Policy Improvement for Perfect Information Additive Reward and Additive Transition Stochastic Games With Discounted and Average Payoffs

Matthew Bourque, T.E.S. Raghavan
Department of Mathematics, Statistics, and Computer Science
University of Illinois at Chicago

July 18, 2012

Abstract

We give a policy improvement algorithm for perfect information and additive reward, additive transition zero-sum two-player stochastic games for both discounted and average payoffs. In the case of discounted payoffs, our algorithm, though identical to one by Raghavan and Syed, has a new proof. The new approach has the distinct advantage of allowing us to prove a similar algorithm for the more difficult case of average payoffs for these same classes of games.

1 Preliminaries

1.1 Introduction

A stochastic game, introduced in a seminal paper of Shapley [1953], is a system which evolves over discrete time steps, representing an ongoing interaction between two players in which each of players' choices at each time step not only determine an immediate payoff for each player, but also influence the actions and payoffs which will be available at the next time step.

The notion of value in a stochastic game depends on how the players evaluate their infinite stream of payoffs. Shapley's paper proved that stochastic games possess a discounted value (when players evaluate their payoff streams by taking a discounted sum). Mertens and Neyman [1982] proved that these games also have an undiscounted or limiting average value (when players evaluate their payoff streams by taking a limiting average).

Stochastic games with rational rewards and transitions need not have rational values [Filar and Vrieze, 1997, example 3.2.1]. It is clear that for such a game, no finite-step algorithm using only arithmetic operations can give an exact answer, and we can only hope for such an algorithm when the game has the order-field property defined in [Parthasarathy and Raghavan, 1981]. Fortunately, many natural classes of games have been shown to possess this property. However, even among games with the order field property, the general existence of finite-step algorithms for computing the value and optimal strategies for zero-sum stochastic games with respect to discounted or undiscounted criteria is an open question. The finite-step algorithms which have been found for some categories of stochastic games are generally of two types: via a linear program, or policy improvement. In this paper we examine policy improvement algorithms for solving two player zero-sum stochastic games of perfect information with respect to the discounted and limiting average criteria. The family of policy improvement algorithms begins with Howard's policy improvement algorithm for discounted MDPs [Howard, 1960]. Policy improvement algorithms are widely recognized as fast in practice for solving MDPs and other related problems [Blackwell, 1962, Veinott, 1966, Cochet-terrasson et al., 1998, Raghavan and Syed, 2003]; indeed, this speed is suggested by the fact that under certain regularity conditions, policy improvement for MDPs is equivalent to Newton's method [Filar and Vrieze, 1997, Section 2.7].

A policy improvement algorithm for zero sum stochastic games of perfect information with respect to the discounted criterion is given by Raghavan and Syed [2003]. We provide an alternate proof for this result based on an intermediate theorem which avoids the induction argument used in their proof and also allows us to prove that a modified version of Raghavan and Syed's algorithm can solve games in which the limiting average criterion is used.

To the best of our knowledge, the first policy improvement algorithm for solving stochastic games of perfect information with average payoffs was given by Cochet-Terrasson and Gaubert [2006]. Their main proof technique

hinges on Kohlberg's theorem on invariant half-lines of nonexpansive piecewise linear transformations [Kohlberg, 1980]. Our work is quite different, following a more strictly game-theoretic approach beginning with Shapley [1953], through Blackwell [1962], Veinott [1966], and Raghavan and Syed [2003]. This has the advantage of yielding an algorithm not only for games of perfect information but also for the more general class of additive reward, additive transition (ARAT) games introduced in [Raghavan et al., 1985]. Although this latter class includes games of perfect information, we nonetheless focus most of the paper on the perfect information case, waiting until Section 4 to extend our results to ARAT games. This is because the perfect information case is a more natural class of games, so that the proofs are somewhat simpler, and we hope the results are more immediately applicable.

1.2 Definition of a stochastic game

A **stochastic game** $\Gamma = \langle S, \mathbf{A}^1, \mathbf{A}^2, \mathbf{R}^1, \mathbf{R}^2, \mathbf{P} \rangle$ comprises

- a finite set of states $S = \{1, 2, \ldots, N\}$;
- a set of finite nonempty action sets $\mathbf{A}^i = \{A^i(s)\}_{s \in S}$ for player $i \in \{1, 2\}$;
- a set of rewards

$$\mathbf{R}^i = \left\{ r^i(s, a^1, a^2) \in \mathbb{R} \middle| a^1 \in A^1(s), a^2 \in A^2(s), s \in S \right\}$$

  for player i, $i \in \{1, 2\}$;
- and a set of Markovian transition probabilities

$$\mathbf{P} = \left\{ p(s' \mid s, a^1, a^2) \in [0, 1] \middle| a^1 \in A^1(s), a^2 \in A^2(s), s, s' \in S \right\},$$

  where $\sum_{s'=1}^{N} p(s' \mid s, a^1, a^2) = 1$ for all $a^1 \in A^1(s), a^2 \in A^2(s)$ for all states s.

We will refer to the players in the game as player 1 and player 2. (When distinct pronouns enhance readability, we will use the convention that player 1 is male and player 2 is female.) We interpret the reward $r^i(s, a^1, a^2)$ as the immediate payoff to player i when player 1 chooses action a^1 and player 2 chooses action a^2 in state s, and interpret $p(s' \mid s, a^1, a^2)$ as the Markovian

probability of transition to state s' in the next time step when the state at the current time step is s, player 1 chooses action a^1 and player 2 chooses action a^2.

We will restrict our attention to zero-sum stochastic games of perfect information, defined below.

Definition 1.1. A stochastic game Γ is a **zero-sum** game if $r^1(s, a^1, a^2) = -r^2(s, a^1, a^2)$ for all actions $a^1 \in A^1(s), a^2 \in A^2(s)$ available in state s.

Definition 1.2. A stochastic game Γ has **perfect information** if at most one of $A^1(s)$ and $A^2(s)$ has more than one element for all states s. In case $A^1(s)$ has more than one element, we well say the state "belongs to" player 1 (and analogously for player 2). In case $A^i(s)$ is a singleton for $i = 1, 2$, we will say that state s is a "chance" state.

For simplicity, when we speak of a "game," we will mean a zero-sum stochastic game of perfect information, unless otherwise specified. We will also refer to a single payoff function $r(s, a^1, s^2) \equiv r^1(s, a^1, a^2)$ with the understanding that, with respect to the discounted or average payoffs corresponding to r, player 1 is a maximizer and player 2 is a minimizer.

1.3 Strategies

A "strategy" for a player is a (possibly randomized) rule for selecting an action at each time step, possibly depending on the current state and the entire history of the game. Strategies which choose an action deterministically in each state, independent of the history of the game, form an important subclass, called pure stationary strategies.

Definition 1.3. A **pure stationary strategy** f for player i in a game Γ is an element of $\times_{s=1}^{N} A^k(s)$, where $f(s)$ represents the action which player i will choose in state s.

We will denote by F (respectively G) the set of pure stationary strategies for player 1 (respectively player 2), with these strategy sets' dependence on the game understood.

Given a pair of strategies $(f, g) \in F \times G$ for the players in a game, Let $P(f, g)$ be the stationary transition matrix whose (s, s') entry is $p\big(s' \mid s, f(s), g(s)\big)$. We also define a column N-vector $r(f, g)$ whose s-th entry is $r\big(s, f(s), g(s)\big)$.

Note that when one player in game fixes his or her strategy, the induced game for the opponent is a Markov decision process (MDP). We will denote by $\Gamma|_f$ the MDP resulting from the game Γ when one player's strategy is fixed at f. If f is a strategy for player 2, then $\Gamma|_f$ is an MDP for player 1 where payoffs are viewed as rewards and player 1 is a maximizer. If f is a strategy for player 1, then $\Gamma|_f$ is an MDP for player 2 where payoffs are viewed as costs and player 2 is a minimizer.

2 Games With discounted payoffs

2.1 Discounted payoffs and discounted value

The discounted payoff with a discount factor $\beta \in [0, 1)$ for a stochastic game models a situation in which players are more concerned with the relatively short term, with β inversely proportional to myopia.

Definition 2.1. For a given discount factor $\beta \in [0, 1)$, and a pure stationary strategy pair (f, g), the ***β*-discounted payoff** is

$$\begin{aligned} v_\beta(f, g) &= \sum_{t=0}^{\infty} \beta^t P^t(f, g) r(f, g) \\ &= \big[I - \beta P(f, g)\big]^{-1} r(f, g). \end{aligned}$$

Definition 2.2. A pure stationary strategy pair (f^*, g^*) for a game Γ is ***β*-optimal** (over pure stationary strategies) for a discount factor $\beta \in [0, 1)$ if there exists a ***β*-discounted value** $v_\beta(\Gamma)$ such that

$$v_\beta(\Gamma) = \max_{f \in F} v_\beta(f, g^*)$$

and

$$v_\beta(\Gamma) = \min_{g \in G} v_\beta(f^*, g).$$

The set of all possible strategies for a stochastic game is much larger than $F \times G$ - it includes strategies which may randomize over the actions in states or which may depend not only on the current state but on the history of play so far. However, zero sum stochastic games of perfect information may be played optimally over pure stationary strategies [Shapley, 1953], and so there is no loss in limiting ourselves to such strategies for these games. We will use

the word "strategy" to refer to a pure stationary strategy unless otherwise specified.

2.2 Policy improvement for discounted games

In this section, we present an alternative proof for the policy improvement algorithm for zero sum stochastic games of perfect information with discounted payoffs due to Raghavan and Syed [2003]. The key advantage of this alternative proof is that it provides the tools to prove a similar algorithm for games with average payoffs, discussed in the next section. Both algorithms can be seen as extensions of the policy improvement algorithm for MDPs introduced in [Howard, 1960] and thoroughly discussed in [Blackwell, 1962]. We begin with some notation.

For a strategy pair (f, g) and a fixed discount factor β, let $G_\beta(s, f \mid g)$ be the (possibly empty) set of actions $a \in A^1(s)$ which satisfy

$$r(s,a) + \beta \sum_{s'=1}^{N} p\big(s' \mid s, a, g(s)\big) v_\beta(s', f, g) > v_\beta(s, f, g), \tag{1}$$

and let $G_\beta(f \mid g)$ be the (possibly empty) set of all pure strategies $f' \neq f$ for player 1 such that, for each state s, $f'(s) \in G_\beta(s, f \mid g)$ or $f'(s) = f(s)$ for all $s \in S$. We define the sets $G_\beta(s, g \mid f)$ and $G_\beta(g \mid f)$ for player 2 similarly with the inequality in (1) reversed.

The following theorem is simply a restatement of [Blackwell, 1962, Theorem 3] in this competitive setting.

Theorem 2.3. *Given a strategy pair (f, g) for a game with discount factor β, if $f' \in G_\beta(f \mid g)$ then $v_\beta(f', g) > v_\beta(f, g)$ and if $G_\beta(f \mid g)$ is empty, then $v_\beta(f, g) \geq v_\beta(f' \mid g)$ for strategies f' for player 1. That is, f is β-optimal for the MDP $\Gamma|_g$.*

Of course, the situation is analogous for player 2 with the inequalities reversed: if $G_\beta(g \mid f)$ is empty, then g is β-optimal for the MDP $\Gamma|_f$. The following corollary is immediate.

Corollary 2.4. *If $G_\beta(f \mid g)$ and $G_\beta(g \mid f)$ are both empty, then (f, g) is a β-optimal strategy pair for the game.*

The following algorithm is almost identical to the one proved in [Raghavan and Syed, 2003]. We will present later on a new proof of its correctness. First,

we will give an intuitive picture of how the algorithm works. We can imagine the following method for two players to try to arrive at an optimal strategy pair. Both players agree on an arbitrary starting policy. Player 1 reasons greedily: with the current strategy of player 2 fixed, he searches until he has found a β-optimal strategy. Now player 2 reasons patiently: having waited for player 1 to find a β-optimal response, she finds a strategy that is guaranteed to be better (though not necessarily the best) against the strategy player 1 just found. Now the chance to improve returns to player 1 and the procedure continues. It turns out that such balance of greedy and patient approaches is guaranteed to end when both players can no longer improve their strategies at all, meaning that they must have found a β-optimal strategy pair. We now present the algorithm formally.

Algorithm 2.5 (Policy Improvement for Discounted Payoffs)**:**

1: Choose an arbitrary initial strategy pair (f^1, g^1), and let $k = 1$.
2: **while** $G_\beta(f^k \mid g^k)$ or $G_\beta(g^k \mid f^k)$ is nonempty **do**
3: Find f for player 1 so that $G_\beta(f \mid g^k)$ is empty. Let $f^{k+1} = f$.
4: **if** $G_\beta(g^k \mid f^{k+1})$ is nonempty **then**
5: Update the strategy for player 2: choose $g^{k+1} \in G_\beta(g^k \mid f^{k+1})$.
6: **else**
7: Let $g^{k+1} = g^k$
8: **end if**
9: Increment k.
10: **end while**
11: When $G_\beta(f^k \mid g^k)$ and $G_\beta(g^k \mid f^k)$ are both empty, return the final strategy pair $(f^*, g^*) = (f^k, g^k)$.

This is more properly a set of algorithms, depending on the exact implementation of the players' choices in lines (3) and (5). The implementation of the algorithm is discussed further in Section 5.

Theorem 2.6. *Algorithm 2.5 will terminate in finite steps, and the returned strategy pair is β-optimal for the game.*

Our proof of the algorithm is based on the following theorem, which can be seen as a justification for the patience of player 2.

Theorem 2.7 (Patience Theorem)**.** *Suppose* (f^0, g^0) *is a strategy pair and* g^1 *a strategy for player 2 for a game* Γ *with a fixed discount factor* β *such that* $G_\beta(f^0 \mid g^0)$ *is empty and* $g^1 \in G_\beta(g^0 \mid f^0)$*. Then* $v_\beta(\Gamma|_{g^0}) > v_\beta(\Gamma|_{g^1})$*.*

In our proof of this patience theorem, and in the rest of the paper, we will write $x > y$ for two comparable N-vectors x and y when $x(s) \geq y(s)$ for all $s \in S$, and the inequality is strict for some s.

Proof. Consider the MDP $\Gamma|_g$ for some $g \in G$. The solution of such an MDP by a linear program is well known [Filar and Vrieze, 1997]. It is the solution to the LP

$$\text{minimize} \sum_{s=1}^{N} \gamma(s)v(s)$$

subject to

$$v(s) \geq r\big(s, a, g(s)\big) + \beta \sum_{s'=1}^{N} p\big(s' \mid s, a, g(s)\big)v(s') \text{ for } a \in A^1(s), s \in S$$

where γ is any positive N vector with $\sum_s \gamma(s) = 1$.

Since $G_\beta(f^0 \mid g^0)$ is empty, f^0 is β-optimal for the MDP $\Gamma|_{g^0}$, and so $v_\beta(f^0, g^0)$ is the optimal solution to the LP corresponding to $\Gamma|_{g^0}$. In particular, it is feasible for this LP. We will show that $v_\beta(f^0, g^0) - \epsilon_X$ is also feasible for the LP corresponding to $\Gamma|_{g^1}$, where ϵ_X is an N-vector whose coordinates are strictly positive for states contained in some nonempty set X (to be defined later), and zero otherwise. This will show that, for any appropriate γ,

$$\sum_{s=1}^{N} \gamma(s)\big(v_\beta(s, f^0, g^0) - \epsilon_X(s)\big) \geq \sum_{s=1}^{N} \gamma(s)v_\beta(s, \Gamma|_{g^1}). \tag{2}$$

From here, the theorem is proved as follows: for any state s, we choose a sequence of positive vectors $\{\gamma_n\}$ with $\sum_t \gamma_n(t) = 1$ for all n, and such that $\gamma_n(s) \to 1$ as $n \to \infty$. Replacing γ with γ_n in (2) and taking a limit as $n \to \infty$ on both sides, yields for all states s, $v_\beta(s, f^0, g^0) - \epsilon_X(s) \geq v_\beta(s, \Gamma|_{g^1})$. Since $\epsilon_X(s)$ is positive for some s and $v_\beta(\Gamma|_{g^0}) = v_\beta(f^0, g^0)$, this gives the theorem. We now proceed to prove that $v_\beta(f^0, g^0) - \epsilon_X$ is feasible for the LP corresponding to $\Gamma|_{g^1}$. We will define ϵ_X appropriately along the way.

In states s where $g^1(s) = g^0(s)$, for any action $a \in A^1(s)$

$$\begin{aligned} v_\beta(s, f^0, g^0) &\geq r\big(s, a, g^0(s)\big) + \beta \sum_{s'=1}^{N} p\big(s' \mid s, a, g^0(s)\big) v_\beta(s', f^0, g^0) \\ &= r\big(s, a, g^1(s)\big) + \beta \sum_{s'=1}^{N} p\big(s' \mid s, a, g^1(s)\big) v_\beta(s', f^0, g^0). \end{aligned} \tag{3}$$

Now let X be the set of states s for which $g^1(s) \neq g^0(s)$. For any such state, we must have $g^1(s) \in G_\beta(s, g^0 \mid f^0)$, and s must be a state for player 2, so that $A^1(s)$ is a singleton. Therefore, by definition of the set $G_\beta(s, g^0 \mid f^0)$,

$$v_\beta(s, f^0, g^0) > r\big(s, a, g^1(s)\big) + \beta \sum_{s'=1}^{N} p\big(s' \mid s, a, g^1(s)\big) v_\beta(s', f^0, g^0) \tag{4}$$

for all $a \in A^1(s)$.

Now strict inequality in (4) for $s \in X$ here means that there is some $\epsilon_s > 0$ such that we can replace $v_\beta(s, f^0, g^0))$ in (4) with $v_\beta(s, f^0, g^0) - \epsilon_s$ with strict inequality maintained.

Let ϵ_X be the N-vector whose entries are ϵ_s in states $s \in X$ and 0 otherwise. By the preceding argument, (3), and (4), $v_\beta(f^0, g^0) - \epsilon_X$ is feasible for the linear program corresponding to $\Gamma|_{g^1}$. This yields (2) and the theorem. □

We now employ this theorem to prove Theorem 2.6, the correctness of Algorithm 2.5.

Proof of Theorem 2.6. If Algorithm 2.5 returns a strategy pair (f^*, g^*), this is because $G_\beta(f^* \mid g^*)$ and $G_\beta(g^* \mid f^*)$ are both empty and so (f^*, g^*) is β-optimal by Corollary 2.4. It therefore suffices to show that the algorithm can never arrive at a strategy pair it has already visited: since there are only a finite number of pure stationary strategies, the algorithm must therefore terminate. To this end, suppose the algorithm does not stop after the kth time throught the while loop, meaning player 2 chooses g^{k+1} from the nonempty set $G_\beta(g^k \mid f^{k+1})$. Since $G_\beta(f^{k+1} \mid g^k)$ is empty, we can apply Theorem 2.7 to conclude that $v_\beta(\Gamma|_{g^k}) > v_\beta(\Gamma|_{g^{k+1}})$. That is, the pure stationary strategies chosen by player 2 define MDPs for player 1 with strictly decreasing β-discounted values, so they cannot recur. The algorithm must terminate, and the theorem is proved. □

3 Games with average payoffs

3.1 Average payoffs and average value

The limiting average payoff for a strategy pair models a situation in which players are concerned only with long term payoffs.

Definition 3.1. For a given pure stationary strategy pair (f, g), the limiting average or Cesaro average payoff (also called simply the average payoff or undiscounted payoff) is

$$v(f,g) = \lim_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} P^t(f,g) r(f,g)$$

Definition 3.2. A strategy pair (f^*, g^*) for a game Γ is **average optimal** for the game if there exists an **average value** $v(\Gamma)$ such that

$$v(\Gamma) = \max_{f\in F} v(f, g^*)$$

and

$$v(\Gamma) = \min_{g\in G} v(f^*, g).$$

Another kind of optimal strategy pair which will interest us is the uniform optimal strategy pairs, which models situations in which players do not necessarily agree on a discount factor for the game, but want strategies which will do well for all discount factors sufficiently close to 1.

Definition 3.3. A pure stationary strategy pair $(\hat{f}, \hat{g})$ is **uniform optimal** for a given game Γ if it is β-optimal for all β sufficiently close to 1.

A uniform optimal strategy for an MDP is average optimal, but as [Blackwell, 1962, example 1] shows, an average optimal strategy may not be β-optimal for any value of β.

Shapley [1953] proved the existence of β-optimal pure stationary strategies for all discount factors β for stochastic games of perfect information. Further, in the case of MDPs, Blackwell [1962] showed that the β-discounted payoff for any fixed pure stationary strategy is a rational function of β, and proved that a pure stationary uniform optimal strategy exists for any MDP. A similar argument applies for any stochastic game with pure stationary

β-optimal strategy pairs, and so we may conclude that games of perfect information admit of a pure stationary uniform optimal pair. As Blackwell showed for MDPs, such a strategy pair also serves as an average optimal strategy pair. This justifies limiting ourselves to pure stationary strategy pairs when defining average and uniform optimal strategy pairs.

3.2 Properties of average payoffs

In this subsection, we collect a few results which will be useful later. We will make extensive use of the following theorem relating β-discounted and average payoffs, which is a restatement of [Blackwell, 1962, Theorem 4, part (a)].

Theorem 3.4. *For a game Γ and strategy pair $(f, g) \in F \times G$*

$$v_\beta(f, g) = \frac{v(f, g)}{1 - \beta} + y(f, g) + \epsilon(f, g, \beta)$$

where $y(f, g)$ is the unique solution to

$$\big(I - P(f, g)\big)y = r(f, g) - v(f, g) \text{ and } P^*(f, g)y = 0$$

and $\epsilon(f, g, \beta) \to 0$ as $\beta \to 1$.

We will follow the notation used in this theorem consistently, using ϵ to denote a function which depends on β and which goes to zero as β increases to 1. Also, we will refer to the vector $y(f, g)$ as the **deviation** of the strategy pair (f, g).

Some results from Markov chain theory will be helpful in proofs or for computation in implementing algorithms. We collect them here; details may be found in Blackwell [1962] and Kemeny and Snell [1960].

Lemma 3.5. *In the case of a pure stationary strategy pair (f, g),*

(a) the average payoff $v(f, g)$ may be computed as

$$v(f, g) = P^*(f, g)r(f, g),$$

and is the unique solution to

$$\big(I - P(f, g)\big)v = 0 \text{ and } P^*(f, g)v = v;$$

where $P^*(f, g)$ is the Cesaro limit of the matrix $P(f, g)$.

(b) The vector $y(f,g)$ may be computed as

$$y(f,g) = D(f,g)r(f,g),$$

where $D(f,g) = \big(I - P(f,g) + P^*(f,g)\big)^{-1} - P^*(f,g)$.

3.3 Policy improvement for games with average payoffs

Our next task is to modify Algorithm 2.5 to solve games with a limiting average criterion. We begin by developing some notation.

Given a strategy pair (f,g) for a game Γ, let $G(s, f \mid g)$ be the (possibly empty) set of actions $a \in A^1(s)$ which satisfy

$$\sum_{s'=1}^{N} p\big(s' \mid s, a, g(s)\big) v(s', f, g) \geq v(s, f, g) \tag{5}$$

and in case (5) holds with equality

$$r\big(s, a, g(s)\big) + \sum_{s'=1}^{N} p\big(s' \mid s, a, g(s)\big) y(s', f, g) > v(s, f, g) + y(s, f, g) \tag{6}$$

Let $G(f \mid g)$ be the (possibly empty) set of all pure strategies $f' \neq f$ for player 1 such that for each state s, $f'(s) \in G(s, f \mid g)$ or $f'(s) = f(s)$. We define the sets $G(s, g \mid f)$ and $G(g \mid f)$ (for player 2) analogously, with the inequalities in (5) and (6) reversed. By a theorem in [Blackwell, 1962], if $G(f \mid g)$ is empty, then f is average optimal for the MDP $\Gamma|_g$.

Given a strategy pair (f,g) for a game Γ, let $H(s, f \mid g)$ be the (possibly empty) set of actions $a \in A^1(s)$ for which (5) and (6) are both equalities and furthermore

$$\sum_{s'=1}^{N} p\big(s' \mid s, a, g(s)\big) z(s', f, g) > y(s, f, g) + z(s, f, g) \tag{7}$$

where $z(f,g)$ is the unique solution to

$$\big(I - P(f,g)\big)z = -y(f,g) \text{ and } P^*(f,g)z = 0.$$

Let $H(f \mid g)$ be the set of pure strategies $f' \neq f$ for the maximizer with $f'(s) \in H(s, f \mid g)$ or $f'(s) = f(s)$ for all s. As before, we define the sets $H(s, g \mid f)$ and $H(g \mid f)$ for player 2 analogously, with the inequality in (7) reversed.

The following theorem is a restatement of [Veinott, 1966, Theorem 6] to the competitive setting.

Theorem 3.6. *For a strategy pair* (f, g)

(a) if $f' \in G(f \mid g)$ *then* $v(f', g) \geq v(f, g)$ *and if* $v(f', g) = v(f, g)$, *then* $y(f', g) > y(f, g)$;

(b) if $f' \in H(f \mid g)$ *then* $v(f', g) = v(f, g)$, $y(f', g) \geq y(f, g)$, *and if* $y(f', g) = y(f, g)$, *then* $z(f', g) > z(f, g)$;

(c) if $G(f \mid g)$ *is empty, then* f *is average optimal for the MDP* $\Gamma|_g$;

(d) if $G(f \mid g) \cup H(f \mid g)$ *is empty, then* $y(f, g) \geq y(f', g)$ *for all* $f' \in F$ *which are average optimal for* $\Gamma|_g$.

Of course, as with Theorem 2.3, a similar result holds with the inequalities reversed when g is a strategy for player 1 and f is a strategy for player 2. In particular, the analog of part (c) states that if $G(g \mid f)$ is empty, then g is average optimal for $\Gamma|_f$. We have an immediate corollary characterizing average optimal strategy pairs for a game.

Corollary 3.7. *If* $G(f \mid g)$ *and* $G(g \mid f)$ *are both empty, then* (f, g) *is a average optimal strategy pair for the game.*

We also include here a technical lemma which will be useful later. It is a direct application of Theorem 3.4, and relates the criteria for membership in $G(s, g \mid f)$ and $G_\beta(s, g \mid f)$ for a state s and a strategy pair (f, g). An analogous lemma holds with the roles of f and g reversed.

Lemma 3.8. *For strategy pair* (f, g) *and a state* s *in a game,* $a \in G_\beta(s, g \mid f)$

if

$$\begin{aligned}\frac{1}{1-\beta}\Big[\sum_{s'=1}^{N} p\big(s' \mid s, f(s), a\big)v(s', f, g) - v(s, f, g)\Big] \\ + \Big[r\big(s, f(s), a\big) + \sum_{s'=1}^{N} p\big(s' \mid s, f(s), a\big)y\big(s', f(s), a\big) \\ - \sum_{s'=1}^{N} p\big(s' \mid s, f(s), a\big)v(s', f, g) - y(s, f, g)\Big] \\ + \epsilon(s, a, f, g, \beta) < 0,\end{aligned} \tag{8}$$

for a function $\epsilon(s, a, f, g, \beta)$ *which goes to zero as* $\beta \nearrow 1$.

Proof. If $a \in G_\beta(s, g \mid f)$, by definition (1) holds. Letting g' be the strategy for player 2 with $g'(t) = g(t)$ for $t \neq s$ and $g'(s) = a$, we rewrite this criterion as

$$\big[r(f, g') + \beta P(f, g')v_\beta(f, g) - v_\beta(f, g)\big]_s < 0.$$

Now, consider the bracketed vector expression above. Using Theorem 3.4, we can rewrite it as

$$\begin{aligned}r(f, g') + \frac{\beta}{1-\beta}P(f, g')v(f, g) + \beta P(f, g')y(f, g) \\ - \frac{1}{1-\beta}v(f, g) - y(f, g) + \beta P(f, g')\epsilon(f, g, \beta) - \epsilon(f, g, \beta),\end{aligned}$$

which can be rearranged to give

$$\begin{aligned}\frac{1}{1-\beta}\big[P(f, g')v(f, g) - v(f, g)\big] \\ + \big[r(f, g') + P(f, g')y(f, g) - P(f, g')v(f, g) - y(f, g)\big] \\ - (1-\beta)P(f, g')y(f, g) + \beta P(f, g')\epsilon(f, g, \beta) - \epsilon(f, g, \beta).\end{aligned}$$

Now the sth coordinate of this vector expression is strictly less than zero, which is exactly (8) with $\epsilon(s, a, f, g, \beta)$ equal to the sth coordinate of the expression in the last line above. □

We are now prepared to present our main result: a policy improvement algorithm for computing optimal strategies for games with average payoffs.

Algorithm 3.9 (Policy Improvement for Average Payoffs)**:**

1: Choose an arbitrary initial strategy pair (f^0, g^0), and let $k = 0$.
2: **while** $G(f^k \mid g^k) \cup H(f^k \mid g^k)$ or $G(g^k \mid f^k)$ is nonempty **do**
3: Choose f for player 1 (the maximizer) so that $G(f \mid g^k) \cup H(f \mid g^k)$ is empty. Let $f^{k+1} = f$.
4: **if** $G(g^k \mid f^{k+1})$ is nonempty **then**
5: Update the strategy for player 2: choose $g^{k+1} \in G(g^k \mid f^{k+1})$.
6: **else**
7: Let $g^{k+1} = g^k$
8: **end if**
9: Increment k.
10: **end while**
11: When $G(f^k \mid g^k) \cup H(f^k \mid g^k)$ and $G(g^k \mid f^k)$ are both empty, return the strategy pair $(f^*, g^*) = (f^k, g^k)$.

As in the discounted case, the algorithm terminates when $G(f^k \mid g^k)$ and $G(g^k \mid f^k)$ are both empty. Hence, by Corollary 3.7, the pair is average optimal and so our goal will be to prove that the algorithm cannot cycle. In the discounted case we were able to show that each improvement by player 2 presents player 1 with a new MDP whose optimal value is strictly smaller than the last. Here, instead of looking directly at the average value of the MDPs for player 1 corresponding to the strategies chosen by player 2 in line 5 of Algorithm 3.9, we will rely on the same discounted result; showing that for β sufficiently close to 1, the β-discounted value of these MDPs is strictly decreasing. Our method will be to show that the sequence of strategies $(f^0, g^0), (f^1, g^1), \ldots$ can be "shadowed" by a sequence for the β-discounted algorithm. That is, we will show that for discount factors sufficiently close to 1, the algorithm for the discounted game can be made to follow a path that is in a sense "very similar to" the path for the algorithm for the limiting average game. The next three lemmas will help us make this precise.

First, we will show that although the strategy that player 1 finds in line 3 may not itself be uniform optimal for $\Gamma|_{g^k}$, it must have the same value and deviation as the uniform optimal strategy for this MDP.

Lemma 3.10. *Given a strategy pair (f, g) for a game Γ, if $G(f \mid g) \cup H(f \mid g)$ is empty then for any f^* uniform optimal for $\Gamma|_g$, $v(f^*, g) = v(f, g)$ and $y(f^*, g) = y(f, g)$.*

Proof. If $G(f \mid g) \cup H(f \mid g)$ is empty then by Theorem 3.6, $v(f,g) \geq v(f^*,g)$ and if $v(f^*,g) = v(f,g)$ then $y(f,g) \geq y(f^*,g)$. Now for all β sufficiently close to 1, f^* is β-optimal for $\Gamma|_g$, so $v_\beta(f^*,g) - v_\beta(f,g) \geq 0$ or, rewriting using Theorem 3.4:

$$\frac{1}{1-\beta}\big[v(f^*,g) - v(f,g)\big] + \big[y(f^*,g) - y(f,g)\big] + \epsilon(f',f,g,\beta) \geq 0.$$

Since the above inequality holds for all β sufficiently close to 1, $v(f^*,g) \geq v(f,g)$, so $v(f^*,g) = v(f,g)$. But then, since $\epsilon(f',f,g,\beta) \to 0$ as $\beta \nearrow 1$, $y(f^*,g) \geq y(f,g)$, and we conclude that $y(f^*,g) = y(f,g)$. □

Now we prove two lemmas (and accomanying corollaries) that establish some relationships between pairs of strategy improvement sets which exist for discount factors sufficiently close to 1.

Lemma 3.11. *Given a strategy pair (f,g), for β sufficiently close to 1, $G(s,g \mid f) \subset G_\beta(s,g \mid f)$.*

Proof. By definition, if $a \in G(s,g \mid f)$ then

$$\sum_{s'=1}^{N} p\big(s' \mid s, f(s), a\big) v(s',f,g) - v(s,f,g) \leq 0 \tag{9}$$

and if equality holds, then

$$r\big(s,f(s),a\big) + \sum_{s'-1}^{N} p\big(s' \mid s, f(s), a\big) y(s',f,g) - v(s,f,g) - y(s,f,g) < 0. \tag{10}$$

Now, for any given β, an action $a \in G_\beta(s,g \mid f)$ when

$$r\big(s,f(s),a\big) + \beta \sum_{s'=1}^{N} p(s' \mid s, f(s), a) v_\beta(s',f,g) - v_\beta\big(s,f(s),a\big) < 0,$$

or equivalently (by Theorem 3.4) when

$$\begin{aligned}\frac{1}{1-\beta}\Bigg[\sum_{s'=1}^{N} p\big(s' \mid s, f(s), a\big)v(s', f, g) - v(s, f, g)\Bigg] \\ + \Bigg[r\big(s, f(s), a\big) + \sum_{s'=1}^{N} p\big(s' \mid s, f(s), a\big)y\big(s', f(s), a\big) \\ - \sum_{s'=1}^{N} p\big(s' \mid s, f(s), a\big)v(s', f, g) - y(s, f, g)\Bigg] \\ + \epsilon(s, a, f, g, \beta) < 0,\end{aligned} \tag{11}$$

where $\epsilon(s, a, f, g, \beta) \to 0$ as $\beta \nearrow 1$. We wish to show that we can choose β close enough to 1 that (11) must hold whenever (9) and (10) hold.

The first bracketed expression in (11) is the left side of the inequality in (9). Consequently, when the inequality in (9) is strict, the strict inequality in (11) holds for β sufficiently close to 1. When equality holds in (9), then the second bracketed expression in (11) is equivalent to the left side of the inequality in (10), and since this is then strict and $\epsilon(s, a, f, g, \beta) \to 0$ as $\beta \nearrow 1$, for β sufficiently close to 1, the strict inequality must hold in (11). □

The following corollary follows from this result and the definition of $G(g \mid f)$ and $G_\beta(g \mid f)$.

Corollary 3.12. *Given a strategy pair (f, g), for a discount factor β sufficiently close to 1, $G(g \mid f) \subseteq G_\beta(g \mid f)$.*

Proof. Suppose that $g' \in G(g \mid f)$. For any state s for player 2 for which $g'(s) \neq g(s)$, choose β close enough to 1 that Lemma 3.11 holds, so $g'(s) \in G_\beta(s, g \mid f)$. □

Lemma 3.13. *Given strategies f^1 and f^2 for player 1 and g for player 2 for a game Γ with $v(f^1, g) = v(f^2, g)$ and $y(f^1, g) = y(f^2, g)$, for all discount factors β sufficiently close to 1 and any state s for player 2, $G_\beta(s, g \mid f^1) = G_\beta(s, g \mid f^2)$.*

Proof. Using Theorem 3.4, we write the criterion for membership in $G_\beta(s, g \mid$

$f^i), i = 1, 2$ as

$$\begin{aligned}\frac{1}{1-\beta}\Big[\sum_{s'=1}^{N} p\big(s' \mid s, f^i(s), a\big)v(s') - v(s)\Big] \\ + \Big[r\big(s, f^i(s), a)\big) + \sum_{s'=1}^{N} p\big(s' \mid s, f^i(s), a\big)y(s') \\ - \sum_{s'=1}^{N} p\big(s' \mid s, f^i(s), a\big)v(s') - y(s)\Big] \\ + \epsilon(s, a, f^i, g, \beta) < 0,\end{aligned} \tag{12}$$

where $v(s) = v(s, f^1, g) = v(s, f^2, g)$ and $y(s) = y(s, f^1, g) = y(s, f^2, g)$.

Choose β close enough to 1 so that for $i = 1, 2$ (12) holds if and only if the first bracketed expression is nonpositive and, if the first bracketed expression is zero, then the second bracketed expression is strictly negative. In particular, with such a β fixed, any action $a \in G_\beta(s, g \mid f^1)$ satisfies

$$\sum_{s'=1`}^{N} p\big(s' \mid s, f^1(s), a\big)v(s') - v(s) \leq 0, \tag{13}$$

and if this holds with equality then

$$r(s, f^1(s), a) + \sum_{s'=1}^{N} p\big(s' \mid s, f^1(s), a\big)y(s') - v(s) - y(s) < 0. \tag{14}$$

When the state s belongs to player 2, then $f^1(s)$ in (13) can be replaced with $f^2(s)$, and if equality holds here then we can also replace f^1 with f^2 in (14), so $a \in G_\beta(s, g \mid f^2)$. □

Corollary 3.14. *Given strategies f^1 and f^2 for player 1 and g for player 2 for a game Γ with $v(f^1, g) = v(f^2, g)$ and $y(f^1, g) = y(f^2, g)$, for all discount factors β sufficiently close to 1 and any state s for player 2, $G_\beta(g \mid f^1) = G_\beta(g \mid f^2)$.*

Proof. Choose β close enough to 1 that Lemma 3.13 holds for all states for player 2, and suppose that $g' \in G_\beta(g \mid f^1)$. By the perfect information property, any state for in which $g'(s) \neq g(s)$, must be for player 2, and we have $g'(s) \in G_\beta(s, g \mid f^1) = G_\beta(s, g \mid f^2)$ by lemma 3.13. □

We now come to our main result. The bulk of the work for this result has already been done in the preceding three lemmas.

Theorem 3.15. *Algorithm (3.9) will terminate, and the returned strategy pair (f^*, g^*) is average optimal for the game Γ.*

Proof. Suppose that the algorithm does not stop after the kth time through the while loop, and consider the strategy pairs (f^k, g^k) after the k-th execution of line 3, so that $G(f^k \mid g^k) \cup H(f^k \mid g^k)$ is empty, and next player 2 will choose $g^{k+1} \in G(g^k \mid f^k)$. We will show the following monoticity claim: for any discount factor β sufficiently close to 1, $v_\beta(\Gamma|_{g^k}) > v_\beta(\Gamma|_{g^{k+1}})$.

Let $\hat{f}^k$ be a uniform optimal strategy for player 1 in the MDP $\Gamma|_{g^k}$. By Lemma 3.10, our choice of f^k means that $v(f^k, g^k) = v\big(\hat{f}^k, g^k\big)$ and $y(f^k, g^k) = y\big(\hat{f}^k, g^k\big)$. Also, although $G_\beta(f^k \mid g^k)$ may be nonempty for all discount factors β, we can choose β close enough to 1 that $G_\beta\big(\hat{f}^k \mid g^k\big)$ is empty. Furthermore we can make sure that β is close enough to 1 that

$$g^{k+1} \in G(g^k \mid f^k) \subseteq G_\beta(g^k \mid f^k) = G_\beta\big(g^k \mid \hat{f}^k\big).$$

Here, the set inclusion follows from the definition of our algorithm, the first subset relation follows from Corollary 3.12, and the set equality follows from Corollary 3.14. Finally Theorem 2.7, the Patience Theorem, proves our monoticity claim. Since there are only a finite number of pure stationary strategies available to player 2, the algorithm must terminate, returning (f^*, g^*). When the algorithm terminates, $G(f^* \mid g^*)$ and $G(g^* \mid f^*)$ are both empty. The average optimality of the pair then follows from Corollary 3.7. □

4 Games with additive rewards and additive transitions

The class of stochastic games with additive rewards and additive transitions was introduced by Raghavan et al. [1985], and includes perfect information games.

Definition 4.1. A stochastic game has the additive reward, additive transition (ARAT) property if for every pair of actions $a^1 \in A^1(s)$ and $a^2 \in A^2(s)$ in every state s, $r(s, a^1, a^2) = r^1(s, a^1) + r^2(s, a^2)$ and $p(s' \mid s, a^1, a^2) = p^1(s' \mid s, a^1) + p^2(s' \mid s, a^2)$ for every state s'.

Our main goal is to prove that Algorithm 2.5 and Algorithm 3.9, for games with discounted and average payoffs, respectively, work not only for games of perfect information, but also in the more general setting of ARAT games. In this section, we will take "game" to mean a two player, zero sum ARAT game unless otherwise specified. In the case of discounted payoffs, our result is already proved by Syed [Syed, 1999]; here we simply provide a new proof which avoids Syed's inductive argument. In the case of average payoffs, our work is entirely new.

A review of the results used to prove the correctness of these algorithms shows that the perfect information property is used in three places: to establish the existence of pure stationary discount optimal, average optimal and uniform optimal strategies for both players; and in proving Theorem 2.7 and Lemma 3.13. The existence of discount, average, and uniform optimal pure stationary strategies for both players in ARAT games is proved in [Raghavan et al., 1985]. In the remainder of this section, we will provide the details for proofs for a theorem and a lemma corresponding to Theorem 2.7 and Lemma 3.13 using the ARAT property.

Theorem 4.2. *Suppose (f^0, g^0) is a strategy pair for a game Γ with a fixed discount factor β such that $G_\beta(f^0 \mid g^0)$ is empty and $g^1 \in G_\beta(g^0 \mid f^0)$. Then $v_\beta(\Gamma|_{g^0}) > v_\beta(\Gamma|_{g^1})$.*

Proof. The idea of the proof is identical to that of Theorem 2.7: with $v^0 = v_\beta(f^0, g^0)$, we must prove that $v^0 - \epsilon_X$ is feasible for the LP corresponding to $\Gamma|_{g^1}$ for some N-vector ϵ_X which is positive in states s contained in some nonempty set X and zero otherwise. We proceed as before, by considering the states s for which $g^1(s) = g^0(s)$ and then s for which $g^1(s) \in G_\beta(s, g^0 \mid f^0)$. These latter states compose the set X.

As before, when $g^1(s) = g^0(s)$,

$$v^0(s) \geq r\big(s, a, g^1(s)\big) + \beta \sum_{s'=1}^{N} p\big(s' \mid s, a, g^1(s)\big) v^0(s') \tag{15}$$

for $a \in A^1(s)$.

Further, for any $s \in X$,

$$v^0(s) > r\big(s, f^0(s), g^1(s)\big) + \beta \sum_{s'=1}^{N} p\big(s' \mid s, f^0(s), g^1(s)\big) v^0(s') \tag{16a}$$

or, by the ARAT property,

$$v^0(s) > \big[r^1\big(s, f^0(s)\big) + r^2\big(s, g^1(s)\big)\big] + \beta \sum_{s'=1}^{N} \big[p^1\big(s' \mid s, f^0(s)\big) + p^2\big(s' \mid s, g^1(s)\big)\big] v^0(s') \quad (16b)$$

We want to show that this inequality holds when we replace $f^0(s)$ by any action $a \in A^1(s)$ in (16a) or, equivalently, in (16b).

Toward this end, for any action $a \in A^1(s)$, since v^0 is optimal (and so feasible) for the LP corresponding to Γ_{g^0} we have

$$v^0(s) \geq \big[r^1\big(s, a\big) + r^2\big(s, g^0(s)\big)\big] + \beta \sum_{s'=1}^{N} \big[p^1(s' \mid s, a) + p^2\big(s' \mid s, g^0(s)\big)\big] v^0(s') \quad (17a)$$

Now the above is an equality when $a = f^0(s)$. Replacing $v^0(s)$ with this equivalent expression, making use of the ARAT property, and taking a difference yields

$$0 \geq \big[r^1\big(s, a\big) - r^1\big(s, f^0(s)\big)\big] + \beta \sum_{s'=1}^{N} \big[p^1(s' \mid s, a) - p^1\big(s' \mid s, f^0(s)\big)\big] v^0(s') \quad (17b)$$

Now summing (16b) and (17b) we have

$$v^0(s) > \big[r^1\big(s, a\big) + r^2\big(s, g^1(s)\big)\big] + \beta \sum_{s'=1}^{N} \big[p^1(s' \mid s, a) + p^2\big(s' \mid s, g^1(s)\big)\big] v^0(s') \quad (18)$$

for $s \in X$, as desired.

Now by the strict inequality in (18), for any $s \in X$ we can choose $\epsilon(s)$ small enough that, for any action a for player 1 in state s, the strict inequality is preserved when we replace $v_\beta(s, f^0, g^0)$ by $v_\beta(s, f^0, g^0) - \epsilon(s)$. We now let $\epsilon_X(s) = \epsilon(s)$ for $s \in X$ and $\epsilon_X(s) = 0$. As before, this argument along with (15) and (18) yield the feasibility of $v_\beta(f^0, g^0) - \epsilon_X$ for the LP corresponding to $\Gamma|_{g^1}$, and the proof is complete. □

Lemma 4.3. *Given strategies f^1 and f^2 for player 1 and g for player 2 with $G(f^1 \mid g)$ empty, $v(f^1, g) = v(f^2, g)$, and $y(f^1, g) = y(f^2, g)$, for all discount factors β sufficiently close to 1 , $G_\beta(s, g \mid f^1) = G_\beta(s, g \mid f^2)$.*

Proof. First, note that the definition of $G(f^i \mid g)$ depends only on the values of $v(f^i, g)$ and $y(f^i, g^i)$, i=1,2. Since these values are the same for f^1 and f^2, we have $G(f^1 \mid g) = G(f^2 \mid g)$, and so the role of these two strategies for player 1 can be exchanged in the statement of the theorem. It therefore suffices to prove $G_\beta(s, g \mid f^1) \subseteq G_\beta(s, g \mid f^2)$ for β sufficiently close to 1.

As in the proof of Lemma 3.13, we can choose β close enough to 1 that an action $a \in G_\beta(s, g \mid f^1)$ must satisfy

$$\sum_{s'=1`}^{N} p\big(s' \mid s, f^1(s), a\big)v(s') - v(s) \le 0, \tag{19}$$

and if this holds with equality then

$$r(s, f^1(s), a) + \sum_{s'=1}^{N} p\big(s' \mid s, f^1(s), a\big)y(s') - v(s) - y(s) < 0. \tag{20}$$

Of course, we cannot use the perfect information property here to replace $f^1(s)$ with $f^2(s)$. Instead note that since $G(f^1 \mid g)$ is empty then for the action $f^2(s)$ we must have

$$\sum_{s'=1}^{N} p\big(s' \mid f^2(s), g(s)\big)v(s') \le v(s) = \sum_{s'=1}^{N} p\big(s' \mid s, f^1(s), g(s)\big)v(s') \tag{21}$$

(where the equality follows from Lemma 3.5, part (a)) and if equality holds throughout (21), then

$$\begin{aligned} &r\big(s, f^2(s), g(s)\big) \\ &\quad + \sum_{s'=1}^{N} p\big(s' \mid s, f^2(s), g(s)\big)y(s') \le v(s) + y(s) \\ &\qquad\qquad = r\big(s, f^1(s), g(s)\big) \\ &\qquad\qquad\quad + \sum_{s'=1}^{N} p\big(s' \mid s, f^1(s), g(s)\big)y(s') \end{aligned} \tag{22}$$

(where the equality follows from the definition of $y(f^1, g)$ in Theorem 3.4).

Now we apply the ARAT property as in the previous proof: taking the difference of the right sides from the left sides of (21) and (22) yields

$$\sum_{s'=1}^{N} \big[p^1\big(s' \mid s, f^2(s)\big) - p^1(s' \mid s, f^1(s))\big]v(s') \le 0. \tag{23}$$

and, if equality holds here, then

$$\big[r^1\big(s, f^2(s)\big) - r^1\big(s, f^1(s)\big)\big] + \sum_{s'=1}^{N} \big[p^1\big(s' \mid s, f^2(s)\big) - p^1\big(s' \mid s, f^1(s)\big)\big] y(s') \leq 0. \tag{24}$$

Now suppose at least one of (19) and (23) holds with a strict inequality. Then summing these two inequalities gives

$$\sum_{s'=1`}^{N} p\big(s' \mid s, f^2(s), a\big) v(s') - v(s) < 0,$$

so that $a \in G_\beta(s, g \mid f^2)$. If both of (19) and (23) are equalities, then summing them gives

$$\sum_{s'=1`}^{N} p\big(s' \mid s, f^2(s), a\big) v(s') - v(s) = 0,$$

and (24) and (14) both hold. Summing these gives

$$r(s, f^2(s), a) + \sum_{s'=1}^{N} p\big(s' \mid s, f^2(s), a\big) y(s') - v(s) - y(s) < 0,$$

so that again $a \in G_\beta(s, g \mid f^2)$. This proves the lemma. □

5 Implementation of the algorithms

Similar considerations apply in implementing Algorithm 2.5 and Algorithm 3.9. We will discuss here the implementation of the latter (for the average case), with some of our observations applying to both.

For a fixed g for player 2, player 1 must find an f such that $G_\beta(f \mid g)$ is empty. This can be achieved with the policy improvement algorithm for MDPs in [Veinott, 1966], which repeatedly improves the strategy for player 1 against a fixed g until arriving at an f for which $G(f \mid g)$ is empty. The proof of this algorithm relies on the monotonicity properties given in Theorem 3.6 when one player's strategy is fixed. For details on policy iteration in MDPs, see Blackwell [1962], Veinott [1966].

When choosing an element of $G(f \mid g)$ or $G(g \mid f)$, one has to decide whether to choose only adjacent strategies (that is, those for which only one action is changed) or to allow choices which make changes in more than one state. The question of which approach is more efficient clearly depends on the game, and can be answered only in a general way by experimentation. In our implementation of both algorithms, we have limited the choices of improvements at each iteration to adjacent strategies for simplicity.

We can compute the average payoff as

$$P^*(f,g)r(f,g)$$

and we also may need $y(f,g) = D(f,g)r(f,g)$ and $z(f,g) = -D^2(f,g)r(f,g)$. In order to compute the Cesaro limit $P^*(f,g)$ we can use use the algorithm in [Fox and Landi, 1968] for computing the ergodic classes of $P(f,g)$.

A prototype implementation of Algorithms 2.5 and 3.9, written in Python, can be found at http://www.math.uic.edu/~mbourque/stochgame.html. The Python module includes various functions supporting these algorithms, including a class for representing stochastic matrices with a method for computing the Cesaro limit of a stochastic matrix, as well as functions for generating random stochastic games and for reading and writing stochastic game data to text files.

References

David Blackwell. Discrete dynamic programming. *The Annals of Mathematical Statistics*, 33(2):719–726, June 1962. ISSN 00034851. URL http://www.jstor.org/stable/2237546. ArticleType: primary_article / Full publication date: Jun., 1962 / Copyright 1962 Institute of Mathematical Statistics.

Jean Cochet-Terrasson and Stphane Gaubert. A policy iteration algorithm for zero-sum stochastic games with mean payoff. *Comptes Rendus Mathematique*, 343(5):377–382, September 2006. ISSN 1631-073X. doi: 10.1016/j.crma.2006.07.011. URL http://www.sciencedirect.com/science/article/B6X1B-4KM46WY-3/2/c67353a1c485fa8b8c24cee471f6b2e4.

Jean Cochet-terrasson, Guy Cohen, Stephane Gaubert, Michael Mc Gettrick, and Jean-pierre Quadrat. Numerical computation of spectral el-

ements in max-plus algebra. 1998. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.36.8156.

Jerzy A. Filar and Koos Vrieze. *Competitive Markov Decision Processes*. Springer, 1997. ISBN 9780387948058.

B. L. Fox and D. M. Landi. An algorithm for identifying the ergodic subchains and transient states of a stochastic matrix. *Commun. ACM*, 11(9):619–621, 1968. doi: 10.1145/364063.364082. URL http://portal.acm.org/citation.cfm?id=364063.364082.

Ronald A. Howard. *Dynamic Programming and Markov Process*. The MIT Press, first edition edition, June 1960. ISBN 0262080095.

John G. Kemeny and James Laurie Snell. *Finite Markov Chains*. Springer, 1960. ISBN 9780387901923.

Elon Kohlberg. Invariant half-lines of nonexpansive piecewise-linear transformations. *Mathematics of Operations Research*, 5(3):366–372, August 1980. ISSN 0364765X. URL http://www.jstor.org/stable/3689443. ArticleType: primary_article / Full publication date: Aug., 1980 / Copyright 1980 INFORMS.

Jean-Francois Mertens and Abraham Neyman. Stochastic games have a value. *Proceedings of the National Academy of Sciences of the United States of America*, 79(6):2145–2146, March 1982. ISSN 0027-8424. PMID: 16593178 PMCID: 346143.

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, March 1981. doi: 10.1007/BF00935250. URL http://dx.doi.org/10.1007/BF00935250.

T. E. S. Raghavan, S. H. Tijs, and O. J. Vrieze. On stochastic games with additive reward and transition structure. *Journal of Optimization Theory and Applications*, 47(4):451–464, December 1985. doi: 10.1007/BF00942191. URL http://dx.doi.org/10.1007/BF00942191.

T.E.S. Raghavan and Zamir Syed. A policy-improvement type algorithm for solving zero-sum two-person stochastic games of perfect information. *Mathematical Programming*, 95(3):513–532, March 2003.

doi: 10.1007/s10107-002-0312-3. URL http://dx.doi.org/10.1007/s10107-002-0312-3.

L. S. Shapley. Stochastic games. *Proceedings of the National Academy of Sciences of the United States of America*, 39(10):1095–1100, October 1953. doi: VL-39. URL http://www.pnas.org/content/39/10/1095.short.

Zamir Uddin Syed. *Algorithms for stochastic games and related topics*. PhD thesis, University of Illinois at Chicago, Chicago, IL, USA, 1999. AAI9941509.

Arthur F. Veinott. On finding optimal policies in discrete dynamic programming with no discounting. *The Annals of Mathematical Statistics*, 37(5): 1284–1294, October 1966. ISSN 00034851. URL http://www.jstor.org/stable/2239082. ArticleType: research-article / Full publication date: Oct., 1966 / Copyright 1966 Institute of Mathematical Statistics.