REVUE FRANÇAISE D'AUTOMATIQUE, D'INFORMATIQUE ET DE RECHERCHE OPÉRATIONNELLE. RECHERCHE OPÉRATIONNELLE

S. GEETHA
K. P. K. NAIR

Markovian assignment decision process

Revue française d'automatique, d'informatique et de recherche opérationnelle. Recherche opérationnelle, tome 26, n° 4 (1992), p. 421-428.

<http://www.numdam.org/item?id=RO_1992__26_4_421_0>

Recherche opérationnelle/Operations Research
(vol. 26, n° 4, 1992, p. 421 à 428)

SHORT NOTE

MARKOVIAN ASSIGNMENT DECISION PROCESS (*)

by S. GEETHA (1) and K. P. K. NAIR (1)

Communicated by S. OSAKI

Abstract. – *A finite-state, discrete-time Markovian decision process, in which, each action in each state is a feasible solution to a state dependent assignment problem, is considered. The objective is to maximize the additive rewards realized by the assignments over an infinite time horizon. In the undiscounted case, the average gain per transition and in the discounted case, the discounted total gain respectively, are maximized. Properties of optimal solutions in the two cases are characterized and finite algorithms are presented.*

Keywords : Markov decision processes; assignment problem; average gain; discounted reward; policy iteration; algorithm; optimal solution.

Résumé. – *Nous considérons un processus de décision markovien, à états finis, à temps discrets, dans lequel chaque action dans chaque état est une solution réalisable d'un problème d'affectation à états dépendants. L'objectif est de maximiser le profit additif dû aux affectations sur un horizon temporel infini. On maximise le gain moyen par transition dans le cas où il n'y a pas d'actualisation, et le gain total actualisé en présence d'actualisation. Pour ces deux cas, nous caractérisons les propriétés des solutions optimales, et nous présentons des algorithmes finis.*

Mots clés : Processus de décision markovien; problème d'affectation; gain moyen; profit actualisé; itération sur politique; algorithme; solution optimale.

1. INTRODUCTION

The finite-state, discrete-time Markovian decision process has been developed by Howard [1]. Here, a dynamic system is observed periodically at time points $t=0, 1, 2, \ldots$, and at each time point, it will be seen in any state i, $i=1, 2, \ldots, N$ of the set S. While in state i, there is a finite set K_i of alternatives of actions available for controlling the system. If action $k \in K_i$ is taken while in state i, the system moves to state j in the next step with probability

(*) Received December 1991, accepted May 1992.
This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada to K. P. K. Nair.
(1) Faculty of Administration, University of New Brunswick, Fredericton, N.B., Canada E3B 5A3.

Recherche opérationnelle/Operations Research, 0399-0559/92/04 421 08/$ 2.80

p_{ij}^k giving rise to a reward of r_i^k. The sequence of rewards are additive and the sum, if discounted or the average, if undiscounted can be optimized by the policy iteration method of Howard [1] or by linear programming as shown by Wolfe and Dantzig [12] and Derman [3]. Optimal policies in these cases are in pure and stationary forms. Also, the optimal value in each of the two cases are unique; however, the value in the discounted case is dependent on the starting state.

The stochastic games of Shapley [4] may be viewed as a generalization of the above decision process to the context of two-person zero-sum games. Similarly, the work of Aggarwal, Chandrasekaran and Nair [5] is a generalization to the context of ratio rewards that has several interesting applications. The ratio Markov decision processes, without and with discounting, have been further generalized to the context of games by the same authors [6, 7].

In the present work, we generalize the basic Markov decision process stated above to the context of the well known assignment problem in mathematical programming. Here, while in state i, the decision involves selecting a feasible solution of a state dependent assignment problem and after realizing a reward dependent on the feasible solution selected, the system moves to state j with a probability that is determined by the feasible solution selected for implementation. Characterizations of an optimal solution to this Markovian decision process, generalized to the context of assignment problem, and finite convergent algorithms for both the undiscounted and discounted cases are presented.

2. DESCRIPTION OF THE MARKOVIAN ASSIGNMENT DECISION PROCESS

A finite state Markovian assignment decision process (MADP) is a generalization of the well known finite state, finite action space, Markovian decision process to the context of the assignment problem of mathematical programming. In this generalization, associated with each state $i \in S$, there is an assignment problem A_i and the reward matrix of size $n_i \times n_i$ associated with A_i is denoted by $R_i = \{ r_i^{ql};\ q, l = 1, 2, \ldots, n_i \}$. Also related to each element r_i^{ql} is a transition probability vector $p_i^{ql} = \{ p_{il}^{ql}, p_{i2}^{ql}, \ldots, p_{ij}^{ql}, \ldots, p_{iN}^{ql} \}$. Now, define x_i^{ql} a zero-one decision variable associated with each element of R_i such that

$$\sum_q x_i^{ql} = 1, \ \forall\, l; \qquad \sum_l x_i^{ql} = 1, \ \forall\, q; \qquad x_i^{ql} = 0 \quad \text{or} \quad 1, \forall\, (q, l) \tag{1}$$

Thus, a zero-one vector x_i obeying (1) is a feasible solution to A_i and let the set of all feasible solutions to A_i, denoted by such vectors, be represented

by X_i. The number of such feasible solutions of A_i is finite and let this set be K_i. For each $k \in K_i$, the vector associated with it be x_i^k. The transition probability vector $p_i^{x_i}$ associated with the feasible solution

$$x_i = \{ x_i^{ql};\ q, l = 1, 2, \ldots, n_i \} \in X_i \text{ is computed as follows}$$

$$p_i^{x_i} = \left\{ p_{ij}^{x_i} = \frac{1}{n_i} \sum_q \sum_l p_{ij}^{ql} x_i^{ql}, j = 1, 2, \ldots, N \right\} \quad \text{for} \quad i = 1, 2, \ldots, N. \tag{2}$$

For $k \in K_i$, the transition probability vector associated with x_i^k is denoted by p_i^k.

Now, given the assignments begin in a specified state $i \in S$ at time period $t = 0$, the evolution of MADP may be represented by

$$\{ i_t, k \in K_{i_t}, r_{i_t}^k, p_{i_t}^k \}, \qquad t = 0, 1, 2, \ldots, \tag{3}$$

where i_t is the state occupied at step or period t, $k \in K_{i_t}$ is the feasible solution of the assignment problem A_{i_t} implemented, $r_{i_t}^k$ is the reward associated with $k \in K_{i_t}$ given by $\sum_q \sum_l r_{i_t}^{ql} x_{i_t}^{ql}$ and $p_{i_t}^k$ is the transition probability vector whose elements are given by $\frac{1}{n_i} \sum_q \sum_l p_{i_t j}^{ql} x_{i_t}^{ql}$ for $j = 1, 2, \ldots, N$. Thus, (3) reveals a sequence of rewards $r_{i_t}^k$, $t = 0, 1, 2 \ldots$ Here, two cases are of interest; in one we consider that the future rewards are discounted by a factor $\beta\ (0 \leqq \beta < 1)$ per period and in the other the rewards are not discounted. If discounting is applied, the relevant objective is maximization of the total discounted reward. If no discounting is applied, obviously, the total reward is unbounded over the infinite time periods and therefore, we maximize the limiting average reward or gain rate per transition. Assuming that the underlying Markov chain is a single ergodic one, the average gain rate would be unique for all starting states. In the discounted case, the total discounted reward will be dependent on the starting state.

3. CHARACTERIZATION OF OPTIMAL POLICIES

In this section, we establish certain properties of optimal policies in the two cases with and without discounting and characterize an optimal policy in each case respectively. A policy is a selection of a feasible solution of the assignment problem in each of the states in S. This characterization would be helpful in developing the respective algorithms.

Firstly, we note that the structure of the MADP is identical to the Markovian decision process of Howard [1]. This follows from the fact that while in state i, there is a finite set K_i of alternatives of feasible solutions to A_i and any choice $k \in K_i$ can be implemented and the associated reward and transition probability vector are r_i^k and p_i^k. The fact that these are not explicitly known is of no consequence as far as the properties of optimal solutions in the discounted and undiscounted cases are concerned, though it may cause some changes in the algorithmic steps. In view of the above, the following theorem holds and the proof is straight forward.

THEOREM 1: *In the discounted and undiscounted MADP, there exist stationry and pure optimal policies.*

In the theorem, stationarity means that at every time the system is in a particular state, the same action (feasible solution to the assignment problem) is applicable for implementation. Also, purity implies that mixture of feasible solutions of the assignment problem in the state is not applied. Thus, Theorem 1 is a direct generalization of the properties used by Howard [1] and later proved rigorously by Blackwell [8].

Let $V_i^{x_i} (i = 1, 2, \ldots, N)$ be the total discounted reward if the system is started in state i under a stationary policy $x = (x_1, x_2, \ldots, x_i, \ldots, x_N)$. Then, $V_i^{x_i}$'s uniquely satisfy the relation

$$V_i^{x_i} = \sum_q \sum_l r_i^{ql} x_i^{ql} + \beta \sum_{j=1}^{N} p_{ij}^{x_i} V_j^{x_i}, \qquad i = 1, 2, \ldots, N. \tag{4}$$

THEOREM 2: *A policy $x^* = (x_1^*, x_2^*, \ldots, x_i^*, \ldots, x_N^*)$ is optimal and the associated V_i^* for $i = 1, 2, \ldots, N$, are maximal, in the discounted case, if and only if,*

$$V_i^* = \sum_q \sum_l r_i^{ql} x_i^{*ql} + \beta \sum_{j=1}^{N} p_{ij}^{x_i^*} V_j^*, \qquad i = 1, 2, \ldots, N \tag{5}$$

and

$$V_i^* = \max_{x_i \in X_i} \left\{ \sum_q \sum_l \left[r_i^{ql} + \beta \sum_{j=1}^{N} p_{ij}^{ql} V_j^* \right] x_i^{ql} \right\}, \qquad i = 1, 2, \ldots, N \tag{6}$$

with maximum attained at $x_i^* \in X_i$.

The validity of theorem 2 is clear from the structural similarity of the current problem to the problem considered by Howard [1] and Blackwell [8].

Thus, Theorem 2 provides a useful characterization of an optimal policy in the discounted case.

As stated earlier, in the undiscounted case, we assume that the underlying Markov chain is an ergodic chain. Here the objective is to maximize the limiting average gain rate per transition and it is independent of the starting state.

For a given policy $x=(x_1, x_2, \ldots, x_i, \ldots, x_N)$, we have

$$g^{x_i}+v_i^{x_i}=\sum_q \sum_l r_i^{ql} x_i^{ql}+\sum_{j=1}^{N} p_{ij}^{x_i} v_j^{x_i}, \qquad i=1,2,\ldots,N \tag{7}$$

In the above, v_i's may be called relative values dependent on the starting state as in the Markov decision process of Howard [1].

THEOREM 3: *A policy $x^*=(x_1^*, x_2^*, \ldots, x_i^*, \ldots, x_N^*)$ is optimal and the associated gain rate g^* is maximal, in the undiscounted case, if and only if,*

$$g^*+v_i^*=\sum_q \sum_l r_i^{ql} x_i^{*ql}+\sum_{j=1}^{N} p_{ij}^{x_i^*} v_j^*, \qquad i=1,2,\ldots,N \tag{8}$$

and

$$g^*+v_i^*=\max_{x_i \in X_i}\left\{\sum_q \sum_l \left[r_i^{ql}+\sum_{j=1}^{N} p_{ij}^{ql} v_j^*\right] x_i^{ql}\right\}, \qquad i=1,2,\ldots,N \tag{9}$$

with the maximum attained at $x_i^* \in X_i$.

The validity of Theorem 3 follows similar to that of Theorem 2. Now, Theorems 2 and 3 respectively lead to the development of the algorithms for computing optimal policies in the respective cases.

4. ALGORITHMS

The algorithms for the respective cases are direct generalizations of those given by Howard [1], to the context of the assignment problem and as such their validation is straight forward and therefore, are not included here. Basically, each algorithm involves repeated application of two steps, one of solving a system of equations (value determination) and the other of solving a set of assignment problems (policy improvement) until convergence is realized. Algorithms for the discounted and undiscounted cases are given below.

Algorithm I (*Discounted Case*)

Step 0: (Initialization)

Set $k=0$, $V_i^{(k)}=0$ for $i=1,2,\ldots,N$. Go to Step 2.

Step 1: (Value determination)

Solve the system

$$V_i^{(k)}=\sum_q \sum_l r_i^{ql}\, x_i^{(k)\,ql}+\beta \frac{1}{n_i} \sum_{j=1}^{N} \sum_q \sum_l p_{ij}^{ql}\, x_i^{(k)\,ql}\, V_j^{(k)}, \qquad i=1,2,\ldots,N,$$

to obtain the unique solution $V_i^{(k)}$, $i=1,2,\ldots,N$. Go to Step 2.

Step 2: (Policy improvement)

Solve the N assignment problems for $i=1,2,\ldots,N$, where the reward matrix of the i-th problem is as follows:

$$\left\{ r_i^{ql}+\beta \sum_{j=1}^{N} p_{ij}^{ql}\, V_j^{(k)} \right\}, q, l=1,2,\ldots,n_i$$

and let the optimal solution be denoted by $x_i^{(k)}$ and the optimal values be G_i^k. If $x_i^{(k)}=x_i^{(k-1)}$ or $G_i^k=V_i^{(k)}$ for all i and $k\geqq 1$, terminate; in this case, $x=\{x_i^{(k)}, i=1,2,\ldots,N\}$ is the optimal policy with values $V_i^{(k)}$, $i=1,2,\ldots,N$.

Otherwise, go to Step 1 with $x_i^{(k)}$, $i=1,2,\ldots,N$, setting $k=k+1$.

Algorithm II (*Undiscounted Case*)

Step 0: (Initialization)

Set $k=0$, $v_i^{(k)}=0$ for $i=1,2,\ldots,N$. Go to Step 2.

Step 1: (Value determination)

By setting $v_N^{(k)}=0$, for the remaining N variables, g and $v_i^{(k)}$, $i=1,2,\ldots,N-1$, solve the system,

$$g+v_i^{(k)}=\sum_q \sum_l r_i^{ql}\, x_i^{(k)\,ql}+\frac{1}{n_i} \sum_{j=1}^{N} \sum_q \sum_l p_{ij}^{ql}\, x_i^{(k)\,ql}\, v_j^{(k)}, \qquad i=1,2,\ldots,N.$$

Go to Step 2.

Step 2: (Policy Improvement)

Solve the N assignment problems for $i=1,2,\ldots,N$, where the reward matrix of the i-th problem is as follows:

$$\left\{ r_i^{ql}+\sum_{j=1}^{N} p_{ij}^{ql}\, v_j^{(k)} \right\}, \qquad q, l=1,2,\ldots,n_i$$

and let the optimal solution be denoted by $x_i^{(k)}$. Let the relative values be $v_i^{(k)}$ and the gain rate be $g^{(k)}$. If $x_i^{(k)} = x_i^{(k-1)}$ for all i or $g^{(k)} = g^{(k-1)}$ for $k \geqq 1$, terminate; in this case, $x = \{ x_i^{(k)}, i = 1, 2, \ldots, N \}$ is the optimal policy with the optimal gain rate $g^{(k)}$. Otherwise, go to Step 1 with $x_i^{(k)}$, $i = 1, 2, \ldots, N$, setting $k = k + 1$.

Validation of the algorithms

The validation of the algorithms follows analogous to those of the algorithms of Howard [1]. The only difference is in the policy improvement operation which involves solving a set of assignment problems instead of finding the maximum of a set of linearly defined quantities in each of the states.

Each algorithm is finite since there are only a finite number of policies in the system and no policy is repeated in the algorithms. Also, the operation in each iteration is polynomial of $0(Nn^3)$.

5. CONCLUSION

In this paper, the well known Markov decision process has been generalized to the context of assignment problem of mathematical programming. The algorithms presented for both discounted and undiscounted cases are finite and the efforts required in each iteration is polynomical of $0(Nn^3)$. The work reveals that the Markov decision process is amenable for generalization to any kind of decision process so long as the process applicable to each state has a well defined structure. Further work in this direction would be of interest both from practical and theoretical points of view. As already known in the case of the basic Markov decision processes, one may develop linear programming algorithms [3, 9] for the above cases as well.

REFERENCES

1. R. Howard, Dynamic Programming and Markov Processes, *John Wiley and Sons*, New York, 1960.
2. P. Wolfe and G. B. Dantzig, Linear Programming in a Markov Chain, *Oper. Res.*, 1962, *10*, p. 702-710.
3. C. Derman, Finite State Markovian Decision Processes, *Academic Press*, New York, 1970.
4. L. S. Shapley, Stochastic Games, *Proceedings* of the National Academy of Sciences, 1953, *39*, p. 1095-1100.

5. V. AGGARWAL, R. CHANDRASEKARAN and K. P. K. NAIR, Markov Ratio Decision Processes, *J. Optimi. Theory Appl.*, 1977, *21*, p. 27-37.
6. V. AGGARWAL, R. CHANDRASEKARAN and K. P. K. NAIR, Discounted Stochastic Ratio Games, *S.I.A.M. J. Algebraic Discr. Meth.*, 1980, *1*, p. 201-210.
7. V. AGGARWAL, R. CHANDRASEKARAN and K. P. K. NAIR, Non-Terminating Stochastic Ratio Games, *R.A.I.R.O. Rech. Opér.*, 1980, *14*, p. 21-30.
8. D. BLACKWELL, Discrete Dynamic Programming, *Ann. of Math. Statist.*, 1962, *33*, p. 719-726.
9. H. MINE and S. OSAKI, Markovian Decision Processes, *American Elsevier Publishing Company, Inc.*, New York, 1970.