 is the average possession of player I. For stationary strategies (f_1, f_2),

$$\lim_{T\to\infty} (T+1)^{-1} \sum_{\tau=0}^{T} \sum_{n=0}^{\tau} r_n(s, f_1, f_2)$$

always exists (cf. Theorem 4.1 below), but for nonstationary strategies this is not true. In definition (3), we could also have taken lim sup instead of lim inf or any convex combination of them in order to define a total reward. We prefer to use the worst case viewpoint of player I. Evidently, whenever $\sum_{\tau=0}^{\infty} r_\tau(s, \pi_1, \pi_2)$ exists, it equals the total reward as defined in (3).

The class of stochastic games with property P1 is closely related to average reward stochastic games as can be seen by the following example of Thuijsman and Vrieze (Ref. 3).

Example 3.4. See Fig. 4. This game, called the bad match, is the total reward analogue of the big match for the average reward, as given in Example 3.3. Strategically, these two games are identical from the viewpoint of player I. Namely, how should he balance between his first and second action in state 1, in order to absorb in a favorable way. The main feature of the big match concerns the nonexistence of ϵ-optimal Markov strategies. Besides, for the big match ϵ-optimal history dependent strategies of a special type exist. The bad match bears the same phenomena with respect to the total rewards.

The bad match has total reward value vector $(0, 0, 2, -2)$ [for all strategies, the average rewards are $(0, 0, 0, 0)$], while the big match has

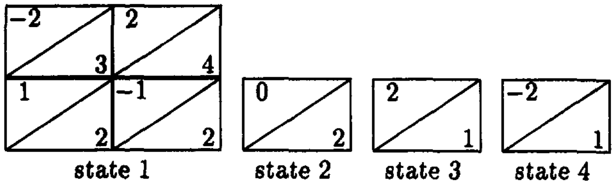


Fig. 4. Example 3.4: The bad match.

average reward vector $(0, 1, -1)$. For both games, an optimal stationary strategy for player II is to play $(1/2, 1/2)$ in state 1, whenever the play is in state 1. Neither for the big match nor for the bad match does player I have optimal strategies. For both games, player I can play $(K+1)^{-1}$-optimal in state 1 by playing the mixed action $(1-(k_\tau+K+1)^{-2}, (k_\tau+K+1)^{-2})$ at the τth visit to state 1, where k_τ denotes the excess number of times that player II chose action 2 over the number of times that player II chose action 1 during the $\tau-1$ previous visits. Notice that, if play starts in state 1 then, as long as player I chooses his first action, play visits state 1 at the even decision times.

4. Reformulation of a Total Reward Game as an Average Reward Game

Every total reward game can be reformulated as an average reward game with countably many states in the following way. Let

$$h_\tau=(s_0, a_{10}, a_{20}, s_1, a_{11}, a_{21}, \dots, s_{\tau-1}, a_{1\tau-1}, a_{2\tau-1}, s_\tau)$$

denote a possible history up to decision time $\tau \geq 1$, and let H_τ be the set of all h_τ's. Observe that $|H_\tau|$ is finite for each τ. The associated average reward game $\tilde{\Gamma}$ to a total reward game Γ is now defined as follows, where tildes refer to the associated game. Let

$$\tilde{S}:=\bigcup_{\tau=0}^{\infty} H_\tau$$

and, for any $\tau=0, 1, 2, \dots$ and for any

$$\tilde{s}=h_\tau=(s_0, a_{10}, a_{20}, \dots, s_{\tau-1}, a_{1\tau-1}, a_{2\tau-1}, s)\in\tilde{S},$$

let

$$\tilde{A}_1(\tilde{s}):=A_1(s), \qquad \tilde{A}_2(\tilde{s}):=A_2(s).$$

Furthermore, let

$$\tilde{r}(\tilde{s}, \tilde{a}_1, \tilde{a}_2):=\sum_{n=0}^{\tau-1} r(s_n, a_{1n}, a_{2n})+r(s, \tilde{a}_1, \tilde{a}_2),$$

$$\tilde{p}(\tilde{t}|\tilde{s}, \tilde{a}_1, \tilde{a}_2):=p(t|s, \tilde{a}_1, \tilde{a}_2), \qquad \text{if } \tilde{t}=h_{\tau+1}=(h_\tau, \tilde{a}_1, \tilde{a}_2, t),$$

$$\tilde{p}(\tilde{t}\,|\tilde{s}, \tilde{a}_1, \tilde{a}_2):=0, \qquad \text{otherwise.}$$

In the game $\tilde{\Gamma}$, states correspond to histories of the game Γ. Observe that, in game $\tilde{\Gamma}$, each state $\tilde{s}$ can only be reached along one path. It can be verified that, for the initial states $\tilde{s}\in H_0\equiv S$, the sets of strategies of the players correspond in a 1 to 1 way with the sets of strategies for the original game; cf. Thuijsman and Vrieze, Ref. 3. Moreover, when we consider strategies for a play that starts in a state $\tilde{s}=h_\tau\in\tilde{S}$, then we do not need to assign actions for states that will never be reached (or we could assign action 1 for all such states). Especially this holds for all states $h_{\tilde{\tau}}$, with $\tilde{\tau}>\tau$, for which the first part of $h_{\tilde{\tau}}$ does not coincide with h_τ. Now, these restricted strategies clearly coincide with the strategies of the original game for starting state $s\in S$.

At each decision time T, for every initial state $s\in H_0$ in $\tilde{\Gamma}$, and for all pairs of corresponding strategies $(\tilde{\pi}_1, \tilde{\pi}_2)$ and (π_1, π_2), it holds that

$$\sum_{\tau=0}^{T} \tilde{r}_\tau(s, \tilde{\pi}_1, \tilde{\pi}_2) = \sum_{\tau=0}^{T} \left[\sum_{n=0}^{\tau-1} r_n(s, \pi_1, \pi_2) + r_\tau(s, \pi_1, \pi_2) \right]$$
$$= \sum_{\tau=0}^{T} \sum_{n=0}^{\tau} r_n(s, \pi_1, \pi_2).$$

Hence,

$$\liminf_{T\to\infty} (T+1)^{-1} \sum_{\tau=0}^{T} \tilde{r}_\tau(s, \tilde{\pi}_1, \tilde{\pi}_2)$$
$$= \liminf_{T\to\infty} (T+1)^{-1} \sum_{\tau=0}^{T} \sum_{n=0}^{\tau} r_n(s, \pi_1, \pi_2). \tag{7}$$

The left-hand side of (7) is the average reward of $(\tilde{\pi}_1, \tilde{\pi}_2)$ in $\tilde{\Gamma}$ for initial state $\tilde{s}_0$, while the right-hand side of (7) is the total reward of (π_1, π_2) in Γ for initial state s_0. Therefore, we have the following theorem.

Theorem 4.1.

(i) The average reward game $\tilde{\Gamma}$ is equivalent to the total reward game Γ for initial states belonging to $S=H_0$.

(ii) In game $\tilde{\Gamma}$, for initial state $\tilde{s}=h_\tau\in H_\tau$ with $s_\tau=s$, the discounted payoff for $(\tilde{\pi}_1, \tilde{\pi}_2)$ is

$$\tilde{\phi}_\beta(s, \tilde{\pi}_1, \tilde{\pi}_2) = \sum_{n=0}^{\tau-1} r(s_n, a_{1n}, a_{2n}) + (1-\beta)^{-1}\phi_\beta(s, \pi_1, \pi_2), \tag{8}$$

where π_1 and π_2 are the unique associates in Γ to $\tilde{\pi}_1$ and $\tilde{\pi}_2$ in $\tilde{\Gamma}$.

Proof. Statement (i) is shown by (7). In game $\tilde{\Gamma}$, for initial state $\tilde{s}=h_\tau$ with $s_\tau=s$ and for strategies $\tilde{\pi}_1$ and $\tilde{\pi}_2$, the expected payoff at decision time T is

$$\tilde{r}_T(\tilde{s}, \tilde{\pi}_1, \tilde{\pi}_2) = \sum_{n=0}^{\tau-1} r(s_n, a_{1n}, a_{2n}) + \sum_{n=0}^{T} r_n(s, \pi_1, \pi_2). \tag{9}$$

Hence, the discounted reward for $\tilde{\pi}_1$ and $\tilde{\pi}_2$ is

$$\begin{aligned}&\tilde{\phi}_\beta(\tilde{s}, \tilde{\pi}_1, \tilde{\pi}_2)\\ &=(1-\beta) \sum_{T=0}^{\infty} \beta^T \left(\sum_{n=0}^{\tau-1} r(s_n, a_{1n}, a_{2n}) + \sum_{n=0}^{T} r_n(s, \pi_1, \pi_2)\right).\end{aligned} \tag{10}$$

If we now exchange the summation order of T and n, the second term of (10) becomes

$$\begin{aligned}&(1-\beta) \sum_{n=0}^{\infty} \sum_{T=n}^{\infty} \beta^T r_n(s, \pi_1, \pi_2)\\ &= \sum_{n=0}^{\infty} \beta^n r_n(s, \pi_1, \pi_2) = (1-\beta)^{-1} \phi_\beta(s, \pi_1, \pi_2).\end{aligned}$$

The first term of (10) obviously equals $\sum_{n=0}^{\tau-1} r(s_n, a_{1n}, a_{2n})$, which completes the proof. □

Corollary 4.1. The β-discounted reward value for initial state $\tilde{s}=h_\tau$ with $s_\tau = s$ in game $\tilde{\Gamma}$ equals

$$\tilde{\phi}_\beta^*(\tilde{s}) = \sum_{n=0}^{\tau-1} r(s_n, a_{1n}, a_{2n}) + (1-\beta)^{-1} \phi_\beta^*(s).$$

Theorem 4.1 shows that a total reward stochastic game is equivalent to an average reward stochastic game with a countable state space. The value existence proof of Mertens and Neyman cannot be applied straightforwardly, though the countable state space is not the bottleneck. From the definition of game $\tilde{\Gamma}$, it can be seen that the immediate rewards may be unbounded. In Section 6, we indicate how the Mertens–Neyman proof can be adapted to this case.

5. Stationary Strategies in Total Reward Games

We now pay attention to stationary strategies. The next theorem is of computational interest.

Theorem 5.1. For a pair of stationary strategies (f_1, f_2), if the total reward is finite, then the following four expressions are equivalent:

(i) $\phi_t(f_1, f_2) = \lim_{T\to\infty} (T+1)^{-1} \sum_{\tau=0}^{T} \sum_{n=0}^{\tau} P^n(f_1, f_2) r(f_1, f_2)$;
(ii) $\phi_t(f_1, f_2) = (I - P(f_1, f_2) + Q(f_1, f_2))^{-1} r(f_1, f_2)$;

(iii) there exists a pair, v, $u \in \mathbb{R}$ satisfying

$$v = r(f_1, f_2) + P(f_1, f_2)v \quad \text{and} \quad u + v = P(f_1, f_2)u,$$

while $\phi_t(f_1, f_2) = v$ for any such pair;

(iv) $\phi_t(f_1, f_2) = \lim_{\beta \uparrow 1}(1-\beta)^{-1}\phi_\beta(f_1, f_2)$.

Here, $P(f_1, f_2)$ is the stochastic transition matrix for (f_1, f_2), i.e., entry (s, t) of $P(f_1, f_2)$ gives the transition probability

$$\sum_{a_1} \sum_{a_2} f_1(s, a_1) f_2(s, a_2) p(t|s, a_1, a_2).$$

Furthermore $Q(f_1, f_2)$ denotes the Cesaro limit of $P(f_1, f_2)$, i.e.,

$$Q(f_1, f_2) = \lim_{T \to \infty} (T+1)^{-1} \sum_{\tau=0}^{T} P^\tau(f_1, f_2).$$

Proof. The proof proceeds as follows: (iv)→(ii)→(iii)→(i)→(iv). The dependence of the different variables on f_1 and f_2 will be suppressed.

(iv)→(ii). From $Qr = 0$ (finite total reward means average reward 0) and

$$\phi_\beta = (1-\beta)(I - \beta P)^{-1} r = (1-\beta) \sum_{\tau=0}^{\infty} \beta^\tau P^\tau r,$$

we derive

$$Q\phi_\beta = 0,$$

since

$$QP = Q.$$

Combined with

$$(I - \beta P)\phi_\beta = (1-\beta)r,$$

this gives

$$(1-\beta)^{-1}\phi_\beta = (I - \beta P + Q)^{-1} r.$$

Since the so-called fundamental matrix $I - P + Q$ is known to be nonsingular, it follows that

$$\lim_{\beta \uparrow 1} (I - \beta P + Q)^{-1} = (I - P + Q)^{-1}.$$

Hence, (ii) follows by taking limits.

(ii)→(iii). First, we discuss the existence of a solution (v, u). Multiplying

$$(I-P+Q)\phi_t=r$$

by Q gives $Q\phi_t=0$. Hence,

$$(I-P)\phi_t=r,$$

showing the first part of (iii). On thc other hand, it is well known [for instance, Vrieze (Ref. 8, Lemma 8.1.3)] that $Q\phi_t=0$ if and only if there exists a vector u with

$$u+\phi_t=Pu,$$

showing the second part of (iii). Second, we discuss the uniqueness of the v-part. If

$$v=r+Pv \quad \text{and} \quad u+v=Pu,$$

then

$$Qu+Qv=QPu=Qu$$

gives

$$Qv=0,$$

and thus

$$v-Pv+Qv=r,$$

which implies

$$v=(I-P+Q)^{-1}r=\phi_t.$$

(iii)→(i). Iterating the first equation of (iii) gives

$$v=\sum_{n=0}^{T} P^n r+P^{\tau+1}v, \qquad \text{for each } \tau=0, 1, 2, \ldots.$$

Taking averages of these expressions leads to

$$v=(T+1)^{-1}\left[\sum_{\tau=0}^{T}\sum_{n=0}^{\tau} P^n r+\sum_{\tau=0}^{T} P^{\tau+1}v\right]. \tag{11}$$

Multiplication of the second equation of (iii) by Q gives $Qv=0$. Hence, by taking limits in (11) and using

$$Q=\lim_{T\to\infty}(T+1)^{-1}\sum_{\tau=0}^{T} P^{\tau},$$

we obtain (i).

(i)→(iv). Here, we just apply the Tauberian theorem,

$$\lim_{T\to\infty} (T+1)^{-1} \sum_{\tau=0}^{T} a_\tau = \lim_{\beta\uparrow 1}(1-\beta) \sum_{\tau=0}^{\infty} \beta^\tau a_\tau,$$

for

$$a_\tau = \sum_{n=0}^{\tau} P^n r,$$

which is bounded by the assumption of finite total reward. In establishing (iv), one should realize that:

$$\begin{aligned}\sum_{\tau=0}^{\infty} \beta^\tau \sum_{n=0}^{\tau} P^n r &= \sum_{n=0}^{\infty} \sum_{\tau=n}^{\infty} \beta^\tau P^n r\\ &=(1-\beta)^{-1} \sum_{n=0}^{\infty} \beta^n P^n r\\ &=(1-\beta)^{-2}\phi_\beta.\end{aligned}$$

□

We finish this section with a characterization of the subclass of games for which both players have optimal stationary strategies with respect to the total reward value. But first we show that the Puiseux series expansion of the discounted value is of a special type, whenever both players have total reward optimal stationary strategies.

Theorem 5.2. If the total reward value ϕ_t^* exists and is finite, and if both players have optimal stationary strategies, then for the Puiseux series

$$\phi_\beta^* = \sum_{k=0}^{\infty} c_k(1-\beta)^{k/M},$$

it holds that

$$c_0 = c_1 = c_2 = \cdots = c_{M-1} = 0,$$

$$\phi_t^* = c_M, \qquad c_{M+1} = c_{M+2} = \cdots = c_{2M-1} = 0.$$

Proof. The fact that $c_0 = c_1 = c_2 = \cdots = c_{M-1} = 0$ is a consequence of property P1 (also see P2), which clearly holds under the assumption of the theorem. Let f_1^* and f_2^* be optimal stationary strategies with respect to the total reward value. Now, let $\bar{f}_2$ be uniform discount optimal for player II in the Markov decision problem that results when player I fixes f_1^*. It is well known [cf. Bewley and Kohlberg (Ref. 7, Corollary 6.5)] that, for a pair of stationary strategies, for all β close to 1, the β-discounted payoff can be written as a power series in $1-\beta$. So,

$$\phi_\beta(f_1^*,\bar{f}_2)=\sum_{k=0}^{\infty} d_k(1-\beta)^k,$$

where d_0 equals the average reward of $(f^*,\bar{f}_2)$. Obviously,

$$\phi_\beta(f_1^*,\bar{f}_2)\leq\phi_\beta^*.$$

On the other hand,

$$\phi_t^*\leq\phi_t(f_1^*,\bar{f}_2)=\lim_{\beta\uparrow 1}(1-\beta)^{-1}\phi_\beta(f_1^*,\bar{f}_2).$$

As a conclusion

$$d_0=0 \quad\text{and}\quad \phi_t^*\leq d_1\leq\lim_{\beta\uparrow 1}(1-\beta)^{-1}\phi_\beta^*=c_M.$$

Similarly, with the aid of f_2^* and an appropriate $\bar{f}_1$, one can prove that

$$\phi_t^*\geq c_M.$$

Then,

$$\phi_t^*=c_M=d_1.$$

Reconsidering $\phi_\beta(f_1^*,\bar{f}_2)\leq\phi_\beta^*$ yields

$$\phi_t^*(1-\beta)+\sum_{k=2}^{\infty} d_k(1-\beta)^k\leq\phi_t^*(1-\beta)+\sum_{k=M+1}^{\infty} c_k(1-\beta)^{k/M},$$

which gives

$$\sum_{k=M+1}^{2M-1} c_k(1-\beta)^{k/M}\geq 0.$$

In a similar way, using f_2^* and $\bar{f}_1$, we derive that

$$\sum_{k=M+1}^{2M-1} c_k(1-\beta)^{k/M}\leq 0,$$

which together with the previous inequality implies

$$c_{M+1}=c_{M+2}=\cdots=c_{2M-1}=0. \qquad \square$$

Theorem 5.3. For a total reward stochastic game the following two statements are equivalent:

(i) the value vector exists and is finite; both players possess optimal stationary strategies;

(ii) the following set of equations has a solution for variables $v, u_1, u_2 \in \mathbb{R}^z$, $\alpha \in \mathbb{R}$:

$$v(s) = \operatorname*{Val}_{A_1(s) \times A_2(s)} \left[r(s, \cdot, \cdot) + \sum_{t=1}^{z} q(t|s, \cdot, \cdot) v(t) \right], \qquad s \in S, \tag{12}$$

$$u_1(s) + v(s) = \operatorname*{Val}_{O_1(s) \times A_2(s)} \left[\alpha r(s, \cdot, \cdot) + \sum_{t=1}^{z} q(t|s, \cdot, \cdot) u_1(t) \right], \qquad s \in S, \tag{13}$$

$$u_2(s) + v(s) = \operatorname*{Val}_{A_1(s) \times O_2(s)} \left[\alpha r(s, \cdot, \cdot) + \sum_{t=1}^{z} q(t|s, \cdot, \cdot) u_2(t) \right] \qquad s \in S. \tag{14}$$

Here $O_1(s)$ and $O_2(s)$, $s \in S$, are the extreme points of the polyhedral sets of optimal strategies for player I and player II, respectively, for the matrix games (12). Furthermore, for all solutions to (12)–(14), v is the same and v is the total reward value. Optimal stationary strategies can be composed by optimal actions for the matrix games (13) for player I and for the matrix games (14) for player II.

Proof. Observe that (i), as well as existence of a solution to (12), imply that property P1 holds.

(ii)→(i). Let v, u_1, u_2, α satisfy (12)–(14), and let $f_1^*(s)$, $s \in S$, be optimal for player I in (13). Then, for any f_2,

$$r(f_1^*, f_2) + P(f_1^*, f_2) v \geq v, \tag{15}$$

$$\alpha r(f_1^*, f_2) + P(f_1^*, f_2) u_1 \geq u_1 + v. \tag{16}$$

We show that $\phi_t(f_1^*, f_2) \geq v$. Multiplication of (15) by $Q(f_1^*, f_2)$ yields

$$Q(f_1^*, f_2) r(f_1^*, f_2) \geq 0.$$

If for a state s we have

$$(Q(f_1^*, f_2) r(f_1^*, f_2))(s) > 0,$$

so positive average reward, then the total reward for that starting state is $\infty > v(s)$. Hence, we can concentrate on the set of states

$$\bar{S} := \{ s \in S | (Q(f_1^*, f_2) r(f_1^*, f_2))(s) = 0 \}.$$

Since $\bar{S}$ is closed with respect to $P(f_1^*, f_2)$, i.e., play never leaves $\bar{S}$, we can assume without loss of generality that $\bar{S}=S$. Then, iteration of (15) gives

$$\sum_{n=0}^{\tau} P^n(f_1^*, f_2)r(f_1^*, f_2) + P^{\tau+1}(f_1^*, f_2)v \geq v.$$

By taking averages, we get, for any T,

$$(T+1)^{-1} \sum_{\tau=0}^{T} \sum_{n=0}^{\tau} P^n(f_1^*, f_2)r(f_1^*, f_2)$$

$$+(T+1)^{-1} \sum_{\tau=0}^{T} P^{\tau+1}(f_1^*, f_2)v \geq v. \tag{17}$$

Multiplication of (16) by $Q(f_1^*, f_2)$ and using

$$Q(f_1^*, f_2)r(f_1^*, f_2)=0$$

gives

$$Q(f_1^*, f_2)v \leq 0.$$

But then, by taking limits in (17), we obtain

$$\phi_t(f_1^*, f_2) \geq v. \tag{18}$$

Similarly, for the stationary strategy f_2^* composed of optimal actions $f_2^*(s)$, $s \in S$, for player II in the matrix games (14) and any strategy f_1 for player I, we have

$$\phi_t(f_1, f_2^*) \leq v. \tag{19}$$

The combination of (18) and (19) shows assertion (i).

(i)→(ii). Let ϕ_t^* be the total reward value vector, and let f_1^* and f_2^* be optimal stationary strategies. In Theorem 5.1, we already showed that

$$\phi_t^* = \lim_{\beta \uparrow 1} (1-\beta)^{-1} \phi_\beta^*.$$

Equation (12) then follows from (5). It remains to show (13) and (14).

Let f_2 be such that the total reward $\phi_t(f_1^*, f_2)$ is finite and hence

$$Q(f_1^*, f_2)r(f_1^*, f_2)=0.$$

From Theorem 5.1 (iii), we deduce that

$$Q(f_1^*, f_2)\phi_t(f_1^*, f_2)=0,$$

and since

$$\phi_t^* \leq \phi_t(f_1^*, f_2),$$

this gives

$$Q(f_1^*,f_2)(-\phi_t^*)\geq 0$$

and

$$Q(f_1^*,f_2)(-\phi_t^*+\alpha r(f_1^*,f_2))\geq 0, \qquad \text{for any } \alpha\in\mathbb{R}.$$

If f_2 is such that $\phi_t(f_1^*,f_2)$ is infinite, then

$$(Q(f_1^*,f_2)r(f_1^*,f_2))(s)>0,$$

since f_1^* is total reward optimal. But then also

$$(Q(f_1^*,f_2)(-\phi_t^*+\alpha r(f_1^*,f_2)))(s)\geq 0, \quad \text{for some appropriate } \alpha\in\mathbb{R}^+. \tag{20}$$

Observe that increasing α in (20) does not violate the inequality. Let α^* be the minimal α, such that (20) holds for all states $s\in S$ and for all pure stationary strategies f_2. Since for the Markov decision problem that results when f_1^* is fixed, with payoff structure $-\phi_t^*(s)+\alpha^* r(s,f_1^*,\cdot)$, player II has an optimal pure stationary strategy, it follows that the minimum of this Markov decision problem is nonnegative. Hence,

$$Q(f_1^*,f_2)(-\phi_t^*+\alpha^* r(f_1^*,f_2))\geq 0, \qquad \text{for all } f_2 \text{ and all } \alpha\geq\alpha^*.$$

Obviously, for f_2^* total reward optimal, we have

$$Q(f_1^*,f_2^*)(-\phi_t^*+\alpha^* r(f_1^*,f_2^*))=0, \qquad \text{for any } \alpha>\alpha^*.$$

Hence, the stochastic game with payoff structure $-\phi_t^*(s)+\alpha^* r(s,\cdot,\cdot)$ defined on the action sets $O_1(s)\times A_2(s)$, $s\in S$, has average reward value vector 0. So, by the already-mentioned Lemma 8.1.3 in Vrieze (Ref. 8), there exists a vector u_1 satisfying Eq. (13). Analogously, the existence of u_2 can be shown. □

6. Existence of Value for Total Reward Stochastic Games

In Section 4, we showed that a total reward stochastic game with finite state and action spaces is equivalent to an average reward stochastic game with infinitely countable many states (corresponding to histories in the original game) and with the same action sets in corresponding states. This equivalence can be used to show that the value of a total reward stochastic game exists.

Theorem 6.1. A total reward stochastic game for which property P1 (or equivalently P2) holds, has a value. ϵ-Optimal strategies can be constructed by playing discounted optimal at every decision time, whereby the discount factor is appropriately adapted after every step.

Our proof is an adaptation of the proof of Mertens and Neyman (Ref. 5) for the existence of the value of average reward stochastic games in the case of finite state and action spaces. However, the proof of Mertens and Neyman consists of several pages of mathematical analysis. We will not repeat that here, but merely indicate the line of the proof and mention the differences.

Sketch of Proof. Let $\epsilon > 0$; let k_0, M, L be sufficiently large constants; let $s_{\tau+1}$ be the state observed at decision time $\tau + 1$. Then, define recursively, for $\tau = 0, 1, 2, \ldots$,

$$\begin{aligned} k_{\tau+1} &:= \max\{L, k_\tau + \tilde{r}(\tilde{s}_\tau, a_{1\tau}, a_{2\tau}) - \tilde{\phi}^*_{\beta_\tau}(\tilde{s}_{\tau+1}) + 4\epsilon\} \\ &= \max\{L, k_\tau - (1-\beta_\tau)^{-1}\phi^*_{\beta_\tau}(s_{\tau+1}) + 4\epsilon\}, \qquad (21) \end{aligned}$$

$$\beta_{\tau+1} := 1 - k_{\tau+1}^{-M/(M-1)}, \qquad (22)$$

$$y_{\tau+1} := (M-1)k_{\tau+1}^{-1/(M-1)}. \qquad (23)$$

Now, player I, at every decision time τ, chooses an optimal action in the matrix game of the Shapley equation (5) for discount factor β_τ. Obviously, for a pair of strategies of the players, a stochastic process evolves with respect to $s_\tau, a_{1\tau}, a_{2\tau}, k_\tau, \beta_\tau, y_\tau$. We denote the stochastic representations by putting bars above the variable. Using Theorem 4.1, it can be shown, along similar lines as in the proof of Mertens and Neyman, that the sequence

$$Y_\tau := \sum_{n=0}^{\tau-1} r(\bar{s}_n, \bar{a}_{1n}, \bar{a}_{2n}) + (1-\bar{\beta}_\tau)^{-1}\phi^*_{\bar{\beta}_\tau}(\bar{s}_\tau) - \bar{y}_\tau,$$

with $\tau = 0, 1, 2, \ldots$, forms a semimartingale. Now, one of two things can happen. Either this semimartingale has finite limit expectation or infinite. If the strategy of player II is such that this expectation is finite, then the analysis of Mertens and Neyman can be followed giving rise to an expected total payoff of at least

$$\lim_{\beta_\tau \to 1} (1-\beta_\tau)^{-1}\phi^*_{\beta_\tau}(s_0) - 8\epsilon.$$

When the strategy of player II is such that this expectation is unbounded, then also the total reward is unbounded, showing that also in this case the constructed strategy is ϵ-optimal. □

Theorem 6.1 and Example 3.4 teach us that, when we are not only interested in the average payoff, but want to play more sensitively by looking also at the behavior of the partial sums, then we can do so, but we generally need to use behavior strategies. This in spite of the fact that reaching the average can be achieved by playing stationary.

Remark 6.1. Recall that we used

$$\liminf_{T\to\infty} (T+1)^{-1} \sum_{\tau=0}^{T} \sum_{n=0}^{\tau} r_n(s, \pi_1, \pi_2),$$

to define the total reward, where the numbers $r_n(s, \pi_1, \pi_2)$ denoted expected payoffs [cf. Eq. (3)]. This is very much different from taking

$$E_{s,\pi_1,\pi_2}\left(\liminf_{T\to\infty} (T+1)^{-1} \sum_{\tau=0}^{T} \sum_{n=0}^{\tau} a_n\right),$$

where E_{s,π_1,π_2} denotes expectation with respect to initial state and strategies and where the numbers a_n are the actual payoffs. Suppose for example that, at each decision time, the payoff is 1 with probability 0.5 and -1 with probability 0.5. Then, our definition would yield a total reward of 0, whereas the alternative definition would yield $-\infty$. Although for average reward stochastic games the value does not change when interchanging expectation and lim inf (cf. Mertens and Neyman, Ref. 5), this is clearly not valid for total reward stochastic games. This phenomenon is related to the fact that for total rewards the partial sums need not be bounded. It is not clear to us whether property P1 is a sufficient condition for the existence of the value for the alternative criterion.

References

1. SHAPLEY, L. S., *Stochastic Games*, Proceedings of the National Academy of Sciences, USA, Vol. 39, pp. 1095–1100, 1953.
2. GILLETTE, D., *Stochastic Games with Zero Stop Probabilities*, Contributions to the Theory of Games, Edited by M. Dresher, A. W. Tucker, and P. Wolfe, Princeton University Press, Vol. 3, pp. 179–187, 1957.
3. THUIJSMAN, F., and VRIEZE, O. J., *The Bad Match, a Total Reward Stochastic Game*, Operations Research Spektrum, Vol. 9, pp. 93–99, 1987.
4. BEWLEY, T., and KOHLBERG, E., *The Asymptotic Theory of Stochastic Games*, Mathematics of Operations Research, Vol. 1, pp. 197–208, 1976.
5. MERTENS, J. F., and NEYMAN, A., *Stochastic Games*, International Journal of Game Theory, Vol. 10, pp. 53–66, 1981.
6. BLACKWELL, D., and FERGUSON, T. S., *The Big Match*, Annals of Mathematical Statistics, Vol. 39, pp. 159–163, 1968.

7. BEWLEY, T., and KOHLBERG, E., *On Stochastic Games with Optimal Stationary Strategies*, Mathematics of Operations Research, Vol. 3, pp. 104–125, 1978.

8. VRIEZE, O. J., *Stochastic Games with Finite State and Action Spaces*, Centre for Mathematics and Computer Science, Amsterdam, Holland, Vol. 33, 1987.