FINITE-STEP ALGORITHMS FOR SINGLE-CONTROLLER AND PERFECT INFORMATION STOCHASTIC GAMES

T.E.S. RAGHAVAN
University of Illinois at Chicago
Chicago, USA

Abstract. After a brief survey of iterative algorithms for general stochastic games, we concentrate on finite-step algorithms for two special classes of stochastic games. They are Single-Controller Stochastic Games and Perfect Information Stochastic Games. In the case of single-controller games, the transition probabilities depend on the actions of the same player in all states. In perfect information stochastic games, one of the players has exactly one action in each state. Single-controller zero-sum games are efficiently solved by linear programming. Non-zero-sum single-controller stochastic games are reducible to linear complementary problems (LCP). In the discounted case they can be modified to fit into the so-called LCPs of Eave's class $\mathcal{L}$. In the undiscounted case the LCP's are reducible to Lemke's copositive plus class. In either case Lemke's algorithm can be used to find a Nash equilibrium. In the case of discounted zero-sum perfect information stochastic games, a policy improvement algorithm is presented. Many other classes of stochastic games with orderfield property still await efficient finite-step algorithms.

1. Introduction

From the point of view of modelling real-life applications of discrete dynamic games as stochastic games, the key issue is, having modelled practical problems as stochastic games, how would one solve for equilibrium payoffs and strategies for such stochastic games? What are some efficient algorithms for stochastic games that can be solved in finite arithmetic steps? Here we report some recent progress in this direction.

2. Zero-Sum Two-Person Stochastic Games with Discounted Payoff

Imagine two players playing one of possibly different matrix games at each stage, and the game at each stage depends on the previous game and the entry selected by the players. Games with just such conceptual structure are called *stochastic games*. They were introduced in a seminal paper by Shapley [44]. Let $A^1, A^2, \ldots, A^N$ be real matrices known to the two players. By *state* s we mean the matrix game A^s. Players start in, say, state s. They play the matrix game A^s. Immediately thereafter, player I receives the payoff from player II and the game moves to A^k with probability $q(k|s, i_s, j_s)$ which depends on the choices i_s, j_s by players I and II in state s. At the next stage they play A^k, and so on. The transition probabilities known to both players are assumed to be Markovian in the sense that the probability of the next game is determined only by the immediate past and not by the entire history. The aim of player I is to get as much as possible. The aim of player II is to lose as little as possible. Of course, a repeated matrix game is a very special case of this game where $A^1 = A^2 = \ldots = A^N$. Since the game never ends it is not clear what is meant by maximizing the payoff. We shall emphasize two particular payoff criteria that are commonly considered in the literature.

In the *discounted payoff*, with $0 \leq \beta < 1$, one takes as payoff

$$\sum_{n=1}^{\infty} \beta^{n-1} r(s_n, i_n, j_n) \tag{1}$$

where $r(s_n, i_n, j_n) = a_{i_n j_n}^{(s_n)}$ = payoff on the n-th time point, where the matrix game A^{s_n} is played, and row i_n and column j_n are chosen there. Under the above criterion, the current rewards are more important than the future prospects.

In the *undiscounted payoff*, or the *limiting average payoff*, also called the *Cesaro average payoff*, one takes as payoff

$$\liminf_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} r(s_n, i_n, j_n). \tag{2}$$

While one may envisage the possibility of developing complex strategies based on all the accumulated history at each time point, in his fundamental paper [44] Shapley showed that β-discounted stochastic games[1] can be

[1]Actually, Shapley considered games with positive stopping probabilities in every instance; however, the analysis of the games introduced in [44] is equivalent to the analysis of the classical discounted stochastic games, and it is the latter class that was studied in subsequent publications.

played optimally using *stationary strategies*. A stationary strategy is one where the players play in a "memoryless" way in the following sense: for each matrix game A^s, player I (player II) selects a fixed probability distribution on the rows (columns) of A^s and no matter how the matrix A^s is reached, the rows (or columns) are chosen according to the fixed probability distribution. Finally, an even simpler strategy is to select for each matrix game A^s a particular row (or column) to be played whenever state s is reached. These are called *pure stationary strategies*, and will be seen to be adequate for some very special classes of stochastic games [43].

3. Iterative Algorithms for Discounted Zero-Sum Games

In his fundamental paper [44], Shapley showed that player I, by using an optimal stationary strategy f^0, can guarantee the expected discounted payoff of $v(s)$, no matter what strategy the opponent adopts. Similarly for an optimal stationary g^0, the expected discounted payoff $\phi_\beta(f, g^0)(s)$ is at most $v(s)$ against all strategies f of player I.

Shapley's proof contained an algorithm to compute approximately the value and optimal stationary strategies. This can be illustrated for the following stochastic game with two states and discount factor $\beta = .5$.

Example 1

$$\begin{array}{cc} s=1 & s=2 \\ \begin{bmatrix} 3/1 & 0/2 \\ 0/2 & 1/1 \end{bmatrix} & \begin{bmatrix} 0/2 \end{bmatrix} \end{array}$$

Here, in state $s = 1$, when players choose row 1 and column 1 or row 2 and column 2, the play remains in the same state. Otherwise it moves to state 2, where the play is permanently absorbed. Imagine God separately promising players I (II) to play for his/her side after their first choice. In state 2, the players have no action. In state 1, player I and player II can choose row 1, column 1 and player I can expect $3 + .5v_1$ where v_1 is what God can get for him/her if He played from the beginning. If they choose row 1 and column 2, player I can expect $0 + 0.5.v_2$. Here v_1 and v_2 are the optimal value starting at states 1 and 2. Clearly $v_2 = 0$ and thus we are led to an auxiliary game with payoffs

$$A(v_1) = \begin{bmatrix} 3 + .5(v_1) & 0 \\ 0 & 1 + .5(v_1) \end{bmatrix}.$$

Playing the original stochastic game optimally corresponds to playing this auxiliary game optimally. Shapley showed that this corresponds to solving the non-linear equation: $v = (v_1, v_2)$ where

$$v_1 = \text{value}\begin{bmatrix} 3+.5(v_1) & 0 \\ 0 & 1+.5(v_1) \end{bmatrix} \quad v_2 = \text{value}\begin{bmatrix} 0+.5(v_2) \end{bmatrix}.$$

The map $\phi : v \to \text{value}[A(v)]$ is a contraction and v is the unique fixed point of this operator. We can therefore solve the game iteratively as follows.

– *Step 1:* $\tau = 1$, $v^1 = 0$. The auxiliary game to solve for v^τ is

$$v^2 = \text{value}\begin{bmatrix} 3+.5(0) & 0 \\ 0 & 1+.5(0) \end{bmatrix} = \frac{3}{4}.$$

– *Step 2:* $\tau = 2$, $v^2 = \frac{3}{4}$. The auxiliary game to solve for v^3 is

$$v^3 = \text{value}\begin{bmatrix} 3+.5(\frac{3}{4}) & 0 \\ 0 & 1+.5(\frac{3}{4}) \end{bmatrix}.$$

The matrix games appearing in the equations defining v^τ $(\tau > 1)$ are completely mixed and the value can be found by the formula for completely mixed games. The simple arithmetic computations yield $v^3 = \frac{297}{304}$ v^4 is approximately 1.0433. Indeed, the iterates v^τ converge as $\tau \to \infty$ to the value $v_1 = \frac{-4+2\sqrt{13}}{3}$ of the discounted game.

In general, solving the discounted stochastic game can be quite slow. More generally, the built-in algorithm of Shapley can be stated as follows.

Algorithm 1 (Shapley [44]).

– *Step 1:* Start with any approximation for the true value $v(s)$ of the stochastic game, say $v^1(s)$, for every state s.
– *Step 2:* Define recursively, for each state s,

$$v^{(n)}(s) = \text{value}\,[a_{ij}^{(s)} + \beta \sum_t q(t \mid s,i,j) v^{n-1}(t)]. \tag{3}$$

It can easily be shown that the above sequence of approximations converges to $v(s)$, the unique fixed point of the non-linear functional equations:

$$v(s) = \text{value}\,[a_{ij}^{(s)} + \beta \sum_t q(t \mid s,i,j) v(t)]. \tag{4}$$

Remark. While near-optimal stationary strategies can be derived from the above scheme when v^n is sufficiently close to v, it should be noted that Shapley's algorithm does not utilize the information contained in the optimal strategies of $A^s(v^n)$'s at each iteration.

The literature on stochastic games now contains a number of iterative algorithms that attempt to improve on the preceding basic scheme of Shapley based on non-linear programming techniques [18], [49], [37], [45], [52], [12].

Algorithm 2 (Hoffman and Karp [18])

– *Step 1:* Set $v^0(s) = 0$ for each state s and $\tau = 0$.
– *Step 2:* Find an optimal strategy for player II in the matrix games $A^s(v^\tau)$ for each state s. Let $g_{\tau+1}$ be one such optimal strategy.
– *Step 3:* Solve the MDP (Markov Decision Process) problem $v^{\tau+1} = \max_f \phi_\beta(f, g_{\tau+1})$.
– *Step 4:* Put $\tau := \tau + 1$ and return to Step 2.

It can be shown that $v^\tau \to v$, the value vector of the stochastic game, as $\tau \to \infty$. Note that this algorithm iterates in both the value space and the strategy space. Since we are moving in both policy and value space, we use past information. The MDP can be solved by linear programming (see [22]).

Algorithm 3 (Pollatschek and Avi-Itzhak [37])

– *Step 1:* Select an arbitrary initial approximation $v^0 = (v^0(1), \ldots, v^0(N))$ to the value vector.
– *Step 2:* At iteration τ, v^τ is known. Solve the N matrix games $A^s(v^\tau)$ for optimal strategies $f^\tau(s), g^\tau(s)$ for players I and II.
– *Step 3:* Set $f^\tau = (f^\tau(1), \ldots, f^\tau(N))$ and $g^\tau = (g^\tau(1), \ldots, g^\tau(N))$. Compute $v^{\tau+1} = [I - \beta Q(f^\tau, g^\tau)]^{-1} r(f^\tau, g^\tau)$.
– *Step 4:* Set $\tau := \tau + 1$ and return to Step 2.

Theorem 1 (Pollatschek and Avi-Itzhak [37]) *The above algorithm converges when*

$$\max_s \{\sum_t [\max_{i,j} q(t \mid s, i, j) - \min_{i,j} q(t \mid s, i, j)]\} \leq \frac{1-\beta}{\beta}.$$

The algorithm of Pollatschek and Avi-Itzhak is closer to the classical Newton-Raphson procedure. For example, if the value of the auxiliary game $\phi(v)$ is differentiable in v with first two partial derivatives in a neighborhood of v, the algorithm reduces to applying Newton's method to solve the equation $\phi(v) - v = 0$. Breton, in her Ph.D. thesis [6], made empirical studies on this algorithm and other algorithms of an iterative nature. With random data and with 15 states and 15 actions in each state, Breton observed that

- The Pollatschek–Avi-Itzhak algorithm is the fastest whenever it converges.
- Shapley's algorithm is better at getting ϵ-optimal strategies than Hoffman and Karp's.
- Hoffman and Karp's algorithm is better at getting ϵ value vector than Shapley's.

There are other algorithms that use fictitious play [13]. They are known to be slow and we know that fictitious play is unsuitable even for ordinary bimatrix games [23].

4. Orderfield Property

The data defining an n-person discounted stochastic game consists of immediate rewards, transition probabilities, and the discount factor. We say that the game has *orderfield property* if all the entries to at least one solution to the problem lie in the smallest ordered subfield $\mathcal{F}$ of reals that contains the data. Since all field elements are generated from the data by finite arithmetic operations, one hopes to solve such games by a finite arithmetic step algorithm. Our first example (Example 1) has no orderfield property. Its value vector $(v_1, v_2) = (\frac{-4+2\sqrt{13}}{3}, 0)$ is irrational. Its unique stationary optimal strategies also have irrational coordinates. From an algorithmic point of view, it therefore becomes important to look for subclasses of stochastic games possessing orderfield property.

Theorem 2 *The following classes of stochastic games possess orderfield property.*

- *Discounted and undiscounted single-controller zero-sum stochastic games. (Here the transition depends upon the actions of the same player in all states.)*
- *Discounted and undiscounted SER-SIT stochastic games. (Here the rewards are separable and the transitions are state-independent.)*
- *Discounted and undiscounted zero-sum switching control stochastic games. (Here the transition depends on the action of at most one player in each state.)*
- *Discounted and undiscounted zero-sum ARAT games. (Here the rewards and transitions are additive.)*
- *Discounted and undiscounted zero-sum games of perfect information. (Here at most one player has more than one action in each state.)*
- *Discounted and undiscounted non-zero-sum single-controller games.*
- *Discounted and undiscounted non-zero-sum games of perfect information.*
- *Discounted and undiscounted non-zero-sum SER-SIT games.*

(See [13].)

Even though many such subclasses of stochastic games do possess orderfield property (see [39], pp. 446-447), we will concentrate on two special classes, namely single-controller stochastic games and perfect information stochastic games. These two classes have been studied extensively in both discounted and Cesaro average payoffs [35], [51], [20], [43], [33],[32],[41], [16], [26], [7], [27], [53], [40].

5. Zero-Sum Two-Person Single-Controller Stochastic Games

Consider the following stochastic game with two states and with immediate rewards.

$$A^1 = \begin{bmatrix} 1 & 2 \\ 5 & 0 \\ 0 & 4 \end{bmatrix} \quad A^2 = \begin{bmatrix} 0 & 3 & 6 \\ 6 & 2 & 0 \end{bmatrix}$$

$$\begin{array}{ccccc} \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ 1 & 2 & 1 & 2 & 1 \end{array}$$

Here the transitions are controlled by player II, the column player. His choice of column determines the transitions. If player II chooses column 2 in state 1, the game moves to state 2. In state 2, if player II chooses column 3, the game moves to state 1. It was shown by Parthasarathy and Raghavan [35] that in this class of games and, more generally, when the transition probability $q(t \mid s, i, j)$ is of the type $q(t \mid s, j)$, the game has orderfield property in both discounted and Cesaro average payoffs. The game can easily be found by linear programming (LP). The reason is quite simple. With a discount β the two auxiliary games are given by

$$A^1(v_1, v_2) = \begin{bmatrix} 1 + \beta v_1 & 2 + \beta v_2 \\ 5 + \beta v_1 & 0 + \beta v_2 \\ 0 + \beta v_1 & 4 + \beta v_2 \end{bmatrix}$$

$$A^2(v_1, v_2) = \begin{bmatrix} 0 + \beta v_1 & 3 + \beta v_2 & 6 + \beta v_1 \\ 6 + \beta v_1 & 2 + \beta v_2 & 0 + \beta v_1 \end{bmatrix}.$$

Every list of variables, $(x_1, x_2, x_3), (\xi_1, \xi_2), (v_1, v_2)$, such that $v_1 + v_2$ maximizes the sum $u_1 + u_2$ subject to $(x_1, x_2, x_3), (\xi_1, \xi_2)$ being a stationary strategy (i.e., $x_1 + x_2 + x_3 = 1, \xi_1 + \xi_2 = 1$, and $x_1, x_2, x_3, \xi_1, \xi_2 \geq 0$) such that (x_1, x_2, x_3) guarantees in $A^1(u_1, u_2)$ a payoff $\geq u_1$ (i.e., $x_1 + 5x_2 + \beta u_1 \geq u_1$ and $2x_1 + 4x_3 + \beta u_2 \geq u_1$) and (ξ_1, ξ_2) guarantees in $A^2(u_1, u_2)$ a payoff $\geq u_2$ (i.e., $6\xi_2 + \beta u_1 \geq u_2$, $3\xi_1 + 2\xi_2 + \beta u_2 \geq u_2$, and $6\xi_1 + \beta u_1 \geq u_2$), consists of a list of optimal stationary strategies of player I and the values v_1 and v_2 of the discounted stochastic games. Therefore, the stationary optimal strategies of player I can be found among optimal solutions of the linear programming problem

$$
\begin{aligned}
&\max\ v_1 + v_2 \\
&\text{subject to} \\
&x_1 + 5x_2 + 0x_3 + \beta v_1 \geq v_1 \\
&2x_1 + 0x_2 + 4x_3 + \beta v_2 \geq v_1 \\
&0\xi_1 + 6\xi_2 + \beta v_1 \geq v_2 \\
&3\xi_1 + 2\xi_2 + \beta v_2 \geq v_2 \\
&6\xi_1 + 0\xi_2 + \beta v_1 \geq v_2 \\
&x_1 + x_2 + x_3 = 1 \\
&\xi_1 + \xi_2 = 1 \\
&x_1, x_2, x_3, \xi_1, \xi_2 \geq 0.
\end{aligned}
$$

In solving a discounted stochastic game, we can always assume that the immediate rewards are positive. Thus we can also assume that the above LP has an optimal solution v_1, v_2 bounded by $\frac{C}{(1-\beta)}$ where C is the maximum immediate payoff over all states. Indeed, the dual to the above LP can be used to construct an optimal stationary strategy $\{(y_1, y_2), (\eta_1, \eta_2, \eta_3)\}$ for player II.

In general terms, the primal in player-II-control games is to find an optimal solution to the LP

$$
\begin{aligned}
&\max \textstyle\sum_t v(t) \\
&\text{subject to} \\
&-\textstyle\sum_i r(s,i,j) f_i(s) - \beta \sum_t q(t \mid s,j) v(t) + v(s) \leq 0 \quad \forall\, j, s \\
&\textstyle\sum_i f_i(s) = 1 \quad \forall\, s \\
&f_i(s) \geq 0 \quad \forall\, , s, i \\
&v(s) \quad \text{arbitrary.}
\end{aligned}
$$

The dual LP is given by

$$
\begin{aligned}
&\min \textstyle\sum_s \theta(s) \\
&\text{subject to} \\
&-\textstyle\sum_j r(s,i,j) y_j(s) + \theta(s) \geq 0 \quad \forall\ i, s \quad (\text{corresponding to variable } f_i(s)) \\
&\textstyle\sum_s \sum_j \{\delta(t/s) - \beta q(t \mid s,j)\} y_j(s) = 1 \quad \forall\,(\text{states } t \text{ corresponding to } v(t)) \\
&y_j(s) \geq 0, \quad \theta(s) \quad \text{arbitrary.}
\end{aligned}
$$

(Here $\delta(t/s)$ is the Kronecker delta.)

From the above we see that $\sum_j y_j(t) > 0 \quad \forall\ t, \text{feasible}\{y_j(s)\}$. Normalizing them will give a stationary strategy for player II. By complementary slackness we can conclude that for an optimal f° for player I, and the stationary g° induced by normalizing an optimal $(y^\circ(s), \theta^\circ(s))$ of the dual, we get $r(f, g^\circ) \leq \theta^\circ(s)/\sum_j y_j(s)$ for all stationary f with equality at

$f = f^\circ$. Thus $r(f, g^\circ) \leq r(f^\circ, g^\circ)$. Premultiplying by the nonnegative matrix $(I - \beta Q(g^\circ))^{-1}$ we get the β-discounted payoff $\phi_\beta(f, g^\circ) \leq \phi_\beta(f^\circ, g^\circ)$. Thus g° is optimal for player II.

In a general two-person zero-sum stochastic game any optimal stationary strategy of player I will depend on the discount factor β. However, for every zero-sum two-person single-controller stochastic game there is a stationary strategy g° for player II (controlling player), which remains optimal for all β sufficiently close to 1. Such a strategy is called a uniform optimal strategy.

Example 2 *Consider the stochastic game with two states given by*

$$A^1 = \begin{array}{c} \begin{bmatrix} 3 & \quad\quad 4 \\ 7 & \quad\quad 0 \end{bmatrix} \\ \begin{matrix} \downarrow & \quad\quad\quad \downarrow \\ 1 & \quad\quad\quad 2 \end{matrix} \end{array} \quad A^2 = \begin{array}{c} \begin{bmatrix} 4 & \quad\quad 1 \\ 0 & \quad\quad 5 \end{bmatrix} \\ \begin{matrix} \downarrow & \quad\quad\quad \downarrow \\ 1 & \quad\quad\quad 2 \end{matrix} \end{array}$$

In the above stochastic game, player II has $(\frac{1}{2}, \frac{1}{2}), (\frac{1}{2}, \frac{1}{2}))$ as the uniform optimal stationary strategy for all β close to 1. However, the unique optimal strategy for player I is $f^\circ(1) = (\frac{7+\beta}{8}, \frac{1-\beta}{8}), f^\circ(2) = (\frac{5-\beta}{8}, \frac{3+\beta}{8})$.

Remark. Solving efficiently for the uniform optimal strategy for player II is still unresolved.

However, for undiscounted single-controller games with Cesaro average payoffs, just the existence of uniform optimal strategies for the controller helps one to solve the problem by a single linear program. This reduction is closely related to an algorithm by Hordijk and Kallenberg [19] for Markov decision processes which in turn is based on a sharp estimate of the Cesaro payoff for MDP via discounted payoff.

Theorem 3 (Blackwell [3]) *Consider an MDP with rewards $r(s, i)$, transitions $p(t/s, i)$ where $i \in A(s)$, the finite action space at state s. Given a stationary policy f, let $Q(f)$ be the Cesaro limit of the stationary transition matrix $P(f) = (p(t/s, f)$. Then the discounted payoff $\phi_\beta(f)$ using f and the Cesaro payoff $\phi(f)$ satisfies*

$$\phi_\beta(f) = \frac{\phi(f)}{1-\beta} + u(f) + e(f, \beta),$$

where $e(f, \beta) \to 0$ as $\beta \uparrow 1$.

One can exploit the above to develop an LP algorithm to solve for undiscounted single-controller games.

Theorem 4 (Hordijk and Kallenberg [20], Vrieze [51]) *Consider a player-II-control stochastic game. Then an optimal (f, ϕ, u) to the following dual* *linear programming problems can be used to find the value and optimal stationary strategies. Any optimal ϕ is the undiscounted value. Any optimal f is an optimal stationary strategy for player I. The LP and its dual are given by*

$$
\begin{aligned}
&\max \textstyle\sum_t \phi(t)\\
&\textit{subject to}\\
&\phi(s) - \textstyle\sum_t q(t \mid s,j)\phi(t) \le 0\\
&\textit{(the associated dual variables are } x_j(s) \quad \forall s,j)\\
&\phi(s) + u(s) - \textstyle\sum_i r(s,i,j) f_i(s) - q(t \mid s,i,j)u(t) \le 0 \quad \forall s,j\\
&\textstyle\sum_i f_i(s) = 1 \quad \forall s\\
&f_i(s) \ge 0, \;\; \phi(s), u(s) \textit{ unrestricted } \forall i,s.
\end{aligned}
$$

$$
\begin{aligned}
&\min \textstyle\sum_s \theta(s)\\
&\textit{subject to}\\
&\textstyle\sum_s \sum_j x_j(s)[\delta(t/s) - q(t \mid s,j)] + \sum_s \sum_j \delta(t/s) y_j(s) = 1 \;\; \forall t\\
&\textstyle\sum_s \sum_j y_j(s)[\delta(t/s) - q(t \mid s,j)] = 0 \;\; \forall t\\
&-\textstyle\sum_j r(s,i,j) y_j(s) + \theta(s) \ge 0 \;\; \forall i,s\\
&x_j(s), y_j(s) \ge 0, \;\; \theta(s) \quad \textit{unrestricted} \quad \forall\, s,j.
\end{aligned}
$$

At an optimal solution for the dual problems let $\sum_j y_j(t) = y_.(t)$, $\sum_j x_j(t) = x_.(t)$. From the above inequalities we have for each t either $y_.(t) > 0$ or $x_.(t) > 0$. We first normalize the vector $(y_1(t), y_2(t), \ldots)$ to get a mixed strategy for each state t. In case $y_.(t) = 0$, we have $x_.(t) > 0$. Normalizing the vector $(x_1(t), x_2(t), \ldots)$, we get a mixed strategy at state t. Such a choice gives an optimal stationary strategy g° for player II. Thus we can solve zero-sum single-controller undiscounted games by a single linear program.

6. Single-Controller Non-Zero-Sum Two-Person Stochastic Games

Fink [14] and independently Takahashi [47] first extended the theorem of Shapley for n-person non-zero-sum discounted stochastic games. They showed that stationary Nash equilibrium strategies exist for these games.

When the transition is controlled by a single player, Parthasarathy and Raghavan [35] showed that these games admit a Nash equilibrium in stationary strategies with orderfield property. They also showed that undiscounted single-controller stochastic games have stationary Nash equilibria and they too possess orderfield property.

Nowak and Raghavan [34] proved the following theorem which contains a recipe for a finite-step algorithm.

Theorem 5 *In a player-II-control game let $f_1, f_2, .., f_m$ be an enumeration of all pure stationary strategies for player I and let $g_1, g_2, .., g_n$ be an enumeration of all pure stationary strategies for player II. Let $(A = \sum_s A(s), \sum_s B(s))$ be an $m \times n$ bimatrix game where*

$$A(s) = (r_1(s, f_i(s), g_j(s)) \qquad B(s) = (\phi_\beta(f_i, g_j)(s).$$

Let (ξ^*, η^*) be a mixed strategy Nash equilibrium point to the bimatrix game (A, B). Then
(a) $f^* = \sum_i \xi_i^* f_i$ and $g^* = \sum_j \eta_j^* g_j$ constitute a Nash equilibrium pair for the discounted game.
(b) For each state s, the equilibrium payoff for player II in the stochastic game is the same as the equilibrium payoff for player II in the bimatrix games $(A(s), B(s))$.
(c) In the case of the undiscounted irreducible player-II-control games, if we replace the above matrix B by the matrix $C = (\sum_s \phi^2(f_i, g_j)$, then any equilibrium point (ξ^*, η^*) of the bimatrix game (A, C) induces as in the discounted case a stationary equilibrium point (f^*, g^*). Further, for the irreducible case, Nash equilibrium payoffs are independent of the starting state.

We will use an example from [34] to illustrate the above algorithm. The stochastic game has three states and each player has two actions at each state. Here the discount factor $\beta = .8$. We take the entries as immediate *penalties*. The players want to minimize their expected discounted penalties.

Example 3

$$\begin{array}{ccc} s=1 & s=2 & s=3 \\ \begin{bmatrix} (6,3) & (0,8) \\ (0,5) & (7,1) \end{bmatrix} & \begin{bmatrix} (0,10) & (9,2) \\ (7,5) & (0,8) \end{bmatrix} & \begin{bmatrix} (3,0) & (0,5) \\ (0,4) & (4,0) \end{bmatrix} \\ \begin{array}{cc} \downarrow & \downarrow \\ 1 & 2 \end{array} & \begin{array}{cc} \downarrow & \downarrow \\ 2 & 3 \end{array} & \begin{array}{cc} \downarrow & \downarrow \\ 3 & 1 \end{array} \end{array}.$$

There will be eight pure stationary strategies for each player that could be lexicographically enumerated as $(111), (112), \ldots, (222)$ with the understanding that (ijk) corresponds to choosing the i-th row in state 1, the j-th row in state 2, and the k-th row in state 3. Similarly, one can define pure stationary strategies for player II. Using the Lemke-Howson algorithm [25] we can get the Nash equilibrium point

$$\begin{aligned} \xi^* &= (0, 0, \frac{192}{1613}, \frac{408}{1613}, 0, 0, 0, \frac{1013}{1613}) \\ \eta^* &= (0, 0, \frac{10}{91}, \frac{39}{91}, 0, 0, \frac{42}{91}, 0). \end{aligned}$$

The stationary strategies f^*, g^* are obtained from ξ^*, η^* by taking the marginal sums. For example, $\xi^*_{112} + \xi^*_{122} + \xi^*_{212} + \xi^*_{222} = f^*_2(3)$. The stationary strategies are given by

$$f^* = \begin{cases} (\frac{600}{1613}, \frac{1013}{1613}) & \text{for s=1} \\ (0,1) & \text{for s=2} \\ (\frac{192}{1613}, \frac{1421}{1613} & \text{for s=3} \end{cases}$$

and

$$g^* = \begin{cases} (\frac{7}{13}, \frac{6}{13}) & \text{for s=1} \\ (0,1) & \text{for s=2} \\ (\frac{4}{7}, \frac{3}{7}) & \text{for s=3.} \end{cases}$$

Remark. Even though the problem is reduced to solving for a Nash equilibrium point of a bimatrix game, the full enumeration of the entire matrix is undesirable. What is desirable is to solve the game via some pivoting algorithm.

7. Discounted Single-Controller Game via Lemke's LCP Algorithm

The linear complementarity problem can be stated as follows. Given a vector $q \in R^n$ and a matrix $M \in R^{n \times n}$, find a vector z such that:

$$w = q + Mz \tag{5}$$

$$z, w \geq 0 \tag{6}$$

$$z^T w = 0. \tag{7}$$

The above system is usually denoted by LCP(q, M). A pair (w, z) of vectors satisfying the above system of inequalities is called a solution to the LCP(q, M). For the literature on Lemke's algorithm to solve LCP(q, M) see [24]. For a recent book on the linear complementarity problem see [8]. It can be shown that the LCP is a generalization of the well-known LP (linear program). In the historic work of Lemke [24], a simplex-like pivoting algorithm to process LCP's is given. Unfortunately, the algorithm does not always find a solution to a given LCP. There are, however, certain classes of matrices M for which Lemke's algorithm will process LCP(q, M).

8. LCP for Discounted Single-Controller Games

For discounted non-zero-sum two-person games where player II alone controls the transitions, [32] gave one such linear complementarity reduction.

The following lemma facilitates such a reduction of the original problem to a linear complementarity problem.

Lemma 1 *Consider the following auxiliary stochastic game with N states controlled by player II with immediate costs $r_\alpha(s,i,j)\quad \alpha=1,2$ to the two players. Player I pays just the immediate cost for the first day and no more, while player II pays the usual β-discounted cost over the infinite horizon. Any Nash equilibrium of the game with payoffs in stationary strategies f,g for players I and II given by*

$$A(s)=r_1(f,g)(s),\quad B(s)=\phi_\beta(f,g)(s),\quad s=1,\ldots,N$$

is also a stationary equilibrium for the single-controller stochastic game.

Proof. For a proof see Lemma 2.3 of [35].

Thus, solving single-controller games is reduced to solving for equilibria of the above games. We are ready to recast this problem as a linear complementarity problem.

Theorem 6 *The pairs $(f^\circ(s),g^\circ(s))$ and $(v_1(s),v_2(s))$ form Nash equilibrium strategies with corresponding equilibrium costs for players I and II iff they satisfy the following system of equations:*

$$u_j(s)-\sum_i r_2(s,i,j)f_i(s)-\sum_t \beta q(t\mid s,j)\phi_\beta(t)+\phi_\beta(s)=0\ \forall\, j,\ \text{and}\ \forall s \tag{8}$$

$$w_i(s)-\sum_j r_1(s,i,j)g_j(s)+v_1(s)=0\ \ \forall\, s\ \text{and}\ \forall\, i \tag{9}$$

$$\theta(s)-\sum_i f_i(s)=-1,\ \ \forall\, s \tag{10}$$

$$\tau(s)-\sum_j g_j(s)=-1\ \ \forall\, s \tag{11}$$

$$w_i(s),f_i(s),u_j(s),g_j(s)\geq 0\ \forall\, i,j,s \tag{12}$$

$$w_i(s).f_i(s)=0,u_j(s).g_j(s)=0,\ \forall i,j,s. \tag{13}$$

$$\theta(s).v_1(s)=0,\tau(s).\phi_\beta(s)=0,\ \forall\, s, \tag{14}$$

where $\phi_\beta(s)$ is the β-discounted equilibrium cost for player II.

For each $f_i(s),g_j(s)$, the complementary slack variables are respectively $w_i(s)$ and $u_j(s)$. Similarly for variables $v_1(s),\phi_\beta(s)$, the corresponding complementary slack variables are $\theta(s)$ and $\tau(s)$. Suppressing s, the above equations can be cast as the LCP

$$\begin{bmatrix} \mathbf{0} & -\mathbf{r_2} & \mathbf{0} & \mathcal{Q} \\ -\mathbf{r_1} & \mathbf{0} & \mathbf{1} & \mathbf{0} \\ -\mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -\mathbf{1} & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} g \\ f \\ v_1 \\ \phi_\beta \end{bmatrix} + \begin{bmatrix} \mathbf{u} \\ \mathbf{w} \\ \tau \\ \theta \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ -\mathbf{1} \\ -\mathbf{1} \end{bmatrix}$$

$$g.u = 0, f.w = 0, v_1.\theta = 0, \phi_\beta.\tau = 0.$$

Lemke's algorithm when applied to the above LCP may terminate in a *secondary ray.* See [8] for its definitions and more details. However, adding the square matrix Λ with all entries unity to the immediate payoff part

$$\begin{bmatrix} 0 & -r_2 \\ r_1 & 0 \end{bmatrix}$$

Mohan et al. [32] showed that the LCP$(q, \bar{M})$ belongs to class $\mathcal{L}$ of Eaves [9], with the following property.

Theorem 7 *LCP$(d, \bar{M})$ has a unique solution when $d > 0$, or when $d = 0$.*

While for this class of matrices Lemke's algorithm with any positive covering vector will compute a unique solution to LCP$(q, \bar{M})$, in our case the q vector is not > 0. However, by Theorem 3.5 of Garcia [15], Lemke's algorithm will process this LCP and hence compute a Nash equilibrium point to the auxiliary game.

9. Non-Zero-Sum Undiscounted Single-Controller Stochastic Games

Let $S = \{1, 2, \ldots, s\}$ be the states and let $A(t) = \{1, 2, \ldots, a_t\}$, $B(t) = \{1, 2, \ldots, b_t\}$ be action spaces at state t for players I and II respectively. Let $r_1(t, a, b), r_2(t, a, b)$ be immediate costs to players I and II at state t when $a \in A(t), b \in B(t)$ are their actions. For any generic states i, j we will denote by $p[i, a, b]_j$ the conditional probability of the game moving from state i to state j when a, b are actions chosen by players I and II at state i. If the game is controlled by player II, then $p[i, a, b]_j = p[i, b]_j$.

In this section we will consider player-II-control games and show that under the limiting average cost criterion, these games can also be solved by a single Lemke-processible LCP.

Consider the induced undiscounted MDP where player I fixes his strategy to a stationary strategy π. When player II chooses action b in state t, the immediate cost incurred is given by $\widetilde{r}_2(t, b) = \sum_{a \in A(t)} \pi_a(t) r_2(t, a, b)$. The transitions of the MDP are the same as those of the original game since π has no influence over them. Using the LP formulation for limiting average MDP's, player II's best reply to π comes as a solution to the following pair of dual LP's:

Primal

$$\text{Maximize} \quad \frac{1}{s} \sum_i \phi(i) \tag{15}$$

subject to

$$\phi(i) - \sum_t p[i,b]_j \phi(j) \le 0 \quad \forall\, i \in S, \ \ \forall\, b \in B(i) \tag{16}$$

$$\phi(i) + u(i) - \sum_j p[i,b]_j u(j) \le \tilde{r}_2(i,b) \quad \forall\, i \in S, \forall\, b \in B(i) \tag{17}$$

$$\phi(i), u(i) \text{ unrestricted} \quad \forall\, i \in S. \tag{18}$$

Dual

$$\text{Minimize} \sum_{i=1}^{s} \sum_{b \in B(i)} \tilde{r}_2(i,b) x_{ib} \tag{19}$$

subject to

$$\sum_{b \in B(j)} x_{jb} + \sum_{b \in B(j)} y_{jb} - \sum_{i=1}^{s} \sum_{b \in B(i)} p[i,b]_j y_{ib} = \frac{1}{s} \ \forall\, j \in S \tag{20}$$

$$\sum_{b \in B(j)} x_{jb} - \sum_{i=1}^{s} \sum_{b \in B(i)} p[i,b]_j x_{ib} = 0 \quad \forall\, j \ \in S \tag{21}$$

$$x_{ib} \ge 0, y_{ib} \ge 0 \quad \forall\, i \in S, \ \forall\, b \in B(i). \tag{22}$$

In this setup we have y_{ib} and x_{ib} complementary to the slack variables of (16) and (17) respectively. Conversely, we have $\phi(j)$ and $u(j)$ complementary slack variables to (20) and (21) respectively. Suppose we have an optimal solution for both programs, say (ϕ^*, u^*, x^*, y^*). Then player II's optimal strategy ρ^* (against π) would be extracted as follows.

$$\rho^*(i,b) = \begin{cases} x^*_{ib} / \sum_{c \in B(i)} x^*_{ic} & \text{when } \sum_{c \in B(i)} x^*_{ic} > 0 \\ y^*_{ib} / \sum_{c \in B(i)} y^*_{ic} & \text{otherwise.} \end{cases}$$

One can verify, using (20), that ρ^* is well defined. Also we have $\phi^*(i) = \phi^2(\pi, \rho^*)$. A key property of this pair of LP's is that those states $i \in S$ for which $\sum_{c \in B(i)} x^*_{ic} = 0$ are *transient* in the Markov chain induced by ρ^*.

Next we make some adjustments to (19)-(22) so that they can be put into LCP form.

- Replace $= 0$ on the right-hand side of (21) by ≥ 0.
  This is still equivalent to the original LP.

- Tighten the constraints on $u(i)$'s.
  Namely, the $u(i)$'s can be restricted to being nonnegative. (Note that for any constant θ changing from $u(i)$ to $u(i)+\theta, \forall\ i$ leaves the objective function of primal LP unchanged.) The complementary slackness of the new (9.7) will be untouched.
- We can assume that all immediate costs are positive.
  This has the effect that any optimal $\phi^* > 0$ in the primal LP. We can replace (20) by the weaker inequality

$$\text{s.t.} \sum_{b\in B(j)} x_{jb} + \sum_{b\in B(j)} y_{jb} - \sum_{i=1}^{s} \sum_{b\in B(i)} p[i,b]_j y_{ib} \geq \frac{1}{s} \quad \forall\ j \in S. \qquad (23)$$

- One additional adjustment is to replace the right-hand side of (17) with

$$\phi(i)+u(i)-\sum_{t} p[i,b]_j u(j) \leq \tilde{r}_2(i,b)+\sum_{i=1}^{s} \sum_{b\in B(i)} x_{ib} \quad \forall\ s \in S, \forall\ j\ . \qquad (24)$$

  Its effect is to introduce complementary variables on the one side, while at an optimal solution all it does is to add 1 to all coordinates of an optimal ϕ.

Next we take care of the player I side with complementary inequalities. Consider the following set of inequalities:

$$\sum_{b\in B(i)} r_1(i,a,b)\rho^*(i,b) \geq v(i)\ \forall\ \ i \in S, \forall\ \ a \in A(i) \qquad (25)$$

$$\sum_{a\in A(i)} \tilde{z}_{ia} \geq 1\ \forall\ i \in S \qquad (26)$$

$$\tilde{z}_{ia},\ v(i) \geq 0\ \forall\ i \in S, \forall\ a \in A(i) \qquad (27)$$

along with the complementary conditions

$$\tilde{z}_{ia}[\sum_{b\in B(i)} r_1(i,a,b)\rho^*(i,b) - v(i)] = 0\ \ \forall\ i \in S, \forall\ a \in A(i) \qquad (28)$$

$$v(i)[\sum_{a\in A(i)} \tilde{z}_{ia} - 1] = 0\ \ \forall\ i \in S. \qquad (29)$$

We can normalize $\tilde{z}_{ia}$'s to get a stationary strategy $\tilde{\pi}$ for player I if in the above equation we have a complementary solution with $v(i) > 0\ \ \forall i$.

This is easily achieved by replacing (25) and (28) with

$$\sum_{i=1}^{s} \sum_{a \in A(i)} \tilde{z}_{ia} + \sum_{b \in B(i)} r_1(i,a,b)\rho^*(i,b) \geq v(i) \ \forall\, i \in S, \forall\, a \in A(i). \tag{30}$$

$$\tilde{z}_{ia}[\sum_{i=1}^{s} \sum_{a \in A(i)} \tilde{z}_{ia} + \sum_{b \in B(i)} r_1(i,a,b)\rho^*(i,b) - v(i)] = 0 \ \forall\, i \in S, \forall\, a \in A(i). \tag{31}$$

We begin by writing equations and inequalities mnemonically where the last ones below are the complementarity conditions.

$w^1_{ia} = \pi_{..} + (r_1 x)_{ia} - v(i) \quad \forall\, i, a$
$w^2_{ib} = (P\phi)_{ib} - \phi(i) \quad \forall\, i, b$
$w^3_{ib} = x_{..} + (\pi.r_2)_{ib} + (Pu)_{ib} - u(i) - \phi(i) \quad \forall\, i, b$
$w^4_j = x_{.j} - (xP)_j$
$w^5_j = -\frac{1}{s} + x_{.j} + y_{.j} - (yP)_j$
$w^6_i = -1 + \pi_{i.}$
All variables are nonnegative
$w^1 \perp \pi;\ w^2 \perp y;\ w^3 \perp x;\ w^4 \perp u;\ w^5 \perp \phi;\ w^6 \perp v.$

Let $z = (\pi, y, x, \phi, u, v)$. The above LCP can be written as $w = Mz + q$ where the matrix M is a partitioned matrix of the type

$$M = \begin{bmatrix} \mathcal{R} & \mathcal{A} \\ -\mathcal{A}^T & 0 \end{bmatrix}$$

where

$$\mathcal{R} = \begin{bmatrix} \mathcal{C}_\infty & 0 & \mathcal{D} \\ 0 & 0 & 0 \\ \mathcal{E} & 0 & \mathcal{C}_\in \end{bmatrix} \text{ and } \mathcal{A} = \begin{bmatrix} 0 & 0 & \mathcal{F} \\ \mathcal{P}_\infty & 0 & 0 \\ \mathcal{G} & \mathcal{P}_\in & 0 \end{bmatrix}$$

where the three-way split is partitioned as $\pi|y|x$ for the rows of $\mathcal{R}$ as well as its columns. We define all entries of $\mathcal{C}_\infty$ and $\mathcal{C}_\in$ to be equal to 1. The (π_{ia}, x_{ib})-th entry of $\mathcal{D}$ is $r_1(i,a,b)$ for $\forall\, i \in S, \forall\, a \in A(i), \forall\, b \in B(i)$. The other entries of $\mathcal{D}$ are 0. The (x_{ib}, π_{ia})-th entry of $\mathcal{E}$ is $r_2(i,a,b)$ for $\forall\, i \in S, \forall\, a \in A(i), \forall\, b \in B(i)$. The other entries of $\mathcal{E}$ are 0. This completes the definition of $\mathcal{R}$.

The $(\pi_{ia}, v(i))$-th entry of $\mathcal{F}$ is -1 for $\forall\, i \in S, \forall\, a \in A(i)$. The rest of the entries of $\mathcal{F}$ are 0. If $i \neq j$ then the $(y_{ib}, \phi(j))$-th entry of $\mathcal{P}_\infty$ is given by $p[i,b]_j$. If $i = j$ then the $(y_{ib}, \phi(j))$-th entry is $p[i,b]_j - 1$. $\mathcal{P}_\infty$ and $\mathcal{P}_\in$ are actually identical. Formally we have that the $(x_{ib}, u(j))$-th entry of the former is the same as the $(y_{ib}, \phi(j))$-th entry of the latter. The $(x_{ib}, \phi(i))$-th entry of $\mathcal{G}$ is -1 for $\forall\, i \in S, \forall\, b \in B(i)$. The other entries of $\mathcal{G}$ are all 0.

To complete the construction of the LCP we need to define the vector q. Like z, q is also an $n \times 1$ vector. We set all the coordinates of q to 0 with the exception of the indices (u, v). Those coordinates of q in u will have value $-\frac{1}{s}$. The coordinates in v will have value -1.

Lemma 2 *Lemke's algorithm will provide a solution to LCP(q, M).*

Proof. It is easy to show show that LCP(q, M) is feasible. Observe

$$M + M^T = \begin{bmatrix} \mathcal{R} + \mathcal{R}^T & 0 \\ 0 & 0 \end{bmatrix}.$$

It is easy to check that the matrix M is copositive plus, that is:

$$\begin{aligned} & z^T M z \geq 0, \forall\, z \geq 0 \\ & z^T M z = 0 \Rightarrow (M + M^T) z = 0, \forall\, z \geq 0. \end{aligned}$$

Thus by a theorem of Lemke [24] (see also [9]), Lemke's algorithm will process the LCP and will terminate.

Indeed, the LCP solution vector $z^* = (\pi^*, y^*, x^*, \phi^*, u^*, v^*)$ supplies a stationary equilibrium strategy for the undiscounted single-controller stochastic games. We can use y^*, x^* to construct an equilibrium stationary strategy ρ^* for player II. We can use π^* to serve as the stationary equilibrium strategy for player I. We can use $\phi^* - 1$ to recover the equilibrium payoff to player II.

10. Discounted Stochastic Games of Perfect Information

Here we consider the special class of discounted stochastic games with *perfect information*. In perfect information games, at each state at most one player has more than one action to choose from his action set. If the player who has one action is the same one in all states then it is the classic Markovian Decision Process (MDP). One can solve the discounted MDP via Howard's policy improvement algorithm [21]. Our task here is to adapt the policy improvement algorithm of the discounted MDP to these games. The existence theorem for perfect information stochastic games imposes a strong combinatorial structure on them. This then serves as a motivation for our algorithm, which is an extension of the Howard–Blackwell [3] policy improvement algorithm for the discounted stochastic game.

Shapley [44] showed that under the discounted payoff criterion, perfect information stochastic games admit optimal pure stationary strategies, for both players. For a pair of pure stationary strategies (f, g) we define as usual $\phi_\beta(f, g)$ to be the vector of expected discounted payoffs, resulting from f and g. For every pair t, s of states we denote by $Q_{t,s}(f, g)$ the probability of transition from state t to state s given the stationary strategies f of player

1 and g of player 2. Since the immediate transition probability $Q_t(f,g) = (Q_{t,s}(f,g))_{s\in S}$ at each state t is determined by the action of at most one player for perfect information games, we can always write either $Q_t(f,g) = Q_t(f)$ or $Q_t(f,g) = Q_t(g)$ as the case may be. In case it is a state with exactly one action for each player, the transitions are given a priori for nature and so in such states t we could even suppress the dependence of Q_t on f or g. Likewise, we write $r(f,g)$ to be the vector indexed by the state space whose t-th component is $r(t,f(t),g(t))$. Just like $Q(f,g)$, the coordinate $r(t,f(t),g(t))$ of $r(f,g)$ does not depend on g (on f) whenever player II (player I) has a single action in state t.

For the discounted MDP there is the *policy improvement* algorithm of Howard [21], which can be used to determine optimal policies. This algorithm starts at an arbitrary policy f_0 and produces a sequence of improvements $f_1, f_2, \ldots, f_k$ until an optimal policy is reached. In the sequence of policies the corresponding values ϕ_β are strictly monotonic and therefore the algorithm must terminate (there are only a finite number of pure stationary policies). Extending the policy improvement algorithm of MDP's to stochastic games was initially attempted by Pollatschek and Avi-Itzhak [37]; however, they were only able to prove that their algorithm terminates for games with a stringent condition on the transitions and the discount factor [51], [49].

We rearrange the states so that player I has more than one action and player II has exactly one action in states $1, \ldots, t_1$ and player I has exactly one action and player II has more than one action in states $t_1+1, \ldots, t_1+t_2$. The rest of the states can likewise be dubbed as states of nature. When a strategy of one player is fixed, we are in a discounted MDP and it is enough to find the best pure stationary strategy among all pure stationary strategies. For a pair of pure stationary strategies (f,g) we write $[(f,g) = (f(1),g(1)),\ldots,(f(s),g(s)]$ where $(f(t),g(t))$ is the pair of actions chosen in state t under (f,g). For any state t at least one of $f(t)$ or $g(t)$ is 1. (The player is essentially a dummy for that state.) An *adjacent improvement of type I* is a new pair of pure stationary strategies (h,g) where:

1. h and f differ in exactly one state, namely there exists $\bar{t}, 1 \le \bar{t} \le t_1$, with $h(\bar{t}) \ne f(\bar{t})$ and $h(\tau) = f(\tau)$ for $\tau \ne \bar{t}$.
2. $\phi_\beta(h,g) \ge \phi_\beta(f,g)$ and $\phi_\beta(h,g)_t > \phi_\beta(f,g)_t$ for some $1 \le t \le s$.

The purpose of the second condition is clearly that player I is better off playing h than f against player II's g. The first condition is an adjacency condition required in our algorithm. It states that h differs from f in exactly one state. Of course we have the corresponding definition for *adjacent improvement of type II*, namely it is a pair (f,h) where:

1. $\exists t', t_1 + 1 \le t' \le t_1 + t_2$, with $g(t') \ne h(t')$ and $g(\tau) = h(\tau)$ for $\tau \ne t'$ such that
$\phi_\beta(f, h) \le \phi_\beta(f, g)$ and $\phi_\beta(f, h)_t < \phi_\beta(f, g)_t$ for some $1 \le t \le s$.

Notice that in both cases we require a *strict improvement* in ϕ_β value in some state. A pair of pure stationary strategies (f', g') will be called an *improvement* of (f, g) if it is a strict but adjacent improvement of either type I or type II. Note that in such a case we would have either $f' = f$ or $g' = g$ depending on the type of improvement.

In our algorithm we start with a pair of pure stationary strategies and generate a sequence of improvements via lexicographic search. That is, we start in state 1 and proceed as follows. We always look for an adjacent improvement of type I for player I. If such an improvement doesn't exist then we search for an improvement of type II (of course, we will not find them in states where player II is a dummy). Now if neither exists then the search moves to state 2 and we repeat the procedure. After an improvement of either type is found, we move to the new pair and begin searching for improvements back from state 1 again. We will prove that such a procedure must terminate in an optimal pair (f^*, g^*).

Algorithm 4

1. Choose arbitrarily a pair of pure stationary strategies (f^0, g^0) (e.g., $f^0(t) = g^0(t) = 1$ for $t = 1, \dots, s$) and set $\alpha = 0$.
2. Search lexicographically for an improvement $(f^{\alpha+1}, g^{\alpha+1})$ of (f^α, g^α) *always looking first for player I* and then only for player II. There are three cases:

   Case 1: An improvement f for player I is found. In this case let $(f^{\alpha+1}, g^{\alpha+1}) = (f, g^\alpha)$ and $\alpha = \alpha + 1$, and repeat step 2.

   Case 2: There are no improvements for player I, but there is an improvement g for player II. In this case let $(f^{\alpha+1}, g^{\alpha+1}) = (f^\alpha, g)$ and $\alpha = \alpha + 1$, and repeat step 2.

   Case 3: There are no improvements. Go to step 3.

3. The pair $(f^*, g^*) = (f^\alpha, g^\alpha)$ is an optimal pure stationary strategy pair for the two players.

Remark. The claim that a lexicographically locally optimal pair is optimal for the stochastic game does not follow directly from local optimality or from MDP. It depends on some intrinsic properties of stochastic games of perfect information and we develop them now.

Remark. In an ordinary matrix game $A = (a_{ij})$ with value v, if $a_{pq} = v$, it does not mean p, q are good pure strategies.

Curiously, however, for the case of stochastic games with perfect information, we have the following.

Lemma 3 *In a zero-sum perfect information stochastic game* Γ*, a pair of pure stationary strategies* (f°, g°) *is optimal if and only if* $\phi_\beta(f^\circ, g^\circ) = \phi_\beta(\Gamma)$*, the value of the stochastic game.*

For what follows we require some notation. Let $t \in S$ be a fixed state. For any $X \subset A(t)$ we let Γ^t_X be the subgame in which only the actions in X are allowed in state t. The corresponding pure stationary strategy sets will be denoted by F^t_X and G^t_X. For the original game Γ we write F and G for the pure stationary strategy sets of players I and II respectively.

When one of the players, say player I, restricts his actions in state t to only those in the set X, while player II has no restrictions at all in any state, we reach the subgame Γ^t_X. The pure stationary strategy space G^t_X for player II for this subgame is the same as G in the original game because player II's strategy is not constrained.

Lemma 4 *For* $t \in S$*,* $X \subset A(t)$*,* $Y \subset A(t)$*,* $X \cap Y = \emptyset$ *we have for each starting state* k*,* $\phi(\Gamma^t_{X\cup Y})(k) = \max\{\phi_\beta(\Gamma^t_X)(k), \phi_\beta(\Gamma^t_Y)(k)\}$*. In fact, as vectors, either* $\phi_\beta(\Gamma^t_X) \geq \phi_\beta(\Gamma^t_Y)$ *or* $\phi_\beta(\Gamma^t_X) \leq \phi_\beta(\Gamma^t_Y)$*.*

An obvious player II analog of the lemma exists using $B(t)$ instead of $A(t)$.

Theorem 8 *The strategy pairs* $(f^\alpha, g^\alpha), \alpha = 0, 1 \ldots$ *obtained at step 2 along the algorithmic path never cycle and hence the algorithm must terminate. The terminal pair* (f^*, g^*) *is locally optimal, in the sense that no adjacent improvement is possible for either player. It is also a globally optimal strategy pair for the stochastic game.*

Remark. Unlike in the policy improvement algorithm of MDP, in our case we cannot expect any monotonicity property of the payoffs along the algorithmic path.

Proof. A proof can be given by an induction on the total number n of actions available for the two players in all states, that is, $n = \sum_{i=1}^{s}(a_i + b_i)$. For details see [41].

11. Undiscounted Simple Stochastic Games

From the point of view of complexity theory, perfect information stochastic games in *undiscounted and total payoffs* have been of interest to computer scientists [7], [17], [27], [28], [53].

Condon [7] studied the so-called *simple stochastic games* (SSG). These are special classes of stochastic games called *recursive games of perfect information* [10]. In recursive games the immediate payoff is 0 at all non-absorbing states. In simple stochastic games, one further assumes that the immediate payoff is 1 just at one absorbing state called the *I-sink* and the immediate payoff is 0 at exactly one absorbing state called the *II-sink*. Depending on the maximizer, or minimizer, or nature (who has two actions in that state), they are called *max*, *min* and *average* states. In average states a coin is tossed to choose one of two available actions. The maximizer prefers the game to get absorbed into I-sink. The minimizer prefers the game to get absorbed into II-sink. Also in these games, the law of motion is exact, in the sense that from, say, a max state, depending on the action of the player, the game moves to another state in unit time with probability 1.

Condon gives a polynomial time value iteration algorithm when the game has exact law of motion and there are no average states. The algorithm is quite simple and intuitive. If from any max state t there is an action taking the game to a state with value 1, we define the value at state t as $v(t) = 1$. If all actions at a max state end in a state with value 0, we define $v(t) = 0$. Similar definitions apply at min states. If from an average state τ one reaches the two states t or $t^{'}$ with values $v(t), v(t^{'})$ respectively, the value satisfies $v(\tau) = \frac{1}{2}v(t) + \frac{1}{2}v(t^{'})$. Since the value is known at the two sinks, the value is determined at all other states by backward induction. She also presents an algorithm for the case when the game terminates with probability one into one of the two sinks. For simple stochastic games a policy iteration-type algorithm is given by Ludwig [27]. He also assumes that for every strategy pair the termination occurs with probability 1 in one of the two sinks. Ludwig's algorithm can be described as follows.

1. Given a simple stochastic game with N states, start with an arbitrary max state s and any pure stationary σ for player I. Consider the substochastic game where player I will follow $\sigma(s)$ when s is reached.
2. Recursively solve for an optimal strategy $\sigma^{'}$ of the substochastic game for player I and extend this to a strategy of the original game by setting $\sigma^{'}(s) = \sigma(s)$.
3. Solve for an optimal $\tau^{'}$ for the MDP for player II (minimizer) with player I's strategy fixed at $\sigma^{'}$.
4. If $\sigma^{'}, \tau^{'}$ is optimal, then stop. Otherwise, change the alternative at state s to the second available alternative for player I and set $\sigma(t) = \sigma^{'}(t)$ for $t \neq s$. Go to step 1 again.

Remark. The problem of solving efficiently for undiscounted value and optimal stationary strategies for zero-sum two-person stochastic games of perfect information is open. So are ARAT undiscounted games [43].

We are unable to prove that our algorithm will not cycle in the undiscounted case. Once it can be solved, one can solve for Nash equilibria for non-zero-sum perfect information stochastic games. Though they may not be stationary, they consist of a pair of stationary strategies, namely a stationary part with a stationary threat [48].

Remark. Switching control games are still open for efficient algorithms even for discounted or undiscounted zero-sum games.

References

1. Bardi, M., Raghavan, T.E.S. and Parthasarathy, T. (1999) *Stochastic and Differential Games, Theory and Numerical Methods*, Bikhauser, Berlin.
2. Bewley, T. and Kohlberg, E. (1978) On stochastic games with stationary optimal strategies, *Mathematics of Operations Research* **2**, 104–125.
3. Blackwell, D. (1962) Discrete dynamic programming, *Annals of Mathematical Statistics* **33**, 719–726.
4. Blackwell, D. (1969) Infinite G_δ games with imperfect information, *Zastosowania Matematyki* **10**, 99–101.
5. Blackwell, D. (1989) Operator solution of infinite G_δ games of imperfect information, in T.W. Anderson, K. Athreya and D.L. Iglehart (eds.), *Probability, Statistics, and Mathematics: Papers in Honor of Samuel Karlin*, Academic Press, New York, pp. 83–87.
6. Breton, M. (1987) Equilibre pour des jeux sequential, Ph.D. thesis, University of Montreal.
7. Condon, A. (1992) The complexity of stochastic games, *Information and Computing* **96**, 203–224.
8. Cottle, R.W., Pang, J.S. and Stone, R.E. (1992) *The Linear Complementary Problem*, Academic Press, Boston.
9. Eaves, B. (1971) Linear complementarity problem, *Management Science* **17**, 612–634.
10. Everett, H. (1957) Recursive games, in M. Dresher, A. W. Tucker and P. Wolfe (eds.), *Contributions to the Theory of Games, Vol. III*, Annals of Mathematics Studies 39, Princeton University Press, Princeton, NJ, pp. 47–78.
11. Filar, J.A. and Raghavan, T.E.S. (1984) A matrix game solution of the single-controller stochastic game, *Mathematics of Operations Research* **9**, 356–362.
12. Filar, J.A. and Schultz, T.A. (1987) Bilinear programming and structured stochastic games, *Journal of Optimization Theory and Applications* **53**, 85–104.
13. Filar, J.A. and Vrieze, O.J. (1996) *Competitive Markov Decision Processes*, Springer-Verlag, Berlin.
14. Fink, A.M. (1964) Equilibrium points of stochastic non-cooperative games, *Journal of Science of the Hiroshima University, Series A-I* **28**, 89–93.
15. Garcia, C. B. (1973) Some classes of matrices in linear complementarity theory, *Mathematical Programming* **5**, 299–310.
16. Gillette, D. (1957) Stochastic games with zero stop probabilities, in M. Dresher, A.W. Tucker and P. Wolfe (eds.), *Contributions to the Theory of Games, Vol. III*, Annals of Mathematics Studies 39, Princeton University Press, Princeton, NJ, pp. 179–188.
17. Gurwich, V.A., Karzanov, A.V. and Khachiyan, L.G. (1988) Cyclic games and an algorithm to find minimax cycle means in directed graphs, *USSR Computational Mathematics and Mathematical Physics* **28**, 85–91.
18. Hoffman, A.J. and Karp, R.M. (1966) On non-terminating stochastic games, *Management Science* **12**, 359–370.
19. Hordijk, A. and Kallenberg, L.C.M. (1979) Linear programming and Markovian decision chains, *Management Science* **25**, 352–362.
20. Hordijk, A. and Kallenberg, L.C.M. (1984) Linear programming and Markov games, in O. Moeschlin and D. Pallaschke (eds.), *Game Theory and Mathematical Economics*, North-Holland, Amsterdam, pp. 307–319.
21. Howard, R.A. (1960) *Dynamic Programming and Markov Processes*, Wiley, New York.

22. Kallenberg, L.C.M. (1983) Linear programming and finite Markovian control problems, Mathematical Centre Tract 148, Centre for Mathematics and Computer Science, Amsterdam.
23. Krishna, V. and Sjöstrom, T. (1998) On the convergence of fictitious play, *Mathematics of Operations Research* **23**, 479–511.
24. Lemke, C. E. (1964) Bimatrix equilibrium points and mathematical programming, *Management Science* **11**, 681–689.
25. Lemke, C.E. and Howson, Jr. J.J. (1964) Equilibrium points of bimatrix games, *Journal of the Society of Industrial and Applied Mathematics* **12**, 413–423.
26. Liggett, T.M. and Lippman, S. A. (1969) Stochastic games with perfect information and time average payoff, *SIAM Review* **11**, 604–607.
27. Ludwig, W. (1995) A subexponential randomized algorithm for the simple stochastic game problem, *Information and Computation* **117**, 151–155.
28. Melekopoglou, M. and Condon, A. (1994) On the complexity of the policy improvement algorithm for Markov decision processes, *ORSA Journal on Computing* **6**, 188–192.
29. Mertens, J.-F. and Neyman, A. (1981) Stochastic games, *International Journal of Game Theory* **10**, 53–56.
30. Mertens, J.-F. and Parthasarathy, T. (1987) Equilibria for discounted stochastic games, CORE Discussion Paper 8750, Université Catholique de Louvain, Louvain-la-Neuve, Belgium (Chapter 10 in this volume).
31. Mertens, J.-F. and Parthasarathy, T. (1991) Non-zero-sum stochastic games, in T.E.S. Raghavan et al., *Stochastic Games and Related Topics*, Kluwer Academic Publishers, Dordrecht, pp. 145–148.
32. Mohan, S.R., Neogy, S.K. and Parthasarathy, T. (1997) Linear complementarity and discounted polystochastic games when one player controls transitions, in M.C. Ferris and J.-S. Pang (eds.), *Complementarity and Variational Problems*, SIAM, Philadelphia, PA, pp. 284–294.
33. Mohan, S.R. and Raghavan, T.E.S. (1987) An algorithm for discounted switching control stochastic games, *OR Spektrum* **9**, 41–45.
34. Nowak, A. S. and Raghavan, T.E.S (1992) A finite-step algorithm via a bimatrix game to a single-controller non-zero-sum stochastic game, *Mathematical Programming* **17**, 519–526.
35. Parthasarathy, T. and Raghavan, T.E.S. (1981) An orderfield property for stochastic games when one player controls transition probabilities, *Journal of Optimization Theory and Applications* **33**, 375–392.
36. Parthasarathy, T., Tijs, S.J. and Vrieze, O.J. (1984) Stochastic games with state independent transitions and separable rewards, in G. Hammer and D. Pallaschke (eds.), *Selected Topics in OR and Mathematical Economics*, Springer-Verlag, Lecture Notes Series 226, pp. 262–271.
37. Pollatschek, M. and Avi-Itzhak, B. (1969) Algorithms for stochastic games with geometrical interpretation, *Management Science* **15**, 399–425.
38. Raghavan, T.E.S., Ferguson, T.S., Parthasarathy, T. and Vrieze, O.J. (eds.) (1990) *Stochastic Games and Related Topics: A Volume in Honor of L.S. Shapley*, Kluwer Academic Publishers, Dordrecht, The Netherlands.
39. Raghavan, T.E.S. and Filar, J.A. (1991) Algorithms for stochastic games - A survey, *Zeitschrift für Operations Research* **35**, 437–472.
40. Raghavan, T.E.S. and Syed, Z. (2002) A policy improvement-type algorithm for solving zero-sum two-person stochastic games of perfect information, *Mathematical Programming*, to appear.
41. Raghavan, T.E.S. and Syed, Z. (2002) An algorithm to solve non-zero-sum undiscounted single-controller stochastic games, *Mathematics of Operations Research*, to appear.
42. Raghavan, T.E.S. and Syed, Z. (2002) A policy improvement-type algorithm for solving zero-sum two-person stochastic games of a special class, *Zeitschrift für Op-*

erations Research, to appear.
43. Raghavan, T.E.S., Tijs, S.J. and Vrieze, O.J. (1986) Stochastic games with additive rewards and additive transitions, *Journal of Optimization Theory and Applications* **47**, 451–464.
44. Shapley, L.S. (1953) Stochastic games, *Proceedings of the National Academy of Sciences of the U.S.A.* **39**, 1095–1100 (Chapter 1 in this volume).
45. Shultz, T.A. (1987) Mathematical programming and stochastic games, Ph.D. thesis, The Johns Hopkins University.
46. Solan, E. (1998), Discounted stochastic games, *Mathematics of Operations Research* **23**, 1010–1021.
47. Takahashi, M. (1964) Equilibrium points of stochastic non-cooperative n-person games, *Journal of Science of the Hiroshima University, Series A-I* **28**, 95–99.
48. Thuijsman, F. and Raghavan, T.E.S (1997) Stochastic games with switching control or ARAT structure, Technical Report M94-06, University of Limburg, Maastricht, The Netherlands.
49. Van der Waal, J. (1977) Discounted Markov games: Successive approximations and stopping times, *International Journal of Game Theory* **6**, 11–22.
50. Vrieze, O.J. (1981) Linear programming and undiscounted stochastic game in which one player controls transitions, *OR Spektrum* **3**, 29–35.
51. Vrieze, O.J. (1983) Stochastic games with finite state and action spaces, University of Nijmegen, Nijmegen, The Netherlands.
52. Vrieze, O.J., Tijs, S.J., Raghavan, T.E.S. and Filar, J.A. (1983) A finite algorithm for switching control stochastic games, *OR Spektrum* **5**, 15–24.
53. Zwick, U. and Paterson, M.S. (1996) The complexity of mean payoff games on graphs, *Theoretical Computer Science* **158**, 343– 359.