

V. AGGARWAL

K. P. K. NAIR

R. CHANDRASEKARAN

Non-terminating stochastic ratio game

RAIRO. Recherche opérationnelle, tome 14, n° 1 (1980), p. 21-30

http://www.numdam.org/item?id=RO_1980__14_1_21_0

© AFCET, 1980, tous droits réservés.

L'accès aux archives de la revue « RAIRO. Recherche opérationnelle » implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
<http://www.numdam.org/>

NON-TERMINATING STOCHASTIC RATIO GAME (*)

by V. AGGARWAL, K. P. K. NAIR ⁽¹⁾ and R. CHANDRASEKARAN ⁽²⁾

Abstract. — In an earlier work by Derman, a finite state Markov Ratio Decision Process was considered with the objective of maximizing the ratio of rewards. We generalize the above mentioned process to a game context resulting in a Stochastic Ratio game. It is shown that in the non-terminating Stochastic ratio game, the players have stationary optimal strategies with an unique value. A convergent algorithm is provided to compute the solution.

Résumé. — Derman a étudié dans un travail antérieur un « Finite state Markov Ratio Decision Process » ayant pour objectif de maximiser un rapport de gains. On étend le processus mentionné ci-dessus pour le placer dans un contexte de jeu, qui conduit à un « Stochastic Ratio Game ». On démontre que dans un tel jeu indéfiniment joué (c'est-à-dire « à horizon infini ») les joueurs ont des stratégies optimales stationnaires avec une valeur unique. On donne un algorithme convergent pour calculer la solution du jeu.

INTRODUCTION

In an earlier work by Derman [1], a finite state Non-terminating Markov Ratio Decision Process (abbreviated by NMRDP) was considered with the objective of maximizing the ratio of rewards over an infinite planning horizon. It has been shown in this process that, an optimal policy is stationary and pure, and the value is unique. Fox [2] has provided an algorithm for computing the solution in the context of an Undiscounted Markov Renewal Program (MRP) but it is applicable to the NMRDP also. Schweitzer [4] has described an iterative procedure for finding a solution of the MRP and has established that a simple data-transformation reduces any MRP or NMRDP into a discrete time Markov Decision Problem, which is equivalent in the sense that it has the same state and policy spaces, and that every policy has the same long run average return per unit time.

In this paper, the finite state NMRDP is generalized to a game context resulting in a Non-terminating Stochastic Ratio Game (NSRG). This game is similar to the game of Hoffman and Karp [3] except that the payoff function is a ratio. Thus, the game proceeds in stages and at each stage, the game is in one of a

(*) Reçu juillet 1978.

⁽¹⁾ University of New Brunswick, Fredericton, N. B., Canada.

⁽²⁾ University of Texas, Richardson, Texas, U.S.A.

finite number of states. Each of the two players observes the current state and then chooses an action from a finite number of alternatives of actions available in that particular state. The players' actions and the current state jointly determine a set of two rewards and transition probabilities to the succeeding state. Thus, there are two sequences of rewards and the payoff function is the limiting ratio of the average rewards per stage over the infinite number of states.

In the stochastic games of Shapley [5] or Hoffman and Karp [3], the game played at each stage is two-person zero-sum type. In NSRG, the game played in each stage is a ratio game. It is shown that the players have stationary optimal strategies and the game has a unique value both of which are independent of the initial probability distribution. A convergent algorithm for computing the solution is presented. It may be possible to generalize the data-transformation approach of Schweitzer [4] for MRP and NMRDP, to introduce data-transformations to reduce NSRG into a non-terminating stochastic game. However, the approach taken here is the generalization of Hoffman and Karp's algorithm [3].

Both theoretical and practical aspects have motivated this work. The generalization considered, simultaneously, achieves also a generalization of MRP [2] to game context. Suppose two players control a MRP. Dependent on the state occupied by the process and the actions taken by the players, the process makes a transition after a random interval of time giving rise to a reward and the time taken for the transition. Thus the sequence of games gives rise to two streams, one of rewards and the other of time intervals respectively. The payoff function defined here is the limiting ratio of the sum of rewards to that of the time intervals. Clearly, NSRG developed in this paper solves the above generalization of MRP also. The generalization has also significance in practical situations of conflict. For example, consider two parties involved in a conflict that continues indefinitely, but may be in different levels or states. Dependent on the state and actions taken, two streams, one of gains and the other of losses, occur. Here the limiting ratio of the sum of gains to that of losses is of interest, and the generalization given in this paper solves this game.

The existence theorem and its proof given in this paper do not follow from the works of Shapley [5] or Hoffman and Karp [3]. An entirely different approach is taken for proving the theorem.

1. DESCRIPTION OF NSRG

NSRG is a sequence of ratio games played in consecutive periods of equal intervals, where each pair of matrices for each ratio game is chosen from a finite set S consisting of N pairs $(A_i, B_i, i=1, \dots, N)$ of matrices. If in a period, the

ratio game is played with matrices (A_i, B_i) , the game is said to be in state i ($i=1, \dots, N$). While in state i , Player I has a finite set C_i of K_i alternatives numbered $1, 2, \dots, k, \dots, K_i$ and similarly Player II has a set D_i of L_i alternatives numbered $1, 2, \dots, l, \dots, L_i$. Thus

$$A_i = \{ a_i^{kl} \mid k \in C_i, l \in D_i \},$$

$$B_i = \{ b_i^{kl} \mid k \in C_i, l \in D_i \}.$$

It is assumed that $B_i > 0$ ($i=1, \dots, N$).

The probabilities of transition to the successive states are given by p_{ij}^{kl} ($i, j \in S, k \in C_i, l \in D_i$), and since the NSRG is assumed to be non-terminating type,

$$\sum_{j=1}^N p_{ij}^{kl} = 1, \quad i \in S, \quad k \in C_i, \quad l \in D_i.$$

It will be assumed that corresponding to any choice of actions by the players, the underlying Markov Chain is irreducible. That is, if mixed strategy vectors x_i and y_i are followed when the game is in state i , and

$$P_{ij}(x_i, y_i) = \sum_{k \in C_i} \sum_{l \in D_i} p_{ij}^{kl} x_i^k y_i^l,$$

are the transition probabilities of a Markov Chain $\{X_n, n=0, 1, \dots\}$, then for any i and j , the probability that state j will ever be reached from the initial state i is 1.

Now, in general given that the game starts in a specified state i_0 , the evolution of NSRG can be represented by the sequence $\{i_n, \Delta_n, \Sigma_n\}, n=0, 1, \dots$, where i_n, Δ_n and Σ_n respectively are the state occupied and actions taken by the Players I and II at period n such that $i_n \in S, \Delta_n \in C_{i_n}$ and $\Sigma_n \in D_{i_n}$. If r_{i_0} denotes the payoff function for NSRG started in state i_0 , then

$$r_{i_0} = \lim_{T \rightarrow \infty} \frac{\sum_{n=0}^T \Delta_n A_{i_n} \Sigma_n / T}{\sum_{n=0}^T \Delta_n B_{i_n} \Sigma_n / T}, \tag{1}$$

where Δ_n and Σ_n are viewed as vectors.

2. RESULTS OF NMRDP

An NMRDP is a special case of NSRG in which the second player is a "dummy" being allowed a fixed strategy in each of the N states. An optimal solution of NMRDP has been characterized by Derman [1] as belonging to the

set of pure stationary strategies. Following Fox [2], its solution can be obtained by solving the following N relations

$$\text{Max}_{k \in C_i} [a_i^k - rb_i^k + \sum_{j \in S} p_{ij}^k s_j - s_i] = 0, \quad i = 1, \dots, N, \quad (2)$$

where $s_i (i = 1, \dots, N)$ are relative values. By setting $s_N = 0$, a unique solution can be obtained. The solution (r, s_1, \dots, s_N) is independent of the initial state in which the process starts.

3. EXISTENCE OF MINIMAX SOLUTION FOR NSRG

Let T be the set of all possible strategies available to Player I which are generated by the sets $C_i (i = 1, \dots, N)$. Let $X (X \subset T)$ denote the set of all stationary strategies available to Player I , where $x \in X$ is an N -tuple of probability vectors such that

$$x = (x_1, x_2, \dots, x_N) \quad \text{where} \quad x_i = (x_i^1, x_i^2, \dots, x_i^{K_i})$$

such that $\sum_{k \in C_i} x_i^k = 1$.

Let the corresponding sets for Player II be, U the set of all strategies; and $Y (Y \subset U)$ the set of all stationary strategies among them.

Let $r_\alpha(t, u)$ denote the ratio for the NSRG starting with an initial probability vector, α , when Players I and II follow $t \in T$ and $u \in U$ strategies respectively. For a fixed strategy $y \in Y$ for Player II , Player I , if y is known to him, faces a Markov Ratio Decision process and the maximal value of this problem is:

$$r_\alpha(y) = \text{Max}_{t \in T} r_\alpha(t, y). \quad (3)$$

Following Derman [1] and Fox [2], it has a stationary optimal solution which is independent of the initial state. Therefore,

$$\begin{aligned} r_\alpha(y) &= \text{Max}_{x \in X} r_\alpha(x, y) \\ &= r(y), \end{aligned} \quad (4)$$

where $r(y)$ can be obtained as the value of a linear program which is continuous in its data. Since X and Y are convex and compact,

$$\text{Min}_{y \in Y} r(y) = r(y^*), \quad (5)$$

exists at atleast one point $y^* \in Y$ with a unique value $r(y^*)$.

Suppose Player II is fixed at y^* and NMRDP corresponding to Player I as in (4) above is solved with

$$a_i^k = \sum_{l \in D_i} a_i^{kl} y_i^{*l}, \quad (6)$$

$$b_i^k = \sum_{l \in D_i} b_i^{kl} y_i^{*l}, \quad (7)$$

$$p_{ij}^k = \sum_{l \in D_i} p_{ij}^{kl} y_i^{*l}. \quad (8)$$

Let the optimal solution be $(r(y^*), s_1^*, \dots, s_N^*)$.

Let S^* be the set of N two-person zero-sum games such that i th game has as its payoff matrix, $Q_i^* = \{q_i^{kl}\} (i=1, \dots, N)$ where

$$q_i^{kl} = (a_i^{kl} - r(y^*)b_i^{kl} + \sum_{j \in S} p_{ij}^{kl} s_j^* - s_i^*), \quad k \in C_i, \quad l \in D_i. \quad (9)$$

LEMMA 1 : *The stationary strategy y^* is optimal for Player II in the set S^* such that y_i^* is optimal in the i th game which is of two-person zero-sum type and whose pay-off matrix is $Q_i^* (i=1, \dots, N)$. Further, the value of each game in the set S^* is zero.*

Proof: The optimal solution $(r(y^*), s_1^*, \dots, s_N^*)$ for NMRDP with a_i^k, b_i^k and p_{ij}^k given in (6), (7) and (8) satisfies (2) which implies

$$\text{Max}_{k \in D_i} \sum_{l \in D_i} (a_i^{kl} - r(y^*)b_i^{kl} + \sum_{j \in S} p_{ij}^{kl} s_j^* - s_i^*) y_i^{*l} = 0, \quad \forall i \in S. \quad (10)$$

Suppose y^* is not optimal in the set S^* . Then there is at least one payoff matrix Q_h^* in which \bar{y}_h is optimal instead of y_h^* , so that for the corresponding game,

$$\text{Max}_{k \in C_h} [\sum_{l \in D_h} a_h^{kl} - r(y^*)b_h^{kl} + \sum_{j \in S} p_{hj}^{kl} s_j^* - s_h^*] \bar{y}_h^l < 0. \quad (11)$$

Now consider Player II being fixed at strategy

$$\bar{y} = (y_1^*, y_2^*, \dots, y_{h-1}^*, y_h^*, \dots, y_N^*)$$

and an NMRDP with respect to player I solved with the corresponding coefficients \bar{a}_i^k, \bar{b}_i^k and $\bar{p}_{ij}^k (k \in C_i, i, j \in S)$ obtained in a manner similar to (6), (7), (8) with \bar{y} . Let its solution obtained from (2) be $r(\bar{y}), \bar{s}_i (i=1, \dots, N)$. It is obvious that $(r(y^*), s_1^*, \dots, s_N^*)$ is feasible for this new program but not optimal since (11) would violate the condition for optimality. Hence $r(\bar{y}) < r(y^*)$, which contradicts y^* being optimal. Thus, y_i^* is optimal in the i th game whose payoff matrix is $Q_i^* (i=1, \dots, N)$.

From (10), it follows then that the value of each game in the set S^* is zero.

Similar results can be established by fixing Player I at \hat{x} , where \hat{x} corresponds to

$$r(\hat{x}) = \text{Max}_{x \in X} \text{Min}_{y \in Y} r_\alpha(x, y). \quad (12)$$

Suppose Player I is fixed at \hat{x} and NMRDP corresponding to Player II is solved and the optimal solution is $(r(\hat{x}), \hat{s}_1, \hat{s}_2, \dots, \hat{s}_N)$.

A set \hat{S} of N 2-person zero-sum games is defined such that the i th game has as its payoff matrix $\hat{Q}_i = \{\hat{q}_i^{kl}\}$ where

$$\hat{q}_i^{kl} = (a_i^{kl} - r(\hat{x}))b_i^{kl} + \sum_{j \in S} p_{ij}^{kl} \hat{s}_j - \hat{s}_i, \quad k \in C_i, \quad l \in D_i.$$

Then, \hat{x} is optimal for Player I in the set \hat{S} such that \hat{x}_i is optimal in the i th game whose payoff matrix is \hat{Q}_i and the value of each game in the set \hat{S} is zero.

LEMMA 2: *The stationary strategies \hat{x} and y^* are such that $r(\hat{x}) = r(y^*) = r(\hat{x}, y^*)$ which can be denoted by r^* and that $\hat{s}_i - s_i^* = C$, a constant ($i = 1, \dots, N$).*

Proof: It is easy to see that $r(y^*) \geq r(\hat{x})$.

Corresponding to N 2-person zero-sum games in set S^* ,

$$\text{Val} \left\{ a_i^{kl} - r(y^*)b_i^{kl} + \sum_{j=1}^N p_{ij}^{kl} s_j^* - s_i^* \right\} = 0, \quad i = 1, \dots, N. \quad (13)$$

Similarly corresponding to games in set \hat{S} ,

$$\text{Val} \left\{ a_i^{kl} - r(\hat{x})b_i^{kl} + \sum_{j=1}^N p_{ij}^{kl} \hat{s}_j - \hat{s}_i \right\} = 0, \quad i = 1, \dots, N. \quad (14)$$

Consider the non-terminating stochastic game (NSG) Γ^* in which the payoff matrices are $[A_i - r(y^*)B_i]$ ($i = 1, \dots, N$). The alternatives of actions for the players and the transition probabilities are as already defined for the NSRG. Hoffman and Karp [3] have shown the existence of stationary optimal strategies for the players in the NSG and that the value is unique. Recalling the characterization of solution in Hoffman and Karp [3], if g^* denotes the value of the game Γ^* , (13) indicates that the parameter $r(y^*)$ has been selected such that the value $g^* = 0$. From lemma 1, y^* is optimal for Player II in game Γ^* .

It can be seen similarly that if $\hat{\Gamma}$ denotes another NSG in which the payoff matrices are $[A_i - r(\hat{x})B_i]$ ($i = 1, \dots, N$) and \hat{g} is its value, then from (14), $\hat{g} = 0$ and \hat{x} is optimal for Player I in the game $\hat{\Gamma}$.

Now suppose $r(y^*) > r(\hat{x})$ then since $B_i > 0$, element-by-element

$$[A_i - r(y^*)B_i] < [A_i - r(\hat{x})B_i], \quad i = 1, \dots, N \quad (15)$$

and the values of the corresponding NSG's Γ^* and $\hat{\Gamma}$ would imply,

$$g^* < \hat{g}, \tag{16}$$

which contradicts the relationship developed from (13) and (14) above.

Hence,

$$r(y^*) = r(\hat{x}) \tag{17}$$

Let the common value of $r(y^*)$ and $r(\hat{x})$ be denoted by r^* , then

$$r(\hat{x}, y^*) \leq r(y^*) = r^* \quad \text{and} \quad r(\hat{x}, y^*) \geq r(\hat{x}) = r^*.$$

Therefore,

$$r(\hat{x}, y^*) = r^* \tag{18}$$

Substituting the common value r^* for $r(y^*)$ and $r(\hat{x})$ in (13) and (14) respectively shows that

$$s_i^* - \hat{s}_i = C, \quad \text{a constant for } i = 1, \dots, N. \tag{19}$$

THEOREM 1: *The NSRG has a minimax solution with a unique value and the players have stationary optimal strategies.*

Proof: Going back to relations (4) and (3),

$$r(y^*) = \text{Max}_{x \in X} r(x, y^*) = \text{Max}_{t \in T} r(t, y^*). \tag{20}$$

Therefore,

$$r_\alpha(t, y^*) \leq r(y^*) = r^* \quad \text{for all } t \in T. \tag{21}$$

Similarly,

$$r_\alpha(\hat{x}, u) \geq r(\hat{x}) = r^* \quad \text{for all } u \in U, \tag{22}$$

where $r^* = r(\hat{x}, y^*)$.

Thus, from (21) and (22) it can be seen that

$$r_\alpha(t, y^*) \leq r(\hat{x}, y^*) \leq r_\alpha(\hat{x}, u), \quad \text{for all } t \in T \text{ and } u \in U \tag{23}$$

which establishes that the pair (\hat{x}, y^*) is optimal among all the strategies (T, U) , r^* is unique and that it is independent of the initial probability vector α .

4. CHARACTERIZATION OF THE SOLUTION OF NSRG

From lemmas 1 and 2:

$$\left. \begin{aligned} \text{Max}_{k \in C_i} \sum_{l \in D_i} \left(a_i^{kl} - r^* b_i^{kl} + \sum_{j=1}^N p_{ij}^{kl} s_j^* - s_i^* \right) y_i^{*l} = 0, \\ i = 1, \dots, N \end{aligned} \right\} \tag{24}$$

and

$$\text{Min } \sum_{l \in D_i} \sum_{k \in C_i} \left(a_i^{kl} - r^* b_i^{kl} + \sum_{j=1}^N p_{ij}^{kl} s_j^* - s_i^* \right) \hat{x}_i^k = 0, \quad \left\{ \begin{array}{l} i=1, \dots, N, \end{array} \right. \quad (25)$$

(24) and (25) together give the characterization of NSRG by the following N relations,

$$\text{Value } G_i = \text{Val} \left\{ a_i^{kl} - r^* b_i^{kl} + \sum_{j=1}^N p_{ij}^{kl} s_j^* - s_i^* \right\} = 0, \quad \left\{ \begin{array}{l} i=1, \dots, N, \end{array} \right. \quad (26)$$

where r^* is unique and one of the s_i^* (say s_N^*) is set to zero.

The stationary optimal strategies (x^*, y^*) are composed of the optimal strategies (x_i^*, y_i^*) in the respective games $G_i (i=1, \dots, N)$.

The above characterization leads to a convergent algorithm for computing the solution and this is presented in the following section.

5. A CONVERGENT ALGORITHM

Step 0: Set iteration $n=0$ and fix Player II at a stationary strategy, $y^{(0)}$.

Step 1: Using stationary strategy $y^{(n)}$ for Player II, find an unique solution $r^{(n)}$, $s_i^{(n)} (i=1, 2, \dots, N)$ of the system,

$$\text{Max } \sum_{k \in C_i} \sum_{l \in D_i} \left(a_i^{kl} - r b_i^{kl} + \sum_{j=1}^N p_{ij}^{kl} s_j - s_i \right) y_i^{(n)l} = 0, \quad \left\{ \begin{array}{l} i=1, \dots, N \end{array} \right. \quad (27)$$

and,

$$s_N = 0, \quad (28)$$

by using the policy iteration algorithm by Fox [2] for the NMRDP.

Step 2: Solve the N two-person zero-sum games whose element in the (k, l) position of the payoff matrix of i th game is

$$\left(a_i^{kl} - r^{(n)} b_i^{kl} + \sum_{j=1}^N p_{ij}^{kl} s_j^{(n)} - s_i^{(n)} \right), \quad (29)$$

to obtain the optimal strategies $x_i^{(n+1)}$, $y_i^{(n+1)}$ and the unique values $g_i^{(n+1)}$ for $i=1, 2, \dots, N$.

Set

$$x^{(n+1)} = (x_1^{(n+1)}, x_2^{(n+1)}, \dots, x_N^{(n+1)})$$

and,

$$y^{(n+1)} = (y_1^{(n+1)}, \dots, y_N^{(n+1)}).$$

Step 3: If $g_i^{(n+1)} = 0$ for all i , then the solution of NSRG is given by $x^* = x^{(n+1)}$, $y^* = y^{(n+1)}$ and $r^* = r^{(n)}$. Stop.

If $g_i^{(n+1)} \neq 0$ for at least one state, then go to step 1 with $n \leftarrow n + 1$. ■

If $g_i^{(n)} = 0$ for all i , then $x^{(n)}$, $y^{(n)}$ and $r^{(n-1)}$ form the the solution of the NSRG since the characterization given by (26) is satisfied.

If however $g_i^{(n)} \neq 0$ for at least one state i , then the following theorem shows that $r^{(n)}$ obtained at the next iteration is a lower value then the current value, $r^{(n-1)}$, and it leads to convergence.

THEOREM 2: *The sequence $r^{(n)}$ ($n=1, 2, \dots$) generated is monotonically decreasing, bounded from below by the unique value, r^* .*

Proof: Clearly, $g_i^{(n)} \leq 0$ for all i since there exists a strategy $y^{(n)}$ which if fixed for Player II gives the maximal payoff zero according to (27)-(28). Therefore,

$$g_i^{(n)} = \text{Max}_{k \in C_i} \sum_{l \in D_i} \left(a_i^{kl} - r^{(n-1)} b_i^{kl} + \sum_{j=1}^N p_{ij}^{kl} s_j^{(n-1)} - s_i^{(n-1)} \right) y_i^{(n)l} \leq 0, \quad \left. \begin{array}{l} i = 1, \dots, N \end{array} \right\} \quad (30)$$

with,

$$s_N^{(n-1)} = 0. \quad (31)$$

Consider the linear programming problem corresponding to (27) and (28) which generates $r^{(n)}$ when Player II is fixed at $y^{(n)}$,

$$\text{Min } r \quad (32)$$

$$\text{S.T. } \left. \begin{array}{l} s_i + r \sum_{l \in D_i} b_i^{kl} y_i^{(n)l} \geq \sum_{l \in D_i} \left(a_i^{kl} + \sum_{j=1}^N p_{ij}^{kl} s_j \right) y_i^{(n)l}, \\ i = 1, \dots, N, \quad k = 1, \dots, K_i, \end{array} \right\} \quad (33)$$

$$s_N = 0, \quad (34)$$

$$r, s_i \quad (i = 1, \dots, N-1) \text{ unrestricted.} \quad (35)$$

It can be seen from (30)-(31) that $r^{(n-1)}$, $s_i^{(n-1)}$ ($i=1, \dots, N$) is a feasible solution of the program (32)-(35).

In view of the non-optimality of the current solution, $g_i^{(n)} < 0$ for at least one state i . Therefore, the optimal value $r^{(n)}$ of the program (32)-(35) has to be smaller than the feasible value $r^{(n-1)}$. Thus

$$r^{(n)} < r^{(n-1)} \quad (36)$$

and it is bounded by r^* as shown in theorem 1.

A similar proof would hold if the algorithm is stated fixing Player *I* instead of Player *II* but the sequence $r^{(n)}$ will be a monotonically increasing one bounded from above by r^* . It may be noticed that the above algorithm and the convergence proof have some similarity to those of Hoffman and Karp [3] for the case of NSG.

REFERENCES

1. C. DERMAN, *On Sequential Decision and Markov Chains*, Management Science, Vol. 9, 1962, pp. 16-24.
2. B. FOX, *Markov Renewal Programming by Linear Fractional Programming*, S.I.A.M. J. Appl. Math., Vol. 14, 1966, pp. 1418-1432.
3. A. J. HOFFMAN and R. M. KARP, *On Non-Terminating Stochastic Games*, Management Science, Vol. 12, 1966, pp. 359-370.
4. P. SCHWEITZER, *Iterative Solution of the Functional Equations of Undiscounted Markov Renewal Programming*, J. Math. Anal. and Appl., Vol. 34, 1971, pp. 495-501.
5. L. SHAPLEY, *Stochastic Games* Proceedings of the National Academy of Science, Vol. 39, 1953, pp. 1095-1100.