JOURNAL OF OPTIMIZATION THEORY AND APPLICATIONS: Vol. 105, No. 2, pp. 371–389, MAY 2000

Almost Stationary ϵ-Equilibria in Zero-Sum Stochastic Games

J. FLESCH,[1] F. THUIJSMAN,[2] AND O. J. VRIEZE[3]

Communicated by G. P. Papavassilopoulos

Abstract. We show the existence of almost stationary ϵ-equilibria, for all $\epsilon > 0$, in zero-sum stochastic games with finite state and action spaces. These are ϵ-equilibria with the property that, if neither player deviates, then stationary strategies are played forever with probability almost 1. The proof is based on the construction of specific stationary strategy pairs, with corresponding rewards equal to the value, which can be supplemented with history-dependent δ-optimal strategies, with small $\delta > 0$, in order to obtain almost stationary ϵ-equilibria.

Key Words. Zero-sum stochastic games, limiting average rewards, equilibria.

1. Introduction

We deal with zero-sum stochastic games with finite state and action spaces. These games model conflict situations in which two players are involved with completely opposite interests. Such a game Γ can be seen as a finite collection of matrices $(M_s)_{s \in S}$, corresponding to states s in the state space S, where entry (i_s, j_s) of M_s consists of a payoff $r(s, i_s, j_s) \in \mathbb{R}$ and $\mathbb{N}$ in the following way. The play starts at stage 1 in an initial state, say in state $s^1 \in S$, where simultaneously and independently both players are to choose an action: player 1 chooses a row $i_{s^1}^1$ of M_{s^1}, while player 2 chooses a column $j_{s^1}^1$ of M_{s^1}. These choices induce an immediate payoff $r(s^1, i_{s^1}^1, j_{s^1}^1)$ to player 1 from player 2. Next, the play moves to a new state according to the probability vector $p(s^1, i_{s^1}^1, j_{s^1}^1)$, say to state s^2. At stage 2, new actions

[1]Postdoctoral Fellow, Department of Mathematics, Maastricht University, Maastricht, Netherlands.
[2]Associate Professor, Department of Mathematics, Maastricht University, Maastricht, Netherlands.
[3]Professor, Department of Mathematics, Maastricht University, Maastricht, Netherlands.

0022-3239/00/0500-0371$18.00/0

$i^2_{s^2}$ and $j^2_{s^2}$ are to be chosen by the players in state s^2. Then, player 1 receives the corresponding payoff $r(s^2, i^2_{s^2}, j^2_{s^2})$ from player 2 and the play moves to some state s^3 according to the probability vector $p(s^2, i^2_{s^2}, j^2_{s^2})$, and so on.

The sequence $(s^1, i^1_{s^1}, j^1_{s^1}, \ldots, s^{n-1}, i^{n-1}_{s^{n-1}}, j^{n-1}_{s^{n-1}}, s^n)$ is called the history up to stage n. The players are assumed to have complete information and perfect recall.

The respective sets of actions in state s will be denoted by I_s and J_s. A mixed action for a player in state s is a probability distribution on the set of his actions in state s. Mixed actions in state s will be denoted by x_s for player 1 and by y_s for player 2; the sets of mixed actions in state s are denoted by X_s and Y_s, respectively. A strategy is a decision rule that prescribes a mixed action in the current state for any past history of the play. Such general strategies, so-called history-dependent strategies, will be denoted by π for player 1 and by σ for player 2. If, for all histories, the mixed actions prescribed by a strategy depend only on the current state, then the strategy is called stationary. Thus, the stationary strategy spaces are simply $X := \times_{s\in S} X_s$ for player 1 and $Y := \times_{s\in S} Y_s$ for player 2. We will use the notations x and y for the stationary strategies of the respective players.

A strategy pair (π, σ) together with an initial state s determines a stochastic process on the payoffs. The sequences of payoffs are evaluated by the limiting average reward,

$$\gamma(s, \pi, \sigma) := \liminf_{N\to\infty} \mathbb{E}_{s\pi\sigma}\left((1/N)\sum_{n=1}^{N} r_n\right) = \liminf_{N\to\infty} \mathbb{E}_{s\pi\sigma}(R_N),$$

where r_n denotes the random variable for the payoff at stage n and R_N denotes the random variable for the average payoff up to stage N. We will also use the vector notation

$$\gamma(\pi, \sigma) := (\gamma(s, \pi, \sigma))_{s\in S}.$$

We assume that player 1 is trying to maximize γ, while player 2 wishes to minimize γ.

For $x_s \in X_s$ and $y_s \in Y_s$, let

$$r(s, x_s, y_s) := \sum_{i_s\in I_s, j_s\in J_s} x_s(i_s) y_s(j_s) \cdot r(s, i_s, j_s),$$

$$p(t|s, x_s, y_s) := \sum_{i_s\in I_s, j_s\in J_s} x_s(i_s) y_s(j_s) \cdot p(t|s, i_s, j_s).$$

For $x \in X$ and $y \in Y$, we will also use the vector notation

$$r(x, y) := (r(s, x_s, y_s))_{s\in S}.$$

A pair of stationary strategies (x, y) determines a Markov chain with transition matrix P_{xy} on S, where entry (s, t) of P_{xy} is $p(t|s, x_s, y_s)$. With respect to this Markov chain, we can speak of transient states and recurrent states, and we can group the recurrent states into minimal closed sets, so-called ergodic sets. As in Ref. 1, let

$$Q_{xy} := \lim_{N\to\infty} (1/N) \sum_{n=1}^{N} (P_{xy})^n.$$

Entry (s, t) of the stochastic matrix Q_{xy}, denoted by $q(t|s, x, y)$, is the expected average number of stages the process is in state t when starting in s. We have

$$\gamma(x, y) = Q_{xy}\gamma(x, y) = Q_{xy}r(x, y). \tag{1}$$

Against a fixed stationary strategy y, there always exists a stationary best reply x of player 1 (cf. Ref. 2); i.e.,

$$\gamma(x, y) \geq \gamma(\pi, y), \qquad \forall \pi.$$

Obviously, a similar statement holds for the best replies of player 2.

In Ref. 3, it is shown that

$$\sup_{\pi} \inf_{\sigma} \gamma(s, \pi, \sigma) = \inf_{\sigma} \sup_{\pi} \gamma(s, \pi, \sigma) =: v_s, \qquad \forall s \in S. \tag{2}$$

Here, $v := (v_s)_{s\in S}$ is called the limiting average value and v satisfies

$$v_s = \mathrm{Val}\,(A_s), \qquad \forall s \in S, \tag{3a}$$

$$\text{where } A_s := \left[\sum_{t\in S} p(t|s, i_s, j_s) v_t \right]_{i_s \in I_s, j_s \in J_s}, \tag{3b}$$

and where Val stands for the matrix game value; see for example Ref. 4, page 112. In a stochastic game, a strategy π of player 1 is called ϵ-optimal, $\epsilon \geq 0$, for initial state $s \in S$ if

$$\gamma(s, \pi, \sigma) \geq v_s - \epsilon, \qquad \forall \sigma.$$

If π is ϵ-optimal for all initial states in S, then π is called ϵ-optimal. 0-optimal strategies will be simply called optimal. For the strategies of player 2, ϵ-optimality is defined analogously. Although for all $\epsilon > 0$, by (2) there exist ϵ-optimal strategies for both players, the famous big match example of Gillette (Ref. 5), examined by Blackwell and Ferguson (Ref. 6), demonstrates that in general the players need not have optimal strategies and that, for achieving ϵ-optimality, history-dependent strategies are indispensable. For $\epsilon \geq 0$, any pair of $\epsilon/2$-optimal strategies (π, σ) for the respective players forms a ϵ-equilibrium, which means that neither player can gain more than

ϵ by unilateral deviation, i.e.,

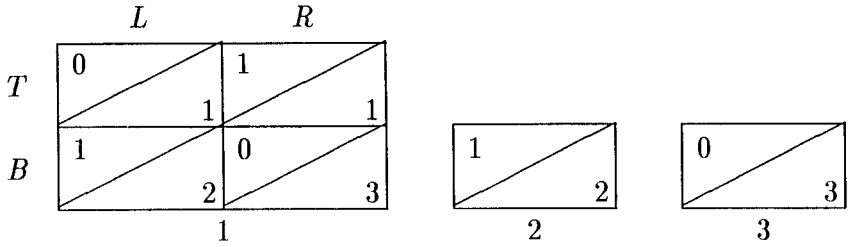
$$\gamma(s, \bar{\pi}, \sigma) - \epsilon \leq \gamma(s, \pi, \sigma) \leqq \gamma(s, \pi, \bar{\sigma}) + \epsilon, \qquad \forall \bar{\pi}, \forall \bar{\sigma}, \forall s \in S.$$

The concept of ϵ-equilibria, as pairs of ϵ-best replies to each other, is used for general stochastic games as well. However, for general nonzero-sum stochastic games, existence of ϵ-equilibria is known only on condition of specifically structured payoff and transition functions. Often, these ϵ-equilibria are stationary or almost stationary. An ϵ-equilibrium is called stationary if it consists of stationary strategies, while it is called almost stationary when it has the following property: if neither player deviates, then with probability almost 1, play develops as if stationary strategies are played at all stages; see Refs. 7–9.

Generally speaking, while playing an ϵ-equilibrium requires both players to play, at each state and stage, new history-dependent mixed actions, the structure of an almost stationary ϵ-equilibrium is substantially simpler. For an almost stationary ϵ-equilibrium, the players play history-independent mixed actions as long as they do not detect any deviation by their opponent. Only in the latter case, which occurs with probability close to 0, the player has to turn to a complex history-dependent strategy. Thus, only in the performance while playing, the almost stationary ϵ-equilibrium is simpler than other history-dependent ϵ-equilibria; in view of computational purposes, it is not any simpler. As we have mentioned above, when dealing with zero-sum stochastic games, the existence of ϵ-equilibria is not an issue here, since any pair of ϵ-optimal strategies yields a 2ϵ-equilibrium.

To illustrate the issue, we now discuss briefly the big match.

Example 1.1. Consider the game represented by



This game, known as the "big match", was introduced in Ref. 5. Player 1 chooses rows and player 2 chooses columns; the number in the up-left corner of an entry is the payoff to player 1 by player 2 whenever this entry is selected; the number in the down-right corner of an entry indicates the

new state where play will continue at the next stage. Of course, the interesting initial state is state 1. In Ref. 6, it is shown that the value of this game is $1/2$, that the stationary strategy $y = (1/2, 1/2)$ is optimal for player 2, while player 1 can only achieve ϵ-optimality by history-dependent strategies. An example of an ϵ-optimal strategy $\pi(\epsilon)$ for player 1 is to play B at stage n when player 2 has chosen Left $L(n)$ times and Right $R(n)$ times, with probability $\epsilon^2(1-\epsilon)^{a(n)}$, where

$$a(n) = \max\{R(n) - L(n), 0\}.$$

This pair of strategies $(\pi(\epsilon), y)$ would be an ϵ-equilibrium that yields precisely $1/2$ to player 1. Instead of achieving $1/2$ through these complicated strategies, the players could play the almost stationary ϵ-equilibrium $(x(\epsilon), y)$, where $x(\epsilon)$ is the following strategy: play Top unless at some stage in the far future you notice that the player 2 action frequencies are not sufficiently close to $(1/2, 1/2)$; in that case, immediately start playing $\pi(\epsilon)$.

In Section 3, we present the construction of stationary strategy pairs satisfying special properties, which enable us to supplement them with history-dependent δ-optimal strategies with small $\delta > 0$, as threat strategies according to the general terminology, to eliminate the profitability of possible deviations from these stationary strategies.

2. Almost Stationary ϵ-Equilibria

First, we define formally the concept of almost stationary ϵ-equilibria.

Definition 2.1. An ϵ-equilibrium (π, σ) is an almost stationary ϵ-equilibrium, $\epsilon \geq 0$, if there exists a pair of stationary strategies (x^*, y^*) such that, when playing (π, σ), the players will use the mixed actions corresponding to the stationary strategies (x^*, y^*) at all stages with probability at least $1 - \epsilon$.

Note that, although ϵ has two different roles in this definition, it will lead to no confusion. Now, we are ready to state the main result of this paper.

Theorem 2.1. In every zero-sum stochastic game, for any $\epsilon > 0$, there exists an almost stationary ϵ-equilibrium.

The proof will be based on the construction of specific stationary strategy pairs with rewards equal to the value. In order to force the players to

play these stationary strategies, as a standard tool, the players will use statistical tests on past action frequencies of their opponents to detect deviations with probability almost 1. If a deviation is detected, then a history-dependent δ-optimal strategy has to be played in the future, where $\delta > 0$ is sufficiently small. The role of these δ-optimal strategies is to rule out the profitability of possible deviations of the players.

2.1. Preliminaries. Before we turn to the construction, we recall some basic concepts from matrix game theory and we derive some preliminary results. For $x \in X$ and $y \in Y$, let

$$V_s(x_s, y_s) := \sum_{t \in S} p(t|s, x_s, y_s) v_t, \qquad V(x, y) := (V_s(x_s, y_s))_{s \in S}.$$

In words, $V_s(x_s, y_s)$ is the expected value after transition from state s with respect to (x_s, y_s). For $s \in S, x_s \in X_s, y_s \in Y_s$, let $L_s(x_s, y_s)$ be the probability that, after transition from state s with respect to (x_s, y_s), the new value v_t is different from v_s, so

$$L_s(x_s, y_s) := \sum_{t \in S, v_t \neq v_s} p(t|s, x_s, y_s);$$

if $v_t = v_s$ for all $t \in S$, then $L_s(x_s, y_s)$ is defined to be equal to 0. Obviously, $V_s(x_s, y_s) \neq v_s$ implies $L_s(x_s, y_s) > 0$. The next lemma states, with respect to (x, y), that if the value does not change in expectation under transitions, then the value is a constant on each set of states that is ergodic with respect to (x, y).

Lemma 2.1. Let $(x, y) \in X \times Y$ satisfy $V(x, y) = v$. Suppose that E is an ergodic set with respect to (x, y). Then, $v_s = v_t$ for all $s, t \in E$, and $L_s(x_s, y_s) = 0$ for all $s \in E$.

Proof. Let

$$\bar{E} := \left\{ s \in E \;\middle|\; v_s = \max_{t \in E} v_t \right\}.$$

Using $V(x, y) = v$ and the fact that E is an ergodic set for (x, y), we obtain

$$\begin{aligned} v_s = V_s(x_s, y_s) &= \sum_{t \in S} p(t|s, x_s, y_s) v_t \\ &= \sum_{t \in E} p(t|s, x_s, y_s) v_t, \qquad \forall s \in \bar{E}; \end{aligned}$$

thus, $\bar{E} \subset E$ is a closed set of states for (x, y). Therefore, $\bar{E} = E$, which implies

$$v_s = v_t, \qquad \text{for all } s, t \in E.$$

Now,

$$L_s(x_s, y_s) = 0, \qquad \text{for all } s \in E,$$

follows from the definition of $L_s(x_s, y_s)$. □

For $s \in S$, let

$$X'_s := \{x_s \in X_s \mid V_s(x_s, y_s) \geq v_s, \forall y_s \in Y_s\}, \qquad X' := \times_{s \in S} X'_s,$$

$$Y'_s := \{y_s \in Y_s \mid V_s(x_s, y_s) \leq v_s, \forall x_s \in X_s\}, \qquad Y' := \times_{s \in S} Y'_s,$$

$$\bar{I}_s := \{i_s \in I_s \mid V_s(i_s, y_s) = v_s, \forall y_s \in Y'_s\}, \qquad \bar{X}_s := \text{conv}(\bar{I}_s), \qquad \bar{X} := \times_{s \in S} \bar{X}_s,$$

$$\bar{J}_s := \{j_s \in J_s \mid V_s(x_s, j_s) = v_s, \forall x_s \in X'_s\}, \qquad \bar{Y}_s := \text{conv}(\bar{J}_s), \qquad \bar{Y} := \times_{s \in S} \bar{Y}_s,$$

where conv stands for the convex hull of a set. The sets X'_s and Y'_s are the respective sets of optimal mixed actions of the players in the matrix game A_s [see (3)], while the elements of $\bar{I}_s$ and $\bar{J}_s$ are so-called equalizers in the matrix game A_s. It is well known in matrix game theory that, for all $s \in S$, the sets X'_s and Y'_s are nonempty polytopes and also that, for all $s \in S$, $x_s \in \text{Int}(X'_s)$, $y_s \in \text{Int}(Y'_s)$, where Int stands for the relative interior of a set, we have

$$\bar{I}_s = \{i_s \in I_s \mid x_s(i_s) > 0\}, \qquad \bar{J}_s = \{j_s \in J_s \mid y_s(j_s) > 0\}. \tag{4}$$

The next lemma provides sufficient conditions for $\bar{X}_s = X'_s$ or $\bar{Y}_s = Y'_s$ in some state s.

Lemma 2.2. Let $s \in S$. If $L_s(x_s, j_s) > 0$ implies $V_s(x_s, j_s) > v_s$ for all $(x_s, j_s) \in X'_s \times J_s$, then $\bar{X}_s = X'_s$. Similarly, if $L_s(i_s, y_s) > 0$ implies $V_s(i_s, y_s) < v_s$ for all $(i_s, y_s) \in I_s \times Y'_s$, then $\bar{Y}_s = Y'_s$.

Proof. We only show the first part; the proof of the second part is similar. By (4), we have $\bar{X}_s \supset X'_s$. It remains to verify that $\bar{X}_s \subset X'_s$. Since X'_s is convex, it is sufficient to show that $i_s \in X'_s$ for all $i_s \in \bar{I}_s$. Take an arbitrary $i_s \in \bar{I}_s$. Using the compactness of X'_s, there exists an $\hat{x}_s \in X'_s$ satisfying

$$\hat{x}_s(i_s) \geq x_s(i_s), \qquad \text{for all } x_s \in X'_s.$$

By the condition, we have that

$$L_s(\hat{x}_s, j_s) > 0 \text{ implies } V_s(\hat{x}_s, j_s) > v_s, \qquad \text{for all } j_s \in J_s;$$

therefore, using $\hat{x}_s \in X'_s$, we obtain that

$$((1 - \lambda) \cdot \hat{x}_s + \lambda \cdot i_s) \in X'_s, \qquad \text{for small } \lambda > 0.$$

By the choice of $\hat{x}_s$, we must have $\hat{x}_s = i_s$, thus $i_s \in X'_s$. □

In Ref. 10, it is shown that, in every zero-sum game, there exists an initial state s_1 in

$$S^{\max} := \left\{ s \in S \,\middle|\, v_s = \max_{t \in S} v_t \right\}$$

for which player 1 has a stationary optimal strategy x^1; similarly, there exists an initial state s_2 in

$$S^{\min} := \left\{ s \in S \,\middle|\, v_s = \min_{t \in S} v_t \right\}$$

for which player 2 has a stationary optimal strategy y^2. Obviously, the strategy x^1 must keep the play in $S^{\max}$ with probability 1 when starting in s^1, and the strategy y^2 must keep the play in $S^{\min}$ with probability 1 when starting in s^2. Hence, if we take stationary best replies y^1 against x^1 and x^2 against y^2, we obtain the following result.

Lemma 2.3. There exist stationary strategy pairs (x^1, y^1), (x^2, y^2) and corresponding ergodic sets E^1, E^2 such that

$$E^1 \subset S^{\max} := \left\{ s \in S \,\middle|\, v_s = \max_{t \in S} v_t \right\}, \qquad \gamma(s, x^1, y^1) = v_s, \qquad \forall s \in E^1,$$

$$E^2 \subset S^{\min} := \left\{ s \in S \,\middle|\, v_s = \min_{t \in S} v_t \right\}, \qquad \gamma(s, x^2, y^2) = v_s, \qquad \forall s \in E^2.$$

Suppose that $E \subset S$ is an ergodic set with respect to $(x', y') \in \mathrm{Int}(X') \times \mathrm{Int}(Y')$ and also that

$$\bar{X}_s = X'_s, \; \bar{Y}_s = Y'_s, \qquad \text{for all } s \in E.$$

Then, we may define a restricted game $\bar{\Gamma}_E$ where the state space is E and the players are restricted to use strategies that prescribe only actions in $\bar{I}_s$ and $\bar{J}_s$, if the play is in state $s \in E$. Obviously, this restricted game $\bar{\Gamma}_E$ is a well-defined stochastic game as well. Let $\bar{v}$ denote the value of the restricted game $\bar{\Gamma}_E$. Observe that, for the original value, by Lemma 2.1, we have

$$v_s = v_t =: v_E, \qquad \text{for all } s, t \in E.$$

The following result follows from Ref. 11.

Lemma 2.4. Suppose that $\bar{v}_s \geq v_E$ for all $s \in E$ or $\bar{v}_s \leq v_E$ for all $s \in E$. Then, there exists a state $s \in E$ such that $\bar{v}_s = v_E$.

Note that the value of the restricted game $\bar{v}_s, s \in E$, does not need to be equal to the original value v_E in all states in E, not even under the above condition, which will be demonstrated by Example 2.1 below.

2.2. Construction. Fix arbitrary $x' \in \mathrm{Int}(X')$ and $y' \in \mathrm{Int}(Y')$. We keep x' and y' fixed for the rest of this section. Let T denote the set of transient states, and let $\mathcal{R}$ be the set of ergodic sets with respect to (x', y'). Since any stationary strategy pair induces at least one ergodic set, we have $\mathcal{R} \neq \emptyset$. Now, we divide $\mathcal{R}$ into three parts. Let

$$\mathcal{R}^1 := \{E \in \mathcal{R} \mid \exists s \in E, \exists (i_s, y_s) \in I_s \times Y'_s : V_s(i_s, y_s) = v_s, L_s(i_s, y_s) > 0\},$$

$$\mathcal{R}^2 := \{E \in \mathcal{R} \setminus \mathcal{R}^1 \mid \exists s \in E, \exists (x_s, j_s) \in X'_s \times J_s : V_s(x_s, j_s) = v_s, L_s(x_s, j_s) > 0\},$$

$$\mathcal{R}^3 := \mathcal{R} \setminus (\mathcal{R}^1 \cup \mathcal{R}^2).$$

Note that the sets $T, \mathcal{R}^1, \mathcal{R}^2, \mathcal{R}^3$ are independent of the particular choices of $x' \in \mathrm{Int}(X')$ and $y' \in \mathrm{Int}(Y')$, because for all such choices the set of actions that get weight 0 is the same; therefore, these choices all yield the same Markov chain structure; the only differences can be found in the positive transition weights. Also note that $\mathcal{R}^1$ is the set of ergodic sets E with respect to (x', y') for which there exists a pair of mixed actions in some state $s \in E$ such that player 1 plays a pure action, player 2 plays optimally in the matrix game A_s, and the expected value after transition equals the original value, but with a positive probability a transition occurs to a state where the value is different. The intuition behind $\mathcal{R}^2$ is analogous. The partition of $\mathcal{R}$ induces naturally the following partition of $S \setminus T$:

$$S^1 := \bigcup_{E \in \mathcal{R}^1} E, \qquad S^2 := \bigcup_{E \in \mathcal{R}^2} E, \qquad S^3 := \bigcup_{E \in \mathcal{R}^3} E.$$

If $\mathcal{R}^1 \cup \mathcal{R}^2 \neq \emptyset$, then by the definitions of $\mathcal{R}^1$ and $\mathcal{R}^2$ there exists a non-empty set $S^* \subset S^1 \cup S^2$, which contains precisely one state from each ergodic set in $\mathcal{R}^1 \cup \mathcal{R}^2$ such that, for all $s \in S^* \cap S^1$, there exists a pair $(i_s^*, y_s^*) \in I_s \times Y'_s$ satisfying

$$V_s(i_s^*, y_s^*) = v_s, \qquad L_s(i_s^*, y_s^*) > 0,$$

and for all $s \in S^* \cap S^2$ there exists a pair $(x_s^*, j_s^*) \in X'_s \times J_s$ satisfying

$$V_s(x_s^*, j_s^*) = v_s, \qquad L_s(x_s^*, j_s^*) > 0.$$

In fact, these states and pairs of mixed actions provide the possibility to leave all the ergodic sets belonging to $\mathcal{R}^1$ and $\mathcal{R}^2$ in such a way that the value does not change in expectation. The construction of strategies with this property (of leaving sets in $\mathcal{R}^1$ and $\mathcal{R}^2$ in a satisfactory way), requires

only an adaptation of (x', y') in states belonging to S^*. Then, for these adapted strategies, the only recurrent states that remain belong to S^3. So, if we can further adapt the strategies such that, in each state of S^3, the players can achieve precisely the value as a reward, then for all initial states the reward equals the value. Therefore, we now turn our attention to R^3.

Using $S^3 \cap (S^1 \cup S^2) = \emptyset$, by Lemma 2.2 we have

$$\bar{X}_s = X'_s, \qquad \bar{Y}_s = Y'_s, \qquad \text{for all } s \in S^3.$$

Assume that $E \in R^3$ (in fact, we will show later that R^3 is always nonempty). As in the preliminaries, we may define a restricted game $\bar{\Gamma}_E$. Clearly, in this restricted game the respective stationary strategy spaces are

$$\bar{X}_E := \times_{s \in E} \bar{X}_s, \qquad \bar{Y}_E := \times_{s \in E} \bar{Y}_s.$$

We use $\bar{v}_s$, $s \in E$, for the value of the restricted game $\bar{\Gamma}_E$. Recall that, for the original value, we have

$$v_s = v_t =: v_E, \qquad \text{for all } s, t \in E.$$

We now show the existence of stationary strategy pairs in $\bar{\Gamma}_E$ with rewards equal to the original value v_E.

Lemma 2.5. Let $E \in R^3$. There exists a stationary strategy pair $(\bar{x}, \bar{y}) \in \bar{X}_E \times \bar{Y}_E$ such that $\gamma(s, \bar{x}, \bar{y}) = v_E$ for all $s \in E$.

Proof. We distinguish two essentially different cases.

Part 1. Assume that

$$\bar{v}_s \geq v_E, \qquad \text{for all } s \in E;$$

the case $\bar{v}_s \leq v_E$, for all $s \in E$, is similar. Lemma 2.4 implies that there exists a state $s \in E$ such that $\bar{v}_s = v_E$. Let

$$E^{\min} := \{t \in E \mid \bar{v}_t = v_E\}.$$

Let $(x^2, y^2) \in \bar{X}_E \times \bar{Y}_E$ and let $E^2 \subset E^{\min}$ in $\bar{\Gamma}_E$ as in Lemma 2.3. So, we have

$$\gamma(s, x^2, y^2) = v_E, \qquad \text{for all } s \in E^2.$$

For $s \in E$, let

$$\bar{x}_s := \begin{cases} x_s^2, & \text{if } s \in E^2, \\ x'_s, & \text{if } s \in E \setminus E^2, \end{cases} \qquad \bar{y}_s := \begin{cases} y_s^2, & \text{if } s \in E^2, \\ y'_s, & \text{if } s \in E \setminus E^2. \end{cases}$$

The only ergodic set for $(\bar{x}, \bar{y}) \in \bar{X}_E \times \bar{Y}_E$ in the restricted game $\bar{\Gamma}_E$ is clearly E^2; hence, for any $s, t \in E$, we have that $q_E(t \mid s, \bar{x}, \bar{y}) > 0$ holds only if $t \in E^2$;

thus, (1) yields

$$\gamma(s, \bar{x}, \bar{y}) = v_E, \qquad \text{for all } s \in E.$$

Part 2. Assume that

$$\min_{s \in E} \bar{v}_s < v_E < \max_{s \in E} \bar{v}_s.$$

Take

$$(x^1, y^1) \in \bar{X}_E \times \bar{Y}_E, \qquad E^1 \subset E^{\max} := \left\{ s \in E \,\middle|\, \bar{v}_s = \max_{t \in E} \bar{v}_t \right\},$$

$$(x^2, y^2) \in \bar{X}_E \times \bar{Y}_E, \qquad E^2 \subset E^{\min} := \left\{ s \in E \,\middle|\, \bar{v}_s = \min_{t \in E} \bar{v}_t \right\},$$

in $\bar{\Gamma}_E$ as in Lemma 2.3. By the assumption, we have $E^1 \cap E^2 = \varnothing$. For $a, b \in (0, 1)$ and $s \in E$, let

$$(x_s^{ab}, y_s^{ab}) := \begin{cases} (a \cdot x_s^1 + (1-a) \cdot x_s', a \cdot y_s^1 + (1-a) \cdot y_s'), & \text{if } s \in E^1, \\ (b \cdot x_s^2 + (1-b) \cdot x_s', b \cdot y_s^2 + (1-b) \cdot y_s'), & \text{if } s \in E^2, \\ (x_s', y_s'), & \text{if } s \in E \setminus (E^1 \cup E^2). \end{cases}$$

Recall that we have fixed $x' \in \mathrm{Int}(X')$ and $y' \in \mathrm{Int}(Y')$, and also that $\bar{X}_s = X_s'$, $\bar{Y}_s = Y_s'$ for all $s \in E$. Notice that

$$x_s^{ab} \in \mathrm{Int}(\bar{X}_s), \qquad y_s^{ab} \in \mathrm{Int}(\bar{Y}_s), \qquad \text{for all } s \in E \text{ and } a, b \in (0, 1);$$

hence, the set E is ergodic for (x^{ab}, y^{ab}) for all $a, b \in (0, 1)$. Notice also that a and b control the respective expected lengths of periods when staying in E^1 and E^2. Since E is ergodic for (x^{ab}, y^{ab}) for all $a, b \in (0, 1)$, we have that

$$q_E(t \mid s_1, x^{ab}, y^{ab}) = q_E(t \mid s_2, x^{ab}, y^{ab}),$$
$$\text{for all } s_1, s_2, t \in E, a, b \in (0, 1);$$

thus, equality (1) implies that

$$\gamma(s, x^{ab}, y^{ab}) = \gamma(t, x^{ab}, y^{ab}) := \gamma_E^{ab}, \qquad \text{for all } s, t \in E, a, b \in (0, 1).$$

We show that there are $a, b \in (0, 1)$ such that $\gamma_E^{ab} = v_E$. Take arbitrary $a', b' \in (0, 1)$. If $\gamma_E^{a'b'} = v_E$, then we are done. So, assume without loss of generality that $\gamma_E^{a'b'} > v_E$ and consider $(x^{a'b}, y^{a'b})$. Observe that, the larger b we take, the more time the play spends in E^2. Thus, one can show that

$$\lim_{b \uparrow 1} \gamma_E^{a'b} = \min_{t \in E} \bar{v}_t < v_E.$$

By the continuity of $q(t|s, x^{ab}, y^{ab})$ and $r(s, x_s^{ab}, y_s^{ab})$ in $a, b \in (0, 1)$, where $s, t \in E$, using (1) we have that γ_E^{ab} is also continuous in $a, b \in (0, 1)$; hence, there is a b such that $\gamma_E^{a'b} = v_E$. □

Now we are ready to complete the construction based on the previously derived results. Recall that we have already fixed a pair of stationary strategies $(x', y') \in \text{Int}(X') \times \text{Int}(Y')$. For all ergodic sets $E \in \mathbf{R}^3$, let $(\bar{x}_s, \bar{y}_s) \in \bar{X}_s \times \bar{Y}_s$, $s \in E$, be as in Lemma 2.5. We define a stationary strategy pair as follows for all $\tau \in (0, 1)$:

$$(x_s^\tau, y_s^\tau) := \begin{cases} (\tau \cdot x_s' + (1-\tau) \cdot i_s^*, y_s^*), & \text{if } s \in S^* \cap S^1, \\ (x_s^*, \tau \cdot y_s' + (1-\tau) \cdot j_s^*), & \text{if } s \in S^* \cap S^2, \\ (\bar{x}_s, \bar{y}_s), & \text{if } s \in S^3, \\ (x_s', y_s'), & \text{otherwise.} \end{cases}$$

The next lemma shows that, for these stationary strategy pairs, the recurrent states all belong to S^3 and the reward equals the value for all initial states.

Lemma 2.6. For all $\tau \in (0, 1)$, we have $\gamma(x^\tau, y^\tau) = v$ and, if F is an ergodic set with respect to (x^τ, y^τ), then $F \subset S^3$.

Proof. Let $\tau \in (0, 1)$. By the definitions, we have

$$V_s(x_s^\tau, y_s^\tau) = v_s, \qquad \text{for all } s.$$

By Lemma 2.1, the value is a constant on each ergodic set for (x^τ, y^τ). By the construction of (x^τ, y^τ), in each state in S^* with positive probability a transition occurs to a state with a different value, so all recurrent states must belong to S^3.

The equality $V(x^\tau, y^\tau) = v$ implies

$$P_{x^\tau y^\tau} v = v;$$

hence, the definition of $Q_{x^\tau y^\tau}$ yields

$$Q_{x^\tau y^\tau} v = v.$$

For any $s \in S$, if $q(t|s, x^\tau, y^\tau) > 0$, then t belongs to an ergodic set with respect to (x^τ, y^τ), so we have $t \in S^3$. Now, the choice of $(\bar{x}_z, \bar{y}_z)$, $z \in S^3$, implies by Lemma 2.5 that

$$\gamma(t, x^\tau, y^\tau) = v_t, \qquad \text{for all } t \in S^3,$$

so applying (1) gives

$$\gamma(s, x^\tau, y^\tau) = \sum_{t \in S} q(t|s, x^\tau, y^\tau) \cdot \gamma(t, x^\tau, y^\tau)$$
$$= \sum_{t \in S^3} q(t|s, x^\tau, y^\tau) \cdot v_t = v_s, \qquad \forall s \in S,$$

which completes the proof. □

We now prove Theorem 2.1. We show that, for any $\epsilon > 0$, the stationary strategy pair (x^τ, y^τ), for sufficiently large $\tau \in (0, 1)$, can be supplemented with history-dependent δ-optimal strategies, for small $\delta > 0$, to obtain an almost stationary ϵ-equilibrium.

Proof of Theorem 2.1. We give only an outline of the proof, since the tools used are standard; see for example Refs. 7 and 12 or, in a more general fashion, Ref. 9. Let $\epsilon > 0$. We will define strategy pairs (π^τ, σ^τ) for all $\tau \in (0, 1)$ so that (π^τ, σ^τ) is an almost stationary ϵ-equilibrium for sufficiently large $\tau \in (0, 1)$. These strategy pairs will be constructed in such a way that, if neither player deviates, then the stationary strategy pair (x^τ, y^τ) is played forever with probability at least τ. In view of Lemma 2.6, this means that the corresponding rewards are converging to the value v as τ tends to 1. Hence, when verifying the ϵ-equilibrium conditions, it suffices to show that, for any initial state $s \in S$, player 1 cannot get more than $v_s + \epsilon/2$ and player 2 cannot decrease the reward below $v_s - \epsilon/2$ by unilateral deviations. The strategies π^τ and σ^τ will be defined analogously, so we focus only on the player 1 strategy π^τ and on the possible deviations of player 2. So, now we define π^τ for $\tau \in (0, 1)$.

The idea is that, while employing stationary strategies, player 1 checks the player 2 behavior during the play by employing statistical tests. These tests are based on the observation that, if player 2 truly uses his stationary strategy, then:

(i) if actions are chosen outside the support of the stationary strategy, they are observed immediately;

(ii) if play remains in the same ergodic set (ergodic w.r.t. these stationary strategies), then the action frequencies of player 2 should converge to the weights of the mixed actions corresponding to this stationary strategy;

(iii) from any transient state (w.r.t. the stationary strategies), the probability of remaining in the transient states longer than n stages converges to 0.

So, if player 2 chooses an action outside the support of y^τ, then player 1 knows for sure that player 2 deviated; if play remains in the set of transient states for longer than some specified number of stages, then player 2 has deviated with probability close to 1; if after some specified number of stages within an ergodic set the player 2 action frequencies are not within some specified range from the theoretical ones, then player 2 has deviated with probability close to 1. On condition that player 2 should play y^τ, if player 1 decides that with probability at least τ player 2 is deviating, then player 1 starts playing a $(1-\tau)$-optimal strategy. Note that these probabilities are conditioned on the initially given stationary strategies.

Deviations of Player 2 outside the Support. If player 2 chooses an action $j_s \in J_s$ in state $s \in S$ with $y_s^\tau(j_s) = 0$, then clearly player 1 notices immediately the deviation, so the inequalities

$$\lim_{\tau \uparrow 1} V_s(x_s^\tau, j_s) \geq v_s, \qquad \text{for all } j_s \in J_s, s \in S,$$

assure that, by choosing any action $j_s \in J_s$ in any state $s \in S$ with $y_s^\tau(j_s) = 0$, the reward is at least

$$V_s(x_s^\tau, j_s) - (1 - \tau) \geq v_s - \epsilon/2,$$

if τ is large enough, using the fact that π^τ prescribes a $(1-\tau)$-optimal strategy afterwards.

Deviations of Player 2 within the Support. If player 2 prescribes only actions which have positive probabilities with respect to y^τ, $\tau \in (0, 1)$, then we divide the set of stages up to the current stage into blocks B^k of consecutive stages as follows: a new block starts at each stage the play enters T, or a new set $E \in R$. In block B^k, the probability that player 1 detects a deviation of player 2 although player 2 truly used y^τ will be at most d^k, where $d^k \in (0, 1)$ for all $k \in \mathbb{N}$ and $\sum_{k=1}^{\infty} d^k \leq 1 - \tau$. The latter inequality will guarantee that the total probability of making this mistake is at most $1-\tau$.

Deviations of Player 2 within Ergodic Sets. If the current block is B^k and the play is in some $E \in R^3$, then player 1 checks the action frequencies of player 2 in E; if the empirical action frequencies are not close enough to the theoretical ones, then player 1 detects a deviation. If the number of stages in block B^k is large enough, then the probability that player 1 detects a deviation although player 2 used y^τ is at most d^k. If the empirical action frequencies are close to the theoretical ones, then the corresponding reward is close to the value. Notice that the play never leaves E if the players use

only actions which are chosen with positive probabilities with respect to the pair (x^τ, y^τ).

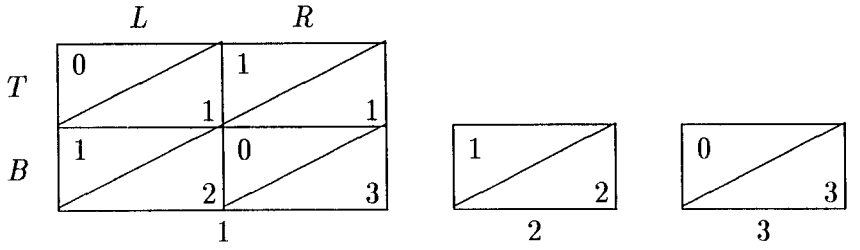
Deviations of Player 2 in Transient States. If the current block is B^k and the play is in T or in some $E \in R^2$, then the play should leave T or E within N^k stages, for large N^k with probability at least $1-d^k$, if player 2 uses y^τ. If the play does not leave T or E within N^k stages, then player 1 detects a deviation of player 2, with probability at least $1-d^k$, so he starts playing a $(1-\tau)$-optimal strategy afterward. Notice that $x_s^\tau \in X_s'$ for all $s \in T \cup S^2$; hence, the play can only leave $T \cup S^2$ in such a way that the value does not decrease in expectation.

Finally, if the current block is B^k and the play is in some $E \in R^1$, then player 1 checks the action frequencies of player 2. This way, player 1 can make sure that the unique state s in $S^* \cap E$ is visited frequently enough and also that the play leaves E via i_s^* and the new value does not differ much from v_s [recall that $V_s(i_s^*, y_s^*) = v_s$]. If player 2 truly uses y^τ, then player 1 detects no deviation with probability at least $1 - d^k$.

We have described how player 1 makes sure that the reward is not much less than the value once the play reaches an ergodic set in R^3 and also that the play reaches eventually an ergodic set in R^3 in such a way that the value does not drop much in expectation; so, by taking a sufficiently large $\tau \in (0, 1)$, the proof is complete. □

2.3. Examples. We provide two examples to illustrate the construction above.

Example 2.1. Once more, consider the game represented by



In other words, we reexamine the "big match"; see Example 1.1. This example shows how the ergodic sets in R^1 and R^2 can be left in such a way that the value does not change much in expectation. The limiting average value is known to be $v = (1/2, 1, 0)$. Following the construction above,

we have

$$X_1' = \{(1,0)\}, \qquad X_2' = X_3' = \{(1)\},$$
$$Y_1' = \text{conv}\{(1/2,1/2),(0,1)\}, \qquad Y_2' = Y_3' = \{(1)\},$$
$$\mathcal{R}^1 = \{\{1\}\}, \qquad S^1 = \{1\},$$
$$\mathcal{R}^2 = \varnothing, \qquad S^2 = \varnothing,$$
$$\mathcal{R}^3 = \{\{2\},\{3\}\}, \qquad S^3 = \{2,3\}.$$

To see that $S^1 = \{1\}$, take

$$S^* = \{1\}, \qquad i_1^* = B, \qquad y_1^* = (1/2, 1/2).$$

As X' is a singleton and states 2 and 3 are trivial, for $\tau \in (0,1)$ we have

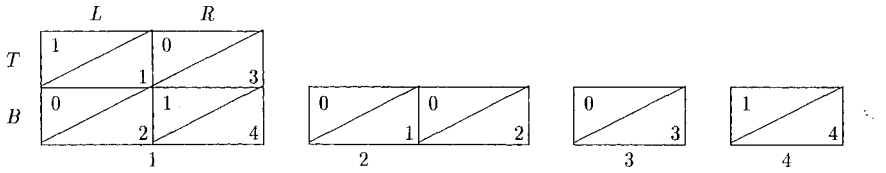
$$x^\tau = ((\tau, 1-\tau),(1),(1)), \qquad y^\tau = ((1/2,1/2),(1),(1)).$$

Clearly,

$$\gamma(x^\tau, y^\tau) = v, \qquad \text{for all } \tau \in (0,1).$$

Note that player 1 has no incentive to deviate from x^τ when playing against y^τ, because any strategy of player 1 would give a reward of $1/2$ against y^τ, $\tau \in (0,1)$. On the other hand, if τ is large, then player 1 is able to check the action frequencies of player 2 in state 1 with a high precision; thus, if the initial state is state 1, then player 1 can make sure that the eventual transitions to state 2 and state 3 will have almost equal probabilities; so, by the choice of a sufficiently large τ, player 2 cannot gain more than an arbitrarily small ϵ by any deviation from y^τ.

Example 2.2. Consider the game represented by



This example clarifies how stationary strategy pairs with rewards equal to the value can be constructed in ergodic sets in $\mathcal{R}^3$; see Lemma 2.5. The notation is the same as in Example 1.1. Notice that states 3 and 4 are trivial. The value of the game is $v = (0,0,0,1)$. To see that $v_1 = v_2 = 0$, take the

stationary δ-optimal strategy

$$y^{\delta} = ((1-\delta, \delta), (0,1), (1), (1))$$

for player 2, where $\delta \in (0,1)$. One can easily check that

$$\gamma(1, x, y^{\delta}) \leq \delta \text{ and } \gamma(2, x, y^{\delta}) = 0, \qquad \text{for all } x \in X;$$

so, using the fact that, against a stationary strategy, there always exists a stationary best reply and the fact that the smallest payoff in the game is zero, we have $v_1 = v_2 = 0$ indeed. Following the construction above, we have

$$X' = X,$$

$$\begin{aligned} &Y_1' = \{(1,0)\}, && Y_s' = Y_s, \quad \forall s = 2,3,4, \\ &R^1 = \varnothing, && S^1 = \varnothing, \\ &R^2 = \varnothing, && S^2 = \varnothing, \\ &R^3 = \{\{1,2\},\{3\},\{4\}\}, && S^3 = \{1,2,3,4\}. \end{aligned}$$

We focus only on the ergodic set $E = \{1,2\}$, as states 3 and 4 are trivial. Define $\bar{\Gamma}_E$ as in the preliminaries, and let $\bar{v}_s, s = 1,2$, denote the value of $\bar{\Gamma}_E$. Clearly,

$$\bar{v}_1 = 1 > 0 = v_1, \qquad \bar{v}_2 = 0 = v_2.$$

Note that

$$\bar{v}_s \geq v_s, \qquad \text{for all } s \in E;$$

thus, Lemma 2.4 assures that $\bar{v}_t = v_t$, for some $t \in E$ [take $t = 2$ here]. Now, the strategies

$$\bar{x} = ((1/2, 1/2), (1)) \in \bar{X}_E, \qquad \bar{y} = ((1,0), (0,1)) \in \bar{Y}_E$$

satisfy

$$\gamma(s, \bar{x}, \bar{y}) = v_s, \qquad \text{for all } s \in E;$$

see proof of Lemma 2.5. Now, for all $\tau \in (0,1)$,

$$x^{\tau} = ((1/2, 1/2), (1), (1), (1)), \qquad y^{\tau} = ((1,0), (0,1), (1), (1))).$$

Clearly,

$$\gamma(x^{\tau}, y^{\tau}) = v, \qquad \text{for all } \tau \in (0,1).$$

Note that player 2 has no incentive to deviate from y^{τ} when playing against x^{τ}. On the other hand, player 2 can check the action frequencies of player 1 in state 1. So, if player 1 decides to play action Top at each stage, which is the only way for player 1 to get a reward higher than 0 when playing

against y^τ from initial state 1, then after finitely many stages player 2 detects the deviation of player 1 with probability almost 1 and starts using the strategy y^δ with small δ. This assures that player 1 cannot get more than an arbitrary small $\epsilon > 0$. Note that the probability that player 1 truly uses x^τ, $\tau \in (0, 1)$, but accidentally chooses action Top for a very long time is small.

3. Concluding Remarks

It is worthwhile to mention that long-term average payoffs are sometimes evaluated by other rewards. The most common alternative rewards are the following:

$$\limsup_{N\to\infty} \mathbb{E}_{s\pi\sigma}(R_N), \qquad \mathbb{E}_{s\pi\sigma}\left(\liminf_{N\to\infty} \mathbf{R}_N\right), \qquad \mathbb{E}_{s\pi\sigma}\left(\limsup_{N\to\infty} R_N\right),$$

where R_N is the random variable for the average payoff up to stage $N \in \mathbb{N}$. All these rewards are known to be equal for stationary strategy pairs. Also, the corresponding values are equal (cf. Ref. 3); hence, for any $\epsilon > 0$, almost stationary ϵ-equilibria exist for these alternative rewards as well. Sometimes ϵ-equilibria, $\epsilon \geq 0$, are expected to be uniform; i.e., a pair (π, σ) is a uniform ϵ-equilibrium, $\epsilon \geq 0$, if $\forall \delta > 0$, $\exists N^\delta$, $\forall N \geq N^\delta$, $\forall \bar{\pi}$, $\forall \bar{\sigma}$ such that

$$(\mathbb{E}_{s\bar{\pi}\sigma}(R_N) - \epsilon) - \delta \leq \mathbb{E}_{s\pi\sigma}(R_N) \leq (\mathbb{E}_{s\pi\bar{\sigma}}(R_N) + \epsilon) + \delta.$$

Since stationary strategies guarantee uniform rewards, and since the players have uniform ϵ-optimal strategies (cf. Ref. 3), almost stationary uniform ϵ-equilibria can be constructed in a similar way.

References

1. Kemeny, J., and Snell, J., *Finite Markov Chains*, Van Nostrand, Princeton, New Jersey, 1960.
2. Hordijk, A., and Kallenberg, L. C. M., *Semi-Markov Strategies in Stochastic Games*, International Journal of Game Theory, Vol. 12, pp. 81–89, 1983.
3. Mertens, J. F., and Neyman, A., *Stochastic Games*, International Journal of Game Theory, Vol. 10, pp. 53–66, 1981.
4. Vrieze, O. J., *Stochastic Games with Finite State and Action Spaces*, Centre for Mathematics and Computer Science, Amsterdam, Holland, 1987.
5. Gillette, D., *Stochastic Games with Zero Stop Probabilities*, Contributions to the Theory of Games III, Edited by M. Dresher, A. W. Tucker, and P. Wolfe, Annals of Mathematical Studies, Princeton University Press, Princeton, New Jersey, Vol. 39, pp. 179–187, 1957.

6. BLACKWELL, D., and FERGUSON, T. S., *The Big Match*, Annals of Mathematical Statistics, Vol. 33, pp. 159–163, 1968.
7. VRIEZE, O. J., and THUIJSMAN, F., *On Equilibria in Repeated Games with Absorbing States*, International Journal of Game Theory, Vol. 18, pp. 293–310, 1989.
8. THUIJSMAN, F., and RAGHAVAN, T. E. S., *Perfect Information Stochastic Games and Related Classes*, International Journal of Game Theory, Vol. 26, pp. 403–408, 1997.
9. THUIJSMAN, F., and VRIEZE, O. J., *The Power of Threats in Stochastic Games*, Stochastic Games and Numerical Methods for Dynamic Games, Edited by M. Bardi et al., Birkhäuser, Boston, Massachusetts, pp. 343–357, 1999.
10. THUIJSMAN, F., and VRIEZE, O. J., *Easy Initial States in Stochastic Games*, Stochastic Games and Related Topics, Edited by T. E. S. Raghaven et al., Kluwer Academic Publishers, Dordrecht, Holland, pp. 85–100, 1991.
11. FLESCH, J., THUIJSMAN, F., and VRIEZE, O. J., *Simplifying Optimal Strategies in Stochastic Games*, SIAM Journal of Control and Optimization, Vol. 36, pp. 1331–1347, 1998.
12. VIEILLE, N., *Solvable States in Stochastic Games*, International Journal of Game Theory, Vol. 21, pp. 395–404, 1993.