Algorithms for Discounted Stochastic Games[1]

S. S. RAO,[2] R. CHANDRASEKARAN,[3] AND K. P. K. NAIR[4]

Communicated by R. A. Howard

Abstract. In this paper, a two-person zero-sum discounted stochastic game with a finite state space is considered. The movement of the game from state to state is jointly controlled by the two players with a finite number of alternatives available to each player in each of the states. We present two convergent algorithms for arriving at minimax strategies for the players and the value of the game. The two algorithms are compared with respect to computational efficiency. Finally, a possible extension to nonzero sum stochastic game is suggested.

1. Introduction

In a stochastic game, the play proceeds in a sequence of steps or transitions assumed to take place at every unit time interval and, at each transition, the play is said to be in some state i chosen from a finite set of states. In each state, a finite number of alternatives is available to each of the players. While in each state there is a zero-sum reward and the game moves to other states probabilistically, both the reward and transition probabilities depend on the actions taken by the players in the state. Shapley (Ref. 1) introduced this concept of stochastic game but with a stop probability in each state, derived a

[1] This research was supported in part by funds allocated to the Department of Operations Research, School of Management, Case Western Reserve University under Contract No. DAHC 19-68-C-0007 (Project Themis) with the U.S. Army Research Office, Durham, North Carolina. The authors thank the referees for their valuable suggestions.

[2] Associate Professor, Department of Operations Research, Case Western Reserve University, Cleveland, Ohio.

[3] Associate Professor, Department of Operations Research, Case Western Reserve University, Cleveland, Ohio.

[4] Associate Professor, Department of Business Administration, University of New Brunswick, Fredericton, New Brunswick, Canada.

set of functional relations that the value of the game must satisfy, and established certain existence theorems. Gillette (Ref. 2) studied the problem without the stop probability assumption and established certain existence theorems in addition to proving continuity of the value of the game over the strategy space. He also considered discounting of future payoffs and obtained certain valuable theorems. Maitra and Parthasarathy (Refs. 3–4) extended the theory of stochastic games to infinite state and action spaces, while Rogers (Ref. 5) and Sobel (Ref. 6) developed the theory of nonzero-sum stochastic games.

A stochastic game with no stop probability as studied by Gillette (Ref. 2) is nonterminating and, for this case, Hoffman and Karp (Ref. 7) developed a convergent algorithm when the future payoffs are not discounted. Subsequently, Pollatschek and Avi-Itzhak (Ref. 8) presented two algorithms, though without proof of convergence for the nonterminating and terminating stochastic games, respectively; however, they have not considered discounted stochastic games. The authors (Ref. 9) have shown that the algorithm for the nonterminating case is indeed convergent.

In a nonterminating stochastic game, if the future payoffs are to be discounted, as yet there is no algorithm in literature for determing the minimax strategies for the players and the value of the game. However, the authors (Ref. 10) have suggested an algorithm for this purpose based on somewhat similar lines as Hoffman and Karp (Ref. 7) developed for the undiscounted case. In the present paper, firstly this algorithm with proof is presented. Secondly, an alternative algorithm and proof based on an entirely different approach are given. Finally, the two algorithms are compared with respect to computational efficiency with the help of a computational experiment, and it is seen that the second algorithm is significantly faster. For achieving the same accuracy, it requires fewer iterations and, moreover, each iteration needs only less time.

2. Stochastic Games with Discounted Payoff

In a two person zero-sum stochastic game with a finite number of states, labeled $1, 2,..., i,..., N$, let the finite number of alternatives or pure strategies in state i available to Players I and II, respectively, be numbered $1, 2,..., k,..., K_i$ and $1, 2,..., l,..., L_i$. The probabilistic movement of the game from state to state takes place at every unit period. While in state i, if the alternatives followed by the players are k and l, respectively, the game moves to state j in one step with probability p_{ij}^{kl}, and Player I receives from Player II a finite reward or payoff q_i^{kl}. Future payoff is discounted by a factor β, $0 \leqslant \beta < 1$, per period. Thus, the players are interested in the effective $1 - \beta$ discounted payoff. Without loss of generality, it is assumed that the quantities q_i^{kl} are nonnegative. From the above description, it is seen that

$$p_{ij}^{kl} \geqslant 0, \qquad i, j = 1, 2,..., N, \qquad k = 1, 2,..., K_i, \qquad l = 1, 2,..., L_i, \quad (1)$$

$$\sum_{j=1}^{N} p_{ij}^{kl} = 1, \qquad i = 1, 2,..., N, \qquad k = 1, 2,..., K_i, \qquad l = 1, 2,..., L_i. \quad (2)$$

Gillette (Ref. 2) in his Theorem 1 states the existence of stationary optimal (in the minimax context) strategies for the players and unique value in a discounted stochastic game. Further, he states that this result could also be obtained from Shapley (Ref. 1) if the stop probabilities in all states are made the same and equal to $1 - \beta$. This property shows that, in a discounted stochastic game, there is no better nonstationary strategy than the optimal strategy among all the stationary strategies. Therefore, while searching for an optimal strategy and the unique value of the game, only stationary strategies need be considered. A stationary strategy x for Player I is an N-tuple of probability vectors such that

$$\begin{gathered} x = (x_1, x_2,..., x_i,..., x_N), \\ x_i = (x_i^1, x_i^2,..., x_i^k,..., x_i^{K_i}), \\ \sum_{k=1}^{K_i} x_i^k = 1, \qquad i = 1, 2,..., N. \end{gathered} \qquad (3)$$

A similar interpretation holds for y, a stationary strategy for Player II. If v_i^*, $i = 1, 2,..., N$, denotes the unique value of the game with discounted payoff, given that the play is started in state i, the functional relations to be satisfied by these values, as seen from Gillette (Ref. 2) as well as Shapley (Ref. 1), are

$$v_i^* = \text{val}\, G_i(v_1^*, v_2^*,..., v_N^*), \qquad i = 1, 2,..., N, \qquad (4)$$

where the (k, l) element of the payoff matrix of game $G_i(')$ is

$$q_i^{kl} + \beta \sum_{j=1}^{N} p_{ij}^{kl} v_j^* \qquad (5)$$

and val $G_i(')$ denotes the value of the game.

3. Algorithms

As outlined in the introduction, two algorithms are presented in this section. Algorithm I is on somewhat similar lines as that of Hoffman and Karp (Ref. 7) for the undiscounted case, while Algorithm II is based on an entirely different approach.

3.1. Algorithm I

Step (1). Choosing a strategy $y^{(0)}$ for Player II, obtain $v_i^{(0)}$, $i = 1, 2,\dots, N$, the unique solution of

$$v_i^{(0)} = \max_{1\leqslant k\leqslant K_i} \sum_{l=1}^{L_i} \left[q_i^{kl} + \beta \sum_{j=1}^{N} p_{ij}^{kl} v_j^{(0)} \right] y_i^{l(0)}, \qquad i = 1, 2,\dots, N.$$

This could be done using the algorithm of Howard (Ref. 11) for Markovian decision processes with discounted rewards.

Step (2). Given $v_i^{(0)}$, $i = 1, 2,\dots, N$, compute $y^{(1)}$ such that $y_i^{(1)}$ is an optimal strategy for Player II in the game

$$G_i^{(1)}(v_1^{(0)}, v_2^{(0)},\dots, v_N^{(0)}),$$

whose payoff matrix has (k, l) element

$$q_i^{kl} + \beta \sum_{j=1}^{N} p_{ij}^{kl} v_j^{(0)}.$$

Repeat the above steps until convergence, which is characterized by

$$v_i^{(r)} = v_i^{(r-1)}, \qquad i = 1, 2,\dots, N;$$

and, when this is true,

$$v_i^{(r)} = v_i{}^*, \qquad y^{(r)} = y^*,$$

and the optimal strategy for Player I is

$$x^{(r)} = x^*,$$

where $x^{(r)}$ is obtained from the solution of the games in Step (2) of rth iteration just as $y^{(r)}$ is got.

Proof. Let $G_i^{(r)}$ denote the value of the game $G_i^{(r)}(\;)$. Since $v_i^{(0)}$, $i = 1, 2,\dots, N$, are the solutions, with Howard's algorithm, of the

Markovian decision process, obtained by fixing the opponent's strategy as $y^{(0)}$, we have

$$v_i^{(0)} \geqslant G_i^{(1)}, \qquad i = 1, 2, \ldots, N.$$

Based on similar arguments, it is seen that

$$v_i^{(r)} \geqslant G_i^{(r+1)}, \qquad i = 1, 2, \ldots, N, \qquad r = 0, 1, 2, \ldots .$$

Obviously, $v_i^{(0)}$ is the payoff obtained with the policy $(y^{(0)}, X^{(0)})$ where $X^{(0)}$ is the solution of Howard's algorithm and $G_i^{(1)}$ is the payoff with the policy $(y^{(1)}, x^{(1)})$ for the first transition and subsequently forever the policy $(y^{(0)}, x^{(0)})$. Since $v_i^{(0)} \geqslant G_i^{(1)}$, $i = 1, 2, \ldots, n$, the payoff $v_i^{(1)}$ obtained if $(y^{(1)}, x^{(1)})$ is followed forever will be smaller. This follows from the work of Blackwell (Ref. 12). Since $(y^{(1)}, x^{(1)})$ will have still smaller payoff, it is seen that

$$v_i^{(1)} \leqslant G_i^{(1)}, \qquad i = 1, 2, \ldots, N.$$

Therefore, $v_i^{(1)}$ is a monotone decreasing sequence for each i. Since a unique solution v_i^*, $i = 1, 2, \ldots, N$, exists, as seen from Shapley (Ref. 1) and Gillette (Ref. 2), these sequences converge to the values v_i^*, $i = 1, 2, \ldots, N$. In the limit, as r tends to infinity, we have

$$v_i^{(r)} = q_i^{(r)} + \beta \sum p_{ij}^{(r)} v_j^{(r)}, \qquad G_i^{(r+1)} = q_i^{(r)} + \beta \sum p_{ij}^{(r)} v_j^{(r)},$$

where

$$q_i^{(r)} = \sum_k \sum_l q_i^{kl} x_i^{k(r)} y_i^{l(r)}, \qquad p_{ij}^{(r)} = \sum_k \sum_l p_{ij}^{kl} x_i^{k(r)} y_i^{l(r)}.$$

Therefore,

$$v_i^{(r)} = G_i^{(r+1)} = v_i^*, \qquad i = 1, 2, \ldots, N.$$

It may be noted that the above algorithm could be described fixing Player I, instead of Player II, at a specified strategy, and the proof would be similar, but Howard's policy iteration will be for minimizing v_i's.

3.2. Algorithm II. It may be noted that, in the description of this algorithm, the total payoff is denoted by V_i, $i = 1, 2, \ldots, N$. Clearly, the unique solution V_i^*, $i = 1, 2, \ldots, N$, will be the same as v_i^*, $i = 1, 2, \ldots, N$, obtained above and would satisfy the relations (4).

Step (1). Assume $V_i^{(0)} = 0$, $i = 1, 2,..., N$, and solve the set of N games

$$g_i^{(1)}(V_1^{(0)},..., V_N^{(0)}), \qquad i = 1, 2,..., N,$$

whose payoff matrix has (k, l) element as

$$q_i^{kl} + \beta \sum_{j=1}^{N} p_{ij}^{kl} V_j^{(0)}$$

to obtain the values $g_i^{(1)}$, $i = 1, 2,..., N$, and the optimal strategies of the game $(x^{(1)}, y^{(1)})$.

Step (2). Solve the system of equations

$$V_i^{(1)} = q_i^{(1)} + \beta \sum_{j=1}^{N} p_{ij}^{(1)} V_j^{(1)},$$

where

$$q_i^{(1)} = \sum_k \sum_l q_i^{kl} x_i^{k(1)} y_i^{l(1)}, \qquad p_{ij}^{(1)} = \sum_k \sum_l p_{ij}^{kl} x_i^{k(1)} y_i^{l(1)}.$$

Step (3). Check whether or not $V_i^{(1)} = g_i^{(1)}$, $i = 1, 2,..., N$. If it is true, the solution is obtained as

$$V_i^{(1)} = g_i^{(1)} = V_i^*, \qquad i = 1, 2,..., N,$$
$$x^{(1)} = x^*, \qquad y^{(1)} = y^*.$$

Otherwise, go to Step (1), with $V_i^{(1)}$, $i = 1, 2,..., N$, obtained in Step (2).

Proof. Gillette (Ref. 2) has established that V_i is continuous in x and y and as seen from Shapley (Ref. 1), the unique solution V_i^*, $i = 1, 2,..., N$, has to satisfy the relations (4). Therefore, if $V_i^{(0)} \leqslant V_i^*$, it is seen that

$$g_i^{(1)} \geqslant V_i^{(0)}, \qquad V_i^{(1)} \leqslant V_i^*, \qquad i = 1, 2,..., N.$$

The above inequalities could be derived from the following considerations also. Since $V_i^{(0)} \leqslant V_i^*$, it follows from the continuity of V_i in x and y that there exists an x such that fixing Player I at x, if Howard's policy iteration (Ref. 11) is carried out minimizing the payoff over y, the resulting payoff is $V_i = V_i^{(0)}$, $i = 1, 2,..., N$. Therefore, the game value is $g_i^{(1)} \geqslant V_i^{(0)}$, $i = 1, 2,..., N$. The second inequality follows from the unique relations (4) to be satisfied by the solution and the fact that V_i is continuous in x and y.

Now, note that $V_i^{(1)}$ is the payoff if the policy $(x^{(1)}, y^{(1)})$ is followed forever, while $g_i^{(1)}$ is the payoff if the policy $(x^{(1)}, y^{(1)})$ is followed just for one transition and after that forever the policy with payoff $V_i^{(0)}$, $i = 1, 2,..., N$. Therefore, since $g_i^{(1)} \geqslant V_i^{(0)}$, $i = 1, 2,..., N$, it is seen from the work of Blackwell (Ref. 12) that

$$V_i^{(1)} \geqslant g_i^{(1)} \geqslant V_i^{(0)}, \qquad i = 1, 2,..., N.$$

Similarly, if $V_i^{(1)} \leqslant V_i^*$, $i = 1, 2,..., N$, we have

$$V_i^{(2)} \geqslant g_i^{(2)} \geqslant V_i^{(1)}, \qquad V_i^{(2)} \leqslant V_i^*, \qquad i = 1, 2,..., N.$$

Since all q_i^{kl} are nonnegative, all V_i^* are nonnegative. The algorithm starts with $V_i^{(0)} = 0$, and obviously $V_i^{(0)} \leqslant V_i^*$; therefore, the sequence generated has the property

$$V_i^{(0)} \leqslant g_i^{(1)} \leqslant V_i^{(1)} \leqslant g_i^{(2)} \leqslant V_i^{(2)} \cdots \leqslant V_i^*.$$

Thus, $V_i^{(r)}$ and $g_i^{(r)}$, $i = 1, 2,..., N$, are monotone increasing sequences bounded from above by the unique value V_i^*, $i = 1, 2,..., N$, of the stochastic game. Hence, in the limit, as r tends to infinity,

$$V_i^{(r)} = q_i^{(r)} + \beta \sum_{j=1}^{N} p_{ij}^{(r)} V_j^{(r)}, \qquad i = 1, 2,..., N,$$

$$g_i^{(r)} = q_i^{(r)} + \beta \sum_{j=1}^{N} p_{ij}^{(r)} V_j^{(r)}, \qquad i = 1, 2,..., N,$$

and hence,

$$g_i^{(r)} = V_i^{(r)} = V_i^*, \qquad i = 1, 2,..., N.$$

It may be noted that the proof justifies the algorithm to start with any set of V_i, $i = 1, 2,..., N$, such that $V_i \leqslant V_i^*$ for each i. However, it is convenient to start the algorithm with $V_i^{(0)} = 0$ for all i provided all q_i^{kl} are nonnegative. Then, obviously $V_i^{(0)} \leqslant V_i^*$ for all i. The nonnegativity restriction on q_i^{kl} does not cause any loss of generality.

4. Comparison of the Algorithms

Having developed two algorithms for discounted stochastic games, a logical step is to make a comparison of the two algorithms with respect to computational efficiency. Indeed, if this could be done based on theoretical considerations, the results would be most valuable. In this direction, only some intuitive arguments are presented, but these are substantiated by a computational experiment.

Every iteration in Algorithm I requires the solution of Howard's algorithm in addition to solving a set of games. Further, Howard's algorithm itself requires a set of linear equations to be solved several times and each time the test quantities have to be obtained and compared. In Algorithm II, every iteration requires the solution of a set of linear equations only once, while the efforts required to solve the set of games remain the same. From these considerations it is seen that each iteration in Algorithm II will require only significantly less time in comparison to Algorithm I. In Algorithm II, the complete results of an iteration is used in the next iteration; while, in Algorithm I, it is used only partially. Therefore, the rate of convergence of Algorithm II would be higher thereby making it a faster one.

4.1. Computational Experiments. Aggarwal (Ref. 13) has conducted an extensive investigation of the relative advantages of the two algorithms by carrying out a number of computational experiments using UNIVAC 1108. His experiments have shown conclusively that Algorithm II is uniformly faster than Algorithm I. The relevant experiment and results are presented here. The problem considered has

$$N = 10, \qquad 8 \leqslant K_i \leqslant 10, \qquad 8 \leqslant L_i \leqslant 10,$$

$$p_{ij}^{kl} > 0, \qquad \sum_{j=1}^{N} p_{ij}^{kl} = 1, \qquad \beta = 0 \cdot 8, \qquad q_i^{kl} = \sum_{j=1}^{N} p_{ij}^{kl} r_{ij},$$

where the quantities r_{ij} are generated from two-digit random number tables.

Table 1. Computational results.

	Algorithm I			Algorithm II		
Iteration number z	Iteration time (seconds)	Cumulative time (seconds)	Number of digits of accuracy	Iteration time (seconds)	Cumulative time (seconds)	Number of digits of accuracy
1	0.62020	0.62020	0	0.60800	0.60800	0
2	0.95580	1.57600	2	0.70980	1.31780	2
3	0.96520	2.54120	3	0.70060	2.01840	6
4	0.96400	3.50520	4	0.70120	2.71960	7
5	0.97060	4.47580	5			
6	0.99350	5.46930	6			
7	1.01580	6.48510	7			

Table 2. Comparison of computation times.

	Algorithm I		Algorithm II		
Number of digits of accuracy	Number of iterations	Time T_1	Number of iterations	Time T_2	T_1/T_2
7	7	6.48510	4	2.71960	2.384

For the purpose of comparing the algorithms, it is essential that both are given the same starting points. This is achieved by choosing $v_i^{(0)} = 0$, $i = 1, 2,..., N$, in Algorithm I and $V_i^{(0)} = 0$, $i = 1, 2,..., N$, in Algorithm II. The results are displayed in Table 1. Comparison of the computation times is shown in Table 2. In the tables, the number of digits of accuracy has the following meaning: for example, if it is five in the case of Algorithm I in the rth iteration, the first five digits of each $v_i^{(r)}$ are the same as the first five digits of the corresponding $G_i^{(r)}$. A similar interpretation holds in the case of Algorithm II. It is seen that Algorithm II requires fewer iterations and, moreover, each iteration takes less time. The computation time for Algorithms I and II are in the proportion 2.3:1. It has been found that, as higher accuracy is demanded, this proportion changes in favor of Algorithm II. Here, accuracy is measured by the number of iterations: the higher the number of iterations, the smaller the deviation of the calculated value from the true value.

5. Discussion and Conclusions

In the operation of military, business, and other systems, several stochastic processes are encountered. Effective management of these processes requires consideration of a competitive aspect inherent in many systems. Stochastic game theory enables effective modeling of these problems. Algorithms available in the literature are not capable of solving the problem if the payoff is discounted. Discounting, indeed, is of practical significance, and this paper provides two convergent algorithms for discounted stochastic games. It is seen, from intuitive considerations as well as from computational experiment, that Algorithm II is significantly faster.

The algorithms have been presented for two-person zero-sum games. However, Algorithm II can be generalized to two-person nonzero-sum stochastic games. This can be done by combining the

Lemke–Howson (Ref. 14) algorithm for a bimatrix game equilibrium points with the necessary and sufficient conditions in the works of Rogers (Ref. 5) and Sobel (Ref. 6). Then, a monotone value sequence results for each of the two players. Further, Algorithm II can also solve a special case of terminating stochastic games. For example, if the same stop probability r is introduced in all states, then put $\beta = 1 - r$, and the solution given by this algorithm will hold for the terminating game. This interesting and useful property of this algorithm results from the relation between the discounted stochastic game and the terminating stochastic game stated in Section 2.

References

1. Shapley, L. S., *Stochastic Games*, Proceedings of the National Academy of Sciences, Vol. 39, pp. 1095–1100, 1953.
2. Gillette, D., *Stochastic Games with Zero Stop Probabilities*, Contributions to the Theory of Games, Vol. III, pp. 179–187, Princeton, New Jersey, 1957.
3. Maitra, A., and Parthasarathy, T., *On Stochastic Games* Journal of Optimization Theory and Applications, Vol. 5, pp. 289–300, 1970.
4. Maitra, A., and Parthasarathy, T., *On Stochastic Games, II*, Journal of Optimization Theory and Applications, Vol. 8, pp. 154–160, 1971.
5. Rogers, P. D., *Non Zero Sum Stochastic Games*, University of California at Berkeley, Operations Research Center, Report No. ORC-69-8, 1969.
6. Sobel, M. J., *Noncooperative Stochastic Games*, Annals of Mathematical Statistics, Vol. 42, pp. 1930–1935, 1971.
7. Hoffman, A. J., and Karp, R. M., *On Nonterminating Stochastic Games*, Management Science (Series A), Vol. 12, pp. 359–370, 1966.
8. Pollatschek, M. A., and Avi-Itzhak, B., *Algorithm for Stochastic Games with Geometrical Interpretation*, Management Science, Vol. 15, pp. 399–415, 1969.
9. Nair, K. P. K., Rao, S. S., and Chandrasekaran, R., *A Faster Algorithm for Nonterminating Stochastic Games*, Case Western Reserve University, Department of Operations Research, Technical Memorandum No. 215, 1971.
10. Chandrasekaran, R., Nair, K. P. K., and Rao, S. S., *Stochastic Games with Semi-Markovian Rewards*, Case Western Reserve University, Department of Operations Research, Technical Memorandum No. 203, 1970.
11. Howard, R., *Dynamic Programming and Markov Processes*, John Wiley and Sons, New York, 1960.
12. Blackwell, D., *Discrete Dynamic Programming*, Annals of Mathematical Statistics, Vol. 33, pp. 719–726, 1962.

13. Aggarwal, V. V., *Algorithms for Discounted Stochastic Games: A Comparison of Efficiency*, Case Western Reserve University, Department of Operations Research, Technical Memorandum No. 243, 1971.

14. Lemke, C. E., and Howson, J. T., *Equilibrium Points of Bimatrix Games*, SIAM Journal on Applied Mathematics, Vol. 12, pp. 413–423, 1964.