Characterization and simplification of optimal strategies in positive stochastic games

J. Appl. Prob. **55,** 728–741 (2018)
*doi:*10.1017/jpr.2018.47

CHARACTERIZATION AND SIMPLIFICATION OF OPTIMAL STRATEGIES IN POSITIVE STOCHASTIC GAMES

JÁNOS FLESCH * AND
ARKADI PREDTETCHINSKI,** *Maastricht University*
WILLIAM SUDDERTH,*** *University of Minnesota*

Abstract

We consider positive zero-sum stochastic games with countable state and action spaces. For each player, we provide a characterization of those strategies that are optimal in every subgame. These characterizations are used to prove two simplification results. We show that if player 2 has an optimal strategy then he/she also has a stationary optimal strategy, and prove the same for player 1 under the assumption that the state space and player 2's action space are finite.

Keywords: Positive; two-person; zero-sum stochastic game; optimal stationary strategy; subgame-optimal strategy; Markov chain; martingale

2010 Mathematics Subject Classification: Primary 91A15; 91A05; 60J10; 60G42

1. Introduction

Zero-sum stochastic games are two-player dynamic games in which the two players have completely opposite interests and in which the actions chosen by the players in each period influence not only the reward in the period but also a state variable. A zero-sum stochastic game is called positive if the reward function is nonnegative and player 1 tries to maximize while player 2 tries to minimize the sum of the rewards during play.

We examine optimal strategies in positive zero-sum stochastic games with countable state and action spaces. For each player, we provide a characterization of those strategies that are optimal in each subgame. By using these characterizations, we prove the following simplification results. We show that if player 2 has an optimal strategy then he/she also has a stationary optimal strategy. We prove the same for player 1 under the restriction that the state space and player 2's action space are finite. Our construction does not require the knowledge of an optimal strategy, only its existence. The results directly transfer to negative zero-sum stochastic games when the roles of the players are reversed.

1.1. Related literature

For an overview of the literature on positive stochastic games and negative stochastic games, we refer the reader to the recent survey by Jaśkiewicz and Nowak (2016). For further reading,

Received 9 October 2017; revision received 20 April 2018.
* Postal address: Department of Quantitative Economics, Maastricht University, PO Box 616, 6200 MD, The Netherlands. Email address: j.flesch@maastrichtuniversity.nl
** Postal address: Department of Economics, Maastricht University, PO Box 616, 6200 MD, The Netherlands. Email address: a.predtetchinski@maastrichtuniversity.nl
*** Postal address: School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA.
Email address: bill@stat.umn.edu

see, in particular, Maitra and Parthasarathy (1971), Parthasarathy (1971), (1973), Frid (1974), Nowak (1985), Nowak and Raghavan (1991), and Maitra and Sudderth (1996).

Simplification results similar to those presented in this paper have appeared in the context of zero-sum stochastic games with average payoff (see Flesch *et al.* (1998)), limsup payoff and liminf payoff (see Flesch *et al.* (2016)), but also in the literature of gambling and dynamic programming. Dubins and Savage (1965) showed that for a (one-person) gambling problem with a finite state space and limsup payoff, the existence of an optimal strategy implies the existence of a stationary optimal strategy. Blackwell (1970) proved the same result for positive dynamic programming with a countable state space. There is also a generalization to a Borel measurable setting by Orkin (1974). In the context of negative dynamic programming, Strauch (1966) presented such a simplification. Recently, Sudderth (2016) showed that for gambling problems with a countable state space and limsup payoff as well as for gambling problems with a finite state space and liminf payoff, the existence of an optimal strategy implies the existence of a Markov optimal strategy. A strategy is called Markov if the prescribed mixed actions depend only on the current state and on the current time period, but not directly on the past states and actions.

The paper is organized as follows. In Section 2 we introduce the model, and in Section 3 we discuss some preliminaries. In Section 4 we summarize the results. The proofs are stated in Section 5. An example and some final remarks are contained in Section 6.

2. The model

2.1. Positive zero-sum stochastic games

We consider positive zero-sum stochastic games with countable state and action spaces. Such a game is played by two players, and has the following properties:

- a nonempty and countable state space S,
- for each state $s \in S$, nonempty and countable action spaces $A(s)$ and $B(s)$ for player 1 and player 2, respectively,
- for each state $s \in S$ and actions $a \in A(s)$, $b \in B(s)$, a probability measure $p(s, a, b) = p(s' \mid s, a, b)_{s' \in S}$ on S, and
- a nonnegative reward function $r\colon Z \to [0, \infty)$, where $Z = \{(s, a, b) \mid s \in S, a \in A(s), b \in B(s)\}$.

The game is played at periods in $\mathbb{N} = \{0, 1, \ldots\}$ and begins in an initial state $s_0 \in S$. At every period $t \in \mathbb{N}$, the play is in a state $s_t \in S$. In this state, player 1 chooses an action $a_t \in A(s_t)$ and simultaneously player 2 chooses an action $b_t \in B(s_t)$. Then, with $z_t = (s_t, a_t, b_t)$, player 1 receives reward $r(z_t)$ from player 2, and state s_{t+1} is drawn in accordance with the probability measure $p(z_t)$. Thus, play induces an infinite sequence $(z_0, z_1, \ldots)$ in Z. The payoff is

$$u(z_0, z_1, \ldots) = \sum_{t=0}^{\infty} r(z_t).$$

The payoff takes values in $[0, \infty]$ and is paid by player 2 to player 1. Player 1's objective is to maximize the expected value of the payoff given by u, and player 2's objective is to minimize it.

2.2. Strategies

The set of histories at period t is denoted by H_t. Thus, $H_0 = S$ and $H_t = Z^t \times S$ for every period $t \geq 1$. Let $H = \bigcup_{t\in\mathbb{N}} H_t$ denote the set of all histories. For each history h, let s_h denote the final state in h.

A mixed action for player 1 in state $s \in S$ is a probability measure $x(s)$ on $A(s)$. Similarly, a mixed action for player 2 in state $s \in S$ is a probability measure $y(s)$ on $B(s)$. The respective sets of mixed actions in state s are denoted by $X(s)$ and $Y(s)$. The support of a mixed action $x(s)$, denoted by $\mathrm{supp}(x(s))$, is the set of actions that are played with positive probability by $x(s)$, i.e. $\mathrm{supp}(x(s)) = \{a \in A(s) \mid x(s)(a) > 0\}$. The support of a mixed action for player 2 is defined similarly.

A strategy for player 1 is a map π that assigns to every history $h \in H$ a mixed action $\pi(h) \in X(s_h)$. Similarly, a strategy for player 2 is a map σ that to each history $h \in H$ assigns a mixed action $\sigma(h) \in Y(s_h)$. The set of strategies is denoted by Π for player 1 and by Σ for player 2. A strategy is called pure if it places probability 1 on one action after each history.

A strategy is called stationary if the assigned mixed actions only depend on the history through its final state. Thus, a stationary strategy for player 1 can be seen as an element x of $X := \times_{s\in S} X(s)$. Similarly, a stationary strategy for player 2 can be seen as an element y of $Y := \times_{s\in S} Y(s)$. A pair of stationary strategies (x, y) induces a Markov chain on the state space S. A nonempty set $E \subseteq S$ is called ergodic with respect to (x, y) if starting in any state in E, the probability that the Markov chain eventually visits every state in E and never leaves E is 1.

An initial state $s \in S$ and a pair of strategies $(\pi, \sigma) \in \Pi \times \Sigma$ determine the distribution $\mathbb{P}_{s,\pi,\sigma}$ of the stochastic process $(z_0, z_1, \ldots)$. We denote the expected payoff $\mathbb{E}_{s,\pi,\sigma}[\sum_{t=0}^{\infty} r(z_t)]$ by $u(s, \pi, \sigma)$.

2.3. Value and optimality

The game is said to have a value for initial state $s \in S$ if

$$\sup_{\pi\in\Pi} \inf_{\sigma\in\Sigma} u(s, \pi, \sigma) = \inf_{\sigma\in\Sigma} \sup_{\pi\in\Pi} u(s, \pi, \sigma).$$

If the value exists for initial state $s \in S$, we denote it by $v(s)$. In that case, for $\varepsilon \geq 0$, a strategy $\pi \in \Pi$ for player 1 is called ε-optimal for initial state s if $u(s, \pi, \sigma) \geq v(s) - \varepsilon$ for every strategy $\sigma \in \Sigma$ for player 2. Similarly, a strategy $\sigma \in \Sigma$ for player 2 is called ε-optimal for initial state s if $u(s, \pi, \sigma) \leq v(s) + \varepsilon$ for every strategy $\pi \in \Pi$ for player 1. A strategy is called ε-optimal if it is ε-optimal for every initial state. Note that if the value exists for every initial state, each player has an ε-optimal strategy for every $\varepsilon > 0$. A 0-optimal strategy is simply called optimal.

As is well known, the value in general does not exist, with a typical example being as follows: there is a state with action space $\mathbb{N}$ for each player, in which the reward function is 1 if player 1's action is greater than player 2's action and is 0 otherwise. From this state, regardless of the chosen actions, the transition is to an absorbing state where the reward is 0.

Since we are interested in optimal strategies, and since we can only speak of optimality if the value exists, we make the following assumption.

Assumption 1. *Value $v(s)$ exists and is finite for every initial state $s \in S$.*

2.4. Subgame optimality

Consider a strategy $\pi \in \Pi$ for player 1 and a sequence $g \in Z^t$ for some $t \in \mathbb{N}$ (for $t = 0$, g is the empty sequence). The continuation strategy $\pi[g]$ is the strategy in Π given, for every history $h \in H$, by $\pi[g](h) = \pi(gh)$, where gh denotes the concatenation of g and h.

A strategy $\pi \in \Pi$ for player 1 is called subgame optimal if $\pi[g]$ is optimal for every $g \in Z^t$ and $t \in \mathbb{N}$. Continuation strategies and subgame-optimal strategies for player 2 are defined analogously.

Note that every subgame-optimal strategy is optimal. The converse holds for stationary strategies: a stationary optimal strategy is always subgame optimal.

For a history h ending with a state s, we can also define in an obvious way the continuation strategy $\pi[h]$ for the state s.

2.5. Negative zero-sum stochastic games

Negative zero-sum stochastic games are defined similarly, but with a nonpositive reward function $r\colon Z \to (-\infty, 0]$.

3. Preliminaries

3.1. The one-day games

For each state $s \in S$, we consider a matrix game $M(s)$ with countable action spaces. This matrix game is fundamental and frequently used in the analysis of stochastic games.

Let $s \in S$. The matrix game $M(s)$ is defined as follows. The sets of actions are $A(s)$ and $B(s)$ for players 1 and 2, respectively, and the payoff for each pair of actions $(a, b) \in A(s) \times B(s)$ is

$$u_M(s, a, b) \coloneqq r(s, a, b) + \sum_{s' \in S} p(s' \mid s, a, b) \cdot v(s').$$

Intuitively, $u_M(s, a, b)$ is the sum of the current reward and the expectation of the value after transition in the original game G, when the players play actions a and b in state s. For mixed actions $x(s) \in X(s)$ and $y(s) \in Y(s)$, as usual, $u_M(s, x(s), y(s))$ denotes the expected payoff in the matrix game $M(s)$.

The following property of the matrix game $M(s)$ is well known, but for completeness we provide a short proof.

Claim A. *Let $\varepsilon \geq 0$. If π is an ε-optimal strategy for player 1 in game G for the initial state s then $u_M(s, \pi(s), y(s)) \geq v(s) - \varepsilon$ holds for every $y(s) \in Y(s)$, where $\pi(s)$ is the mixed action that π prescribes for the initial state s. Similarly, if σ is an ε-optimal strategy for player 2 in game G for the initial state s then $u_M(s, x(s), \sigma(s)) \leq v(s) + \varepsilon$ holds for every $x(s) \in X(s)$.*

Proof. We only prove the first part of the claim, the proof of the second part being similar. Let π be an ε-optimal strategy for player 1 in game G for the initial state $s = s_0$. Take any $y(s) \in Y(s)$. Let $\delta > 0$, and let σ be a strategy for player 2 such that $\sigma(s) = y(s)$ and let the continuation strategy $\sigma[s, a, b]$ beδ-optimal for every $a \in A(s)$ and $b \in B(s)$. Then

$$\begin{aligned} v(s) - \varepsilon &\leq u(s, \pi, \sigma) \\ &= \mathbb{E}_{s,\pi,\sigma}[r(s_0, a_0, b_0)] + \mathbb{E}_{s,\pi,\sigma}\left[\sum_{t=1}^{\infty} r(s_t, a_t, b_t)\right] \\ &\leq \mathbb{E}_{s,\pi,\sigma}[r(s_0, a_0, b_0)] + \mathbb{E}_{s,\pi,\sigma}[v(s_1) + \delta] \end{aligned}$$

$$= \mathbb{E}_{s,\pi,\sigma}[r(s_0, a_0, b_0) + v(s_1)] + \delta$$
$$= u_M(s, \pi(s), y(s)) + \delta.$$

Since $\delta > 0$ is arbitrary, we have proved $u_M(s, \pi(s), y(s)) \geq v(s) - \varepsilon$. □

Since each player has an ε-optimal strategy in game G for every $\varepsilon > 0$, it follows from Claim A that the value of the matrix game $M(s)$ exists and is equal to $v(s)$, i.e.

$$v(s) = \sup_{x(s)\in X(s)} \inf_{y(s)\in Y(s)} u_M(s, x(s), y(s)) = \inf_{y(s)\in Y(s)} \sup_{x(s)\in X(s)} u_M(s, x(s), y(s)).$$

In the matrix game $M(s)$, we define

$$X^*(s) := \{x(s) \in X(s) \mid u_M(s, x(s), y(s)) \geq v(s) \text{ for all } y(s) \in Y(s)\},$$
$$Y^*(s) := \{y(s) \in Y(s) \mid u_M(s, x(s), y(s)) \leq v(s) \text{ for all } x(s) \in X(s)\}.$$

The sets $X^*(s)$ and $Y^*(s)$ consist of all mixed actions for player 1 and player 2, respectively, that are optimal in the matrix game $M(s)$. The set $X^*(s)$ is convex. Furthermore, it is nonempty if $A(s)$ is finite, but it may be empty if $A(s)$ is infinite. Similar properties hold for the set $Y^*(s)$.

Whenever $X^*(s) \neq \varnothing$, we define

$$A^*(s) := \{a \in A(s) \mid x(s)(a) > 0 \text{ for some } x(s) \in X^*(s)\} = \bigcup_{x(s)\in X^*(s)} \operatorname{supp}(x(s)),$$

and

$$X^{**}(s) := \{x(s) \in X^*(s) \mid x(s)(a) > 0 \text{ for all } a \in A^*(s)\}$$
$$= \{x(s) \in X^*(s) \mid \operatorname{supp}(x(s)) = A^*(s)\}.$$

The set $A^*(s)$ consists of all actions in $A(s)$ that are used by some mixed action in $X^*(s)$, and the set $X^{**}(s)$ consists of all mixed actions in $X^*(s)$ which put positive probability on each such action.

Claim B. *If $X^*(s)$ is nonempty then the set $X^{**}(s)$ is also nonempty.*

Indeed, suppose that $X^*(s)$ is nonempty, then $A^*(s)$ is nonempty. Since $A^*(s)$ is countable, we can list its elements in a finite or countably infinite sequence $a_1, a_2, \ldots$.

Proof of Claim B. Assume that the sequence is infinite. The proof for the finite case is similar. For each action a_n, choose a mixed action $x_n(s) \in X^*(s)$ such that $x_n(s)(a_n) > 0$. Then define the mixed action $x^*(s)$ by setting

$$x^*(s)(a) = \sum_{n=1}^{\infty} \frac{1}{2^n} x_n(s)(a) \quad \text{for all } a \in A(s).$$

Clearly, $x^*(s)(a) > 0$ for all $a \in A^*(s)$. Also, $x^*(s) \in X^*(s)$ since, for all $y(s) \in Y(s)$,

$$u_M(s, x^*(s), y(s)) = \sum_{n=1}^{\infty} \frac{1}{2^n} u_M(s, x_n(s), y(s)) \geq \sum_{n=1}^{\infty} \frac{1}{2^n} v(s) = v(s).$$

Hence, $x^*(s) \in X^{**}(s)$ and $X^{**}(s)$ is nonempty. □

3.2. Specific types of strategies

Define $X^{**} = \times_{s\in S} X^{**}(s)$. Thus, the set X^{**} consists of all stationary strategies for player 1 that use a mixed action in $X^{**}(s)$ in every state $s \in S$. These stationary strategies are called *maximally mixed* strategies.

In our results and analysis, a crucial role is played by the so-called *locally optimal* strategies. For player 1, we call a strategy π locally optimal if, for every history $h \in H$, we have $\pi(h) \in X^*(s_h)$. Intuitively, π is locally optimal if it only uses mixed actions optimal in the corresponding matrix games. Clearly, every maximally mixed strategy for player 1 is locally optimal. Locally optimal strategies are defined similarly for player 2.

Example 1. As mentioned earlier, it is known that in a positive zero-sum stochastic game, even if the state and action spaces are finite, player 1 may have no optimal strategy; see Kumar and Shiau (1981), Maitra and Sudderth (1996), and Jaśkiewicz and Nowak (2016). Consider the following game.

	L	R
T	0	1
B	1	0

In this game, there is only one nontrivial state. In this state, player 1's actions are T and B, and player 2's actions are L and R. The rewards for the corresponding action combinations are given in the matrix. The transitions are as follows: if action combination (T, L) is chosen then the state remains the same, but after any other action combination transition occurs to an absorbing state where the reward is equal to 0. Clearly, player 1 can guarantee an expected payoff of $1-\varepsilon$, for any $\varepsilon \in (0, 1)$, by playing the stationary strategy $(1-\varepsilon, \varepsilon)$. Thus, the value is equal to 1 for the nontrivial state. Yet, player 1 has no optimal strategy.

The stationary strategy $(1, 0)$ is the unique locally optimal strategy for player 1 and, therefore, also the unique maximally mixed strategy for him/her. This means that a locally optimal strategy or a maximally mixed strategy for player 1 is not necessarily optimal. An additional and crucial property is needed to ensure optimality, and this property appears as condition (ii) in Theorem 1.

4. Results

In this section we discuss our results. Recall Assumption 1 that we imposed on the positive zero-sum stochastic game.

We start with a characterization of the strategies for player 1 that are optimal in each subgame.

Theorem 1. (Characterization of subgame-optimal strategies for player 1.) *Consider a positive zero-sum stochastic game. A strategy π for player 1 is subgame-optimal if and only if the following two conditions are satisfied:*

1. *π is locally optimal, and*
2. *for every history h and every strategy σ for player 2, either*
   (a) $\lim_{k\to\infty} \mathbb{E}_{s_h,\pi[h],\sigma[h]}[v(s_k)] = 0$,
   (b) $u(s_h, \pi[h], \sigma[h]) = \infty$.

The first condition is quite intuitive. The second condition requires that, in each subgame, the value of the state converges to 0 with probability 1, which guarantees that π accumulates good rewards for player 1, or that the expected payoff is infinite. In view of our assumption that the value of the game is finite for every initial state, the expected payoff will be infinite only if player 2 chooses a bad strategy.

The two conditions on strategy π are analogous to the conditions Dubins and Savage (1965) called 'thrifty' and 'equalizing' in their study of one-person gambling problems. We refer the reader to Blackwell (1970) and Puterman (1994) for the connection of these two properties and optimal strategies in Markov decision processes.

Based on the above characterization, we will prove the following simplification result for player 1.

Theorem 2. (Simplification of optimal strategies for player 1.) *Consider a positive zero-sum stochastic game. Assume that the state space S and player 2's action space $B(s)$ in every state $s \in S$ are finite. If player 1 has an optimal strategy then he/she also has a stationary optimal strategy.*

In the proof we show that, under the assumptions of Theorem 2, player 1 has a maximally mixed stationary strategy, and each such strategy is optimal. Our construction does not require the knowledge of an optimal strategy, only its existence.

If player 2 is a dummy with only one action available at each state, the positive stochastic game becomes a positive dynamic programming problem. For such problems, Blackwell (1970) proved the result corresponding to Theorem 2 for countably infinite state spaces. However, the claim of Theorem 2 is not valid in general for positive stochastic games if S is countably infinite, as we demonstrate in Example 2 of Section 6. In that game, player 1 has an optimal strategy but no subgame-optimal strategy, and, in particular, no stationary optimal strategy. We do not know whether the claim of Theorem 2 remains valid in games in which the action space for player 2 is countably infinite.

Now we turn to player 2 and provide a simple characterization of strategies of player 2 that are optimal in each subgame. This result for player 2 is related to the work of Strauch (1966) on negative dynamic programming.

Theorem 3. (Characterization of subgame-optimal strategies for player 2.) *Consider a positive zero-sum stochastic game. A strategy σ for player 2 is subgame-optimal if and only if σ is locally optimal.*

As we shall see, the following result is an easy corollary.

Theorem 4. (Simplification of optimal strategies for player 2.) *Consider a positive zero-sum stochastic game. If player 2 has an optimal strategy then he/she also has a stationary optimal strategy.*

Our results directly translate to negative zero-sum stochastic games. Indeed, consider a negative zero-sum stochastic game. In this game, player 1 maximizes the sum of the nonpositive rewards $\sum_{t=0}^{\infty} r_t$. This is equivalent to minimizing $\sum_{t=0}^{\infty}(-r_t)$. Since $-r$ is a nonnegative function, player 1 can be seen as the minimizing player in a positive zero-sum stochastic game. Similarly, player 2 can be seen as the maximizing player in a positive zero-sum stochastic game.

5. Proofs

In the following proofs we make use of two processes. We define the process $\{R_n\}_{n=0}^{\infty}$ by setting

$$R_n = \sum_{k=0}^{n} r(s_k, a_k, b_k)$$

and the process $\{Q_n\}_{n=0}^{\infty}$ by setting

$$Q_0 = v(s_0), \qquad Q_n = R_{n-1} + v(s_n) \quad \text{for all } n \geq 1,$$

where s_k, a_k, b_k denote the state and the actions at period k, respectively.

Lemma 1. 1. *Assume that strategy π for player 1 is locally optimal. Then, for any state $s \in S$ and any strategy σ for player 2, the process $\{Q_n\}_{n=0}^{\infty}$ is a submartingale with respect to $\mathbb{P}_{s,\pi,\sigma}$.*

2. *Assume that the strategy σ for player 2 is locally optimal. Then, for any state $s \in S$ and any strategy π for player 1, the process $\{Q_n\}_{n=0}^{\infty}$ is a supermartingale with respect to $\mathbb{P}_{s,\pi,\sigma}$.*

Proof. We only prove the first statement, since the proof of the second statement is similar. So, assume that strategy π for player 1 is locally optimal. Take any state $s = s_0$ and any strategy σ for player 2. Let $n \in \mathbb{N}$ and $h_n = (s_0, a_0, b_0, \ldots, s_n)$. As π is locally optimal, $\pi(h_n)$ is optimal in $M(s_n)$ and, hence,

$$\mathbb{E}_{s,\pi,\sigma}[r(s_n, a_n, b_n) + v(s_{n+1}) \mid h_n] \geq v(s_n).$$

This implies

$$\begin{aligned}\mathbb{E}_{s,\pi,\sigma}[Q_{n+1} \mid h_n] &= \mathbb{E}_{s,\pi,\sigma}[R_n + v(s_{n+1}) \mid h_n] \\ &= R_{n-1} + \mathbb{E}_{s,\pi,\sigma}[r(s_n, a_n, b_n) + v(s_{n+1}) \mid h_n] \\ &\geq R_{n-1} + v(s_n) \\ &= Q_n.\end{aligned}$$

Thus, $\{Q_n\}_{n=0}^{\infty}$ is a submartingale with respect to $\mathbb{P}_{s,\pi,\sigma}$. □

Lemma 2. *Consider an initial state $s \in S$, a strategy π for player 1, and a strategy σ for player 2. Assume that π is locally optimal and*

$$\lim_{n\to\infty} \mathbb{E}_{s,\pi,\sigma}[v(s_n)] = 0. \tag{1}$$

Then $u(s, \pi, \sigma) \geq v(s)$.

Proof. In view of Lemma 1, the process $\{Q_n\}_{n=0}^{\infty}$ is a submartingale with respect to $\mathbb{P}_{s,\pi,\sigma}$. Also $\{R_n\}_{n=0}^{\infty}$ converges to $R_\infty = \sum_{k=0}^{\infty} r(s_k, a_k, b_k)$. From the submartingale property, (1), and the monotone convergence theorem it follows that

$$v(s) = Q_0 \leq \lim_{n\to\infty} \mathbb{E}_{s,\pi,\sigma}[Q_n] = \lim_{n\to\infty} \mathbb{E}_{s,\pi,\sigma}[R_{n-1} + v(s_n)] = \mathbb{E}_{s,\pi,\sigma}[R_\infty] = u(s, \pi, \sigma),$$

which completes the proof. □

Lemma 3. *Assume that π is a subgame-optimal strategy for player 1. Suppose that for some initial state $s = s_0 \in S$, strategy σ for player 2, positive integer n, and $\lambda > 0$, the inequality $\mathbb{E}_{s,\pi,\sigma}[v(s_n)] > \lambda$ holds. Then there is a positive integer $m > n$ such that $\mathbb{E}_{s,\pi,\sigma}[\sum_{t=n}^{m} r_t] > \lambda$.*

Proof. By the subgame-optimality of π, it holds that

$$\mathbb{E}_{s_n,\pi[h_n],\sigma[h_n]}\left[\sum_{t=0}^{\infty} r_t\right] \geq v(s_n)$$

for every history $h_n = (s_0, a_0, b_0, \ldots, s_{n-1}, a_{n-1}, b_{n-1}, s_n)$. Hence,

$$\mathbb{E}_{s,\pi,\sigma}\left[\sum_{t=n}^{\infty} r_t\right] = \mathbb{E}_{s,\pi,\sigma}\left[\mathbb{E}_{s_n,\pi[h_n],\sigma[h_n]}\left[\sum_{t=0}^{\infty} r_t\right]\right] \geq \mathbb{E}_{s,\pi,\sigma}[v(s_n)].$$

Since $\sum_{t=n}^{m} r_t$ increases to $\sum_{t=n}^{\infty} r_t$ as $m \to \infty$, the conclusion follows from the monotone convergence theorem. □

Proof of Theorem 1. The proof comprises two parts.

Part 1. Assume that conditions 1 and 2 of Theorem 1 hold. Then, for every history h and every strategy σ for player 2, the inequality $u(s_h, \pi[h], \sigma[h]) \geq v(s_h)$ follows from Lemma 2 under condition 2(a) and is obvious under condition 2(b). Hence, π is subgame-optimal.

Part 2. Now assume that π is subgame-optimal for player 1. Condition 1 of Theorem 1 follows directly from Claim A. It remains to check that condition 2 also holds.

We assume without loss of generality that $h = \varnothing$. Suppose that, by way of contradiction, condition 2 fails. Then there exists a state $s \in S$, a strategy σ for player 2, and a positive number λ such that

$$u(s, \pi, \sigma) < \infty \quad \text{and} \quad \limsup_{k\to\infty} \mathbb{E}_{s,\pi,\sigma}[v(s_k)] > \lambda. \tag{2}$$

By (2), there exists a positive integer n_1 such that $\mathbb{E}_{s,\pi,\sigma}[v(s_{n_1})] > \lambda$. So, by Lemma 3, there exists $m_1 > n_1$ such that $\mathbb{E}_{s,\pi,\sigma}[\sum_{t=n_1}^{m_1} r_t] > \lambda$. By (2), again there exists $n_2 > m_1$ such that $\mathbb{E}_{s,\pi,\sigma}[v(s_{n_2})] > \lambda$, and by Lemma 3, there exists $m_2 > n_2$ such that $\mathbb{E}_{s,\pi,\sigma}[\sum_{t=n_2}^{m_2} r_t] > \lambda$.

We continue in this way choosing sequences $\{n_k\}$ and $\{m_k\}$ so that, for all $k \in \mathbb{N}$, $n_k < m_k < n_{k+1}$ and $\mathbb{E}_{s,\pi,\sigma}[\sum_{t=n_k}^{m_k} r_t] > \lambda$. But then

$$u(s, \pi, \sigma) = \mathbb{E}_{s,\pi,\sigma}\left[\sum_{t=0}^{\infty} r_t\right] \geq \lambda + \lambda + \cdots = \infty,$$

which is in contradiction to (2). □

Proof of Theorem 2. Assume that the state space S and player 2's action space $B(s)$, for every $s \in S$, are finite, and assume that player 1 has an optimal strategy π. Then, for every state s, we have $\pi(s) \in X^*(s)$ by Claim A. By Claim B, the set $X^{**}(s)$ is nonempty for every state $s \in S$. This implies that player 1 has a maximally mixed stationary strategy $x \in X^{**}$. We now prove that x is optimal.

Since S and $B(s)$ for every $s \in S$ are finite, player 2 has a stationary best response y to x. Indeed, when playing against x, player 2's best responses can be found by solving a negative Markov decision problem with finite state and action spaces, and in such problems the player has a stationary optimal strategy (see Strauch (1966) or Puterman (1994)).

Therefore, it suffices to show that, for every state $s \in S$,

$$u(s, x, y) \geq v(s). \tag{3}$$

Note that, since y is a best response to x,

$$u(s, x, y) = \inf_{\sigma \in \Sigma} u(s, x, \sigma) \leq v(s) < \infty \tag{4}$$

by Assumption 1. To prove (3), we will apply Lemma 2. So we need to prove that x is locally optimal and that

$$\lim_{n \to \infty} \mathbb{E}_{s,x,y}[v(s_n)] = 0. \tag{5}$$

As $x(s) \in X^{**}(s) \subseteq X^{*}(s)$ for every state $s \in S$, strategy x is locally optimal. Thus, it suffices to prove (5). Note that, since the state space S is finite, the value function v is uniformly bounded. Thus, (5) is a consequence of the dominated convergence theorem and the equality

$$\mathbb{P}_{s,x,y}\left[\lim_{n \to \infty} v(s_n) = 0\right] = 1. \tag{6}$$

Now we prove (6). The pair (x, y) of stationary strategies induces a Markov chain on the finite state space S. Hence, under (s, x, y), an ergodic set will be reached almost surely. Let E be an ergodic set with respect to (x, y). We need to show that

$$v(s') = 0 \quad \text{for each } s' \in E. \tag{7}$$

The proof of (7) comprises three steps.

Step 1. It holds that $r(s', a, b) = 0$ for each $s' \in E$, each $a \in \operatorname{supp}(x(s'))$, and each $b \in \operatorname{supp}(y(s'))$.

This follows as a result of (4) and the fact that, starting from any state in E, every state in E is visited infinitely often with probability 1.

Step 2. The value is constant on the set E.

Let $v_E = \max_{s' \in E} v(s')$ and $E' = \{s' \in E \mid v(s') = v_E\}$. Suppose that the initial state s_0 of the Markov chain induced by (x, y) is an element of $E' \subseteq E$. It follows from the inclusion $x(s_0) \in X^{*}(s_0)$ and step 1 that

$$v_E = v(s_0) \leq \mathbb{E}_{s_0,x,y}[r(s_0, a_0, b_0) + v(s_1)] = \mathbb{E}_{s_0,x,y}[v(s_1)].$$

Now, under (s_0, x, y), state s_1 belongs to the ergodic set E almost surely and, for the expectation to have the maximum value v_E, state s_1 must be in E' almost surely. Iterating this argument, we can conclude that if the initial state is in E', the play under (s_0, x, y) never leaves the set E', almost surely. Since $E' \subseteq E$ and E is an ergodic set, we have $E' = E$ and, consequently, $v(s') = v_E$ for every $s' \in E$. That is, the value is a constant on E.

The intuition behind the third step is as follows: suppose that player 1 plays the optimal strategy π, that player 2 plays y, and that the initial state is some state $s_0 \in E$. Then, at every period, π only places a positive probability on actions that also receive positive probability from x. Hence, the play under (s_0, π, y) stays inside E almost surely. Since the value is constant on E, in any subgame that is reached with a positive probability, the continuation strategy of π has to be optimal. Formally, our third step is as follows.

Step 3. Let $s_0 \in E$. Then under (s_0, π, y) it holds almost surely that for every $n \in \mathbb{N}$:

- $r_n = 0$, $s_n \in E$, and $v(s_n) = v_E$,
- the continuation strategy $\pi[z_0, \ldots, z_n]$ is optimal for the state s_{n+1}.

Since π is optimal and, in particular, optimal for the initial state $s_0 \in E$, the initial mixed action $\pi(s_0)$ belongs to $X^*(s_0)$. Since x is maximally mixed, every action that is used by $\pi(s_0)$ with positive probability is also used by $x(s_0)$ with positive probability, i.e. $\mathrm{supp}(\pi(s_0)) \subseteq \mathrm{supp}(x(s_0))$. Thus, under (s_0, π, y), we have $r_0 = 0, s_1 \in E$, and $v(s_1) = v_E$ with probability 1.

We now argue that, for every $h_1 = (s_0, a_0, b_0, s_1) = (z_0, s_1)$ that can arise with positive probability under (s_0, π, y), the continuation strategy $\pi[z_0]$ must be optimal for state s_1. Suppose that, to have a contradiction, there is a history $h_1' = (s_0, a_0', b_0', s_1') = (z_0', s_1')$ that has probability $p > 0$ and $\pi[z_0']$ is not optimal for state s_1'. Then there must exist $\varepsilon > 0$ and a strategy σ' such that $u(s_1', \pi[z_0'], \sigma') < v_E - \varepsilon$. Let $\delta > 0$ and consider a strategy σ^* for player 2 that has $\sigma^*(s_0) = y(s_0)$, continues after z_0' with $\sigma^*[z_0'] = \sigma'$, and continues after every $z_0 \neq z_0'$ with a δ-optimal strategy for player 2. Then, since $r_0 = 0$,

$$u(s_0, \pi, \sigma^*) = \mathbb{E}_{s_0,\pi,y}[u(s_1, \pi[z_0], \sigma^*[z_0])] \leq p(v_E - \varepsilon) + (1 - p)(v_E + \delta) < v_E$$

for $\delta < p\varepsilon/(1 - p)$, which is in contradiction to the optimality of π.

Now since $\pi[z_0]$ is optimal for state s_1 almost surely under (s_0, π, y), the initial mixed actions $\pi[z_0](s_1)$ must belong to $X^*(s_1)$ and, therefore, belong to $\mathrm{supp}(x(s_1))$. Hence, $r_1 = 0$, $s_2 \in E$, and $v(s_2) = v_E$. By an argument similar to that in the previous paragraph, we find that, almost surely under (s_0, π, y), the continuation strategy $\pi[z_0, z_1]$ is optimal for state s_2.

The result follows by iterating the argument.

Step 3 clearly implies that for $s_0 \in E$, we have $u(s_0, \pi, y) = 0$. But since π is optimal, $v(s_0) = 0$. Therefore, (7) is valid, and the proof of (3) is complete. □

Proof of Theorem 3. Consider a strategy σ for player 2.

First assume that σ is subgame optimal. Then, for every history h, the continuation strategy $\sigma[h]$ is optimal and, hence, $\sigma(h) \in Y^*(s_h)$ by Claim A. Hence, σ is locally optimal.

Now assume that σ is locally optimal. We will show that σ is optimal for the initial state s. The proof for a general history h is the same since the continuation strategy $\sigma[h]$ is also locally optimal in the subgame at h.

Take any state $s = s_0 \in S$ and any strategy π for player 1. In view of Lemma 1, the process $\{Q_n\}_{n=0}^{\infty}$ is a supermartingale with respect to $\mathbb{P}_{s,\pi,\sigma}$. By the supermartingale property, the monotone convergence theorem, and the fact that $v(s_n) \geq 0$ for every $n \in \mathbb{N}$, we obtain

$$\begin{aligned} v(s) &= Q_0 \\ &\geq \lim_{n\to\infty} \mathbb{E}_{s,\pi,\sigma}[Q_n] \\ &= \lim_{n\to\infty} \mathbb{E}_{s,\pi,\sigma}[R_{n-1} + v(s_n)] \\ &\geq \lim_{n\to\infty} \mathbb{E}_{s,\pi,\sigma}[R_{n-1}] \\ &= \mathbb{E}_{s,\pi,\sigma}[R_\infty] \\ &= u(s, \pi, \sigma). \end{aligned}$$

Thus, σ is optimal for the initial state s, as desired. □

Proof of Theorem 4. Assume that player 2 has an optimal strategy σ. Then, for every state $s \in S$, we have $\sigma(s) \in Y^*(s)$ by Claim A. So we consider the stationary strategy y for player 2 that uses the mixed action $\sigma(s)$ in every state $s \in S$, i.e. $y = (\sigma(s))_{s\in S}$. The strategy y is then locally optimal and, hence, by Theorem 3 it is also optimal. □

6. Concluding remarks

6.1. A counter example to the claim of Theorem 2 when the state space is infinite

The following game is a variant of an example in Flesch *et al.* (2016). In this game, player 1 has an optimal strategy, but he/she has no subgame-optimal strategy and, in particular, no stationary optimal strategy. We do not know whether the result of Theorem 2 holds when the state space is finite and actions sets are allowed to be infinite.

Example 2. Consider the following game. The state space is

$$S = \{(0, c), (1, c), \ldots\} \cup \{(0, s), (1, s), \ldots\} \cup \{s^*\}.$$

The states can be described as follows.

- In each state (n, c), where n is even, player 1 has two actions c and s, and player 2 has only one action. If player 1 chooses action c then the reward is 0 and the play moves to state $(n + 1, c)$. If player 1 chooses action s then the reward is 0 and the play moves to state (n, s).
- In each state (n, c), where n is odd, player 2 has two actions c and s, and player 1 has only one action. If player 2 chooses action c then the reward is 0 and the play moves to state $(n + 1, c)$. If player 2 chooses action s then the reward is 0 and the play moves to state (n, s).
- In each state (n, s), the reward is $n/(n + 1)$ and the play moves to state s^*.
- State s^* is absorbing with reward 0.

Note that this game has perfect information, i.e. in every state only one player has more than one action (we can assume that each player has only one action in states (n, s) and s^*). In fact, this game is equivalent to the centipede game presented in Figure 1.

Take an initial state (n, c), where n is even. Player 1 is the active player at this state. We can easily verify for this initial state that

- the value is $(n + 1)/(n + 2)$,
- it is optimal for player 1 to play action c in state (n, c) and action s in state $(n + 2, c)$,
- it is optimal for player 2 to play action s in state $(n + 1, c)$.

Note that any optimal strategy of player 1 requires him/her to play action c in state (n, c) with probability 1.

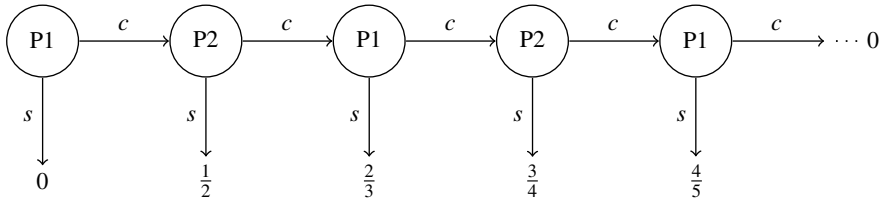

FIGURE 1: The centipede game for player 1 (P1) and player 2 (P2) with actions c and s.

Consequently, player 1 has an optimal strategy in the game. However, player 1 has no subgame-optimal strategy, as playing action c in every state (n, c), where n is even, only yields payoff 0 against the strategy of player 2 that always chooses action c.

6.2. An alternative way of attempting to prove Theorem 2

Flesch *et al.* (2016) showed that in a zero-sum stochastic game with finite state and action spaces, with respect to the limsup payoff, if player 1 has an optimal strategy then he/she also has a stationary optimal strategy. This suggests an alternative, albeit less direct, approach to proving Theorem 2.

Consider a positive zero-sum stochastic game G with a finite state space S. Assume that player 1 has an optimal strategy π. There is a natural way to transform game G into game G^* with the limsup payoff.

Let R denote the set of numbers that can arise as a finite sum of rewards in game G. For example, if rewards 0, 3, and 7 are possible in G, then R consists of $0, 3, 6 = 3+3, 7, 9 = 3+3+3, 10 = 3+7$, and so on. The set R is countable, as G has countably many different payoffs.

We now consider game G^*: the state space is $R \times S$, which is countable. The interpretation of state (r, s) is that during play in game G we are now in state s and the sum of the past rewards is r. So, in state (r, s), the set of actions is $A(s)$ for player 1 and $B(s)$ for player 2, i.e. the sets of actions in state s in the original game G. The transitions are defined according to the above interpretation. The reward in state (r, s) is defined to be r. The payoff is the limsup payoff u^*, defined for each play $p^* = ((r_0, s_0), a_0, b_0, (r_1, s_1), a_1, b_1, \ldots)$ as

$$u^*(p^*) = \limsup_{t\to\infty} r_t.$$

The strategy π for player 1 in game G has a corresponding strategy π^* in game G^*, which remains optimal. It is likely that the proof of Flesch *et al.* (2016) can be extended to this case to show that player 1 has an optimal strategy in G^*. In game G^*, under any pair of stationary strategies, we have the following important observations.

- If E is an ergodic set then E is finite. Indeed, if (r, s) and (r', s') are both in E then either state will be visited from the other one, so $r = r'$. As we assumed that S is finite, E is also finite.
- With probability 1, an ergodic set is eventually reached, as a result of Assumption 1.
- Against any fixed stationary strategy of player 1, for any $\varepsilon > 0$, player 2 has a stationary ε-best response, provided that player 2's action set is finite.
- For every $r, r' \in R$ and $s \in S$, the one-day matrix games in states (r, s) and (r', s) are identical up to adding a constant and, hence, player 1 has the same optimal mixed actions in these matrix games, say $X^*(s)$.

6.3. Extensions

It seems likely that Theorems 1, 3, and 4 can be extended to a Borel measurable setting such as in Nowak (1985). No such extension is possible for Theorem 2 as we see from Example 2.

Acknowledgements

The authors are grateful to Hugo Gimbert for our discussions on the subject. The support for the visit of William Sudderth to Maastricht provided by a travel grant (number 040.11.495) of the Netherlands Organisation for Scientific Research (NWO) is acknowledged.

References

Blackwell, D. (1970). On stationary policies. *J. R. Statist. Soc.* **133,** 33–37.

Dubins, L. E. and Savage, L. J. (1965). *How to Gamble If You Must: Inequalities for Stochastic Processes*. McGraw-Hill, New York.

Flesch, J., Predtetchinski, A. and Sudderth, W. (2018). Simplifying optimal strategies in limsup and liminf stochastic games. To appear in *Discrete Appl. Math.* Available at https://doi.org/10.1016/j.dam.2018.05.038.

Flesch, J., Thuijsman, F. and Vrieze, O. J. (1998). Simplifying optimal strategies in stochastic games. *SIAM J. Control Optimization* **36,** 1331–1347.

Frid, E. B. (1973). On stochastic games. *Theory Prob. Appl.* **18,** 389–393.

Jaśkiewicz, A. and Nowak, A. S. (2016). Zero-sum stochastic games. In *Handbook of Dynamic Game Theory*, Springer, Cham, pp. 215–279.

Kumarm, P. R. and Shiau, T. H. (1981). Existence of value and randomized strategies in zero-sum discrete-time stochastic dynamic games. *SIAM J. Control Optimization* **19,** 617–634.

Maitra, A. and Parthasarathy, T. (1971). On stochastic games. II. *J. Optimization Theory Appl.* **8,** 154–160.

Maitra, A. P. and Sudderth, W. D. (1996). *Discrete Gambling and Stochastic Games.* Springer, New York.

Nowak, A. S. (1985). Universally measurable strategies in zero-sum stochastic games. *Ann. Prob.* **13,** 269–287.

Nowak, A. S. and Raghavan, T. E. S. (1991). Positive stochastic games and a theorem of Ornstein. In *Stochastic Games and Related Topics*, Springer, Dordrecht, pp. 127–134.

Parthasarathy, T. (1971). Discounted and positive stochastic games. *Bull. Amer. Math. Soc.* **77,** 134–136.

Parthasarathy, T. (1973). Discounted, positive, and noncooperative stochastic games. *Internat. J. Game Theory* **2,** 25–37.

Puterman, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley, New York.

Strauch, R. E. (1966). Negative dynamic programming. *Ann. Math. Statist.* **37,** 871–890.

Sudderth, W. (2016). Optimal Markov strategies. Preprint.