Proceedings of the 28th Conference
on Decision and Control
Tampa, Florida • December 1989

A SAMPLE-PATH APPROACH TO STOCHASTIC GAMES

Melike Baykal-Gursoy
Industrial Engineering Department
Rutgers University
Piscataway, NJ 08854

Abstract

Two-person zero-sum stochastic games with finite state and action spaces are considered. The expected average payoff criterion is used for multichain structures. In the special case that only one player controls the transitions, it is shown that the optimal stationary policies and the value of the game can be obtained from the optimal solutions to a pair of dual linear programs. A decomposition algorithm is given which produces such optimal stationary policies for both players. In the case that both players control the transitions, a generalized game is obtained, the solution of which gives the optimal policies.

1 Introduction

In this paper we investigate the optimal policies for a two-person zero-sum stochastic game (SG) that was first introduced by Shapley [17]. The game is played sequentially. At each epoch, the game is in one of finitely many states and each player observes the current state and chooses one of finitely many actions. The state of the game and the pair of actions determines:(i) a payoff to be made by player II to player I; (ii) the probability distribution over the states of the game, which provides the transition probabilities to the next state of the game. Stochastic games generalize Markov decision processes, in that MDPs may be viewed as SGs in which one of the players has only one action in each state.

An objective function is defined depending on the evaluation of the payoffs. In a game, player I tries to maximize its gain while player II tries to minimize its loss. Shapley [17] considered the game with discounted payoffs and proved that this game has a value and that both players have optimal stationary policies. Hoffman and Karp [10] studied the long-run average payoff criterion and proved that the game has a value if the transition matrix of each pure policy pair is irreducible. If R_m denotes the payoff at epoch $m \in \mathcal{N}_+$, then the long-run average expected payoff to player I is defined as

$$\phi(\mathbf{u},\mathbf{v}) := \liminf_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} E_{\mathbf{u},\mathbf{v}}[R_m],$$

where the expectation is taken with respect to each player's policy. The problem of which games have a value is solved by Mertens and Neyman [14] who show that every stochastic game has a value, i.e.,

$$\sup_{\mathbf{u}} \inf_{\mathbf{v}} \phi(\mathbf{u},\mathbf{v}) = \inf_{\mathbf{v}} \sup_{\mathbf{u}} \phi(\mathbf{u},\mathbf{v}).$$

The average expected payoff problem has also been studied by other authors [1], [5], [11]. In general, there do not exist optimal policies for both players [3]. Bewley and Kohlberg [2] give sufficient conditions for the existence of stationary optimal policies for both players. They show that if SG is unichain or if only one player controls the transitions, then there exists optimal stationary policies for both players. Kallenberg [12] study the case when one player controls the transitions and give an algorithm to locate the optimal stationary policies. This case has also been studied by other authors [6],[7], [8], [9], [15].

We consider the *expected average criterion* (Bewley and Kohlberg define it as limit average criterion [2]),

$$\psi(\mathbf{u},\mathbf{v}) := E_{\mathbf{u},\mathbf{v}}[\liminf_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} R_m]. \qquad (1)$$

Bewley and Kohlberg [2] prove that if there exist optimal stationary policies for the long-run average expected payoff criterion, then these policies are optimal for the limit average criterion. In this paper first, we show in the special case that only one player controls the transitions that optimal stationary policies and the value of the game can be obtained from optimal solutions of a pair of dual linear programs. This problem is considered for multichain SGs. The decomposition approach is used in developing the algorithm to locate optimal policies. For the general case that both players control the transitions, we obtain a generalized game. If a solution exists for this game we can locate optimal stationary policies for both players in the unichain case.

Section 2 presents the stochastic game model with some notations and definitions. In section 3 we study the case of one player controlling the transitions and give a decomposition algorithm to locate the optimal policies. In section 4 we discuss the games with both players controlling the transitions.

2 Stochastic Game Model

Let $\{X_m\}$ denote the state process of the two-person zero-sum game, taking values from a finite state space $\mathcal{S}$. After observing the state of the game at epoch m each player chooses an action from a finite set of actions. Let $\{A_m\}$ and $\{B_m\}$ be the sequence of actions taken by player I and player II, respectively. Let $\mathcal{A}$ and $\mathcal{B}$ denote the set of available actions for player I and II, respectively.

Player II pays player I a payoff $R_m = r(X_m, A_m, B_m)$ at each epoch m. At any epoch if the system is in state $x \in \mathcal{S}$, player I choses action $a \in \mathcal{A}$ and player II choses action $b \in \mathcal{B}$, a payoff of $r(x,a,b)$ is earned. It is assumed that the payoffs are nonnegative and finite. By the time homogeneity assumption the next state of the game depends only on the present state and actions. In particular, when the system is in state $x \in \mathcal{S}$ at epoch m and actions $a \in \mathcal{A}$ and $b \in \mathcal{B}$ are chosen by player I and II, respectively, then the state at epoch $m+1$ is $y \in \mathcal{S}$ with transition probabilities P_{xaby}, statisfying

$$\sum_{y} P_{xaby} = 1, \;\; P_{xaby} \geq 0, \;\; \forall\, x,y \in \mathcal{S}\,,\; a \in \mathcal{A}\,,\; b \in \mathcal{B}\,.$$

P_{xaby} is refered to as the law of motion and assumed to be known to each player.

The probability space that supports the process $\{X_m, A_m, B_m : m \in \mathcal{N}_+\}$ is defined as follows. The underlying sample-space is $\Omega := \{\mathcal{S} \times \mathcal{A} \times \mathcal{B}\}^\infty$, so that a typical realization $\omega \in \Omega$ is represented by $\omega := (x_1, a_1, b_1, x_2, a_2, b_2, \ldots)$. Let Ω be equipped with the σ-algebra $\mathcal{F}$ generated by the random variables $\{X_m, A_m, B_m : m \in \mathcal{N}_+\}$.

A *decision rule* $\mathbf{u}^m$ (respectively $\mathbf{v}^m$) at epoch m for player I (respectively II) is a mapping from $\{\mathcal{S} \times \mathcal{A} \times \mathcal{B}\}^{m-1} \times \mathcal{S}$ to the set of all probability measures on $\mathcal{A}$ (respectively $\mathcal{B}$). Let $u_a^m(x_1,\ldots,x_m)$ (respectively $v_b^m(x_1,\ldots,x_m)$) denote the conditional probability of choosing action a (respectively b) at epoch m given the past history $(x_1,\ldots,x_m)$. A policy for player I (respectively II) is denoted by $\mathbf{u}$ (respectively $\mathbf{v}$).

Let $\xi \in \mathcal{S}$ be a fixed initial state known to the players. Policies chosen by each player $\mathbf{u}$ and $\mathbf{v}$ induce a probability measure $P_{\mathbf{u},\mathbf{v}}$ on $(\Omega, \mathcal{F})$ through the following equations:

$$P_{\mathbf{u},\mathbf{v}}\{X_1 = \xi\} = 1,$$

$$P_{\mathbf{u},\mathbf{v}}\{A_m = a | X_1 = x_1, \ldots, X_m = x\} = \mathbf{u}_a^m(x_1, \ldots, x),$$

$$P_{\mathbf{u},\mathbf{v}}\{B_m = b | X_1 = x_1, \ldots, X_m = x\} = \mathbf{v}_b^m(x_1, \ldots, x),$$

$$P_{\mathbf{u},\mathbf{v}}\{X_{m+1} = y | X_1 = x_1, \ldots, X_m = x, A_m = a, B_m = b\} = P_{xaby}.$$

Stationary policies $\mathbf{f}$ and $\mathbf{h}$ for player I and II, respectively, are vectors with components

$$u_a^m(x_1, \ldots, x) = f_{xa} \ \forall\, m \in \mathcal{N}_+,$$

$$v_b^m(x_1, \ldots, x) = h_{xb} \ \forall\, m \in \mathcal{N}_+.$$

Let C^1 and $C_S^1(C^2, C_S^2)$ denote the set of all policies and stationary policies for player I (II). Under policies $\mathbf{f} \in C_S^1$ and $\mathbf{h} \in C_S^2$ for player I and II respectively, the state process $\{X_m\}$ is a Markov chain with transition probabilities

$$P_{xy}(\mathbf{f}, \mathbf{h}) = \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{B}} P_{xaby} f_{xa} h_{xb}.$$

Define the random variables,

$$W_n(x, a, b) := \frac{1}{n} \sum_{m=1}^{n} \mathbf{1}(X_m = x, A_m = a, B_m = b).$$

Under stationary policies $\mathbf{f}$ and $\mathbf{h}$ the process $\{Y_m = (X_m, A_m, B_m)\}_{m=1}^{\infty}$ is also a Markov chain, thus,

$$W(x, a, b) = \lim_{n \to \infty} \frac{1}{n} \sum_{m=1}^{n} \mathbf{1}(X_m = x, A_m = a, B_m = b)$$

exists $P_{\mathbf{f},\mathbf{h}}$-almost surely for all $x \in \mathcal{S}$, $a \in \mathcal{A}$ and $b \in \mathcal{B}$.

A stochastic game is said to be *unichain* if the Markov chain induced by each pure policy pair $(\mathbf{g}_1, \mathbf{g}_2)$ is unichain. Let $P(\mathbf{f}, \mathbf{h})$ denote the transition matrix induced by stationary policies $\mathbf{f}$ and $\mathbf{h}$ for player I and II respectively. If $P(\mathbf{f}, \mathbf{h})$ is unichain, then there exists a unique probability vector $\pi(\mathbf{f}, \mathbf{h}) = \{\pi_x(\mathbf{f}, \mathbf{h}) : \ x \in \mathcal{S}\}$ independent of initial state.

The long-run average reward of player I is given by

$$R = \liminf_{n \to \infty} \frac{1}{n} \sum_{m=1}^{n} r(X_m, A_m, B_m).$$

If $\{W_n(x, a, b)\}$ converges for all $x \in \mathcal{S}$, $a \in \mathcal{A}$ and $b \in \mathcal{B}$ to random variable $\{W(x, a, b)\}$ $P_{\mathbf{u},\mathbf{v}}$-almost surely, then

$$R = \sum_{x,a,b} r(x, a, b) W(x, a, b), \tag{2}$$

$P_{\mathbf{u},\mathbf{v}}$-almost surely.

Player I maximizes his expected reward while player II minimizes his expected loss. Hence, the problem is to find policies $\mathbf{u}^*$ and $\mathbf{v}^*$ such that

$$\psi(\mathbf{u}, \mathbf{v}^*) \le \psi(\mathbf{u}^*, \mathbf{v}^*) \le \psi(\mathbf{u}^*, \mathbf{v}),$$

where $\psi(\mathbf{u}, \mathbf{v})$ is the expected average reward under player I's policy $\mathbf{u}$ and player II's policy $\mathbf{v}$, i.e.,

$$\psi(\mathbf{u}, \mathbf{v}) := E_{\mathbf{u},\mathbf{v}}[R].$$

If $\mathbf{u}^*$,$\mathbf{v}^*$ satisfy the above equation then $\mathbf{u}^*$ and $\mathbf{v}^*$ are called optimal policies for player I and player II, respectively. And $\psi(\mathbf{u}^*, \mathbf{v}^*)$ is called the value of the game. It is straightforward to prove the following proposition.

Proposition 1 *Let $\mathbf{f} \in C_S^1$ and $\mathbf{h} \in C_S^2$ and let $\mathcal{R}^1(\mathbf{f}, \mathbf{h}), \ldots, \mathcal{R}^p(\mathbf{f}, \mathbf{h})$ be the recurrent classes induced by $P(\mathbf{f}, \mathbf{h})$. Denote $\{\pi_x^i(\mathbf{f}, \mathbf{h}) : x \in \mathcal{R}^i(\mathbf{f}, \mathbf{h})\}$ for the equilibrium probability vector associated with class i, $i = 1, \ldots, p$. Let*

$$\tau := \min\{m : m \in \mathcal{N}_+, X_m \in \cup_{i=1}^{p} \mathcal{R}^i(\mathbf{f}, \mathbf{h})\}.$$

Then,

$$\psi(\mathbf{f}, \mathbf{h}) = \sum_{i=1}^{p} P_{\mathbf{f},\mathbf{h}}\{X_\tau \in \mathcal{R}^i(\mathbf{f}, \mathbf{h})\} \sum_{x \in \mathcal{R}^i(\mathbf{f},\mathbf{h})} \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{B}} r(x, a, b) \pi_x^i(\mathbf{f}, \mathbf{h}) f_{xa} h_{xb}.$$

If the game is unichain, then

$$\psi(\mathbf{f}, \mathbf{h}) = \sum_{x \in \mathcal{S}} \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{B}} r(x, a, b) \pi_x(\mathbf{f}, \mathbf{h}) f_{xa} h_{xb}.$$

3 One Player Controlling the Transition Probabilities

We will first investigate a special case, where the transition probabilities are assumed to be independent of the player II's actions. We consider the problem of finding the value of the game and locating the optimal stationary policies for both players. Throughout this section the following assumption holds.

Assumption: The transition probabilities P_{xaby} do not depend on b for all x, y, a, i.e.,

$$P_{\mathbf{u},\mathbf{v}}\{X_{m+1} = y | X_1 = x_1, \ldots, X_m = x, A_m = a, B_m = b\} = P_{xay}.$$

This condition implies that transitions are influenced by player I's policy $\mathbf{u}$ and the components of the transition matrix under a stationary policy are given as

$$P_{xy}(\mathbf{f}, \mathbf{h}) = \sum_{a,b} P_{xay} f_{xa} h_{xb} = P_{xy}(\mathbf{f}),$$

which is independent of second player's policy.

Remark: For $\mathbf{f} \in C_S^1$ and $\mathbf{h} \in C_S^2$ and the recurrent classes induced by $P(\mathbf{f})$, $\mathcal{R}^1(\mathbf{f}), \ldots, \mathcal{R}^p(\mathbf{f})$, the *expected average payoff* is

$$\psi(\mathbf{f}, \mathbf{h}) = \sum_{i=1}^{p} P_{\mathbf{f}}\{X_\tau \in \mathcal{R}^i(\mathbf{f})\} \sum_{x \in \mathcal{R}^i(\mathbf{f})} \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{B}} r(x, a, b) \pi_x^i(\mathbf{f}) f_{xa} h_{xb}.$$

If the game is unichain, then

$$\psi(\mathbf{f}, \mathbf{h}) = \sum_{x \in} \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{B}} r(x, a, b) \pi_x(\mathbf{f}) f_{xa} h_{xb}. \square$$

By the above assumption, one can define the strongly communicating class depending only on the law of motion $\{P_{xay}\}$ as follows:

Definition 1 *A set of states C is a strongly communicating class if*

(i) There exists a stationary policy $\mathbf{f}$ such that C is a recurrent class for the associated probability transition matrix $P(\mathbf{f})$,

(ii) C is not a proper subset of some set D which also satisfies (i).

The optimal policies for player I and II shall be constructed using this decomposition approach. The stochastic game is first restricted to each of the strongly communicating classes $C^1, \ldots, C^p$. Each restricted game corresponds to a pair of dual linear programs. Based on the value of the restricted game an aggregated

game is constructed. Since the maximizing player controls the transitions, the aggregated game corresponds to a maximum average reward MDP. An optimal stationary policy for player I is then obtained by combining the optimal policy for the aggregated game with the optimal solutions for the restricted game. The optimal policy for player II is obtained from the restricted game.

To this end, we first define the restricted game.

3.1 The Restricted Game

For each $i = 1,\ldots,s$ and $x \in \mathcal{C}^i$, define the set $\mathcal{F}_x = \{a \in \mathcal{A} : P_{xay} = 0 \text{ for all } y \notin \mathcal{C}^i\}$ as a subset of $\mathcal{A}$, player I's actions. Then the restricted game SG-i is defined as follows: (i) the state space is $\mathcal{C}^i$; (ii) the action space for player I is $\mathcal{F}_x$ for $x \in \mathcal{C}^i$, the action space for player II is $\mathcal{B}$; (iii) P_{xay} and $r(x,a,b)$ are restricted to the state space $\mathcal{C}^i$ and action spaces $\mathcal{F}_x$, $\mathcal{B}$.

For a fixed SG-i and a fixed initial state $\xi \in \mathcal{C}^i$, each pair of policies $(\mathbf{u},\mathbf{v})$ induces a probability measure $P^{i,\xi}_{\mathbf{u},\mathbf{v}}$ on $(\Omega,\mathcal{F})$. The corresponding expected average payoff for SG-i is given by

$$\psi^i(\mathbf{u},\mathbf{v}) := E^i_{\mathbf{u},\mathbf{v}}[\liminf_{n\to\infty} \frac{1}{n}\sum_{m=1}^{n} r(X_m,A_m,B_m)].$$

For each SG-i, consider the following pair of dual linear programs.

Program Q^i_1

$$T^i := \max \sum_{x\in\mathcal{C}^i} \gamma_x \tag{3}$$

$$s.t. \sum_{x\in\mathcal{C}^i}\sum_{a\in\mathcal{F}^i}(\delta_{xy} - P_{xay})z_{xa} = 0, \ \forall y \in \mathcal{C}^i \tag{4}$$

$$\sum_{x\in\mathcal{C}^i}\sum_{a\in\mathcal{F}^i} z_{xa} = 1 \tag{5}$$

$$-\sum_{a\in\mathcal{F}^i} r(x,a,b)z_{xa} + \gamma_x \le 0, \ \forall b \in \mathcal{B}, \ x \in \mathcal{C}^i \tag{6}$$

$$z_{xa} \ge 0, \ \forall x \in \mathcal{C}^i, \ a \in \mathcal{F}^i \tag{7}$$

Program Q^i_2

$$U^i := \min \varphi \tag{8}$$

$$s.t. \varphi + \sum_y(\delta_{xy} - P_{xay})t_y - \sum_b r(x,a,b)s_x(b) \ge 0, \tag{9}$$

$$\sum_b s_x(b) = 1, \ \forall x \in \mathcal{C}^i \tag{10}$$

$$s_x(b) \ge 0, \ \forall x \in \mathcal{C}^i, \ b \in \mathcal{B} \tag{11}$$

Remark: Program Q^i_1 maximizes $\sum_{x\in\mathcal{C}^i}\sum_{a\in\mathcal{F}_x} r(x,a,b)z_{xa}$.

Theorem 1 *Let $\mathbf{f}^*$ denote a stationary optimal policy for player I and let $\mathbf{h}^*$ denote a stationary policy for player II. For each $i = 1,\ldots,s$ and $\mathbf{u} \in C^1$, $\mathbf{v} \in C^2$*

$$P_{\mathbf{u},\mathbf{h}^*}\{\liminf_{n\to\infty}\frac{1}{n}\sum_{m=1}^{n} r(X_m,A_m,B_m) \le T^i|\Phi_i\} = 1,$$

$$P_{\mathbf{f}^*,\mathbf{v}}\{\limsup_{n\to\infty}\frac{1}{n}\sum_{m=1}^{n} r(X_m,A_m,B_m) \ge U^i|\Phi_i\} = 1,$$

Furthermore, for a stationary policy $\mathbf{h}$ for player II we have

$$P_{\mathbf{f}^*,\mathbf{h}}\{\lim_{n\to\infty}\frac{1}{n}\sum_{m=1}^{n} r(X_m,A_m,B_m) \ge U^i|\Phi_i\} = 1.$$

Proof: Let

$$W_n(x,a,b) := \frac{1}{n}\sum_{m=1}^{n} \mathbf{1}(X_m = x, A_m = a, B_m = b),$$

$$Z_n(x,a) := \frac{1}{n}\sum_{m=1}^{n} \mathbf{1}(X_m = x, A_m = a).$$

By compactness properties there is a subsequence $\{N_k(\omega)\}$ along which the limits $W(x,a,b;\omega)$ and $Z(x,a;\omega)$ exist,

$$\lim_{k\to\infty} W_{N_k}(x,a,b) = W(x,a,b) \ge 0, \ \forall x \in \mathcal{S}, \ a \in \mathcal{A}, \ b \in \mathcal{B},$$
$$\lim_{k\to\infty} Z_{N_k}(x,a) = Z(x,a) \ge 0, \ \forall x \in \mathcal{S}, \ a \in \mathcal{A}.$$

Let

$$S_x(b) = \frac{W(x,a,b)}{Z(x,a)}.$$

Fix a policy $\mathbf{u} \in C^1$ and $i = \{1,\ldots,s\}$. Assume that player II uses his optimal policy $\mathbf{h}^* \in C^2_S$. Clearly

$$\sum_b h^*_{xb} = 1, \ h^*_{xb} \ge 0, \ \forall x \in \mathcal{S}, \ b \in \mathcal{B}.$$

Let Γ be the set of all sample paths $\omega = (x_1,a_1,b_1,\ldots,)$ that satisfy

(i) $a_n \in \mathcal{F}_{x_n}$ for all $n \ge N$, for some $N \in \mathcal{N}$,

(ii) $\sum_{x\in\mathcal{S}}\sum_{a\in\mathcal{A}} P_{xay}Z(x,a;\omega) = \sum_{a\in\mathcal{A}} Z(y,a;\omega)$ for all $y \in \mathcal{S}$

(iii) $W(x,a,b;\omega) = h^*_{xb}Z(x,a;\omega)$

By the Strong Law of Large Numbers for Martingale Differences (e.g., see [13]) and the property that $P_{\mathbf{u}}\{A_m \in \mathcal{F}_{X_m} \ a.a.\} = 1$ for all $\mathbf{u} \in C^1$,

$$P_{\mathbf{u}}(\Gamma) = 1.$$

Since $\mathbf{u}$ and $\mathbf{h}^*$ are independent and

$$P_{\mathbf{h}^*}\{S_x(b) = h^*_{xb}\} = 1,$$

we have,

$$P_{\mathbf{u},\mathbf{h}^*}(\Gamma) = 1.$$

We want to show that all sample paths in the intersection of Γ with the set $\Phi_i = \{X_m \in \mathcal{C}^i \ a.a.\}$ satisfy

$$\liminf_{n\to\infty}\frac{1}{n}\sum_{m}^{n} r(X_m,A_m,B_m) \le T^i,$$

i.e.,

$$\Phi_i \cap \Gamma \subset \{\liminf_{n\to\infty}\frac{1}{n}\sum_{m=1}^{n} r(X_m,A_m,B_m) \le T^i\}.$$

On the set $\Phi_i \cap \Gamma$

$$W(x,a,b) = 0 \ \forall x \notin \mathcal{C}^i, \ a \notin \mathcal{F}_x$$
$$Z(x,a) = 0 \ \forall x \notin \mathcal{C}^i, \ a \notin \mathcal{F}_x,$$

giving

$$\sum_{x\in\mathcal{C}^i}\sum_{a\in\mathcal{F}_x} Z(x,a) = 1,$$

$$\sum_{x\in\mathcal{C}^i}\sum_{a\in\mathcal{F}_x} P_{xay}Z(x,a) = \sum_{a\in\mathcal{F}_y} Z(y,a), \quad \forall y \in \mathcal{C}^i.$$

Thus, $Z(x,a)$ satisfies primal conditions over the set $\Phi_i \cap \Gamma$. Hence from the primal program (9) we have

$$\sum_{x\in\mathcal{C}^i}\sum_{a\in\mathcal{F}_x}\sum_{b\in\mathcal{B}} r(x,a,b)Z(x,a)h^*_{xb} \le T^i, \tag{12}$$

on $\Phi_i \cap \Gamma$. On the other hand

$$\liminf_{n\to\infty}\frac{1}{n}\sum_{m=1}^{n} r(X_m, \ A_m \ , B_m) \le \lim_{k\to\infty}\frac{1}{N_k}\sum_{m=1}^{N_k} r(X_m,A_m,B_m)$$

$$= \sum_{x\in C^i}\sum_{a\in\mathcal{F}_x}\sum_{b\in\mathcal{B}} r(x,a,b)Z(x,a)h^*_{xb}$$

holds over the set $\Phi_i \cap \Gamma$. When combined with (12) gives the result.

Now fix a policy $\mathbf{v} \in C^2$ for player II and $i = \{1,\ldots,s\}$. Assume that player I uses his optimal policy $\mathbf{f}^* \in C^1_S$. Let Γ denote the set of all sample paths ω that statisfy

(*i*) $a_n \in \mathcal{F}_{x_n}$ for all $n \geq N$, for some $N \in \mathcal{N}$.

(*ii*) $\sum_{x\in\mathcal{S}} \sum_{a\in\mathcal{A}} P_{xay} Z(x,a;\omega) = \sum_{a\in\mathcal{A}} Z(y,a;\omega)$ for all $y \in \mathcal{S}$.

(*iii*) $W(x,a,b;\omega) = \pi_x(\mathbf{f}^*) f^*_{xa} S_x(b;\omega)$

Clearly

$$P_{\mathbf{f}^*,\mathbf{v}}(\Gamma) = 1.$$

On the set $\Phi_i \cap \Gamma$

$$\begin{aligned} W(x,a,b) &= 0, \ \forall\, x \notin C^i,\ a \notin \mathcal{F}_x \\ S_x(b) &= 0, \ \forall\, x \notin C^i, \end{aligned}$$

giving

$$\sum_{b\in\mathcal{B}} S_x(b) = 1 \ \forall\, x \in C^i.$$

Since $z^*_{xa} = \pi_x(\mathbf{f}^*) f^*_{xa}$ satisfies the condition (4) and $S_x(b)$ satisfies dual condition (13) we have from (12),

$$\begin{aligned} \sum_{x,a,b} r(x,a,b) z^*_{xa} S_x(b) &\leq \varphi + \sum_{x,a,y} (\delta_{xy} - P_{xay}) z^*_{xa} t_y \\ &= \varphi. \end{aligned}$$

Thus, dual program minimizes $\sum_{x,a,b} r(x,a,b) z^*_{xa} S_x(b)$, giving

$$\sum_{x\in C^i}\sum_{a\in\mathcal{F}_x}\sum_{b\in\mathcal{B}} r(x,a,b) z^*_{xa} S_x(b) \geq U^i$$

on the set $\Phi_i \cap \Gamma$. The result is obtained by considering the following equations.

$$\begin{aligned} \limsup_{n\to\infty} \frac{1}{n}\sum_{m=1}^{n} r(X_m,A_m,B_m) &\geq \lim_{k\to\infty} \frac{1}{N_k}\sum_{m=1}^{N_k} r(X_m,A_m,B_m) \\ &= \sum_{x\in C^i}\sum_{a\in\mathcal{F}_x}\sum_{b\in\mathcal{B}} r(x,a,b) z^*_{xa} S_x(b). \end{aligned}$$

When second player uses stationary policies $\mathbf{h}$, then the limit below exists also on the set $\Phi_i \cap \Gamma$

$$\lim_{n\to\infty} \frac{1}{n}\sum_{m=1}^{n} r(X_m,A_m,B_m) = \sum_{x\in C^i}\sum_{a\in\mathcal{F}_x}\sum_{b\in\mathcal{B}} r(x,a,b) z^*_{xa} S_x(b).$$

Hence, giving

$$P_{\mathbf{f}^*,\mathbf{h}}\{\lim_{n\to\infty} \frac{1}{n}\sum_{m=1}^{n} r(X_m,A_m,B_m) \geq U^i\} = 1 \square$$

Corollary 1 *Let* $\mathbf{f}^*$ *denote the stationary optimal policy for player I, and let* $\mathbf{h}^*$ *denote the stationary optimal policy for player II. For each SG-i, initial state* $\xi \in C^i$, *policies* $\mathbf{u} \in C^1$ *and* $\mathbf{v} \in C^2$, $\mathbf{h} \in C^2$

$$P^{\xi}_{\mathbf{u},\mathbf{h}^*}\{\liminf_{n\to\infty} \frac{1}{n}\sum_{m=1}^{n} r(X_m,A_m,B_m) \leq T^i\} = 1,$$

$$P^{\xi}_{\mathbf{f}^*,\mathbf{v}}\{\limsup_{n\to\infty} \frac{1}{n}\sum_{m=1}^{n} r(X_m,A_m,B_m) \geq U^i\} = 1,$$

$$P^{\xi}_{\mathbf{f}^*,\mathbf{h}}\{\lim_{n\to\infty} \frac{1}{n}\sum_{m=1}^{n} r(X_m,A_m,B_m) \geq U^i\} = 1,$$

holds.

Using the definitions of $z_x = \sum_a z_{xa}$ and $I^i_x = \{x \in C^i : z_x > 0\}$, one can construct stationary policies for SG-i by an algorithm similar to the one given in [12]

Algorithm 1:

1. Find optimal solutions $\mathbf{z}^* = \{z^*_{xa}\}$ and $\mathbf{s}^* = \{s^*_x(b)\}$ for the above pair of programs.
2. Define $h^{*i}_{xb} := s^*_x(b)$ if $x \in C^i$.
3. Define $f^{*i}_{xa} := \frac{z^*_{xa}}{z^*_x}$ if $x \in I^i_x$.
4. Set $E = I^i_x$.
5. While $E \neq C^i$, Do:
   - Choose a triple (x, a_x, y) such that $x \in C^i - E$, $y \in E$, $a \in \mathcal{F}_x$ and $P_{xa_xy} > 0$.
   - Set $f^{*i}_{xa_x} = 1$; and Set $E = E \cup \{x\}$.
6. Continue.

Theorem 2 *The stationary policies* $\mathbf{f}^{*i}$ *and* $\mathbf{h}^{*i}$ *are optimal for SG-i, for all* $\xi \in C^i$.

Proof: For a unichain transition matrix

$$z^*_{xa} = \pi_x(\mathbf{f}^{*i}) f^{*i}_{xa}, \quad \forall\, x \in C^i,\ a \in \mathcal{F}_x.$$

Also $s^*_x(b) = h^{*i}_{xb}$, for all $x \in C^i$. Since z^*_{xa} and $s^*_x(b)$ are optimal for the pair of mathematical programs, respectively, $T^i = U^i$ holds, giving

$$\begin{aligned} T^i &= \sum_{x\in C^i}\sum_{a\in\mathcal{F}_x}\sum_{b\in\mathcal{B}} r(x,a,b) z^*_{xa} s^*_x(b) \\ &= \sum_{x\in C^i}\sum_{a\in\mathcal{F}_x}\sum_{b\in\mathcal{B}} r(x,a,b) \pi_x(\mathbf{f}^{*i}) f^{*i}_{xa} h^{*i}_{xb} \\ &= \psi^{i,\xi}(\mathbf{f}^{*i},\mathbf{h}^{*i}) \end{aligned}$$

By Corollary 1 we have

$$\psi^{i,\xi}(\mathbf{u},\mathbf{h}^{*i}) \leq T^i,$$

and

$$\psi^{i,\xi}(\mathbf{f}^{*i},\mathbf{h}) \geq T^i.$$

We have

$$\psi^{i,\xi}(\mathbf{u},\mathbf{h}^{*i}) \leq \psi^{i,\xi}(\mathbf{f}^{*i},\mathbf{h}^{*i}) \leq \psi^{i,\xi}(\mathbf{f}^{*i},\mathbf{h}). \tag{13}$$

Since if one of the players uses a stationary policy, the problem for the other player becomes a MDP and it is well known that in this case there exist a stationary optimal policy, thus implying that it is enough to consider the optimization problem over the stationary policies [12]. Hence (13) implies the result$\square$

3.2 The Aggregated SG

Consider the aggregated game, where there is one state corresponding to each strongly communicating class C^i plus states corresponding to the transient states T. For each state $i = 1,\ldots,s$, the action θ is available, which keeps the system in state i with probability 1. Since only the maximizing player controls the transition probabilities, the actions of the form (x,a) are also available, so that the original game moves to state x and player I chooses action $a \notin \mathcal{F}_x$. Thus the aggregated SG is defined so that the state space is $\bar{\mathcal{S}} = \{1,\ldots,s+t\}$ where t denotes the cardinality of the set $\mathcal{T}$, the state-dependent action spaces for player I are

$$\mathcal{A}_{\bar{s}} = \{\theta\} \cup \{(x,a) : x \in C^i, a \notin \mathcal{F}_x\}\ i \in \bar{S}\ , 1 \le i \le s$$
$$\mathcal{A}_{\bar{s}} = \mathcal{A}\ s+1 \le i \le s+t$$

the action space for player II is $\mathcal{B}$, and the law of motion is defined accordingly. The *average reward for the aggregated game* given that the initial state is $i \in \bar{S}$ is defined as,

$$\bar{\beta}_i(\mathbf{u}) := E^i_{\mathbf{u}}[\liminf_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} \sum_{k=1}^{s} T^k\, 1(X_m = k)],$$

and the *optimal payoff for the aggregated game* is

$$\bar{\beta}_i = \sup_{u \in \mathcal{C}} \bar{\beta}_i(\mathbf{u}).$$

An optimal pure policy $\mathbf{g}$ can be found for the aggregated game. The optimal pure policy $\mathbf{g}$ of player I moves the game through state space $\bar{S}$. Hence, if $g(i) = \theta$ then, it is best to play the restricted game SG-i once it has entered the strongly communicating class C^i. But if $g(i) = (x,a)$, then player I moves the original game to state x and then chooses action a.

We can construct the stationary policies as follows. The stationary policy $\mathbf{h}$ for player II is obtained by setting

$$h^*_{xb} = \begin{cases} h^{*i}_{xb} & \text{if } x \in C^i, \\ 1 & \text{for arbitrary } b \text{ for } x \in \mathcal{T}. \end{cases}$$

And stationary policy $\mathbf{f}^*$ for player I can be obtained from the algorithm given in [1]. The following Lemma is stated without proof, since the proof is essentially the same as the proof of Lemma 3.11 in [17]

Lemma 1 *The stationary policy $\mathbf{f}^*$ constructed by the algorithm is optimal for the intermediate problem. Moreover, if i is a recurrent state under the pure policy $\mathbf{g}$ in the aggregated SG, then C^i is closed and contains exactly one recurrent class under $P(\mathbf{f}^*)$; if $i \notin H$, then*

$$P_{\mathbf{f}^*}(\Phi_i) = 0.$$

Theorem 3 *The stationary policies $\mathbf{f}^*$ and $\mathbf{h}^*$ are optimal for the original problem.*

Proof: From Proposition 1

$$E_{\mathbf{f}^*,\mathbf{h}^*}[R] = \sum_{i\in H} P^{\xi}_{\mathbf{f}^*}\{X_\tau \in C^i\} \sum_{x\in C^i} \sum_{a\in\mathcal{A}} \sum_{b\in\mathcal{B}} r(x,a,b)$$

$$\pi_s(\mathbf{f}^{*i}, \mathbf{h}^{*i}) f^{*i}_{xa} h^{*i}_{xb} = \sum_{i\in H} P^{\xi}_{\mathbf{f}^*}\{X_\tau \in C^i\} T^i = \sum_{i=1}^{p} T^i P^{\xi}_{\mathbf{f}^*}\{\Phi_i\}.$$

But from Lemma 1 and Theorem 1

$$\begin{aligned} E_{\mathbf{u},\mathbf{h}^*}[R] &= \sum_{i=1}^{p} E^{\xi}_{\mathbf{u},\mathbf{h}^*}[\liminf_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} r(X_m, A_m, B_m) | \Phi_i] P^{\xi}_{\mathbf{u}}\{\Phi_i\} \\ &\le \sum_{i=1}^{p} T^i P^{\xi}_{\mathbf{u}}\{\Phi_i\} \le E_{\mathbf{f}^*,\mathbf{h}^*}[R]. \end{aligned}$$

On the other hand,

$$\begin{aligned} E_{\mathbf{f}^*,\mathbf{h}}[R] &= \sum_{i=1}^{p} E^{\xi}_{\mathbf{f}^*,\mathbf{h}}[\liminf_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} r(X_m, A_m, B_m) |\ \Phi_i] P^{\xi}_{\mathbf{f}^*}\{\Phi_i\} \\ &\ge \sum_{i=1}^{p} T^i P^{\xi}_{\mathbf{f}^*}\{\Phi_i\} \square \end{aligned}$$

4 General Stochastic Games

In this section we investigate the stochastic game where both players control the transition probabilities. We assume that the Markov chain generated by each stationary policy pair $(\mathbf{f},\mathbf{h})$ is irreducible, i.e., there is only one recurrent class with no transient states.

Assume that player I uses a stationary policy $\mathbf{f}$, then the payoff earned from player II is

$$r_{xb}(\mathbf{f}) := \sum_a r(x,a,b) f_{xa},$$

and the transition probabilities determined by player II is

$$P_{xby}(\mathbf{f}) := \sum_a P_{xaby} f_{xa}.$$

Then, player II will try to minimize his loss, which gives rise to the following mathematical program

$$\begin{aligned} &\min \sum_{x,b} r_{xb}(\mathbf{f}) v_{xb} \\ s.t.\ &\sum_{x,b} P_{xby}(\mathbf{f}) v_{xb} = \sum_b v_{yb},\ \ \forall\, y \in \mathcal{S} \\ &\sum_{x,b} v_{xb} = 1, \\ &v_{xb} \ge 0,\ \ \forall\, x \in \mathcal{S}\ ,\ b \in \mathcal{B}\ . \end{aligned}$$

Now let

$$f_{xa} = \frac{z_{xa}}{\sum_a z_{xa}}$$

with $z_{xa} \ge 0$, and $\sum_a z_{xa} = \sum_b w_{xb}$ for all $x \in \mathcal{S}$. And consider the following minimax problem.

$$\begin{aligned} T &:= \max_{\mathbf{z}} \min_{\mathbf{v}} \sum_{x,a,b} r(x,a,b) \frac{z_{xa}}{\sum_a z_{xa}} v_{xb} \\ &= \min_{\mathbf{v}} \max_{\mathbf{z}} \sum_{x,a,b} r(x,a,b) \frac{z_{xa}}{\sum_a z_{xa}} v_{xb} && (14) \\ &\quad \sum_{x,a,b} P_{xaby} \frac{z_{xa}}{\sum_a z_{xa}} v_{xb} = \sum_b v_{xb},\ \ \forall\, y \in \mathcal{S} && (15) \\ &\quad \sum_a z_{xa} = \sum_b v_{xb},\ \ \forall\, x \in \mathcal{S} && (16) \\ &\quad \sum_{x,b} v_{xb} = 1, && (17) \\ &\quad v_{xb} \ge 0,\ \ z_{xa} \ge 0,\ \ \forall\, x \in \mathcal{S}\ ,\ a \in \mathcal{A}\ ,\ b \in \mathcal{B} && (18) \end{aligned}$$

Along the same lines of proof of Theorem 1, one can prove

Theorem 4 *Let $\mathbf{f}^* \in C^1_S$ and $\mathbf{h}^* \in C^2_S$ be optimal policies for player I and II, respectively. Then for all policies $\mathbf{f} \in C^1_S$ and $\mathbf{h} \in C^2_S$*

$$P_{\mathbf{f}^*,\mathbf{h}}\{\limsup_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} r(X_m, A_m, B_m) \ge T\} = 1,$$

$$P_{\mathbf{f},\mathbf{h}^*}\{\liminf_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} r(X_m, A_m, B_m) \le T\} = 1.$$

Theorem 5 *Suppose that $\{z^*_{xa}\}$ and $\{v^*_{xb}\}$ are solutions to the static game. If $\mathbf{f}^*$ and $\mathbf{h}^*$ are obtained through the transformations*

$$f^*_{xa} = \begin{cases} \frac{z^*(x,a)}{z^*(x)} \text{ if } z^*(x) = \sum_a z^*(x,a) > 0 \\ \text{arbitrary but } f^*_{xa} > 0, \forall a \in \mathcal{A}, \text{ otherwise} \end{cases}$$

$$h^*_{xb} = \begin{cases} \frac{v^*(x,b)}{v^*(x)} \text{ if } v^*(x) = \sum_b v^*(x,b) > 0 \\ \text{arbitrary but } h^*_{xb} > 0, \forall b \in B, \text{ otherwise} \end{cases}$$

then they are optimal for the stochastic game.

Proof: From the constraints we have

$$v^*(x) = z^*(x),$$

$$\sum_x v^*(x) = 1,$$

$$v^*(y) = \sum_{x,a,b} P_{xaby} v^*(x,b) f^*_{xa} = \sum_{x,b} P_{xby}(\mathbf{f}^*) h^*_{xb} v^*(x)$$
$$= \sum_x P_{xy}(\mathbf{f}^*, \mathbf{h}^*) v^*(x).$$

Since there exist a unique probability vector associated with $P(\mathbf{f}^*, \mathbf{h}^*)$, these equations imply that

$$\pi_x(\mathbf{f}^*, \mathbf{h}^*) = v^*(x), \quad \forall\, x \in \mathcal{S}\,.$$

Thus, we have

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} r(X_m, A_m, B_m) = \sum_{x,a,b} r(x,a,b) \pi_x(\mathbf{f}^*, \mathbf{h}^*) f^*_{xa} h^*_{xb}$$

$$= \sum_{x,a,b} r(x,a,b) \frac{v^*(x,b) z^*(x,a)}{v^*(x)} = T$$

The result follows from Theorem 1□

Thus, we can obtain optimal stationary policies for both players from the static game, if this minimax problem has a solution.

Remark: Note that in this minimax problem the policies of players are not independent. A game with this added constraint is called *generalized game*[4].

References

[1] M. Baykal-Gursoy and K. W. Ross. Variance-penalized Markov decision processes. *submitted.*

[2] T. Bewley and E. Kohlberg. The asymptotic theory of stochastic games. *Mathematics of Operations Research*, 1:197–208, 1976.

[3] T. Bewley and E. Kohlberg. On stochastic games with stationary optimal strategies. *Mathematics of Operation Research*, 3:104–125, 1978.

[4] D. Blackwell and T.S. Ferguson. The big match. *Annals of Mathematical Statistics*, 39:159–163, 1968.

[5] K.C. Border. *Fixed Point Theorems with Applications to Economics and Game Theory.* Cambridge Univ. Press, 1985.

[6] A. Federgruen. successive approximation methods in undiscounted stochastic games. *Operations Research*, 28:794–809, 1980.

[7] J.A. Filar. Ordered field property for stochastic games when the player who controls transitions changes from state to state. *Journal of Optimization Theory and Appl.*, 34:503–515, 1981.

[8] J.A. Filar. On stochastic equilibria of a single-controller stochastic game. *Mathematical Programming*, 30:313–325, 1984.

[9] J.A. Filar. Quadratic programming and the single-controller stochastic game. *Journal of Math. Analysis and Appl.*, 113:136–147, 1986.

[10] J.A. Filar and T.A. Schultz. Nonlinear programming and stationary stategies in stochastic games. *Mathematical Programming*, pages 243–247, 1986.

[11] A.J. Hoffman and R.M. Karp. On nonterminating stochastic games. *Management Science*, 12(5):359–370, 1966.

[12] F. Thuijsman J.A. Filar, T.A. Schultz and D.J. Vrieze. Nonlinear programming and stationary equilibria in stochastic games. Technical report, UMBC, August 1987.

[13] L.C.M. Kallenberg. *Linear Programming and Finite Markovian Control Problems*, volume 148. Mathematical Centre Tracts, Amsterdam, 1983.

[14] M. Loeve. *Probability Theory*, volume 2. Springer-Verlag, New York, 1978.

[15] J.F. Mertens and A. Newman. Stochastic games. *Int. Journal of Game Theory*, 10(2):53–66, 1981.

[16] T. Parthasarathy and T.E.S. Raghavan. An orderfield property for stochastic games when one player controls transition probabilities. *Journal of Optimization Theory and Applications*, 33(3):375–392, 1981.

[17] K.W. Ross and R. Varadarajan. The decomposition of time average markov decision processes: Theory, algorithms and applications. *submitted.*

[18] L.S. Shapley. Stochastic games. *Proc. Nat. Acad. Sci. U.S.A.*, 39:1095–1100, 1953.