A Policy Improvement Algorithm for Some Classes of Stochastic Games

BY

MATTHEW BOURQUE
B.S., Oberlin College, 1998
M.S., University of Illinois at Chicago, 2009

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Mathematics
in the Graduate College of the
University of Illinois at Chicago, 2013

Chicago, Illinois

Defense Committee:

T.E.S. Raghavan, Chair and Advisor
Joel Brown, Biological Sciences
Shmuel Friedland; Mathematics, Statistics, and Computer Science
Stéphane Gaubert; INRIA Saclay Île-de-France and CMAP, École Polytechnique
Jan Verschelde; Mathematics, Statistics, and Computer Science
Jie Yang; Mathematics, Statistics, and Computer Science

I dedicate this thesis to my parents, for their steadfast love and support of me always, and in particular as I have made my way through, out of, and back into the academic world.

ACKNOWLEDGMENTS

Without the support of my wife, Araidia, this thesis, and indeed all of my graduate work, would have been impossible. She has bravely joined me on the roller coaster ride of graduate education, and has helped bring me through with love, support, advice, and patience. My deep gratitude also goes to my advisor, T.E.S. Raghavan, who kindled my interest in game theory. He is truly a model scholar: brilliant, focused, and generous with his time and talent. His spirit of adventure and energy in research, his love of mathematics and game theory, and his respect for the human and social values which underlie any vibrant research community have been inspirational to me in my work at UIC, and will continue to inform me in the future. I also extend my thanks to the members of my thesis committee, Prof. Joel Brown, Prof. Shmuel Friedland, Prof. Jan Verschelde, Prof. Jie Yang, and Prof. Stéphane Gaubert, for their helpful guidance. Special thanks are due to Prof. Gaubert for taking the time to visit Chicago to serve on the committee, and for his invaluable help in locating an error in a draft of the thesis.

MJB

TABLE OF CONTENTS

LIST OF ABBREVIATIONS AND NOTATION

ARAT	additive rewards, additive transitions
F, G	pure stationary strategy sets for player 1 and player 2
$G_\beta(f \mid g)$	set of discounted improvement strategies, see Section 3.2 (see Section 2.1 for the MDP case)
$G(f \mid g)$	set of average improvement strategies, see Section 3.3 (see Section 2.2.2 for the MDP case)
$H(f \mid g)$	set of one-optimal improvement strategies, see Section 3.3 (see Section 2.2.3 for the MDP case)
MDP	Markov decision process
N	largest state index in a game or MDP
$r(s, a, b)$	immediate reward in state s when players choose actions $a \in A^1(s)$ and $b \in A^2(s)$
$\boldsymbol{r}(f, g)$	immedate reward vector induced by pure stationary strategies f and g
$p(s' \mid s, a, b)$	probability of transition from state s to state s' when players choose actions $a \in A^1(s)$ and $b \in A^2(s)$

LIST OF ABBREVIATIONS AND NOTATION (Continued)

$P(f, g)$ (one-step) transition matrix induced by a pure stationary strategy pair (f, g)

$P^{(t)}(\pi, \rho)$ t-step transition matrix induced by a pure stationary strategy pair (π, ρ)

$P^*(f, g)$ Cesaro limit of the transition matrix $P(f, g)$ for a pure stationary strategy pair (f, g), see Section 2.2.1

$v_\beta(s, \pi, \rho)$ β-discounted payoff using strategy pair (π, ρ) starting in state s

$\boldsymbol{v}_\beta(\pi, \rho)$ β-discounted payoff vector using strategy pair (π, ρ), indexed by starting state

$v(s, \pi, \rho)$ average payoff using strategy pair (π, ρ) starting in state s

$\boldsymbol{v}(\pi, \rho)$ average payoff vector using stratey pair (π, ρ), indexed by starting state

$\boldsymbol{y}(f, g)$ deviation vector associated with a pure stationary strategy pair (f, g). See Section 2.2.1

$\Gamma|_g$ The MDP for one player in a stochastic game Γ induced by the other player fixing a strategy g

SUMMARY

In Chapter 1 and Chapter 2 we survey existing results: we define Markov decision processes, strategies, and payoffs, and give the results underlying policy improvement algorithms for discounted and average-payoff Markov decision processes.

In Chapter 3, we define zero-sum stochastic games with additive rewards and additive transitions and present new results: a new proof for an existing policy improvement algorithm for such games with discounted payoffs, and a proof for a new policy improvement algorithm for additive reward, additive transition zero-sum two-player stochastic games for both discounted and average payoffs. This class of games includes perfect information zero-sum games.

Chapter 4 presents some numerical results from a Python implementation of our algorithm, and discusses possible future directions for this work.

CHAPTER 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 proved that zero-sum stochastic games possess a discounted value (when players evaluate their payoff streams by taking a discounted sum), and 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 by Parthasarathy and Raghavan (1981). Fortunately, many natural classes of games have been shown to possess this property, including the additive reward additive transition (ARAT) games, introduced by Raghavan, Tijs, and

Vrieze (1985), which are the focus of our work. ARAT games include the class of perfect information games. However, even among zero-sum games with the order field property, the general existence of finite-step algorithms for computing the value and optimal strategies with respect to discounted or undiscounted criteria is an open question. The finite-step algorithms which have been found for some categories of zero-sum 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 with additive rewards and additive transitions with respect to the discounted and limiting average criteria. The family of policy improvement algorithms begins with Howard's policy improvement algorithm for discounted Markov decision processes (MDPs) (Howard, 1960). Policy improvement algorithms are widely recognized as quite 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 discounted MDPs is equivalent to Newton's method (Filar and Tolwinski, 1991).

Raghavan and Syed establish a policy improvement algorithm for zero sum stochastic games of perfect information with respect to the discounted criterion (Raghavan and Syed, 2003). Syed extends the proof of correctness of the algorithm to additive reward additive transition games with discounted payoffs (Syed, 1999). The methods used in these proofs do not extend directly to the case of average payoffs. Instead, we prove a result conjectured by Raghavan and Syed, which provides an alternate proof for the algorithm for the discounted case. Furthermore, we extend this proof to our main result, a policy improvement algorithm for ARAT games with average payoffs.

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-Terrason and Gaubert (2006), with a more thorough treatment in a later paper (Akian et al., 2012). The main proof technique in these papers hinges on Kohlberg's theorem on invariant half-lines of nonexpansive piecewise linear transformations (Kohlberg, 1980). Though producing a similar algorithm, our work is completely independent of these results, and the techniques are quite different, following a more purely game-theoretic approach beginning with Shapley (1953), making use of results of Blackwell (1962), Veinott (1966), and Raghavan and Syed (2003). Avrachenkov et. al. (2010) apply Horkdijk and Kallenberg's (1985) parametric linear programming method for computing uniform optimal strategies for MDPs to compute uniform optimal strategy pairs for perfect information stochastic games.

1.2 Markov decision processes

Stochastic games generalize Markov decision processes, and our discussion of stochastic games will depend on several results on this simpler case. In the remainder of this chapter, we define Markov decision processes and related concepts.

A **Markov decision process** or MDP $\mathbf{M} = \langle S, \mathbf{A}, \mathbf{R}, \mathbf{P} \rangle$ comprises

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

- a set of rewards

$$\mathbf{R} = \{r(s,a) \in \mathbb{R} | a \in A(s), s \in S\}$$

- and a set of stationary Markovian transition probabilities

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

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

The process unfolds in discrete time steps indexed by natural numbers. A realization of the process begins at time step $t = 0$ with a specified starting state s_0. At time step t, a decision maker chooses an action $a \in A(s_t)$. The next state s_{t+1} is selected with probability $p(s_{t+1} \mid s_t, a)$, and t is incremented. We assume process continues indefinitely, and that the decision maker's choice of actions at time step t may be functions of the entire history of the game prior to time t. (That is, the decision maker remembers all her previous decisions; this is equivalent to the property of perfect recall for an extensive game.) We interpret the reward $r(s,a)$ as the immediate payoff to the decision maker when she chooses action a in state s.

Example 1.2.1. The graph shown in Figure 1 defines an MDP in which nodes correspond to states, and the decision maker's action set in each state consists of the directed arcs which originate at that state. We may identify the actions by the state to which they lead; for example, $A(0) = \{0,1\}$. Rewards are indicated next to each arc; for example $r(1,0)$ is the immediate reward resulting from action 0 in state 1, which is -1.

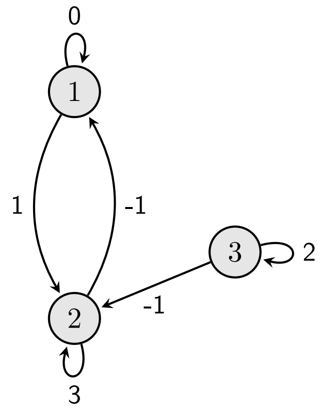

Figure 1. An MDP defined by a graph.

1.3 Strategies

A "strategy" for an MDP captures the notion of a rule which tells the decision maker what to do for every possible situation in the game. We now give a formal definition of the two classes of strategies we will study in detail: Markov strategies and pure stationary strategies.

Definition 1.3.1. A **decision rule** f for the decision maker in an MDP is an element of $\times_{s=1}^{N} A(s)$, where $f(s)$ represents the action chosen in state s.

Definition 1.3.2. A **Markov strategy** is a sequence $\{f_t\}_{t=0}^{\infty}$ of decision rules, where $f_t(s_t)$ is the action chosen at time step t when the current state is s_t.

Definition 1.3.3. A **pure stationary strategy** is a Markov strategy for which $f_t = f$ for some fixed decision rule f for every time step t.

The key characteristic of a Markov strategy is that it does not depend on the history of the game except through the current state and current time step. A pure stationary strategy depends on the history of the game only through the current state.

We will abuse notation by using f to represent both a decision rule and the pure stationary strategy $\{f, f, \dots\}$ composed of that fixed decision rule when there is little risk of confusion. We will denote by F the set of pure stationary strategies.

Given a decision rule f, let $P(f, g)$ be the stationary transition matrix whose (s, s') entry is $p\big(s' \mid s, f(s)\big)$. We also define a column N-vector $\boldsymbol{r}(f)$ whose s-th entry is $r\big(s, f(s)\big)$. Given a Markov strategy $\pi = \big(\{f_t\}\big)$ we write

$$P^{(t)}(\pi) = \begin{cases} I & \text{if } t = 0 \\ P(f_0)P(f_1)\cdots P(f_{t-1}) & \text{if } t > 0 \end{cases},$$

for the t-step transition matrix, where I is the $n \times n$ identity matrix, so that the (s, s')th entry of $P^{(t)}(\pi)$ gives the probability that $s_t = s'$ when $s_0 = s$ and the decision maker uses a Markov strategy π. Note that for a pure stationary strategy f, we have $P^{(t)}(f) = P^t(f)$, the tth power of the matrix $P(f)$.

1.4 Payoffs and optimality

The outcome of a Markov decision process can be characterized by the stochastic process $\{R_t\}_{t=0}^{\infty}$, defined by the decision maker's strategy π and the starting state s, where R_t is interpreted as the random reward received at time step t when the MDP starts in state s. In order to evaluate a choice of strategy, we wish to associate a single scalar payoff to this process. We discuss two ways of doing this: discounted and average payoffs.

1.4.1 Discounted payoffs

The discounted payoff with a discount factor $\beta \in [0, 1)$ for an MDP models a situation in which the decision maker is more concerned with the relatively short term, with β inversely proportional to myopia. In the following, let $\mathrm{E}_{s,\pi}(R_t)$ be the expected value of R_t when the process starts in state s and the decision maker uses strategy π.

Definition 1.4.1. For a given discount factor $\beta \in [0, 1)$, a starting state s, and Markov strategy $\pi = \{f_t\}$, the ***β*-discounted payoff** associated with this strategy is

$$v_\beta(s, \pi) = \sum_{t=0}^{\infty} \beta^t \, \mathrm{E}_{s,\pi}(R_t).$$

Let $\boldsymbol{v}_\beta(\pi)$ be the discounted reward vector whose s-th entry is $v_\beta(s, \pi)$, and observe that

$$\boldsymbol{v}_\beta(\pi) = \sum_{t=0}^{\infty} \beta^t P^{(t)}(\pi) \boldsymbol{r}(f_t).$$

for any Markov strategy π. Further, the powers of a stochastic matrix are themselves stochastic, so the entries of $P^t(f)$ are bounded between 0 and 1 for all t for any decision rule f. Therefore, for any decision rule f, we can write

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

Definition 1.4.2. A Markov strategy π^* is ***β*-optimal** for an MDP with discount factor $\beta \in [0, 1)$ if

$$v_\beta(s, \pi^*) \geq v_\beta(s, \pi)$$

for all starting states s and all Markov strategies π. We will also write this as $\boldsymbol{v}_\beta(\pi^*) \geq \boldsymbol{v}_\beta(\pi)$.

For a Markov decision process $\mathbf{M}$ with a discount factor $\beta \in [0, 1)$, the discounted value of $\mathbf{M}$ is $\boldsymbol{v}_\beta(\mathbf{M}) = \boldsymbol{v}_\beta(\pi^*)$, where π^* is a β-optimal Markov strategy for $\mathbf{M}$.

1.4.2 Average payoffs

The limiting average payoff for a strategy models a situation in which the decision maker is concerned only with long term average payoffs.

Definition 1.4.3. For a given Markov strategy π and starting state s, the limiting average or Cesaro average payoff (also called simply the average payoff or undiscounted payoff) is

$$v(s, \pi) = \liminf_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} E_{s,\pi}(R_t).$$

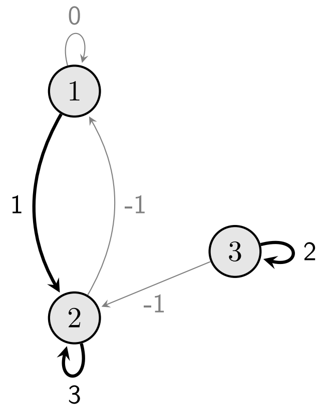

Figure 2. A “greedy” strategy for the MDP of Figure 1.

Equivalently, $\boldsymbol{v}(\pi)$ for any Markov strategy is given by

$$\boldsymbol{v}(\pi) = \liminf_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} P^{(t)}(\pi)\boldsymbol{r}(f_t).$$

Definition 1.4.4. A strategy π^* is **average optimal** for an MDP if

$$v(s,\pi^*) \geq v(s,\pi)$$

for any starting state s and any Markov strategy π.

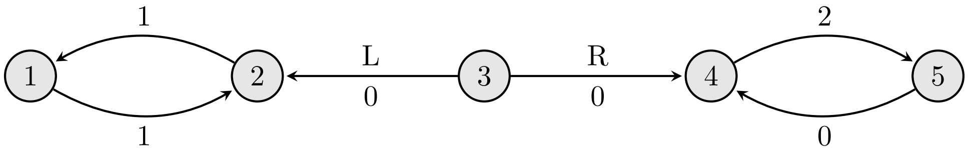

Figure 3. An MDP with two pure stationary strategies.

Example 1.4.5. A choice of pure stationary strategy is illustrated in Figure 2: the action chosen in each state is indicated by the heavy edges. With this greedy strategy as f, the discounted payoff vector is

$$v_\beta(f) = \big[1 + \frac{3\beta}{1-\beta}, \frac{3}{1-\beta}, \frac{2}{1-\beta}\big]^T, \quad 0 \le \beta < 1$$

and the average payoff vector is

$$v(f) = [3, 3, 2]^T.$$

1.4.3 Uniform optimal strategies

Another kind of optimality which will interest us is uniform optimality, which models situations in which the decision maker seeks strategies which will do well for all discount factors sufficiently close to 1.

Definition 1.4.6. A Markov strategy π is **uniform optimal** for an MDP if it is β-optimal for all β sufficiently close to 1.

Example 1.4.7. The MDP given in Figure 3 has a choice for the decision maker only in state 3, and has only two pure stationary strategies, depending on whether the decision maker chooses the action which goes left from 3 to 2 or right from 3 to 4. Call these, respectively, L and R. Although both strategies have the same average payoff of 1 beginning in any state, R is the unique uniform optimal strategy when starting in state 3:

$$v_\beta(3, L) = \frac{\beta}{1-\beta} < \frac{2\beta}{1-\beta^2} = v_\beta(3, R)$$

for all $0 \leq \beta < 1$.

1.4.4 1-optimal strategies

Blackwell defines the notion of a 1-optimal strategy (Blackwell, 1962) (referred to by him as "nearly optimal;" the term "1-optimal" comes from (Veinott, 1966)).

Definition 1.4.8. A strategy $\hat{\pi}$ for a Markov decision process $\mathbf{M}$ is **1-optimal** if

$$\lim_{\beta \nearrow 1} \big[\boldsymbol{v}_\beta(\hat{\pi}) - \boldsymbol{v}_\beta(\mathbf{M})\big] = \mathbf{0}.$$

CHAPTER 2

POLICY IMPROVEMENT FOR MARKOV DECISION PROCESSES

In this chapter we describe a policy improvement algorithm for solving MDPs, and give some known results underlying this class of algorithms which will be useful as we extend it to cover certain classes of stochastic games. Policy improvement for MDPs was introduced in (Howard, 1960) with further treatment by Blackwell (1962), and relies on two ingredients: given an existing pure stationary strategy f and a payoff criterion, there must be an easy computational test to determine whether f' which changes the action choice in at least one state will improve the payoff; and there must be a computationally simple test for optimality. In fact, this test will be essentially the same as the test for an improvement: if we find no one-state action change which improves f, we will be able to conclude that f is in fact optimal. The "policy" in policy improvement derives from the literature on MDPs, where the term is used to describe what we call strategies.

2.1 Discounted payoff MDPs

In order to be self-contained, we restate in this section the policy improvement algorithm for discounted MDPs and its proof by Blackwell. For the remainder of this chapter, we fix a Markov decision process $\mathbf{M}$.

Given a pure stationary strategy strategy f, a fixed discount factor β, and a starting state s, let $G_\beta(s, f)$ be the (possibly empty) set of actions $a \in A(s)$ which satisfy

$$r(s,a) + \beta \sum_{s'=1}^{N} p\big(s' \mid s,a\big) v_\beta(s', f) > v_\beta(s, f), \tag{2.1}$$

and let $G_\beta(f)$ be the (possibly empty) set of all pure strategies $f' \neq f$ such that, for each state s, $f'(s) \in G_\beta(s, f)$ or $f'(s) = f(s)$ for all $s \in S$. The following theorem of Blackwell shows that we may think of $G_\beta(f)$ as a set of strategies which improve on f by having a stricly larger discounted payoff in some state. Importantly, the computation to identify members of $G_\beta(f)$ does not require actually computing the discounted payoff of the candidates.

In the following proof, and in the rest of the thesis, when comparing vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, we use the notation $\boldsymbol{x} > \boldsymbol{y}$ to mean that $x(s) \geq y(s)$ for all coordinates s, and $\boldsymbol{x} \neq \boldsymbol{y}$.

Theorem 2.1.1 (Blackwell). *If $f' \in G_\beta(f)$ then $\boldsymbol{v}_\beta(f) < \boldsymbol{v}_\beta(f')$. If $G_\beta(f)$ is empty, then $\boldsymbol{v}_\beta(f) \geq \boldsymbol{v}_\beta(\sigma)$ for any Markov strategy σ.*

Proof. Note that the self-map of $\mathbb{R}^N$ given by $L_f(\boldsymbol{u}) = \boldsymbol{r}(f) + \beta P(f)\boldsymbol{u}$ is order-preserving for any decision rule f; that is if $\boldsymbol{u} \leq \boldsymbol{v}$ then $L_f(\boldsymbol{u}) \leq L_f(\boldsymbol{v})$.

If $f' \in G_\beta(f)$, then

$$\boldsymbol{v}_\beta(f) < \boldsymbol{r}(f') + \beta P(f')\boldsymbol{v}_\beta(f)$$

and a simple inductive argument shows that $\boldsymbol{v}_\beta(f) < \sum_{t=0}^{T-1} \beta^t P^t(f')\boldsymbol{r}(f') + \beta^T P^T(f')\boldsymbol{v}_\beta(f)$ for any finite $T > 0$. Taking the limit as T goes to infinity yields $\boldsymbol{v}_\beta(f) < \boldsymbol{v}_\beta(f')$.

Now suppose that $G_\beta(f)$ is empty, and take any Markov strategy $\sigma = \{f_0, f_1, \dots\}$. For any finite T we have

$$\begin{aligned}
\boldsymbol{v}_\beta(f) &\geq \boldsymbol{r}(f_T) + \beta P(f_T)\boldsymbol{v}_\beta(f) \\
&\geq \boldsymbol{r}(f_{T-1}) + \beta P(f_{T-1})\boldsymbol{r}(f_T) + \beta^2 P(f_{T-1})P(f_T)\boldsymbol{v}_\beta(f) \\
&\vdots \\
&\geq \sum_{t=0}^{T-1} P^{(t)}(\sigma)\boldsymbol{r}(f_{t+1}) + \beta^T P^{(T)}\boldsymbol{v}_\beta(f).
\end{aligned}$$

Taking the limit as T goes to infinity yields $\boldsymbol{v}_\beta(f) \geq \boldsymbol{v}_\beta(\sigma)$. □

This theorem gives Algorithm 1 for computing a discount-optimal strategy for an MDP.

Algorithm 1 Policy improvement for discounted MDPs

```
Choose an arbitrary $f^0 \in F$ and let $k = 0$
while $G_\beta(f^k)$ is nonempty do
    Choose $f^{k+1} \in G_\beta(f^k)$
    Increment $k$
end while
When $G_\beta(f^k)$ is empty, return $f^k$.
```

Theorem 2.1.2 (Blackwell, 1962). *Algorithm 1 will halt and return a β-optimal strategy for any MDP.*

Proof. For any MDP, if the algorithm halts in finite steps, it returns an f such that $G_\beta(f)$ is empty, so by Theorem 2.1.1, f is β-optimal. There are only finitely many pure stationary strategies, so in order to show that the algorithm halts for any MDP, it suffices to prove that the algorithm cannot cycle. But by Theorem 2.1.1, $\boldsymbol{v}_\beta(f^{k+1}) > \boldsymbol{v}_\beta(f^k)$ for every k, so there can be no cycling. □

Corollary 2.1.3 (Blackwell, 1962)**.** *For any discount factor $\beta \in [0, 1)$, any Markov decision process has a β-optimal pure stationary strategy.*

2.1.1 Sufficiency of pure stationary strategies

We have only given definitions for a limited set of possible strategies; namely, pure Markov strategies and the pure stationary subset of Markov strategies. In fact these are all that are needed to play optimally in the MDP model we have described. Here we outline some results supporting this claim.

Aumann's extension of Kuhn's theorem to infinite extensive games (Aumann, 1964) implies that, since our MDP model can be viewed as a 1-player game with perfect recall, for any mixed strategy, there is a behavioral strategy which induces the same distribution over outcomes, so that the decision maker can achieve the same results with either kind of strategy. Derman and Strauch (Derman and Strauch, 1966) further show that players may restrict further to the class of Markov strategies.

Theorem 2.1.2 constitutes a constructive proof that every MDP has a pure stationary strategy which is discount optimal among all Markov strategies. Blackwell (Blackwell, 1962, Theorem 5) states that any MDP has a pure stationary uniform optimal strategy, and such a strategy is also average optimal. Therefore, in the next section, where we will discuss an extension of the policy improvement algorithm for computing average optimal strategies, we will only deal with pure stationary strategies.

2.2 Average payoff MDPs

2.2.1 Properties of average payoffs for pure stationary strategies

In this subsection, we collect a few results regarding pure stationary strategies which will be useful later. We begin with a fundamental result from Markov chain theory.

Theorem 2.2.1. *For any $n \times n$ stochastic matrix P,*

(a) the Cesaro limit $P^* = \lim_{T\to\infty} \frac{1}{T+1}\sum_{t=0}^{T} P^t$ exists; and

(b) the matrix $I - P + P^*$ is nonsingular, and the deviation matrix

$$\begin{aligned} D &= \left[I - P + P^*\right]^{-1} - P^* \\ &= \lim_{\beta\nearrow 1}\sum_{t=0}^{\infty}\beta^t(P^t - P^*) \end{aligned}$$

exists.

Furthermore, the following identities hold for P, P^*, and D:

- $P^*P = PP^* = P^*P^* = P^*$,
- $P^*D = DP^* = \mathbf{0}$, *and*
- $(I - P)D = D(I - P) = I - P^*$.

Proof. See (Blackwell, 1962) and (Kemeny and Snell, 1960). □

We will make extensive use of the following theorem (Blackwell, 1962, Theorem 4, part (a)) relating β-discounted and average payoffs.

Theorem 2.2.2. *For a pure stationary strategy $f \in F$ for an MDP,*

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

where $\boldsymbol{y}(f) = D(f)\boldsymbol{r}(f)$ and $\boldsymbol{\epsilon}(f,\beta)$ is a function which goes to $\mathbf{0} \in \mathbb{R}^N$ as $\beta \to 1$.

We will follow the notation used in this theorem consistently, using $\boldsymbol{\epsilon}$ to denote the $\mathbb{R}^N$-valued function which depends on β and f, and which goes to zero in all coordinates as β increases to 1. Also, we will refer to the vector $\boldsymbol{y}(f)$ as the **deviation** associated with the strategy f.

Let us say a few words about the crucial significance of this theorem. For two pure stationary strategies f and g, say that f uniformly dominates g for a starting state s if $v_\beta(s,f) > v_\beta(s,g)$ for all discount factors β sufficiently close to 1. Notice that Theorem 2.2.2 implies that f uniformly dominates g in starting state s if the value and deviation of f lexicographically dominate the value and deviation of g in state s. By lexicographic domination, we mean $v(s,f) \geq v(s,g)$ and, if these average payoffs are equal, then $y(s,f) > y(s,g)$.

Corollary 2.2.3. *If a strategy $f \in F$ is uniform optimal, then it is average optimal.*

Proof. Suppose that $f \in F$ is not average optimal. Then there is some f' with $\boldsymbol{v}(f') > \boldsymbol{v}(f)$. Then by Theorem 2.2.2, for β sufficiently close to 1 we must have $\boldsymbol{v}_\beta(f') > \boldsymbol{v}_\beta(f)$, and so f is not uniform optimal. □

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 2.2.4. *Given a pure stationary strategy $f \in F$,*

(a) the average payoff $\boldsymbol{v}(f)$ may be computed as

$$\boldsymbol{v}(f) = P^*(f)\boldsymbol{r}(f),$$

and is the unique solution to

$$\big(I - P(f)\big)v = \boldsymbol{0} \text{ and } P^*(f)v = v.$$

(b) The vector $\boldsymbol{y}(f) = D(f)\boldsymbol{r}(f)$ is the unique solution to

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

2.2.2 Policy improvement for average payoff MDPs

Next we define sets which are the average-payoff analogues of $G_\beta(s,f)$ and $G_\beta(f)$ for a given strategy f for a Markov decision process $\mathbf{M}$. Let $G(s,f)$ be the (possibly empty) set of actions $a \in A(s)$ which satisfy

$$\sum_{s'=1}^{N} p\big(s' \mid s,a\big)v(s',f) \geq v(s,f) \tag{2.2}$$

and in case (2.2) holds with equality,

$$r\big(s,a\big) + \sum_{s'=1}^{N} p\big(s' \mid s,a\big)y(s',f) > v(s,f) + y(s,f). \tag{2.3}$$

Let $G(f)$ be the (possibly empty) set of all pure strategies $f' \neq f$ for player 1 such that for each state s, $f'(s) \in G(s,f)$ or $f'(s) = f(s)$. The following theorem from Blackwell is the basis for a policy improvement algorithm for MDPs with average payoffs.

Theorem 2.2.5 (Blackwell, 1962, Theorem 4, parts (b) and (e)). *For a strategy f for a Markov decision process,*

(a) if $f' \in G(f)$ then $\boldsymbol{v}_\beta(f') > \boldsymbol{v}_\beta(f)$ for all $\beta \in [0,1)$ sufficiently close to 1; and

(b) if $G(f)$ is empty, then f is average optimal for the Markov decision process.

Theorem 2.2.2 is the basis of another, perhaps more satisfying, interpretation of the process of improvement.

Corollary 2.2.6 (Veinott, 1966, Corollary 1). *If $f' \in G(f)$ then $\boldsymbol{v}(f') \geq \boldsymbol{v}(f)$ and if $\boldsymbol{v}(f') = \boldsymbol{v}(f)$, then $\boldsymbol{y}(f') > \boldsymbol{y}(f)$.*

We now have the basis for Algorithm 1-B for computing an average optimal pure stationary strategy for an MDP: we simply replace the set $G_\beta(f^k)$ with $G(f^k)$ in line 3 of Algorithm 1. Now every time the algorithm chooses $f^{k+1} \in G(f^k)$, we have a strict increase in the associated discounted value for a sufficiently large discount factor (or, equivalently by 2.2.6, a lexicographic increase in the average payoff and deviation). Hence, as before, the algorithm cannot cycle, and this suffices to prove that it must halt and return an average optimal strategy.

Theorem 2.2.7. *If the set $G_\beta(f^k)$ is replaced with the set $G(f^k)$ in Algorithm 1, the resulting Algorithm 1-B will halt in finite steps and return an average optimal strategy.*

2.2.3 Computing 1-optimal strategies for MDPs

Veinott shows that a further extension to Algorithm 1 can be used to compute a 1-optimal strategy for an MDP (Veinott, 1966). Given a strategy f for a Markov decision process $\mathbf{M}$, let $H(s, f)$ be the (possibly empty) set of actions $a \in A^1(s)$ for which (2.2) and (2.3) are both equalities and furthermore

$$\sum_{s'=1}^{N} p(s' \mid s, a) z(s', f) > y(s, f) + z(s, f) \tag{2.4}$$

where $\boldsymbol{z}(f) = D(f)\big(-\boldsymbol{y}(f)\big)$ is the unique solution to

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

Let $H(f)$ be the set of pure strategies $f' \neq f$ for player 1 with $f'(s) \in H(s, f)$ or $f'(s) = f(s)$ for all s.

Theorem 2.2.8. *For a strategy f for a Markov decision process $\mathbf{M}$,*

(a) if $f' \in H(f)$ then $\boldsymbol{v}(f') = \boldsymbol{v}(f)$, $\boldsymbol{y}(f') \geq \boldsymbol{y}(f)$, and if $\boldsymbol{y}(f') = \boldsymbol{y}(f)$, then $\boldsymbol{z}(f') > \boldsymbol{z}(f)$;

(b) if $G(f) \cup H(f)$ is empty, then f is 1-optimal for $\mathbf{M}$.

Proof. See (Veinott, 1966). □

Theorem 2.2.9. *Algorithm 1-V, in which the set $G_\beta(f^k)$ is replaced with the set $G(f^k) \cup H(f^k)$ in Algorithm 1, will halt in finite steps and return a 1-optimal strategy.*

Proof. By Theorem 2.2.5 and Theorem 2.2.8, the vectors $\boldsymbol{v}(f^{k+1})$, $\boldsymbol{y}(f^{k+1})$, and $\boldsymbol{z}(f^{k+1})$ lexicographically dominate $\boldsymbol{v}(f^k)$, $\boldsymbol{y}(f^k)$, and $\boldsymbol{z}(f^k)$, so the algorithm cannot cycle. This guarantees that it will reach some $\hat{f}$ for which $G(\hat{f}) \cup H(\hat{f})$ is empty, and which by Theorem 2.2.8 is 1-optimal. □

Example 2.2.10. Figure 4 shows a two state MDP with three actions in state 1 and one action in state 2. The top left value in each box represents the immediate payoff for the corresponding action, and the bottom right tuple represents the transition probability. The decision maker has three pure stationary strategies, namely, choosing the top, middle, or bottom action in state 1, which we will identify as T, M, and B, respectively.

In this example, we fix the order T, M, B for the strategies, and show that Algorithm 1-B results in changing from T to M, and that Algorithm 1-V changes from T to M to B. Brute

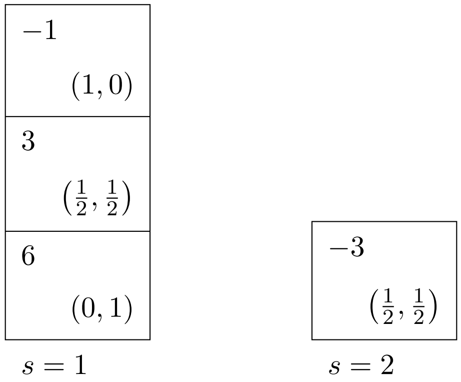

Figure 4. A two-state MDP.

force calculations confirm that both M and B are average optimal strategies, and that B is the unique 1-optimal (and hence unique uniform optimal) for this MDP.

If we begin by choosing strategy T, with reward vector $\boldsymbol{r}(T) = (-1,-3)^T$ and transition and Cesaro limit matrices

$$P(T) = \begin{bmatrix} 1 & 0 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} \text{ and } P^*(T) = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$$

we may compute $\boldsymbol{v}(T) = P^*(T)\boldsymbol{r}(T) = (-1,-1)^T$ and $\boldsymbol{y}(T)$ as the unique solution to the system

$$\begin{aligned} \big(I - P(T)\big)\boldsymbol{y} &= \boldsymbol{r}(T) - \boldsymbol{v}(T) \\ P^*(T)\boldsymbol{y} &= \boldsymbol{0} \end{aligned}$$

which gives $\boldsymbol{y}(T) = (0, -4)^T$.

We find that $M \in G(T)$, because 2.2 holds with equality since $\boldsymbol{v}(T)$ has the same value in all coordinates, and

$$\begin{aligned}\big[r(M) + P(M)\boldsymbol{y}(T)\big]_1 &= 3 + \frac{1}{2}(0) + \frac{1}{2}(-4) \\ &= 1 \\ &> -1 = v(1, T).\end{aligned}$$

We may now compute $\boldsymbol{v}(M) = (0, 0)^T$ and $\boldsymbol{y}(M) = (3, -3)^T$, and so checking strategy B gives

$$P(B)\boldsymbol{v}(M) = \boldsymbol{v}(M)$$

and

$$r(B) + P(B)\boldsymbol{y}(M) = \boldsymbol{v}(M) + \boldsymbol{y}(M)$$

so $G(M)$ is empty, and M is an average optimal strategy.

In order to compute a 1-optimal strategy, we must also check whether $B \in H(M)$. We compute $\boldsymbol{z}(M)$ as the unique solution of $\big(I - P(M)\big)\boldsymbol{z} = -\boldsymbol{y}(M)$ and $P^*(M)\boldsymbol{z} = 0$, which gives $\boldsymbol{z}(M) = (-3, 3)^T$. We find

$$\begin{aligned}\big[P(B)\boldsymbol{z}(M)\big]_1 &= z(2, M) \\ &= 3 \\ &> -3 + 3 = z(1, M) + y(1, M),\end{aligned}$$

and so $B \in H(M)$. Since we have checked every strategy along the way, we conclude that $G(B) \cup H(B)$ is empty, and B is a 1-optimal strategy. Since it is a unique 1-optimal strategy, it must also be uniform optimal for this MDP.

CHAPTER 3

POLICY IMPROVEMENT FOR STOCHASTIC GAMES

This chapter contains the new results in this thesis: a new proof of the correctness of a policy improvement algorithm for discounted stochastic games with additive rewards and additive transitions, and a proof of a further extension of policy improvement to average-payoff games of this class.

3.1 Stochastic Games

3.1.1 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 $S = \{1, 2, \dots, N\}$;
- 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.)

Play of the game proceeds in discrete time steps indexed by natural numbers. A play of the game begins with time step $t = 0$ with a specified starting state s_0. At time step t, players choose actions $a^1 \in A^1(s_t)$ and $a^2 \in A^2(s_t)$. The next state s_{t+1} is selected with Markovian probability $p(s_{t+1} \mid s_t, a^1, a^2)$, t is incremented. We assume game continues indefinitely, and that players' choices of actions at time step t may be functions of the entire history of the game prior to time t (that is, players remember everything that has happened so far). 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.

We will restrict our attention to zero-sum stochastic games with additive reward and additive transitions (ARAT games), defined below.

Definition 3.1.1. 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.

For simplicity, since this paper deals exclusively with zero-sum games, we will use a single payoff function $r(s, a^1, a^2) \equiv r^1(s, a^1, a^2)$ 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.

Definition 3.1.2. 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', with $p^i(s' \mid s, a^i) \geq 0$ for all actions $a^i \in A^i(s)$, for all states $s, s' \in S$, and for both players $i = 1, 2$.

Definition 3.1.3. 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 a perfect information game, we may think of each state as "belonging" to one of the players (the one who alone has a choice of action in that state), or as a "chance" state in which neither player has a choice of action.

For simplicity, when we speak of a "game," we will mean a zero-sum stochastic game with the ARAT property, unless otherwise specified. Note that the class of ARAT stochastic games includes the perfect information stochastic games.

Example 3.1.4. A directed graph with weighted arcs such as in Figure 1 induces an ARAT game in which the states of the game are identified with nodes of the graph, and both players have the same action set at each node, which consists of the directed arcs orginating at that node. We will identify an action in such a game by the node to which the action leads. For example, a directed arc $s \to t$ in a graph will be referred to as action t available in state s in the induced ARAT game.

It remains to define the rewards and transitions in such a game. Suppose a graph contains arcs $s \to t$ and $s \to t'$, with weights w_{st} and $w_{st'}$, giving actions $t, t' \in A^i(s)$ for $i = 1, 2$. When the player 1 chooses action t and player 2 action t' in state s, the immediate reward is the

	1	2
1	0 $(1,0,0)$	$\frac{1}{2}$ $(\frac{1}{2},\frac{1}{2},0)$
2	$\frac{1}{2}$ $(\frac{1}{2},\frac{1}{2},0)$	1 $(0,1,0)$

$s = 1$

	1	2
1	-1 $(1,0,0)$	1 $(\frac{1}{2},\frac{1}{2},0)$
2	1 $(\frac{1}{2},\frac{1}{2},0)$	3 $(0,1,0)$

$s = 2$

	2	3
2	-1 $(0,1,0)$	$\frac{1}{2}$ $(0,\frac{1}{2},\frac{1}{2})$
3	$\frac{1}{2}$ $(0,\frac{1}{2},\frac{1}{2})$	2 $(0,0,1)$

$s = 3$

Figure 5. An ARAT game induced by the graph in Figure 1.

convex sum $p_s w_{st} + (1 - p_s) w_{st'}$ for some fixed weight $p_s \in [0, 1]$ for state s. The transition probability to state t is p_s, to state t' is $1 - p_s$, and the transition probability to any other state is 0.

We may interpret such a game as one in which the players secretly choose their actions, and a referee implements player 1's action with probability p_s, and player 2's with probability $1 - p_s$, where p_s is fixed for each state s. The game induced by a graph in this way is an ARAT game, but not all ARAT games are of this type.

For a concrete example, we fix $p_s = \frac{1}{2}$ for all $s \in 1, 2, 3$ for the graph in Figure 1. We may then represent the game as in Figure 5, where each state is represented in a matrix game format. For example, in state 1, each player's actions are identified with the choice of staying in state 1 or moving to state 2. Actions for player 1 are represented by rows in the state 1 matrix game, and actions for player 2 are represented by columns. When both players choose action 1 (staying in state 1),the immediate reward is 0 and the game remains in state 1 with probability 1. When player 1 chooses action 1 and player 2 chooses action 2, the expected immediate reward is $\frac{1}{2}$ and the game remains in state 1 and transitions to state 2 with equal probability.

Alternatively, we may generate a perfect information game from the graph in Figure 1 by assigning each node to a player, meaning that only the player to whom the node is assigned may choose an action in that state. For example, nodes 1 and 2 may be for player 1, and node 3 for player 2. This is illustrated in matrix game notation in Figure 6.

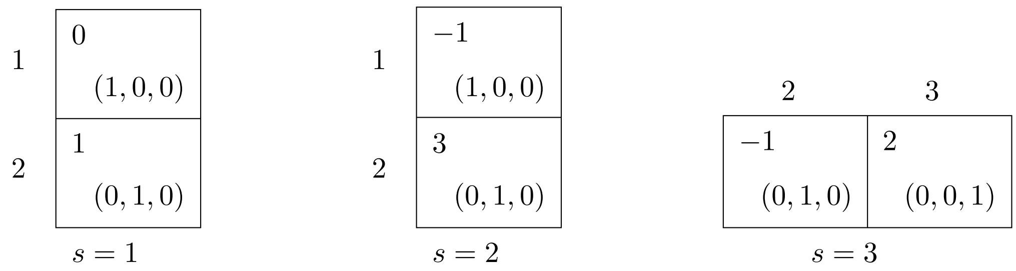

Figure 6. A perfect information game induced by the graph in Figure 1.

3.1.2 Strategy pairs for stochastic games

For stochastic games, we will deal with strategy pairs. We have seen that for MDPs, pure stationary strategies suffice. This is not the case in general for stochastic games. Since our stochastic game model preserves the perfect recall condition, Aumann has shown (Aumann, 1964) that players may limit to behavioral strategies for any stochastic game. However, in the average case, optimal play may require randomized strategies which depend on the entire history of the game (see, for example the Big Match (Blackwell and Ferguson, 1968)). This is one reason for limiting the class of games we consider: in the special case of ARAT games players can play optimally using pure stationary strategies (Raghavan et al., 1985). The sufficiency of pure stationary strategies follows in the average payoff case by an argument we will discuss in Section 3.1.4.

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. The remainder of this subsection details natural extensions of our notation for MDPs to stochastic games.

Given a pair of decision rules (f, g) with f for player 1 and g for player 2 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 $\boldsymbol{r}(f, g)$ whose s-th entry is $r\big(s, f(s), g(s)\big)$. Given a pair of Markov strategies $(\pi, \rho) = \big(\{f_t\}, \{g_t\}\big)$ we write

$$P^{(t)}(\pi, \rho) = \begin{cases} I_n & \text{if } t = 0 \\ P(f_0, g_0)P(f_1, g_1) \cdots P(f_{t-1}, g_{t-1}) & \text{if } t > 0 \end{cases},$$

where I_n is the $n \times n$ identity matrix. Note that for a pure stationary strategy pair (f, g), we have $P^{(t)}(f, g) = P^t(f, g)$.

When one player in game fixes his or her strategy, we may think of the induced "one-player" game for the opponent as an 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.

3.1.3 Optimal strategy pairs for stochastic games

Definition 3.1.5. For a given discount factor $\beta \in [0,1)$, and a pure stationary strategy pair (f,g), the **$\boldsymbol{\beta}$-discounted payoff** is

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

and we write $v_\beta(s,f,g)$ for the sth coordinate of $\boldsymbol{v}_\beta(f,g)$.

Definition 3.1.6. A pure stationary strategy pair (f^*,g^*) for a game Γ is **$\boldsymbol{\beta}$-optimal** (over pure stationary strategies) for a discount factor $\beta \in [0,1)$ if there exists a **$\boldsymbol{\beta}$-discounted value** $\boldsymbol{v}_\beta(\Gamma)$ such that

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

and

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

in which case $\boldsymbol{v}_\beta(\Gamma) = \boldsymbol{v}_\beta(f^*,g^*)$.

Definition 3.1.7. 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

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

and we write $v(s, f, g)$ for the sth coordinate of $\boldsymbol{v}(f, g)$.

Definition 3.1.8. A strategy pair (f^*, g^*) for a game Γ is **average optimal** for the game if there exists an **average value** $\boldsymbol{v}(\Gamma)$ such that

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

and

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

in which case $\boldsymbol{v}(\Gamma) = \boldsymbol{v}(f^*, g^*)$.

We also extend the defintion of uniform optimality to stochastic games.

Definition 3.1.9. 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.

It will be useful to think of discount and average optimality as being characterized by mutual optimality for two induced MDPs, as described in the following lemma.

Lemma 3.1.10. *A strategy pair (f, g) is β-optimal for a game Γ with discount factor β if and only if f is β-optimal for $\Gamma|_g$ and g is β-optimal for $\Gamma|_f$. This also holds when "β-optimal" is replaced with "average optimal."*

3.1.4 Sufficiency of pure stationary strategy pairs for ARAT games

A theorem analogous to Blackwell's for the existence of pure stationary discount optimal strategies for MDPs (Blackwell, 1962, Theorem 4) also applies to any stochastic game with pure stationary β-optimal strategy pairs.

Theorem 3.1.11. *Any stochastic game Γ which has a pure stationary β-optimal strategy pair for all discount factors $\beta \in [0,1)$ has a pure stationary uniform optimal strategy pair.*

Proof. There are only finitely many pure stationary strategy pairs, and there is an optimal pure stationary strategy pair for each β. Therefore, we can find some sequence of discount factors with 1 as a limit point and a pure stationary strategy pair (f^*, g^*) β-optimal for each discount factor β in the sequence. Furthermore, for any pure stationary strategy pair (f, g), the representation

$$\boldsymbol{v}_\beta(f,g) = \left[I - \beta P(f,g)\right]^{-1} \boldsymbol{r}(f,g)$$

shows that each coordinate of $\boldsymbol{v}_\beta(f,g)$ is a rational function in β. For any pure stationary strategies f and g for player 1 and 2, we must have

$$\boldsymbol{v}_\beta(f^*,g) \geq \boldsymbol{v}_\beta(f^*,g^*) \geq \boldsymbol{v}_\beta(f,g^*) \tag{3.1}$$

for all β sufficiently close to 1. If not, suppose (without loss of generality) that we have a sequence of βs with limit point 1 such that $\boldsymbol{v}_\beta(f^*,g) < \boldsymbol{v}_\beta(f^*,g^*)$. Since we also have $\boldsymbol{v}_\beta(f^*,g^*) \geq \boldsymbol{v}_\beta(f^*,g)$ for a sequence of discount factors with limit point 1, it must be that for some coordinate s, the rational functions $\boldsymbol{v}_\beta(s,f^*,g)$ and $\boldsymbol{v}_\beta(s,f^*,g^*)$ intersect infinitely often, which is impossible. Therefore, (3.1) holds for any f and g for the two players, so the pure stationary pair (f^*, g^*) is uniform optimal. □

Shapley shows that perfect information games meet the hypothesis of this theorem (Shapley, 1953), and the same has been shown for ARAT games (Raghavan et al., 1985). Therefore we may conclude that these classes of games possess stationary uniform optimal strategy pairs and so by Corollary 2.2.3 and Lemma 3.1.10, they also have pure stationary average optimal strategy pairs. This justifies limiting ourselves to pure stationary strategy pairs for ARAT games as well as MDPs.

3.2 Policy improvement for discounted payoff ARAT games

In this section, we prove a result conjectured by Raghavan and Syed (Raghavan and Syed, 2003) which allows for a somewhat simpler proof of the policy improvement algorithm for discounted stochastic games of perfect information than the inductive proof used in (Raghavan and Syed, 2003) or for ARAT games in (Syed, 1999).

First, we will define some notation. Let $G_\beta(s, f \mid g)$ for a game Γ be equal to $G_\beta(s, f)$ for the Markov decision process $\Gamma|_g$ as defined in 2.1. For clarity, we will write out the criterion explicitly: an action $a \in A^1(s)$ is an element of $G_\beta(s, f \mid g)$ if

$$r\big(s, a, g(s)\big) + \beta \sum_{s'}^{N} p\big(s' \mid s, a, g(s)\big) v_\beta\big(s', a, g(s)\big) > v_\beta\big(s, a, g(s)\big). \tag{3.2}$$

We also define $G_\beta(f \mid g)$ as the set of all strategies $f' \in F$ such that whenever $f'(s) \neq f(s)$, $f'(s) \in G_\beta(s, f \mid g)$, and $f \neq f'$. The sets $G_\beta(s, g \mid f)$ and $G_\beta(f \mid f)$ for player 2 are defined similarly, with the direction of the inequality reversed. These sets defined for each pure stationary strategy pair (f, g) in the game are the basis for Algorithm 2.

Algorithm 2 Policy improvement for discounted ARAT games

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

One can view Algorithm 2 as solving a sequence of discounted MDPs for player 1. In their paper, Raghavan and Syed observed that, in computational examples, the discounted value of these MDPs decreases monotonically when the payoff vectors are compared coordinatewise, and conjectured that this holds in general. In the rest of this subsection we prove that this conjecture is correct, and use it to provide an alternate proof for this Algorithm 2. We call this result the Patience Theorem, because it provides an intuitive justification of the “patient” approach the algorithm takes to improving strategies for player 2.

Theorem 3.2.1 (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 $\boldsymbol{v}_\beta(\Gamma|_{g^0}) > \boldsymbol{v}_\beta(\Gamma|_{g^1})$.*

Proof. Consider the Markov decision process $\Gamma|_g$ for some $g \in G$. The solution of such an MDP by a linear program is well known (see, for example, 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 Markov decision process $\Gamma|_{g^0}$, and so $\boldsymbol{v}_\beta(f^0, g^0)$ is the optimal solution to the LP corresponding to $\Gamma|_{g^0}$. In particular, it is feasible for this LP. We will show that $\boldsymbol{v}_\beta(f^0, g^0) - \boldsymbol{\epsilon}_X$ is also feasible for the LP corresponding to $\Gamma|_{g^1}$, where $\boldsymbol{\epsilon}_X$ 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,

$$\sum_{s=1}^{N} \gamma(s)\big(v_\beta(s, f^0, g^0) - \boldsymbol{\epsilon}_X(s)\big) \geq \sum_{s=1}^{N} \gamma(s) v_\beta(s, \Gamma|_{g^1}). \tag{3.3}$$

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 (3.3) and taking a limit as $n \to \infty$ on both sides, yields for all states s, $v_\beta(s, f^0, g^0) - \boldsymbol{\epsilon}_X(s) \geq v_\beta(s, \Gamma|_{g^1})$. Since $\boldsymbol{\epsilon}_X(s)$ is positive for some s and $\boldsymbol{v}_\beta(\Gamma|_{g^0}) = \boldsymbol{v}_\beta(f^0, g^0)$, this gives the theorem. We now proceed to prove that $\boldsymbol{v}_\beta(f^0, g^0) - \boldsymbol{\epsilon}_X$ is feasible for the LP corresponding to $\Gamma|_{g^1}$. We will define $\boldsymbol{\epsilon}_X$ appropriately along the way.

Let $v^0 = v_\beta(f^0, g^0)$. Let X be the set of states s for which $g^1(s) \in G_\beta(s, g^0 \mid f^0)$.

When $g^1(s) = g^0(s)$ (that is, for $s \in X^C$),

$$v^0(s) \geq r\big(s, a, g^1(s)\big) + \beta \sum_{s'=1}^{N} p\big(s' \mid s, a, g^1(s)\big) v^0(s') \tag{3.4}$$

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

For any $s \in X$,

$$v^0(s) > r\big(s, f^0(s), g^1(s)\big) + \beta \sum_{s'=1}^{N} p\big(s' \mid s, f^0(s), g^1(s)\big) v^0(s') \tag{3.5a}$$

or, by the ARAT property,

$$v^0(s) > \big[r^1\big(s, f^0(s)\big) + r^2\big(s, g^1(s)\big)\big] + \beta \sum_{s'=1}^{N} \big[p^1\big(s' \mid s, f^0(s)\big) + p^2\big(s' \mid s, g^1(s)\big)\big] v^0(s') \tag{3.5b}$$

We want to show that this inequality holds when we replace $f^0(s)$ by any action $a \in A^1(s)$ in (3.5a) or, equivalently, in (3.5b).

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

$$v^0(s) \geq \big[r^1\big(s,a\big) + r^2\big(s, g^0(s)\big)\big] + \beta \sum_{s'=1}^{N} \big[p^1(s' \mid s,a) + p^2\big(s' \mid s, g^0(s)\big)\big] v^0(s') \tag{3.6a}$$

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

$$0 \geq \big[r^1\big(s,a\big) - r^1\big(s, f^0(s)\big)\big] + \beta \sum_{s'=1}^{N} \big[p^1(s' \mid s,a) - p^1\big(s' \mid s, f^0(s)\big)\big] v^0(s') \tag{3.6b}$$

Now summing (3.5b) and (3.6b) we have

$$v^0(s) > \big[r^1\big(s,a\big) + r^2\big(s, g^1(s)\big)\big] + \beta \sum_{s'=1}^{N} \big[p^1(s' \mid s,a) + p^2\big(s' \mid s, g^1(s)\big)\big] v^0(s') \tag{3.7}$$

for $s \in X$, as desired.

Now by the strict inequality in (3.7), 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 $\epsilon_X(s) = 0$ otherwise. This argument, along with (3.4) and (3.7) yield the feasibility of $\boldsymbol{v}_\beta(f^0, g^0) - \boldsymbol{\epsilon}_X$ for the LP corresponding to $\Gamma|_{g^1}$, and the proof is complete. □

Theorem 3.2.2. *Algorithm 2 will terminate in finite steps, and the returned strategy pair is β-optimal for the game.*

Proof. If Algorithm 2 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 Theorem 2.1.1 and Lemma 3.1.10. 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 through the while loop starting in line 2, 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 3.2.1 to conclude that $\boldsymbol{v}_\beta(\Gamma|_{g^k}) > \boldsymbol{v}_\beta(\Gamma|_{g^{k+1}})$. 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. □

3.3 Average payoff ARAT games

We are now prepared to present our main result: Algorithm 3 is a policy improvement algorithm for computing optimal strategies for games with average payoffs. We first define sets for stochastic games corresponding to $G(f)$ and $H(f)$ for a strategy f for an MDP, as defined in Section 2.2.2 and Section 2.2.3. The new sets $G(f \mid g)$ and $H(f \mid g)$ extend these sets to a strategy pair (f, g) for a stochastic game. For a strategy pair (f, g) for a stochastic game Γ, we define $G(f \mid g)$ to be equal to the set $G(f)$ for f in the MDP $\Gamma|_g$, and define $H(f \mid g)$ to be equal to the set $H(f)$ for the MDP $\Gamma|_g$.

For clarity and for reference in the rest of the thesis, we rewrite the conditions using our stochastic game notation. A strategy f' is an element of $G(f \mid g)$ if for any state s for which $f(s) \neq f'(s)$, we have

$$\sum_{s'=1}^{N} p\big(s' \mid s, f'(s), g(s)\big) v(s', f, g) \geq v(s, f, g), \tag{3.8}$$

and whenever 3.8 holds with equality for such $f'(s)$,

$$r\big(s, f'(s), g(s)\big) + \sum_{s'=1}^{N} p\big(s' \mid s, f'(s), g(s)\big) y(s', f, g) > v(s, f, g) + y(s, f, g). \tag{3.9}$$

A strategy f' is an element of $H(f \mid g)$ when 3.8 and 3.9 both hold with the inequalities relaced with equalities for all states s and for any states for which $f'(s) \neq f(s)$ then

$$\sum_{s'=1}^{N} p\big(s' \mid s, f'(s), g(s)\big) z(s', f, g) > y(s, f, g) + z(s, f, g). \tag{3.10}$$

We define $G(g \mid f)$ and $H(g \mid f)$ similarly, with the inequalities in (3.8), (3.9), and (3.10) reversed. With these definitions, we are prepared to state our policy improvement algorithm for average payoff stochastic games.

For any game Γ and starting strategy pair (f^0, g^0) for which the algorithm terminates, the resulting pair (f^*, g^*) is such that the sets $G(f^* \mid g^*)$ and $G(g^* \mid f^*)$ are both empty. Hence, by Theorem 2.2.5 and Lemma 3.1.10, the pair is average optimal. Since there are only finitely many pure stationary strategy pairs, our goal will be to prove that the algorithm cannot cycle.

Algorithm 3 Policy improvement for average payoff ARAT games

```
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 $(f^*, g^*) = (f^k, g^k)$.
```

In the discounted case this can be done by showing that each improvement by player 2 presents player 1 with a new MDP whose optimal value is strictly smaller than the last. The proof for the average case bears some similarity to the proof of the algorithm for average-payoff MDPs. Recall that in that case, Theorem 2.2.5 shows that each new strategy chosen by the algorithm has a strictly larger discounted payoff for all discount factors sufficiently close to 1 - we may say that each new strategy chosen by the algorithm for MDPs is uniformly greater. We will show here that each new MDP solved by player 1 in line 3 of Algorithm 3 has a uniformly smaller discounted value. Much of the work for the proof is done in the following three lemmas. First, we show that any average improvement chosen by player 2 against a fixed strategy for player 1 in fact constitutes a uniform improvement.

Lemma 3.3.1. *Given a strategy pair (f, g), for β sufficiently close to 1, $G(g \mid f) \subset G_\beta(g \mid f)$.*

Proof. Take any $g' \in G(g \mid f)$. And fix any state s with $g'(s) \neq g(s)$. For such a state, we must have

$$\big[P(f,g')\boldsymbol{v}(f,g) - \boldsymbol{v}(f,g)\big]_s \leq 0 \tag{3.11}$$

and if this holds with equality, then

$$\big[\boldsymbol{r}(f,g') + P(f,g')\boldsymbol{y}(f,g) - \boldsymbol{v}(f,g) - \boldsymbol{y}(f,g)\big]_s < 0. \tag{3.12}$$

Now for any β, $g' \in G_\beta(g \mid f)$ precisely when for our state s,

$$\big[\boldsymbol{r}(f,g') + \beta P(f,g')\boldsymbol{v}_\beta(f,g) - \boldsymbol{v}_\beta(f,g)\big]_s < 0. \tag{3.13}$$

But by Theorem 2.2.2 we have

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

where $\boldsymbol{\epsilon}(\beta,f,g)$ goes to zero in all coordinates at β increases to 1. Using this, we may write (3.13) as

$$\begin{aligned}\Big[\boldsymbol{r}(f,g') + \beta P(f,g')\frac{\boldsymbol{v}(f,g)}{1-\beta} + \beta P(f,g')\boldsymbol{y}(f,g) + \beta P(f,g')\boldsymbol{\epsilon}(\beta,f,g,g')& \\ - \frac{\boldsymbol{v}(f,g)}{1-\beta} - \boldsymbol{y}(f,g) - \boldsymbol{\epsilon}(\beta,f,g)\Big]_s &< 0.\end{aligned}$$

Now taking $\frac{\beta}{1-\beta} = \left(\frac{1}{1-\beta} - 1\right)$ and $\beta = 1 - (1-\beta)$, we may gather three groups of terms: a group of terms with a factor of $\frac{1}{1-\beta}$, a group of terms with no factor involving β, and a group of terms which approach zero as β approaches 1. In this way, we rewrite (3.13) as

$$\begin{aligned}\frac{1}{1-\beta}\big[\boldsymbol{r}(f,g') + P(f,g')\boldsymbol{v}(f,g) - \boldsymbol{v}(f,g)\big]_s& \\ + \big[\boldsymbol{r}(f,g') + P(f,g')\boldsymbol{y}(f,g) - P(f,g')\boldsymbol{v}(f,g) - \boldsymbol{y}(f,g')\big]_s& \\ + \big[(\beta-1)P(f,g')\boldsymbol{y}(f,g) + \beta p(f,g')\boldsymbol{\epsilon}(\beta,f,g,g') - \boldsymbol{\epsilon}(\beta,f,g')\big]_s < 0.& \qquad (3.14)\end{aligned}$$

Now if (3.11) holds strictly, then we may choose β near enough to 1 that (3.14) holds. If (3.11) holds with equality, then $\big[P(f,g')\boldsymbol{v}(f,g)\big]_s = v(s,f,g)$ and (3.12) is strict, and so we again choose β near enough to 1 that the negative value of the second bracketed expression in (3.14) guarantees the negative value of the entire expression as the third bracketed expression approaches zero.

This argument holds for any state s with $g'(s) \neq g(s)$, so for all $\beta < 1$ near enough to 1, we have $g' \in G_\beta(g \mid f)$. □

Next, we will show that although the 1-optimal 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 3.3.2. *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 $\Gamma|_g$, $\boldsymbol{v}(f^*,g) = \boldsymbol{v}(f,g)$ and $\boldsymbol{y}(f^*,g) = \boldsymbol{y}(f,g)$.*

Proof. If $G(f \mid g) \cup H(f \mid g)$ is empty then by Theorem 2.2.5, $\boldsymbol{v}(f, g) \geq \boldsymbol{v}(f^*, g)$ and if $\boldsymbol{v}(f^*, g) = \boldsymbol{v}(f, g)$ then $\boldsymbol{y}(f, g) \geq \boldsymbol{y}(f^*, g)$. Now for all β sufficiently close to 1, f^* is β-optimal for $\Gamma|_g$, so $\boldsymbol{v}_\beta(f^*, g) - \boldsymbol{v}_\beta(f, g) \geq \boldsymbol{0}$ or, rewriting using Theorem 2.2.2:

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

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

The last lemma shows that although in fact the choice of strategy for player 2 in line (5) is made against a 1-optimal strategy, the same choice could have been made against a stronger uniform optimal strategy for player 1.

Lemma 3.3.3. *Let f^1 and f^2 be pure stationary strategies for player 1 and g a pure stationary strategy for player 2 for a game Γ with $\boldsymbol{v}(f^1, g) = \boldsymbol{v}(f^2, g)$ and $\boldsymbol{y}(f^1, g) = \boldsymbol{y}(f^2, g)$. Then $G(g \mid f^1) = G(g \mid f^2, g)$.*

Proof. Since the roles of f^1 and f^2 are completely interchangeable, it suffices to show that $G(g \mid f^1) \subseteq G(g \mid f^2)$. Write $\boldsymbol{v}$ for the common value of $\boldsymbol{v}(f^i, g)$, $i = 1, 2$ and $\boldsymbol{y}$ for the common value of $\boldsymbol{y}(f^i, g)$, $i = 1, 2$.

By the definition of the set $G(g \mid f^1)$, for a strategy g' contained in this set and any state s with $g'(s) \neq g(s)$, we must have

$$\sum_{t=1}^{N} p\bigl(t \mid s, f^1(s), g'(s)\bigr) v(t) \leq v(s) \tag{3.15}$$

and, if this holds with equality,

$$r\bigl(s, f^1(s), a\bigr) + \sum_{t=1}^{N} p\bigl(t \mid s, f(s), g'(s)\bigr) y(t) < v(s) + y(s) \tag{3.16}$$

Now observe that by Lemma (2.2.4), we have

$$P(f^i, g)\boldsymbol{v} = \boldsymbol{v} \text{ and } r(f^i, g) + P(f^i, g)\boldsymbol{y} = \boldsymbol{v} + \boldsymbol{y}$$

for $i = 1, 2$, and so

$$P(f^1, g)\boldsymbol{v} = P(f^2, g)\boldsymbol{v} \tag{3.17}$$

and

$$r(f^1, g) + P(f^1, g)\boldsymbol{y} = r(f^2, g) + P(f^2, g)\boldsymbol{y}. \tag{3.18}$$

Now fix an s with $g'(s) \neq g(s)$. The sth line of (3.17) is

$$\sum_{s'=1}^{N} p\bigl(s' \mid s, f^1(s), g(s)\bigr) v(s') = \sum_{s'=1}^{N} p\bigl(s' \mid s, f^2(s), g(s)\bigr) v(s')$$

Using the ARAT property of the game,

$$\begin{aligned}\sum_{s'=1}^{N} p^1\big(s' \mid s, f^1(s)\big)v(s') + \sum_{s'=1}^{N} p^2\big(s' \mid s, g(s)\big)v(s')\\ = \sum_{s'=1}^{N} p^1\big(s' \mid s, f^2(s)\big)v(s') + \sum_{s'=1}^{N} p^2\big(s' \mid s, g(s)\big)v(s')\end{aligned}$$

and so

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

and we can add $\sum_{s'=1}^{N} p^2\big(s' \mid s, g'(s)\big)v(s')$ to both sides of this equation, yielding

$$\sum_{s'=1}^{N} p\big(s' \mid s, f^1(s), g'(s)\big)v(s') = \sum_{s'=1}^{N} p\big(s' \mid s, f^2(s), g'(s)\big)v(s') \tag{3.19}$$

Similarly, from the sth line of (3.18), we obtain

$$\begin{aligned}r\big(s, f^1(s), g'(s)\big) + \sum_{s'=1}^{n} p\big(s' \mid s, f^1(s), g'(s)\big)y(s')\\ = r\big(s, f^2(s), g'(s)\big) + \sum_{s'=1}^{n} p\big(s' \mid s, f^2(s), g'(s)\big)y(s')\end{aligned}$$

from which, by the ARAT property, we have

$$
\begin{aligned}
&r^1\big(s, f^1(s)\big) + r^2\big(s, g'(s)\big) + \sum_{s'=1}^{n} p^1\big(s' \mid s, f^1(s)\big) y(s') + \sum_{s'=1}^{n} p^2\big(s' \mid s, g'(s)\big) y(s') \\
&\quad = r^1\big(s, f^2(s)\big) + r^2\big(s, g'(s)\big) + \sum_{s'=1}^{n} p^1\big(s' \mid s, f^2(s)\big) y(s') + \sum_{s'=1}^{n} p^2\big(s' \mid s, g'(s)\big) y(s')
\end{aligned}
$$

Now we may eliminate the terms contributed by player 2's choice of $g'(s)$, and replace them with $r^2(s, a)$ and $\sum_{s'=1}^{n} p^2\big(s' \mid s, a\big) y(s')$ for any $a \in A^2(s)$, yielding

$$
\begin{aligned}
&r(s, f^1(s), a) + \sum_{s'=1}^{n} p\big(s' \mid s, f^1(s), a\big) y(s') \\
&\qquad\qquad = r(s, f^2(s), a) + \sum_{s'=1}^{n} p\big(s' \mid s, f^2(s), a\big) y(s') \qquad (3.20)
\end{aligned}
$$

By virtue of (3.19) and (3.20), we can replace all superscripts of 1 with 2 in (3.15) and (3.16) for any s with $g'(s) \neq g(s)$. This is precisely the requirement for g' to be an element of $G(g \mid f^2)$. This shows $G(g \mid f^1) \subseteq G(g \mid f^2)$, and so the lemma. □

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 3.3.4. *Algorithm 3 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+1}, g^k) after the $k+1$-st execution of line 3, so that $G(f^{k+1} \mid g^k) \cup H(f^{k+1} \mid g^k)$ is empty, and next the algorithm will choose $g^{k+1} \in G(g^k \mid f^{k+1})$ for player 2. We will show the following monotonicity claim: for any discount factor β sufficiently close to 1, $\boldsymbol{v}_\beta(\Gamma|_{g^k}) > \boldsymbol{v}_\beta(\Gamma|_{g^{k+1}})$.

Let $\hat{f}^{k+1}$ be a uniform optimal strategy for player 1 in the Markov decision process $\Gamma|_{g^k}$. This will not, in general, be the strategy actually chosen by the algorithm. However, by Lemma 3.3.2, the algorithm's choice of a 1-optimal f^{k+1} has the same average value and deviation as $\hat{f}^{k+1}$ against g^k. Therefore, by Lemma 3.3.3, $G(g^k \mid f^{k+1})$ and $G(g^k \mid \hat{f}^{k+1})$ are in fact the *same set*, so g^{k+1} is an improvement for g^k against any uniform optimal choice for player 1. Finally, by Lemma 3.3.1, $G(g^k \mid \hat{f}^{k+1}) \subseteq G_\beta(g^k \mid \hat{f}^{k+1})$ for all $\beta < 1$ sufficiently close to 1. Putting this all together, we have that for all $\beta < 1$ sufficiently close to 1

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

Therefore, since $g^{k+1} \in G_\beta(g^k \mid f^{k+1})$ for all $\beta < 1$ sufficiently close to 1, the Patience Theorem (Theorem 3.2.1) proves our monotonicity claim. Since there are only a finite number of pure stationary strategies available to player 2, this monotonicity demands that 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 Lemma 3.1.10.

□

Example 3.3.5. We now illustrate the algorithm on the ARAT game in Example 3.1.4. We represent players' pure stationary strategies by 3-tuples giving the choice of action in each state. Recall that an action is identified with the node to which it leads. Let us start with $f^0 = (2,1,2)$ for player 1 and $g^0 = (1,1,3)$ for player 2.

$$P(f^0, g^0) = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ 1 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix} \text{ and } P^*(f^0, g^0) = \begin{bmatrix} \frac{2}{3} & \frac{1}{3} & 0 \\ \frac{2}{3} & \frac{1}{3} & 0 \\ \frac{2}{3} & \frac{1}{3} & 0 \end{bmatrix}$$

and solving for the average payoff and deviation yields $\boldsymbol{v}(f^0, g^0) = (0,0,0)^T$ and $\boldsymbol{y}(f^0, g^0) = (\frac{1}{3}, -\frac{2}{3}, \frac{1}{3})^T$. First we find a 1-optimal strategy for player 1 for $\Gamma|_{g^0}$. With $f^1 = (2,2,3)$, we have

$$P(f^1, g^0)\boldsymbol{v}(f^0, g^0) = \boldsymbol{v}(f^0, g^0)$$

and

$$\left[\boldsymbol{r}(f^1, g^0) + P(f^1, g^0)\boldsymbol{v}(f^0, g^0)\right]_s > v(s, f^0, g^0) + y(s, f^0 g^0)$$

for $s = 2, 3$, and $G(f^1 \mid g^0)$ is empty. Now we find an element of $G(g^0 \mid f^1)$. In fact the only element of this set is $g^1 = (1,1,2)$. Furthemore, $G(f^1 \mid g^1)$ is empty, so the algorithm halts with the average optimal pair (f^1, g^1), with $\boldsymbol{v}(f^1, g^1) = \left(\frac{3}{4}, \frac{3}{4}, \frac{3}{4}\right)^T$ and $\boldsymbol{y}(f^1, g^1) = \left(-\frac{1}{4}, \frac{1}{4}, -\frac{1}{4}\right)^T$.

CHAPTER 4

IMPLEMENTING THE ALGORITHM AND FURTHER WORK

4.1 A Python implementation

A prototype implementation of Algorithm 3, along with its discounted-case analogue, written in Python, can be found at `http://www.math.uic.edu/~mbourque/stochgame.html`. The Python module makes use of the matrix object in NumPy (Oliphant, 2006), and extends this to a class for stochastic matrices with methods for computing the Cesaro limit of a stochastic matrix and deviation matrix. These computations require computing the ergodic classes of the stochastic matrix, which is done using an implementation of the algorithm of (Fox and Landi, 1968). The central element of the module is the StochGame class, which represents a two player stochastic game, and a Strategy class representing pure stationary strategy pairs.

The module also contains a function for generating random two-player stochastic games of arbitrary size in which rewards are uniformly distributed in an interval centered around zero and transitions are drawn from a Dirichlet distribution. Finally, it includes functions for reading and writing stochastic game descriptions to text files.

4.1.1 Numerical results

Figure 4.1.1 displays the results of a applying Algorithm 3 to randomly generated ARAT games. Each game in ths simulaton has five actions available for each player. The number of states varies from 5 to 50. Ten games of each size were solved. Figure 4.1.1 shows the

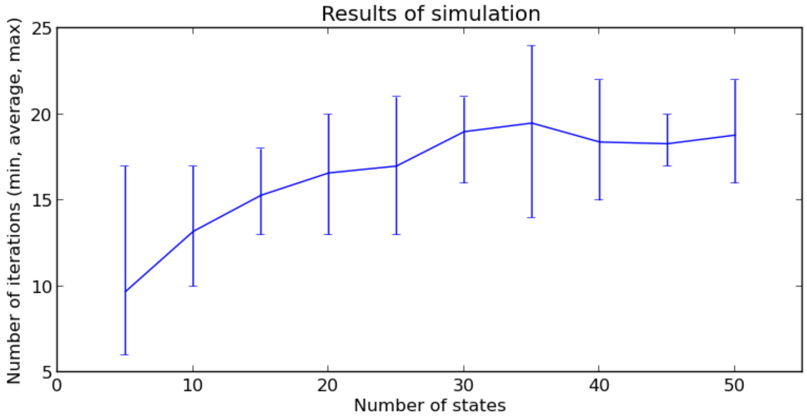

Figure 7. Maximum, minimum, and average number of iterations required for solving ARAT games with 5 actions for each player and from 5 - 50 states.

average number of iterations required (counting all iterations for player 1 and player 2), with bars displaying the maximum and minimum number of iterations required at each games size.

4.2 Further work

Given a game Γ, Algorithm 3 produces a solution (f^*, g^*) which is imbalanced, in the sense that f^* is 1-optimal for $\Gamma|_{g^*}$, but g^* is only guaranteed to be average optimal for $\Gamma|_{f^*}$, a weaker condition. Two possible resolutions to this imbalance are available. One is finding some way to reduce the work done for computing the strategy for player 1 at each iteration, resulting in mutually average-optimal strategies. The other is strengthening the algorithm to find a pair of mutually 1-optimal strategies. Such a pair may be termed 1-optimal for the game. As far as we are aware, there has been no work on 1-optimal strategies for stochastic games. We give here a definition and some basic properties.

Definition 4.2.1. A strategy pair (f^*, g^*) is 1-optimal if

$$\lim_{\beta \nearrow 1} \big(\boldsymbol{v}_\beta(f, g^*) - \boldsymbol{v}_\beta(\Gamma) \big) \leq \mathbf{0}$$

for all pure stationary strategies f for player 1 and

$$\lim_{\beta \nearrow 1} \big(\boldsymbol{v}_\beta(f^*, g) - \boldsymbol{v}_\beta(\Gamma) \big) \geq \mathbf{0}$$

for all pure stationary strategies g for the player 2.

We collect a few simple facts about 1-optimal strategies in the following lemma.

Lemma 4.2.2. *Every two player zero sum stochastic game of perfect information has at least one 1-optimal strategy pair. For any 1-optimal strategy pair (f^*, g^*),*

(a) (f^*, g^*) is average optimal; and

(b) if (h^*, k^*) is another 1-optimal strategy pair for the game, then

$\lim_{\beta \nearrow 1} \big(\boldsymbol{v}_\beta(f^*, g^*) - \boldsymbol{v}_\beta(h^*, k^*) \big) = 0$*, and*

(c) $\boldsymbol{x}(f^*, g^*) = \boldsymbol{x}(h^*, k^*)$ and $\boldsymbol{y}(f^*, g^*) = \boldsymbol{y}(h^*, k^*)$.

Proof. The existence of a 1-optimal strategy follows from the observation that a uniform optimal strategy is 1-optimal. To prove part a, suppose that (f^*, g^*) is 1-optimal but not average optimal: one of the player can improve their average payoff by changing strategies. Suppose that this is the maximizer, and that $x(f^0, g^*) \geq x(f^*, g^*)$ with a strict equality in some state for some strategy f^0 for the maximizer. Then, for any fixed β we write, by Theorem 2.2.2,

$$\begin{aligned}\big[\boldsymbol{v}_\beta(f^0, g^*) - \boldsymbol{v}_\beta(f^*, g^*)\big]_s = \\ \left[\frac{\boldsymbol{x}(f^0, g^*) - \boldsymbol{x}(f^*, g^*)}{1-\beta} + \boldsymbol{y}(f^0, g^*) - \boldsymbol{y}(f^*, g^*) + \boldsymbol{\epsilon}(f^*, g^*, f^0, \beta)\right]_s.\end{aligned} \tag{4.1}$$

Since $\big[\boldsymbol{x}(f^0, g^*) - \boldsymbol{x}(f^*, g^*)\big]_s \geq \boldsymbol{0}$ with a strict inequality for some s and $\big[\boldsymbol{y}(f^0, g^*) - \boldsymbol{y}(f^*, g^*)\big]_s$ is bounded, the left hand side of (4.1) grows without bound as $\beta \nearrow 1$. Now

$$\boldsymbol{v}_\beta(f^0, g^*) - \boldsymbol{v}_\beta(\Gamma) = \big[\boldsymbol{v}_\beta(f^0, g^*) - \boldsymbol{v}_\beta(f^*, g^*)\big] + \big[\boldsymbol{v}_\beta(f^*, g^*) - \boldsymbol{v}_\beta(\Gamma)\big],$$

and the first bracketed expression has some coordinate which grows without bound as $\beta \nearrow 1$, while the second bracketed expression coverges to zero. But this contradicts the definition of 1-optimality for (f^*, g^*).

Part b follows directly from writing, for any β

$$\boldsymbol{v}_\beta(f^*, g^*) - \boldsymbol{v}_\beta(h^*, k^*) = \big[\boldsymbol{v}_\beta(f^*, g^*) - \boldsymbol{v}_\beta(\Gamma)\big] - \big[\boldsymbol{v}_\beta(h^*, k^*) - \boldsymbol{v}_\beta(\Gamma)\big],$$

and the 1-optimality of these strategy pairs means both bracketed expressions converge to zero.

To prove part c, we use Theorem 2.2.2 again to write, for any $0 \leq \beta < 1$,

$$\boldsymbol{v}_\beta(f^*, g^*) - \boldsymbol{v}_\beta(h^*, k^*) = \frac{\boldsymbol{x}(f^0, g^*) - \boldsymbol{x}(f^*, g^*)}{1-\beta} + \boldsymbol{y}(f^*, g^*) - \boldsymbol{y}(h^*, k^*) + \boldsymbol{\epsilon}(f^*, g^*, h^*, k^*, \beta).$$

Since the above expression converges to zero, we must have $\boldsymbol{x}(f^*, g^*) = \boldsymbol{x}(h^*, k^*)$ and $\boldsymbol{y}(f^*, g^*) = \boldsymbol{y}(h^*, k^*)$. □

The previous lemma justifies the following definitions of the deviation of a two-person zero sum stochastic game of perfect information.

Definition 4.2.3. If (f^*, g^*) is a 1-optimal strategy pair for Γ, a two person zero sum stochastic game of perfect information, then deviation of the game is $\boldsymbol{y}(\Gamma) = \boldsymbol{y}(f^*, g^*)$.

4.3 Conclusion

The central new result in this thesis is the policy improvement algorithm for zero-sum ARAT stochastic games with average payoffs, Algorithm 3, which we have proved using method based on a new result regarding policy improvement for such games with discounted payoffs, the Patience Theorem. In simulations, the algorithm has shown itself to be quite fast, never checking more than 25 pure stationary strategy pairs, even with games with 50 states and 5 actions for each player for a total of 5^{50} pure stationary strategies for each player.

The maturity of any applied field is partly measured by its success in solving real-world problems. Stochastic games have arguably received less attention than they deserve from practicioners in areas that have made use of other tools from game theory such as economics and evolutionary biology. One possible reason for this is the difficulty in collecting acccurate data for describing a complex stochastic game model, but another is the difficulty in computing solutions for games, even in cases where optimal strategies are known to exist. We hope that this work will contribute to the application of stochastic games to practical problems.

CITED LITERATURE

Akian, M., Cochet-Terrasson, J., Detournay, S., and Gaubert, S.: Policy iteration algorithm for zero-sum multichain stochastic games with mean payoff and perfect information. arXiv:1208.0446, August 2012.

Aumann, R. J.: Mixed and behavior strategies in infinte extensive games. In Advances in Game Theory, eds. M. Dresher and R. J. Aumann, pages 627–650. Princeton University Press, 1964.

Avrachenkov, K., Cottatellucci, L., and Maggi, L.: Algorithms for uniform optimal strategies in two-player zero-sum stochastic games with perfect information. Operations Research Letters, 40(1):56–60, 2012.

Blackwell, D.: Discrete dynamic programming. The Annals of Mathematical Statistics, 33(2):719–726, June 1962.

Blackwell, D. and Ferguson, T. S.: The big match. The Annals of Mathematical Statistics, 39(1):159–163, February 1968.

Cochet-terrasson, J., Cohen, G., Gaubert, S., Gettrick, M. M., and Quadrat, J.-p.: Numerical computation of spectral elements in max-plus algebra. 1998.

Cochet-Terrasson, J. and Gaubert, S.: A policy iteration algorithm for zero-sum stochastic games with mean payoff. Comptes Rendus Mathematique, 343(5):377–382, September 2006.

Derman, C. and Strauch, R. E.: A note on memoryless rules for controlling sequential control processes. The Annals of Mathematical Statistics, 37(1):276–278, February 1966.

Filar, J. A. and Tolwinski, B.: On the algorithm of pollatschek and avi-ltzhak. In Stochastic Games And Related Topics, eds. T. E. S. Raghavan, T. S. Ferguson, T. Parthasarathy, O. J. Vrieze, W. Leinfellner, G. Eberlein, and S. H. Tijs, volume 7 of Theory and Decision Library, pages 59–70. Springer Netherlands, 1991.

Filar, J. A. and Vrieze, K.: Competitive Markov Decision Processes. Springer, 1997.

Fox, B. L. and Landi, D. M.: An algorithm for identifying the ergodic subchains and transient states of a stochastic matrix. *Commun. ACM*, 11(9):619–621, 1968.

Hordijk, A., Dekker, R., and Kallenberg, L. C. M.: Sensitivity-analysis in discounted markovian decision problems. *OR Spektrum*, 7(3):143–151, September 1985.

Howard, R. A.: *Dynamic Programming and Markov Process*. The MIT Press, first edition edition, June 1960.

Kemeny, J. G. and Snell, J. L.: *Finite Markov Chains*. Springer, 1960.

Kohlberg, E.: Invariant half-lines of nonexpansive piecewise-linear transformations. *Mathematics of Operations Research*, 5(3):366–372, August 1980.

Mertens, J.-F. and Neyman, A.: Stochastic games have a value. *Proceedings of the National Academy of Sciences of the United States of America*, 79(6):2145–2146, March 1982. PMID: 16593178 PMCID: 346143.

Oliphant, T.: *Guide to NumPy*. Provo, UT, Brigham Young University, March 2006.

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

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

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

Shapley, L. S.: Stochastic games. *Proceedings of the National Academy of Sciences of the United States of America*, 39(10):1095–1100, October 1953.

Syed, Z. U.: Algorithms for stochastic games and related topics. Doctoral dissertation, University of Illinois at Chicago, Chicago, IL, USA, 1999. AAI9941509.

Veinott, A. F.: On finding optimal policies in discrete dynamic programming with no discounting. *The Annals of Mathematical Statistics*, 37(5):1284–1294, October 1966.

VITA

Education

BS in Mathematics	Oberlin College, Oberlin, Ohio, May 1998
MS in Mathematics	University of Illinois at Chicago, May 2009 (with concentration in Statistics)
PhD in Mathematics	University of Illinois at Chicago, July 2013

Teaching Experience

Spring 2013	Lecturer and instructor of record for Stat 101 large lecture, University of Illinois at Chicago
2008 - 2012	Emerging Scholars Program instructor at UIC
2007 - 2012	Teaching assistant for mathematics discussion sections at UIC

Talks

Dec 2012	*A policy improvement algorithm for stochastic games of perfect information with average payoffs*, Internation Conference on Game Theory and Operations Research, Hyderabad, India

April 2010 *Stochastic Games: Discounted Competitive Markov Decision Processes*, Fourth Annual Graduate Student Probability Conference, Duke University, Durham, NC

April 2009 *Intervention Analysis in a Water Usage Study* (with Xuejing Wang) UIC Statistics Seminar

Awards

July 2010 Graduate student travel award for Fourth International Congress of the Game Theory Society

April 2009 Data Analysis Award for statistical consulting, UIC Statistics program