OR Spektrum (1987) 9:93–99

OR Spektrum

The Bad Match; A Total Reward Stochastic Game

F. Thuijsman* and O. J. Vrieze

Department of Mathematics, University of Nijmegen, Toernooiveld, NL-6525 ED Nijmegen, The Netherlands

Received September 24, 1985 / Accepted in revised form February 17, 1987

Summary. For two person zero sum stochastic games we introduce a new criterion for evaluating streams of payoffs. When the players use this criterion we call such games total reward stochastic games. It is unknown whether total reward stochastic games, with the property that the average value is zero for each initial state, always have a value. We examine an example of such a total reward stochastic game in which one of the players can play ϵ-optimal only by using history dependent strategies.

Zusammenfassung. Für stochastische Zwei-Personen-Null-Summen-Spiele wird ein neues Kriterium zur Bewertung der Auszahlungsströme eingeführt, das Gesamt-Gewinn-Kriterium. Es ist bisher unbekannt, ob stochastische Spiele, deren Wert bezüglich des Durchschnittsgewinn-Kriteriums gleich Null ist, bezüglich des Gesamt-Gewinn-Kriteriums einen "Wert" besitzen. Es wird ein Beispiel untersucht, in dem ein Spieler nur ϵ-optimal spielen kann, wenn er von der Vorgeschichte abhängige Strategien benutzt.

1. Introduction

We consider a two person stochastic game. Such a game is played in stages. At each stage the game is in one of finitely many states. Both players observe the current state and independently choose an action out of a finite set. The pair of actions at stage $n \in \mathbb{N} = \{1, 2, 3, \ldots\}$ together with the current state determine a payoff r_n to be made by player II to player I. Furthermore the current state and the action choices determine a probability function according to which the system moves to a next state.

A player's strategy is a specification of a probability distribution at each stage over his available actions, conditional on the history of the game up to that stage. By history we mean the sequence of past states and actions up to that stage. Any pair of strategies determines an expected payoff $E[r_n]$ at stage n, for each $n \in \mathbb{N}$.

There are different ways of evaluating such a stream of expected payoffs. The value of the game depends on which evaluation is being used.

Shapley [6] proved that the β-discounted game, i.e. the game with "evaluation" $\Sigma_{n=1}^{\infty} \beta^{n-1} E[r_n]$, for $\beta \in [0, 1)$ has a value and that both players possess optimal stationary strategies, i.e. strategies independent of history and stage. (Here E stands for the expectation sign.)

It was unknown for many years whether average stochastic games, i.e. games with "evaluation" $\liminf_{t\to\infty} \frac{1}{t} \Sigma_{n=1}^{t} E[r_n]$, always have a value. The fact that average stochastic games do have a value has been proved by Mertens and Neyman [4]. At the same time they showed, that the evaluation rule $E\left[\liminf_{t\to\infty} \frac{1}{t} \Sigma_{n=1}^{t} r_n\right]$ leads to the same value. (Evidently also for the discounted criterion the evaluation $E[\Sigma_{n=1}^{\infty} \beta^{n-1} r_n]$ leads to the same discounted value.)

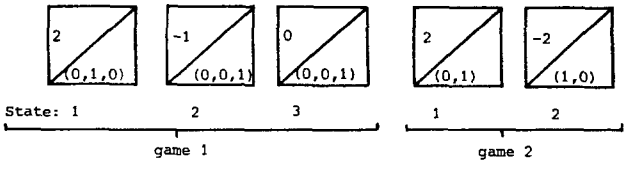
Before that, Blackwell and Ferguson [2] have shown, that history dependent strategies are essential for average stochastic games. They have analysed "the big match", an average stochastic game introduced by Gillette [3], and proved that this game has a value. However, one of the players has no history independent average ϵ-optimal strategy in the big match.

* This research is supported by the Netherlands Organization for the Advancement of Pure Research (ZWO, project 10-64-10)

The aim of this paper is to introduce a third evaluation rule, called the total reward criterion. This criterion turns out to be a refinement in addition to the average criterion. We will give an example showing, that, like in "the big match", also for this criterion players may need to use history dependent strategies. In the last section we show that a total reward stochastic game can be viewed as an average stochastic game, however the set of states has to be enlarged to an infinite set.

2. The Total Reward Evaluation

Observe that for the discounted evaluation rule early incomes have the main impact on the discounted reward, while for the average criterion future returns determine the average reward. Several objections can be raised against the average evaluation rule. Consider, for example, the following two trivial games. In every state both players have only one action and the transitions are deterministic.



(Player I is the row player, player II the column player; a cell r/p means payoff r to player I, p is the transition probability vector.)

For game 1, clearly the average value equals (0, 0, 0). However player I would prefer to start in state 1 (getting total reward 1) and player II would prefer to start in state 2 (paying total reward -1). Likewise in game 2 the average value equals (0, 0), but also here player I likes to start in state 1, thus owning half of the time two units and half of the time zero units. And player II likes to start in state 2, being due half of the time minus two units and half of the time zero units.

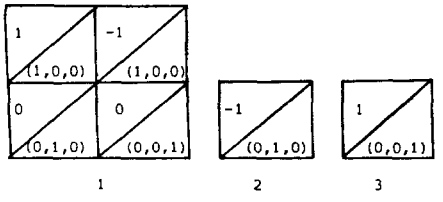
This phenomenon asks for a more sensitive evaluation rule, which we call the total reward criterion and which is defined by the "evaluation" $\liminf_{T\to\infty} \frac{1}{T} \Sigma_{t=1}^{T} \Sigma_{n=1}^{t} E[r_n]$.

The choice of this evaluation rule may need some explanation. Observe that $\Sigma_{n=1}^{t} E[r_n]$ are partial sums for every $t \in \mathbb{N}$. Then $\frac{1}{T} \Sigma_{t=1}^{T} \Sigma_{n=1}^{t} E[r_n]$ is the average of the first T partial sums.

Speaking of total rewards we would like to evaluate a stream $(E[r_1], E[r_2], \ldots)$ by $\Sigma_{n=1}^{\infty} E[r_n]$. However this sum may have more than one limit point in $\mathbb{R} \cup \{+\infty, -\infty\}$. See for example game 2 above.

The next evaluation one can think of is the Cesaro-limit of the row of partial sums, i.e. $\lim_{T\to\infty} \frac{1}{T} \Sigma_{t=1}^{T} \Sigma_{n=1}^{t} E[r_n]$. For instance, it sounds fair that for game 2, starting in state 1, the stream of payoffs, giving partial sums $(2, 0, 2, 0, \ldots)$ is evaluated as 1, since 1 is the average possession of player I. It can be shown that for stationary strategies this limit always exists in $\mathbb{R} \cup \{+\infty, -\infty\}$. However, for non-stationary strategies this need not be the case. Therefore we choose $\liminf_{T\to\infty} \frac{1}{T} \Sigma_{t=1}^{T} \Sigma_{n=1}^{t} E[r_n]$ as evaluation. We also might have chosen "lim sup" or any convex combination of "lim inf" and "lim sup" as evaluation. In case $\Sigma_{n=1}^{\infty} E[r_n]$ or $\lim_{T\to\infty} \frac{1}{T} \Sigma_{t=1}^{T} \Sigma_{n=1}^{t} E[r_n]$ exists in $\mathbb{R} \cup \{+\infty, -\infty\}$ it equals $\liminf_{T\to\infty} \frac{1}{T} \Sigma_{t=1}^{T} \Sigma_{n=1}^{t} E[r_n]$.

We now mention some properties of the total reward evaluation rule. Evidently, if the average value is unequal to zero in some state, then for such a state the total reward value exists and equals $+\infty$ or $-\infty$ dependent on a positive or negative sign of that average value. If for some state the average value equals 0 and if one player has no average optimal strategy, then the total reward value does not need to exist as is demonstrated by the following example (the "big match" of Blackwell and Ferguson [2]).

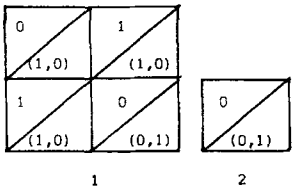


For state 1 the average value equals 0. It is well known that player I has no average optimal strategy in this game. So for every strategy π_1 for player I there exists a strategy π_2 for player II such that the corresponding average reward is $\liminf_{t\to\infty} \frac{1}{t} \Sigma_{n=1}^{t} E[r_n] < 0$. Hence for the corresponding total reward it holds that $\liminf_{T\to\infty} \frac{1}{T} \Sigma_{t=1}^{T} \Sigma_{n=1}^{t} E[r_n] = -\infty$. Furthermore it can easily be verified that the stationary strategy $\pi_2^* = \left(\frac{1}{2}, \frac{1}{2}\right)^{\infty}$ for player II yields an expected payoff 0 at each stage, no matter what choices player I makes. Consequently, the total reward value does not exist since $\sup_{\pi_1} \inf_{\pi_2} v(\pi_1, \pi_2) = -\infty \neq 0 = \inf_{\pi_2} \sup_{\pi_1} v(\pi_1, \pi_2)$.

(Throughout this paper π_1 and π_2 are strategies for player I and player II respectively and $v(\pi_1, \pi_2)$ is the corresponding total reward.)

For this example the nonexistence of the total reward value is a consequence of the fact that not for each state the average value equals 0. For games for which the average value equals zero for all states and for which in addition at least one player has no average optimal stationary strategy for some state, the total reward value, so it exists, is plus or minus infinity.

The following example illustrates this fact.



For this game the average value equals $(0, 0)$ and the total reward value $(\infty, 0)$. This can be seen as follows. Starting in state 1, player I can be sure that the total reward is ∞. To acquire this he could use the strategy π_1^* defined by: if the system is in state 1 at stage n, then choose the first row with probability $1 - \frac{1}{n+1}$, and the second row with probability $\frac{1}{n+1}$. Hence for state 1 the total reward value is ∞. Although player II has no average optimal stationary strategy, he does have an average optimal Markov strategy. (At the first stage he should play optimal in the one-stage game, at the next two stages he should play optimal in the two-stage game, at the next three stages he should play optimal in the three-stage game, etc., cf. Vrieze [7].) Here a (semi-)Markov strategy is a strategy in which the action choice at each stage solely depends (on the initial state,) on the current state and on the stage.

So in the above example we have the curious phenomenon that on the one hand player II can assure himself an average reward 0, but on the other hand he cannot prevent player I from obtaining total reward ∞. The above examples illustrate that, only under the next assumption *A*, the total reward value, if it exists, can be finite.

Assumption A:

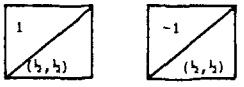
(i) The average value equals 0 for all initial states.

(ii) Both players possess average optimal stationary strategies.

We do not know, whether the total reward value always exists under Assumption A. We do know that $\sup_{\pi_1} \inf_{\pi_2} v(\pi_1, \pi_2)$ and $\inf_{\pi_2} \sup_{\pi_1} v(\pi_1, \pi_2)$ are both finite in this case (the average optimal stationary strategies assure each of the players a finite total reward) and we have good indications that these quantities are equal under Assumption A. In a subsequent paper we will show besides other results, that, under Assumption A, if both players possess total reward ϵ-optimal stationary strategies, then the total reward value equals the limit for β tending to 1 of the β-discounted values. Notice, that, under Assumption A, this limit also equals the constant term x_0 in the solution of the limit discount equation (cf. Bewley and Kohlberg [1]). It is known that the average value also appears in the solution of the limit discount equation (as the leading coefficient x_{-1}). We expect that under Assumption A, x_0 will always be the total reward value but we did not yet find a proof for this. If for a game, satisfying Assumption A, the total reward value exists, then obviously any total reward ϵ-optimal strategy is optimal with regard to the average criterion. In this sense the total reward criterion can be interpreted as a sensitive criterion in addition to the average criterion.

We like to conclude this section with the following remark, which was communicated to us by Neyman [5].

It is not clear whether for total rewards the expression $E\left[\liminf_{T\to\infty} \frac{1}{T} \Sigma_{t=1}^{T} \Sigma_{n=1}^{t} r_n\right]$ makes any sense. Consider the following example.

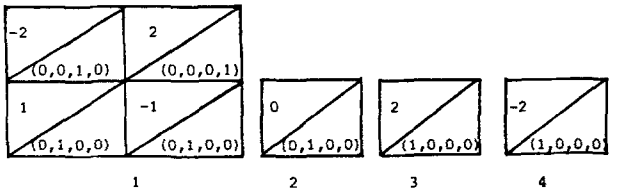


Then $\liminf_{T\to\infty} \frac{1}{T} \Sigma_{t=1}^{T} \Sigma_{n=1}^{t} E[r_n] = 0$, while

$E\left[\liminf_{T\to\infty} \frac{1}{T} \Sigma_{t=1}^{T} \Sigma_{n=1}^{t} r_n\right] = -\infty$, as for every realisation of the random walk it holds that $\liminf_{T\to\infty} \frac{1}{T} \Sigma_{t=1}^{T} \Sigma_{n=1}^{t} r_n = -\infty$.

3. The Model of the Bad Match

By the bad match we mean the following game:



The initial state is state 1. The players only have to make a decision at stages where the system is in state 1. Evidently if the system starts in state 1 at stage 1, then the players only need to make a decision at the odd stages. Those stages we call decision epochs. Notice, if player I chooses action 2 on any decision epoch, the system immediately moves to state 2, where it will remain forever without any nonzero payoff. Hence the game may be viewed as being terminated as soon as the system reaches state 2. It can easily be seen for this game that the average value equals zero for all states.

4. A Solution to the Bad Match

We start by defining a history dependent strategy for player I. First we define the function $p: \{0, 1, 2, \ldots\} \to [0, 1]$ by $p(m) = 1/(m+1)^2$. Let N be a nonnegative integer. We define the history dependent strategy π_1^N for player I by: having observed the choices player II made at the first n decision epochs (hence the first n odd stages), say $b_1, b_2, \ldots, b_n \in \{1, 2\}, n \geqslant 0$, calculate the excess k_n of 2's over 1's among $b_1, b_2, \ldots, b_n$ and choose on decision epoch $n+1$ action 2 with probability $p(k_n + N)$ (and action 1 with probability $1 - p(k_n + N)$). This strategy appears to be total reward $1/(N+1)$ − optimal for player I. We now state our main theorem, the proof of which can be found in the appendix.

Theorem

(i) The total reward value of the bad match is zero (for initial state 1).

(ii) Player II can play total reward optimal by choosing action 1 with probability 1/2 and action 2 with probability 1/2 at each decision epoch.

(iii) For each nonnegative integer N strategy π_1^N is total reward $1/(N+1)$-optimal for player I.

(iv) Player I has no total reward ϵ-optimal history independent strategy for $\epsilon > 0$ sufficiently small.

(v) Player I has no total reward optimal strategy.

5. Conclusions

The bad match underlines that there is an analogy between total reward stochastic games and average stochastic games.

From the above game it appears that non-Markovian strategies seem to play a similar role in total reward games as in average stochastic games. This relationship is narrowed by the fact that each total reward game can be associated with an average stochastic game, in such a way that the total rewards in the original game equal the average rewards in the associated game for each pair of strategies.

Let for the original game Γ, S be the state space, A_s and B_s the action spaces in state $s \in S$, $r(s, a, b)$ the immediate payoff and $p(s, a, b)$ the transition probability vector for $s \in S$, $a \in A_s$, $b \in B_s$.

Next, let for $n = 2, 3, \ldots$

$H_n := \{(s_1, a_1, b_1, s_2, a_2, b_2, \ldots, s_{n-1}, a_{n-1}, b_{n-1});$

$s_k \in S, a_k \in A_{s_k}, b_k \in B_{s_k}$, for $k = 1, 2, \ldots, n-1\}$ and let

$H_1 = \{1\}$.

So H_n consists of the finite set of histories that could have occurred up to stage n in the original game.

Let $S_n := H_n \times S$, $n = 1, 2, \ldots$

Now the associated game can be defined. The variables are denoted with quotation-marks.

Define the game Γ' by:

$$S' := \bigcup_{n=1}^{\infty} S_n,$$

for $s' = (h_n, s) \in S_n$ let $A'_{s'} := A_s$ and $B'_{s'} := B_s$,
payoffs: $r'(s', a', b') := \Sigma_{k=1}^{n-1} r(s_k, a_k, b_k) + r(s, a, b)$,
when $s' = (h_n, s), a' = a, b' = b$
and $h_n = (s_1, a_1, b_1, \ldots, s_{n-1}, a_{n-1}, b_{n-1})$,
transitions: $p'(t'|s', a', b') := p(t|s, a, b)$
if $s' = (h_n, s), a' = a, b' = b$ and $t' = (h_n, s, a, b, t)$;
else $p'(t'|s', a', b') = 0$.

In the game Γ' histories are translated into states at each stage. It can be verified that the sets of strategies coincide for the original game and game Γ'. Notice that in Γ' each state s' can be reached along exactly one history path. At each stage T, for each pair of strategies and for every initial state $(1, s) \in \{1\} \times S$ in Γ' it holds that

$$\Sigma_{t=1}^{T} E[r'_t] = \Sigma_{t=1}^{T} E[\Sigma_{n=1}^{t-1} r(s_n, a_n, b_n) + r(s_t, a_t, b_t)]$$
$$= \Sigma_{t=1}^{T} \Sigma_{n=1}^{t} E[r(s_n, a_n, b_n)].$$

Hence $\liminf_{T\to\infty} \frac{1}{T} \Sigma_{t=1}^{T} E[r'_t]$

$$= \liminf_{T\to\infty} \frac{1}{T} \Sigma_{t=1}^{T} \Sigma_{n=1}^{t} E[r(s_n, a_n, b_n)],$$

showing the equivalency of the average criterion for

game Γ' with the total reward criterion for the original game. The technique of Mertens and Neyman [4] cannot be applied straightforward to Γ'. This is due to the fact that, even under Assumption A, the immediate payoffs in Γ' are not uniformly bounded. Realisations $(s_1, a_1, b_1, s_2, a_2, b_2, \ldots)$ may occur for which $\Sigma_{k=1}^{n} r(s_k, a_k, b_k)$ becomes arbitrary large (or small). Whether the Mertens and Neyman technique can be adapted to our case is not yet known to us.

Concerning the discounted criterion, it can be verified, that for game Γ' with regard to the discounted evaluation rule it holds, that for initial state $s' = (s_1, a_1, b_1, \ldots, s_{n-1}, a_{n-1}, b_{n-1}, s)$:

$$v'_\beta(\pi'_1, \pi'_2)(s') = \frac{1}{1-\beta}\,(\Sigma_{k=1}^{n-1} r(s_k, a_k, b_k) + v_\beta(\pi_1, \pi_2)(s)).$$

So, if the Mertens and Neyman technique could be applied, in the sense that for Γ' the average value equals the limit for β tending to 1 of $(1-\beta)$ times the β-discounted value, then the above equality shows that this limit would, for initial state $(1, s)$, equal the limit for β tending to 1 of the β-discounted value for the original game. Hence this would imply that the total reward value of the original game equals the limit of the β-discounted values of the original game for discount factor β tending to 1.

Appendix

In this appendix we give a proof of the theorem in Sect. 4.

Lemma 1. *Let π_2^* be the stationary strategy for player II defined by choosing action 1 with probability 1/2 and choosing action 2 with probability 1/2 at each decision epoch. Then $v(\pi_1, \pi_2^*) = 0$ for any strategy π_1 of player I.*

Proof. Whatever choices player I makes, at each stage the expected payoff is 0. Namely, in state 1 at each decision epoch the expected payoff is zero. At each other stage the system is in state 3 or state 4, with the same probability, giving an expected payoff zero. When the system has moved to state 2 no nonzero payoffs occur anymore. □

The following corollary is immediate.

Corollary 2. $\inf_{\pi_2} \sup_{\pi_1} v(\pi_1, \pi_2) \leqslant 0$.

The next lemma states that, restricting to (semi-) Markov strategies, player I can only assure himself a total reward of at most -1.

Lemma 3. *Using a history independent strategy, player I cannot guarantee himself more than -1.*

Proof. Let π_{1M} be a history independent strategy for player I. We consider two cases.

In case 1, suppose that the probability that player I will ever choose action 2, is zero. Then player I chooses action 1 with probability 1 at each decision epoch. Strategy $\hat{\pi}_2$, described by always choosing action 1, leads to $v(\pi_{1M}, \hat{\pi}_2) = -1$ (the average of the alternating partial sums -2 and 0).

In case 2, suppose that the probability that player I will ever choose action 2 is $\epsilon > 0$. Then for each $\delta \in (0, \epsilon)$ there is a $N_\delta \in \mathbb{N}$ such that the probability that player I will choose action 2 before decision epoch N_δ is larger than $\epsilon - \delta$. For each $\delta \in (0, \epsilon)$ define strategy π_2^δ for player II by: at decision epochs $1, 2, \ldots, N_\delta$ choose action 2 and choose action 1 always thereafter. One can verify

$$v(\pi_{1M}, \pi_2^\delta) \leqslant (\epsilon - \delta) \cdot (-1) + \delta \cdot (1) + (1 - \epsilon) \cdot (-1)$$
$$= -1 + 2\delta.$$

Since player II can choose δ as small as he wants, the proof is completed. □

We will show that the strategy π_1^N (defined in Sect. 4) is a total reward $\frac{1}{N+1}$ − optimal strategy for player I. So we have to show $v(\pi_1^N, \pi_2) \geqslant \frac{1}{N+1}$ for all strategies π_2 for player II. To do this we fix an arbitrary strategy π_2 for player II. The random variables defined below are supposed to correspond to π_1^N and this π_2. Let the random variable X denote the number of decision epochs before player I chooses action 2. For each $m \in \mathbb{N}$ define the event $K(m)$ by $K(m) := \{X \geqslant m$, or $X < m$ and $b_{X+1} = 1\}$. So $K(m)$ is the event that at decision epoch m: either player I has not yet chosen his second row, or he did choose his second row and was lucky in receiving 1 unit. In other words $K(m)$ is the event that the total reward up to decision epoch m is non-negative.

Let $P_N(K(m))$ be the probability that $K(m)$ occurs. In the following lemma we relate the probability of $K(m+1)$ under π_1^N with the probability of $K(m)$ under π_1^{N-1} and π_1^{N+1} respectively.

Note that $P_N(K(m+1)) = P_N(X = 0 \text{ and } b_1 = 1) + P_N(X \geqslant m+1$, or $1 \leqslant X < m+1$ and $b_{X+1} = 1 | b_1 = 1) + P_N(X \geqslant m+1$, or $1 \leqslant X < m+1$ and $b_{X+1} = 1 | b_1 = 2)$.

Lemma 4

(i) $P_N(X \geq m+1, \text{ or } 1 \leq X < m+1 \text{ and } b_{X+1} = 1 | b_1 = 1) = (1-p(N))P_{N-1}(K(m))$.

(ii) $P_N(X \geq m+1, \text{ or } 1 \leq X < m+1 \text{ and } b_{X+1} = 1 | b_1 = 2) = (1-p(N))P_{N+1}(K(m))$.

Proof. We only proof *(i)* as the proof of *(ii)* is similar. Observe that in the left hand side of *(i)* the event $X = 0$ is excluded. Given $b_1 = 1$, making a choice at decision epoch $n+1$, with some history $(1, b_2, \ldots, b_n)$ according to π_1^N yields the same (randomized) choice as making a decision on decision epoch n with history $(b_2, b_3, \ldots, b_n)$ according to π_1^{N-1}. At every decision epoch, when the system is in state 1, the game can be considered as starting again. At decision epoch 1, player I, playing π_1^N chooses action 1 with probability $1-p(N)$, which then equals the survival chance at this epoch. Hence, given $b_1 = 1$, using π_1^N gives the same stochastic process as initially choosing action 1 with probability $1-p(N)$ and using π_1^{N-1} thereafter. So *(i)* holds. □

Consequently, $P_N(K_{m+1}) = P_N(X = 0, b_1 = 1) + (1-p(N))P_{N-1}(K(m)) + (1-p(N))P_{N+1}(K(m))$. The following lemma states that for all m and N the probability that the total reward up to decision epoch m is nonnegative, is at least $N/2(N+1)$.

Lemma 5. *$P_N(K(m)) \geq N/2(N+1)$ for all $m \in \mathbb{N}$ and all nonnegative N.*

Proof. The proof proceeds by induction.

(a) Take $m = 1$.

If $b_1 = 1$, then $P_N(X \geq 1 | b_1 = 1) = 1-p(N)$ and

$P_N(X < 1 \text{ and } b_{X+1} = 1 | b_1 = 1) = p(N)$.

So $P_N(K(1) | b_1 = 1) = 1 \geq N/2(N+1)$.

If $b_1 = 2$, then $P_N(K(1) | b_1 = 2) = P_N(X \geq 1 | b_1 = 2) = 1-p(N) \geq N/2(N+1)$. Let p_1 be the probability with which player II will choose his action 1 at decision epoch 1, then

$$P_N(K(1)) = p_1 P_N(K(1) | b_1 = 1) + (1-p_1)P_N(K(1) | b_1 = 2) \geq p_1(N/2(N+1)) + (1-p_1)(N/2(N+1)) = N/2(N+1).$$

(b) Now suppose $P_N(K(m)) \geq N/2(N+1)$ for some $m \in \mathbb{N}$ and all nonnegative N. Then in view of Lemma 4:

$$P_N(K(m+1) | b_1 = 1) = P_N(X = 0 \text{ and } b_1 = 1 | b_1 = 1) + P_N(X \geq m+1, \text{ or } 1 \leq X < m+1 \text{ and } b_{X+1} = 1 | b_1 = 1) = p(N) + (1-p(N))P_{N-1}(K(m)) \geq p(N) + (1-p(N))(N-1)/2N = N/2(N+1).$$

Also in view of Lemma 4:

$$P_N(K(m+1) | b_1 = 2) = P_N(X = 0 \text{ and } b_1 = 1 | b_1 = 2) + P_N(X \geq m+1, \text{ or } 1 \leq X < m+1 \text{ and } b_{X+1} = 1 | b_1 = 2) = 0 + (1-p(N))P_{N+1}(K(m)) \geq (1-p(N))(N+1)/2(N+2) = N/2(N+1).$$

Hence $P_N(K(m+1)) = p_1 P_N(K(m+1) | b_1 = 1) + (1-p_1)P_N(K(m+1) | b_1 = 2) \geq N/2(N+1)$, which shows that the lemma is also true for $m+1$. □

The next lemma demonstrates that π_1^N guarantees a total reward of at least $-1/(N+1)$ for each π_2 for which player I, using strategy π_1^N, will choose his second row with probability 1.

Lemma 6. *If* $\lim_{m \to \infty} P_N(X \geq m) = 0$, *then* $v(\pi_1^N, \pi_2) \geq -1/(N+1)$.

Proof. Since by definition

$$P_N(K(m)) = P_N(X \geq m) + P_N(X < m \text{ and } b_{X+1} = 1),$$

we derive from Lemma 5 in view of the assumption of Lemma 6:

$$\lim_{m \to \infty} P_N(K(m)) = P_N(b_{X+1} = 1) \geq N/2(N+1).$$

Player I will surely choose action 2 some time and the total reward is solely determined whether player II plays his first or his second action at that stochastic moment $X+1$ (payoffs until decision epoch $X+1$ sum up to zero). Then

$$v(\pi_1^N, \pi_2) = P_N(b_{X+1} = 1) \cdot 1 + P_N(b_{X+1} = 2) \cdot (-1) = 2P_N(b_{X+1} = 1) - 1 \geq N/(N+1) - 1 = -1/(N+1).$$ □

Notice that, if for a certain n, $k_n = -N$, then player I will choose action 2 with probability 1 at decision epoch $n+1$, which moves the system to state 2 where we can consider the game to be finished. Therefore $k_n \geq -N$ as long as the game has survived.

Lemma 7. *For any realisation of the stochastic process associated to π_1^N and π_2, for which player I never chooses action 2, it holds that the corresponding total reward is at least zero.*

Proof. Since in this case $k_n > -N$ for every $n \in \mathbb{N}$, we have $\Sigma_{t=1}^{T} \Sigma_{n=1}^{t} r_n > -2N$ for every $T \in \mathbb{N}$ from which follows that $\liminf_{T\to\infty} \frac{1}{T} \Sigma_{t=1}^{T} \Sigma_{n=1}^{t} r_n \geq 0$. □

Let $\lambda(m) := P_N(X < m \text{ and } b_{X+1} = 1)$ and $\mu(m) := P_N(X < m \text{ and } b_{X+1} = 2)$. Since $\lambda(m)$ and $\mu(m)$, $m = 1, 2, \ldots$, are bounded monotone increasing sequences we can define $\lambda = \lim_{m\to\infty} \lambda(m)$ and $\mu = \lim_{m\to\infty} \mu(m)$. The next lemma states that π_1^N is $\frac{1}{N+1}$ − good against π_2, when the probability that player I never chooses action 2 is positive.

Lemma 8. *If* $\lim_{N\to\infty} P_N(X \geq m) > 0$, *then* $v(\pi_1^N, \pi_2) \geq -1/(N+1)$.

Proof. The probability that player I will ever choose action 2 equals $\lambda + \mu$ and $1 - \lambda - \mu$ is the probability that the system never reaches state 2. Then by Lemma 7 and the definitions of λ and μ:

$$v(\pi_1^N, \pi_2) \geq \lambda \cdot 1 + \mu \cdot (-1) + (1 - \lambda - \mu) \cdot 0 = \lambda - \mu.$$

If we prove $\lambda - \mu \geq -1/(N+1)$ the proof is done.

Now for $m \in \mathbb{N}$ define strategy π_2^m for player II by: use π_2 up to decision epoch m, choose action 1 with probability 1/2 and action 2 with probability 1/2 at each decision epoch thereafter.

Observe that π_2^m will give rise to sequences $(b_1, b_2, \ldots)$ such that with probability 1 $k_n = -N$ for some n. Then the condition of Lemma 6 applies (where P_N now refers to π_1^N and π_2^m). Hence $v(\pi_1^N, \pi_2^m) \geq -1/(N+1)$ for all $m \in \mathbb{N}$. On the other hand: stopping before m contributes $\lambda(m) \cdot 1 + \mu(m) \cdot (-1)$ to the total reward and stopping on m or thereafter contributes $(1 - \lambda(m) - \mu(m)) \cdot 0$ (cf. Lemma 1). Summing up leads to $\lambda(m) - \mu(m) \geq -1/(N+1)$ for each $m \in \mathbb{N}$. So $\lambda - \mu \geq -1/(N+1)$, which finishes the proof. □

An immediate consequence of the Lemma's 6 and 8 is the following:

Corollary 9. $\sup_{\pi_1} \inf_{\pi_2} v(\pi_1, \pi_2) \geq 0$.

Now we have enough tools to prove our main theorem, stated in Sect. 4.

Proof. Part *(i)* follows from the Corollaries 2 and 9.

Part *(ii)* follows from Lemma 1.

Part *(iii)* follows from the Lemma's 6 and 8.

Part *(iv)* follows from Lemma 3.

We now prove *(v)*. Let π_1 be any strategy for player I.

Case 1. Suppose there is some strategy π_2 for player II such that for some decision epoch n a history h_{n-1} up to stage n occurs with positive probability, say q_n, such that for this history π_1 prescribes player I to choose his second action with positive probability. Let, for this π_2, m be the first decision epoch for which player I's second action can occur with positive probability, say $\epsilon > 0$. Now, define π_2^m as: up to decision epoch m, π_2^m equals π_2, at decision epoch m choose action 2 and always after decision epoch m choose action 1 with probability 1/2. It can be verified that $v(\pi_1, \pi_2^m) = -\epsilon q_m < 0$.

Case 2. If such a strategy π_2 does not exist, then apparently player I always chooses action 1. If $\hat{\pi}_2$ is the strategy for player II defined by always choosing action 1, then $v(\pi_1, \hat{\pi}_2) = -1$. □

References

1. Bewley T, Kohlberg E (1976) The asymptotic theory of stochastic games. Math Oper Res 1:197–208
2. Blackwell D, Ferguson TS (1968) The big match. Ann Math Statist 39:159–163
3. Gillette D (1957) Stochastic games with zero stop probabilities. In: Dresher M, Tucker AW, Wolfe P (eds) Contributions to the theory of games 3. Annals of Mathematics Studies, Princeton University Press, Princeton, pp 179–187
4. Mertens JF, Neyman A (1981) Stochastic games. Int J Game Theory 10:53–66
5. Neyman A (1986) Private communication
6. Shapley LS (1953) Stochastic games. Proc Natl Acad Sci USA 39:1095–1100
7. Vrieze OJ (1983) Stochastic games with finite state and action spaces. PhD dissertation, Math Centre, Amsterdam