

A Policy-Improvement Type Algorithm for Solving Zero-Sum Two-Person Stochastic Games of a Special Class

T. E. S. Raghavan^{1,2} and Zamir Syed³

Abstract

We give a policy-improvement type algorithm to locate an optimal pure stationary strategy for discounted stochastic games with additive reward and additive transition structure.

Keywords: Stochastic games, MDP, ARAT games, and Policy Improvement

1. Introduction

Discounted stochastic games were first introduced by Shapley [1953]. In a stochastic game Γ , we have a finite set of states $S = \{1, 2, \dots, s\}$, and for each state $t \in S$ there are two finite sets $A(t) = \{1, 2, \dots, a_t\}$ and $B(t) = \{1, 2, \dots, b_t\}$ called the action sets for players I and II respectively. For each triple (t, a, b) with $a \in A(t)$ and $b \in B(t)$ there is an immediate reward $r(t, a, b)$ as well as a probability distribution $p(t, a, b)$ on the set S . Given an initial starting state $t_0 \in S$, the game is played as follows. The players simultaneously choose actions $a^0 \in A(t_0)$ and $b^0 \in B(t_0)$ resulting in the payment $r(t_0, a^0, b^0)$ to player I by player II. The system moves to a new state t_1 according to $p(t_0, a^0, b^0)$ and the players again choose actions $a^1 \in A(t_1)$ and $b^1 \in B(t_1)$. Accordingly the payment $r(t_1, a^1, b^1)$ is made to player I by player II and the game moves to a new state t_2 according to $p(t_1, a^1, b^1)$ and so on. The game continues infinitely and the rewards $r(t_i, a^i, b^i)$ are recorded. A general strategy for a player would be a function from the

¹Department of Mathematics, Statistics and Computer Science, University of Illinois at Chicago; e-mail: ter@uic.edu

²Partially Funded by NSF Grant DMS 930-1052 and DMS 970-4951

³Department of Mathematics, Statistics and Computer Science, University of Illinois at Chicago; e-mail: ztatti@uic.edu

set of all possible histories into the set of probability distributions over the player's action space. A general strategy can therefore be very complicated but nevertheless, given a pair of strategies (π, ρ) for both players, we can evaluate the expected β -discounted value:

$$\phi_\beta(\pi, \rho)(t_0) = \sum_{n=0}^{\infty} \beta^n r_n(t_0, \pi, \rho)$$

where t_0 is the starting state and $r_n(t_0, \pi, \rho)$ is the expected reward (to player I) at the n th stage when the players are using π and ρ . Under this payoff we can define an equilibrium pair of strategies to be a pair (π^*, ρ^*) such that:

$$\phi_\beta(\pi, \rho^*) \leq \phi_\beta(\pi^*, \rho^*) \leq \phi_\beta(\pi^*, \rho)$$

for all π and ρ . We write $\phi_\beta(\Gamma) = \phi_\beta(\pi^*, \rho^*)$. A strategy is said to be *stationary* if it only depends on the current state. Shapley[1953] showed that under the discounted payoff criterion, there always exists an equilibrium pair in stationary strategies. If a stationary strategy is non-randomized in every state then it is called *pure stationary*.

2. ARAT Games

A stochastic game is of the ARAT (additive reward and additive transition) class if the reward and transition functions can be written as

$$r(t, a, b) = r^1(t, a) + r^2(t, b) \tag{1}$$

$$p(t, a, b) = p^1(t, a) + p^2(t, b) \tag{2}$$

for all triples (t, a, b) with $t \in S$, $a \in A(t)$, and $b \in B(t)$. In Tijs, Raghavan, and Vrieze[1985] it is shown that such games possess pure stationary equilibrium strategies.

For player I a pure stationary strategy is simply a function $f : S \rightarrow \cup_{i=1}^s A(i)$ with $f(t) \in A(t)$ for all t , that is, in state t player I always chooses the action $f(t)$. Similarly a function $g : S \rightarrow \cup_{i=1}^s B(i)$ with $g(t) \in B(t)$ for all t is a pure stationary strategy for player II. For a pair of pure stationary strategies (f, g) we write $[(f^1, g^1), \dots, (f^s, g^s)]$ where (f^k, g^k) is the pair of actions chosen in state k under (f, g) . We write $r(f, g)$ to be a vector, indexed by the state space, with $[r(f, g)]_t = r(t, f^t, g^t)$. Also we write $Q(f, g)$ for the

matrix whose i th row is $p(i, f^i, g^i)$. $Q(f, g)$ is simply the transition matrix induced by f and g on the state space. Because of the ARAT condition we can appropriately split $Q(f, g) = Q^1(f) + Q^2(g)$ and $r(f, g) = r^1(f) + r^2(g)$ into each player's part.

Because of such structure in the parameters, it is possible to use the policy-improvement procedures of markov decision problems for solving ARAT stochastic games.

3. Algorithm

Policy iteration procedures (Blackwell[1962]) for MDPs basically proceed by improving on an initial policy thus generating a sequence of policies, each one better than the previous. A problem encountered in using policy-improvement in stochastic games is that it is not clear as to what an "improvement" is. In an MDP there is only one player who is trying to optimize payoff. Thus, if the MDP is a maximization problem, an improvement f' over a policy f would simply entail having the payoff of f' greater than that of f . To find such a policy f' is a trivial calculation, and that is what makes policy improvement a desirable method. In zero-sum stochastic games there are two players, player I trying to maximize and player II trying minimize. Furthermore, in such a scenario we are dealing with pairs of strategies rather than a single policy. Fortunately this duality allows us to speak of the concept of individual improvement with respect to each player.

Given a pair of pure stationary strategies (f, g) , a new pair of pure stationary strategies (h, g) is called an *improvement for player I* if

1. $\exists k, 1 \leq k \leq s$, with $h^k \neq f^k$ and $h^i = f^i$ for $i \neq k$
2. $\phi_\beta(h, g) > \phi_\beta(f, g)$

The first condition is an adjacency condition required for termination in our algorithm, and the second condition is self-explanatory. In the first condition we say that h differs from f in state k . Of course we have the analogous definition for an *improvement for player II*. Namely it is a pair (f, h) where:

1. $\exists k, 1 \leq k \leq s$, with $g^k \neq h^k$ and $g^i = h^i$ for $i \neq k$
2. $\phi_\beta(f, h) < \phi_\beta(f, g)$

Clearly if a pair has no improvements for either player then it is a saddlepoint. The algorithm that follows will start with any initial pair of pure stationary strategies (f_0, g_0) and generate a sequence of improvements $(f_1, g_1), (f_2, g_2), \dots$. In order to ensure termination we require a *lexicographic* search for improvements. The procedure for finding a lexicographic improvement (f_{i+1}, g_{i+1}) of (f_i, g_i) will be to start at state 1 and look for an improvement for player I. If there are none then a search for an improvement for player II is done. If there are still none then we proceed to state 2 and repeat the same procedure. More precisely we can say that a pair (f', g') is a *lexicographic* improvement of the pair (f, g) if exactly one of the following holds:

1. It is an improvement for player I, f' differs from f in state k , and there are no improvements for either player of (f, g) in states $1, 2, \dots, k-1$.
2. It is an improvement for player II, g' differs from g in state k , there are no improvements for either player of (f, g) in states $1, 2, \dots, k-1$, and no improvements for player I in state k .

Clearly a pair of pure stationary strategies is an equilibrium pair if and only if it has no lexicographic improvements.

Remark: When writing computer programs for searching for an improvement, lexicographic improvement is the natural method which arises.

Algorithm:

1. Choose an initial pair of pure stationary strategies (f_0, g_0) arbitrarily (e.g. $f_0^k = g_0^k = 1$ for $k = 1, \dots, s$) and set $\tau = 0$.
2. Compute $\phi_\beta(f_\tau, g_\tau)$
3. Search for a lexicographic improvement $(f_{\tau+1}, g_{\tau+1})$ of (f_τ, g_τ) . There are two cases:
 - Case 1: A lexicographic improvement is found. In this case let $\tau = \tau+1$ and go to step 2.
 - Case 2: There are no lexicographic improvements. Go to step 4.
4. The pair $(f^*, g^*) = (f_\tau, g_\tau)$ is a saddlepoint.

By repeated use of a few lemmas we will show that the algorithm terminates with a saddlepoint pair (f^*, g^*) . The proof will be based on an assumption concerning player II's condition in the game. If this condition is assumed for player I the proof is the same except with reverse notations. Thus we will only give the former.

Given an ARAT stochastic game Γ , let F and G denote the sets of pure stationary strategies available to players I and II respectively. For a subset $X \subset B_t$ we write Γ_X^t to be the subgame of Γ in which only the actions in X are allowed to player II in state t . Let F_X^t and G_X^t be the corresponding pure stationary strategy sets of Γ_X^t . We will omit the superscript t and write Γ_X , F_X , and G_X when the reduced state is understood. In some cases when we deal with a singleton $X = \{i\}$ we will write Γ_i .

Lemma 1: Let X and Y be a partition of B_t , i.e. $X \cup Y = B_t$ and $X \cap Y = \emptyset$. Then $\phi_\beta(\Gamma) = \min\{\phi_\beta(\Gamma_X), \phi_\beta(\Gamma_Y)\}$.

Proof: By assumption Γ is ARAT so that Γ_X and Γ_Y are ARAT as well. Therefore, all the aforementioned games possess equilibria in pure stationary strategies. It is also clear that $F = F_X = F_Y$ and $G = G_X \cup G_Y$. Hence we can write:

$$\begin{aligned} \phi_\beta(\Gamma) &= \min_{g \in G} \max_{f \in F} \phi_\beta(f, g) \\ &= \min\{\min_{g \in G_X} \max_{f \in F} \phi_\beta(f, g), \min_{g \in G_Y} \max_{f \in F} \phi_\beta(f, g)\} \\ &= \min\{\min_{g \in G_X} \max_{f \in F_X} \phi_\beta(f, g), \min_{g \in G_Y} \max_{f \in F_Y} \phi_\beta(f, g)\} \\ &= \min\{\phi_\beta(\Gamma_X), \phi_\beta(\Gamma_Y)\} \end{aligned}$$

.

◇

The next lemma is the heart of the algorithm. It uses a theorem of Vrieze & Tijs[1980] for which some notation is required.

Given a finite stochastic game Γ and an s by 1 vector v , a collection of matrix games $M_t(v)$, $t = 1, \dots, s$ are induced. We define:

$$[M_t(v)]_{ij} = r(t, i, j) + \beta p(t, i, j)v$$

so that $M_t(v)$ as an a_t by b_t matrix. When considering a stationary strategy σ for player I in Γ , we are dealing with a collection of probability distributions over the action sets A_1, \dots, A_s . Each probability distribution can now be viewed as a collection of strategies for the row players of the games $M_1(v), M_2(v), \dots, M_s(v)$.

Theorem 1: A stationary strategy σ (ρ) for player I (II) is optimal in the game Γ if and only if it is optimal for the matrix games $M_t(\phi_\beta(\Gamma))$, $t = 1, \dots, s$.

Proof: See Vrieze[1983]. ◇

Lemma 2: If (f, g) is a pure stationary equilibrium pair for the reduced game Γ_i and if $\phi_\beta(\Gamma) = \phi_\beta(\Gamma_i)$ then (f, g) is an equilibrium pair for Γ as well.

Proof: The strategy g being optimal for player II in Γ_i implies that $\forall f' \in F_i = F$, $\phi_\beta(f', g) \leq \phi_\beta(\Gamma_i) = \phi_\beta(\Gamma)$. This shows that g is optimal for II in Γ as well. Let $v = \phi_\beta(\Gamma) = \phi_\beta(\Gamma_i)$. In view of theorem 1, to show that f is optimal for player I in Γ requires showing that f chooses the optimal action in the matrix games $M_t(v)$, $t = 1, \dots, s$, induced by Γ . We will write $M'_t(v)$ for the matrix games induces by Γ_i using the same vector v .

Since f is optimal for player I in the reduced game Γ_i , it chooses the optimal action in $M'_t(v)$ for $t = 1, 2, \dots, s$. As $M'_t(v) = M_t(v)$ for $t = 1, \dots, s - 1$ we only need to show that f chooses the optimal action of $M_s(v)$. Because Γ is ARAT we can write:

$$[M_s(v)]_{ij} = r^1(s, i) + \beta p^1(s, i)v + r^2(s, j) + \beta p^2(s, j)v = m_i^1 + m_j^2$$

For such a matrix game, it is easy to check that a best reply of player I to any single strategy of player II is actually a best reply to all strategies of player II. Since f^s is an optimal action of $M'_s(v)$, it is a best reply to g^s . Thus we can conclude that f chooses an optimal action in $M_s(v)$. This proves that (f, g) is an equilibrium pair of Γ as well. ◇

Remark: The structure of the matrix $M_s(v)$ described in the last paragraph of the proof explains why ARAT games possess pure stationary equilibrium.

Theorem 2: The algorithm terminates with an equilibrium pair (f^*, g^*) of pure stationary strategies.

Proof: We proceed by induction on the quantity $n = \sum_{t=1}^s (a_t + b_t)$. The smallest this can be is 2 in which case there is nothing to prove. Assume that the theorem holds for $n = 2, \dots, k - 1$ and suppose $n = k$. If $|a_t| = |b_t| = 1$ for all t then again there is nothing to prove so assume that at least one player has more than one action in some state. Let j be the the largest state in which a player has more than one action. In state j suppose that player II has more than one action. It is possible that player II have only one action in state j in which case player I must have more than one action. As mentioned

before, the proof of this latter case is almost identical to that of the former. Therefore we will only present a proof for the case of player II having more than one action in state j .

Let $c_1 = g_0^j$, i.e. the action chosen by g_0 in state j , and let $X_1 = \{c_1\}$. By the induction hypothesis we have that the algorithm will generate a sequence $(f_0, g_0), (f_1, g_1), \dots, (f_{m_1}, g_{m_1})$ with (f_{m_1}, g_{m_1}) an equilibrium pair of $\Gamma_{X_1}^j$ (from now on we omit the superscript j). If (f_{m_1}, g_{m_1}) is an equilibrium pair of Γ as well then the algorithm terminates with $(f^*, g^*) = (f_{m_1}, g_{m_1})$. From lemma 1 and lemma 2 we know that this would happen if and only if $\phi_\beta(f_{m_1}, g_{m_1}) = \phi_\beta(\Gamma)$. Otherwise, the only possible improvement of (f_{m_1}, g_{m_1}) will be for player II to change action in state j (and in this case we would have $\phi_\beta(f_{m_1}, g_{m_1}) < \phi_\beta(\Gamma)$). From here the algorithm would improve to a pair (f_{m_1+1}, g_{m_1+1}) with $g_{m_1+1}^j \neq c_1$ and continue the lexicographic improvement. Let $c_2 = g_{m_1+1}^j$ and let $X_2 = \{c_1, c_2\}$.

Again by the induction hypothesis we have that the algorithm will reach an equilibrium pair (f_{m_2}, g_{m_2}) of the subgame Γ_{X_2} . If $\phi_\beta(f_{m_2}, g_{m_2}) = \phi_\beta(\Gamma)$ then we have reached an equilibrium point of Γ . Otherwise the only improvement is for player II to change action in state j . The difference now is that player II cannot improve to the action c_1 . The reason for this is that (f_{m_2}, g_{m_2}) is an equilibrium pair of Γ_{X_2} . We get this by using lemma 1 (with $X = X_1$ and $Y = c_2$) and lemma 2 (with $i = c_2$). Repeating this same line of argumentation, it is easy to see that the algorithm will never revisit an action of player II in state j . Thus the sequences $c_1, c_2, \dots, X_1, X_2, \dots$, and $(f_{m_1}, g_{m_1}), (f_{m_2}, g_{m_2}), \dots$ are generated. As state j has only a finite number of actions, the algorithm must reach an action c_q with $\phi_\beta(\Gamma_{c_q}) = \phi_\beta(\Gamma)$ (lemma 1) so that $(f^*, g^*) = (f_{m_q}, g_{m_q})$ is an equilibrium pair of the game Γ (lemma 2). \diamond

Remark: The proof in the case of only player I having more than one action in state j requires using the player I analogs of lemmas 1 and 2. In lemma 1 one need only employ the "max-min" version of an equilibrium pair.

4. Computation

One of the computational advantages of our algorithm comes from the adjacency condition of an improvement. Given a pair of pure stationary strategies

(f, g) , the computation of $\phi_\beta(f, g)$ involves solving the system:

$$[I - \beta Q(f, g)]\phi = r(f, g)$$

However, in step 3 of the algorithm, after finding an improvement pair, it is not necessary to solve another complete system. Instead one might do the following: First compute the inverse $[I - \beta Q(f_0, g_0)]^{-1}$ and use it to compute $\phi_\beta(f_0, g_0) = [I - \beta Q(f_0, g_0)]^{-1}r(f_0, g_0)$. Now any improvement pair (f_1, g_1) would differ from (f_0, g_0) in exactly one state. Thus $Q(f_1, g_1)$ would differ from $Q(f_0, g_0)$ in exactly one row. This allows us to compute the inverse $[I - \beta Q(f_1, g_1)]^{-1}$ by a single matrix multiplication and a column pivot to obtain $\phi_\beta(f_1, g_1) = [I - \beta Q(f_1, g_1)]^{-1}r(f_1, g_1)$.

An Example:

In the following example we have three states. In each state player I plays the rows and player II the columns. Each block contains the immediate reward followed by a probability vector. For instance, in state 2 if player I choose action 2 and player II chooses action 1, then the immediate reward paid to I by II is 7 and the system moves to a new state via the probability vector $(.50, .25, .25)$.

State 1:

	1	2
1	4/(.50, .25, .25)	6/(.75, .25, .00)
2	3/(.00, .75, .25)	5/(.25, .75, .00)
3	7/(.00, .25, .75)	9/(.25, .25, .50)

State 2:

	1	2	3
1	5/(.75, .25, .00)	4/(.25, .75, .00)	4/(.25, .25, .50)
2	7/(.50, .25, .25)	6/(.50, .25, .25)	6/(.00, .25, .75)

State 3:

	1	2	3
1	2/(.25, .50, .25)	6/(.50, .25, .25)	4/(.25, .00, .75)

One can check that this game is indeed ARAT. We will use the notation $[(a_1, b_1), (a_2, b_2), (a_3, b_3)]$ to refer to the pair of strategies (f, g) with $f^i = a_i$ and $g^i = b_i$. The algorithm used $[(1, 1), (1, 1), (1, 1)]$ as the initial pair (f_0, g_0) and generated the following sequence:

τ	(f_τ, g_τ)	$\phi_\beta(f_\tau, g_\tau)$
0	$[(1, 1), (1, 1), (1, 1)]$	(39.3, 40.7, 37.6)
1	$[(3, 1), (1, 1), (1, 1)]$	(47.7, 48.0, 44.3)
2	$[(3, 1), (2, 1), (1, 1)]$	(51.5, 53.0, 48.3)
3	$[(3, 1), (2, 2), (1, 1)]$	(49.7, 50.5, 46.3)
4	$[(3, 1), (2, 3), (1, 1)]$	(43.4, 42.4, 39.8)

Thus the optimal pair of strategies is given by $[(3, 1), (2, 3), (1, 1)]$.

References

D. Blackwell [1962] : *Discrete Dynamic Programming*. Annals of Mathematical Statistics **33**, 719-726.

L.S. Shapley [1953]: *Stochastic Games* Proceedings of the National Academy of Sciences U.S.A. **39**, 1095-1100.

S.H. Tijs, T.E.S. Raghavan, and O.J. Vrieze [1985]: *On Stochastic Games with Additive Reward and Transition Structure* Journal of Optimization Theory and Applications, **47**, 451-464.

O.J. Vrieze [1983]: *Stochastic Games with Finite State and Action Spaces*, (Ph.D Thesis), Free University, Math. Centrum, Amsterdam.

O.J. Vrieze and S.H. Tijs [1980]: *Relations Between Game Parameters, Value, and Optimal Strategy Spaces in Stochastic Games and Construction of Games with Given Solution* Journal of Optimization Theory and Applications, **31**, 501-513.