数理解析研究所講究録
第835巻 1993年 165-172

Markov Bidecision Processes

岩本 誠一 (Seiichi IWAMOTO)

九州大学 経済学部 経済工学科

1 Introduction

In this paper we consider the Markov decision processes on finite state and action spaces, which have the multiplicatively additive reward system. Our multilpicative additivity does not come from the usual 'discount factor' but from a new notion "discount function" . The discount function depends on the current stage, state and action. Furthermore, it may also take negative values. Our monotonicity becomes either monotone nonincreasingness or monotone nondecreasingness according as nonpositiveness or nonnegativeness. On the basis of both the usual recursiveness and our monotonicity, we must consider simultaneously both maximum and minimum subproblems.

2 Bimax Theorem

Throughout the paper we use , for two sets of real values $\{a, b\}$ and $\{a_1, a_2, \ldots, a_n\}$, the following notations for their maxima and minima, respectively:

$$a \vee b = \max\{a, b\}, \qquad a \wedge b = \min\{a, b\}$$

$$\bigvee_{i=1}^{n} a_i = \max\{a_1, a_2, \ldots, a_n\}, \quad \bigwedge_{i=1}^{n} a_i = \min\{a_1, a_2, \ldots, a_n\}.$$

Let X, Y and Z be three nonempty sets. Let X be the disjoint union of X^-, X^+:

$$X = X^- + X^+.$$

Theorem 1 *Let $g : X \times R^1 \to R^1$ satisfy* (i) *for $x \in X^-$ $g(x;\cdot) : R^1 \to R^1$ is nonincreasing, and* (ii) *for $x \in X^+$ $g(x;\cdot) : R^1 \to R^1$ is nondecreasing. Let $h : X \times Y \times Z \to R^1$ be a real-valued function. If the right-hand two-stage optima in (1),(2) exist, then the left-hand simultaneous optima $\mathrm{Max}_{x,y,z} g(x; h(x; y, z))$ and $\min_{x,y,z} g(x; h(x; y, z))$ exist, and the following two equalities hold, respectively:*

$$\begin{aligned} \mathrm{Max}_{x,y,z} g(x; h(x; y, z)) \;=\; & \mathrm{Max}_{x \in X^-} g(x; \min_{y,z} h(x; y, z)) \\ & \vee \mathrm{Max}_{x \in X^+} g(x; \mathrm{Max}_{y,z} h(x; y, z)) \end{aligned} \tag{1}$$

$$\begin{aligned}\min_{x,y,z} g(x; h(x; y, z)) \;=\; & \min_{x\in X^-} g(x; \mathrm{Max}_{y,z} h(x; y, z)) \\ & \vee \min_{x\in X^+} g(x; \min_{y,z} h(x; y, z)) \qquad (2)\end{aligned}$$

where, unless specified, the suffices x, y and z range over X, Y and Z, respectively.

Corollary 1 *Let*

$$g(x; h) = r(x) + \beta(x)h : X \times R^1 \to R^1$$

$$h(x; y, z) = U(y)p(x) + V(z)q(x) : X \times Y \times Z \to R^1$$

where

$$r,\ \beta : R^1 \to R^1, \quad U : Y \to R^1, \quad V : Z \to R^1$$

and

$$p(x) + q(x) = 1, \quad p(x) \geqq 0 \ \ for \ \ x \in X.$$

If the right-hand two-stage optima in (3),(4) exist, then the left-hand simultaneous optima exist, and both are equal, respectively:

$$\begin{aligned}& \mathrm{Max}_{x,y,z}[r(x) + \beta(x)[U(y)p(x) + V(z)q(x)]] \\ = \;& \mathrm{Max}_{x\in X^-}[r(x) + \beta(x)\min_{y,z}[U(y)p(x) + V(z)q(x)]] \\ & \vee \mathrm{Max}_{x\in X^+}[r(x) + \beta(x)\mathrm{Max}_{y,z}[U(y)p(x) + V(z)q(x)]] \qquad (3)\end{aligned}$$

$$\begin{aligned}& \min_{x,y,z}[r(x) + \beta(x)[U(y)p(x) + V(z)q(x)]] \\ = \;& \min_{x\in X^-}[r(x) + \beta(x)\mathrm{Max}_{y,z}[U(y)p(x) + V(z)q(x)]] \\ & \vee \min_{x\in X^+}[r(x) + \beta(x)\min_{y,z}[U(y)p(x) + V(z)q(x)]]. \qquad (4)\end{aligned}$$

3 Finite-stage Markov Bidecision Processes

A finite-stage Markov bidecision process (BDP) is specified by a five-tuple:

$$\mathbf{B} = (\ \mathrm{Opt},\ \{S_n\}_1^{N+1},\ \{A_n\}_1^N,\ (\{r_n\}_1^N,\ \{\beta_n\}_1^N,\ k),\ \{q_n\}_1^N\ )$$

where

(*i*) N is a positive integer, *total number of stages.* The subscript n ranges $1 \leqq n \leqq N (\mathrm{or} N+1)$. It specifies the current number of stage.

(*ii*) S_n is a nonempty finite set, *n-th state space.* Its elements $s_n, s_n^* \in S_n$ are called *n-th states.* s_1 is an initial state. s_{N+1} is a terminal state.

(*iii*) A_n is a nonempty finite set, *n-th action space.* Let $A_n(s_n) \subset A_n$ be a nonempty subset, *n-th feasible action space at state s_n.* Its elements $a_n, a_n^* \in A_n(s_n)$ are called *n-th actions at state s_n.*

(*iv*) $r_n : S_n \times A_n \to R^1$ is an *n-th reward function.*

(*v*) $\beta_n : S_n \times A_n \to R^1$ is an *n-th discount function.* We don't always assume that

$\beta_n(s_n, a_n) \geqq 0$. The constant discount function $\beta_n(s_n, a_n) = \beta \geqq 0$ has been called the *discount factor*, as has been frequently used.

(vi) $k : S_{N+1} \to R^1$ is a *terminal reward function.* The three-tuple $(\{r_n\}_1^N, \{\beta_n\}_1^N, k)$ is called a *reward system.*

(vii) $q_n = q_n(s_{n+1} | s_n, a_n)$ is an *n-th Markov transition law* from s_n onto S_{n+1} depending on the current action a_n. When the system is in state s_n on stage n and action a_n is chosen, the next state will become s_{n+1} with probability $q_n(s_{n+1} | s_n, a_n) \geqq 0$. We write $s_{n+1} \sim q_n(\cdot | s_n, a_n)$.

(viii) Opt denotes either Max or min, *optimizer.* It means that BDP **B** represents the stochastic optimization problem:

$$\begin{aligned} \text{Opt} \sum_{s_1 \in S_1} \sum_{s_2 \in S_2} \cdots \sum_{s_{N+1} \in S_{N+1}} & I_1(s_1, a_1, s_2, a_2, ..., s_N, a_N, s_{N+1}) \\ & \times q_1(s_2 | s_1, a_1) q_2(s_3 | s_2, a_2) ... q_N(s_{N+1} | s_N, a_N) \\ \text{s.t. (i)} \quad & s_{n+1} \sim q_n(\cdot | s_n, a_n) \quad 1 \leq n \leq N \\ \text{(ii)} \quad & a_n \in A_n(s_n) \quad 1 \leq n \leq N. \end{aligned} \tag{5}$$

Here

$$\begin{aligned} & I_n(s_n, a_n, s_{n+1}, a_{n+1}, ..., s_N, a_N, s_{N+1}) \\ = \; & r_n + \beta_n r_{n+1} + \beta_n \beta_{n+1} r_{n+2} + ... + \beta_n \beta_{n+1} ... \beta_{N-1} r_N + \beta_n \beta_{n+1} ... \beta_N k \end{aligned}$$

where

$$r_n = r_n(s_n, a_n), \quad \beta_n = \beta_n(s_n, a_n), \quad k = k(s_{N+1}).$$

Let

$$\nu_n : S_1 \times S_2 \times \cdots \times S_n \to A_n \quad \nu_n(s_1, s_2, ..., s_n) \in A_n(s_n) \quad s_n \in S_n$$

be an *n-th (not necessarily Markovian) decision function.* A sequence $\nu = \{\nu_1, \nu_2, ..., \nu_N\}$ is called *strategy.* Thus let E^ν be the expectation (integral) operator on $S_2 \times S_3 \times \cdots \times S_{N+1}$ with an initial state $s_1 \in S_1$.

Therefore the problem (5) is rewritten as follows.

$$\text{Opt} \; E^\nu[I_1(s_1, a_1, s_2, a_2, ..., s_N, a_N, s_{N+1}) \mid (i), (ii) \quad 1 \leq n \leq N] \tag{6}$$

The aim of BDP **B** is to find two optimal strategies in the following sense. A strategy $\lambda = \{\lambda_1, \lambda_2, ..., \lambda_N\}$ is called *maximal* if for each $s_1 \in S_1$ λ attains the maximum value of Maximum Problem (6). Another strategy $\mu = \{\mu_1, \mu_2, ..., \mu_N\}$ is called *minimal* if for each $s_1 \in S_1$ μ attains the minimum value of minimum Problem (6).

A *policy* is an ordered set of N decision functions $\pi = \{\pi_1, \pi_2, ..., \pi_N\}$ where $\pi_n : S_n \to A_n$ with the feasibility $\pi_n(s_n) \in A_n(s_n)$ $s_n \in S_n$ is an *n-th (Markovian) decision function.*

Given BDP **B**, we for each $n (1 \leqq n \leqq N)$ define the following maximum and minimum problems together with maximum and minimum values, respectively:

$$U^{N-n+1}(s_n) = \text{Max } E^{\nu}[I_n(s_n, a_n, ..., s_N, a_N, s_{N+1}) \mid \text{(i)}, \text{(ii)}]$$

$$u^{N-n+1}(s_n) = \min E^{\nu}[I_n(s_n, a_n, ..., s_N, a_N, s_{N+1}) \mid \text{(i)}, \text{(ii)}]$$

Here

$$\begin{aligned} &\text{(i)} \quad s_{m+1} \sim q_m(\cdot \mid s_m, a_m) \quad n \leq m \leq N \\ &\text{(ii)} \quad a_m \in A_m(s_m) \quad n \leq m \leq N \end{aligned}$$

and

$$\nu = \{\nu_m, \nu_{m+1}, ..., \nu_N\}$$

where

$$\nu_m : S_n \times S_{n+1} \times \cdots \times S_m \to A_m \quad \nu_m(s_n, s_{n+1}, ..., s_m) \in A_m(s_m) \quad s_m \in S_m.$$

Then we have the following system of two alternate-recursive formulae between *maximum reward functions* $\{U^0, U^1, ..., U^N\}$ and *minimum ones* $\{u^0, u^1, ..., u^N\}$.

Theorem 2 *(Bicursive Formula)*

$$U^{N-n+1}(s) = \text{Max}_{a;-} T_n(s, a; u^{N-n}) \vee \text{Max}_{a;+} T_n(s, a; U^{N-n}) \quad s \in S_n \tag{7}$$

$$u^{N-n+1}(s) = \min_{a;-} T_n(s, a; U^{N-n}) \wedge \min_{a;+} T_n(s, a; u^{N-n}) \quad s \in S_n \tag{8}$$

$$U^0(s) = u^0(s) = k(s) \qquad s \in S_{N+1}$$

where

$$T_n(s, a; w) = r_n(s, a) + \beta_n(s, a) \sum_{t \in S} w(t) q(t \mid s, a)$$

$$\hat{A}_n(s) = \{a \in A_n(s) \mid \beta_n(s, a) \leqq 0\}$$

$$A_n^*(s) = \{a \in A_n(s) \mid \beta_n(s, a) > 0\}$$

and

$$a;-, \quad a;+ \quad \textit{denote} \quad a \in \hat{A}_n(s), \quad a \in A_n^*(s), \quad \textit{respectively.}$$

Further letting

$$\pi_n^*(s_n) = \text{an } a \text{ which attains the maximum of (7)}$$

$$\hat{\sigma}_n(s_n) = \text{an } a \text{ which attains the minimum of (8)}$$

we have the 'maximal' n-th decision function π_n^* and the 'minimal' n-th decision function $\hat{\sigma}_n$ in the remaining $(N - n + 1)$ satges.

4 Infinite-stage Bidecision Processes

We consider an infinite-horizon *stationary* Markov bi-decision process on finite state and action spaces:

$$\mathbf{B} = (\text{Opt},\ \{S_n\}_1^\infty,\ \{A_n\}_1^\infty,\ (\{r_n\}_1^\infty,\ \{\beta_n\}_1^\infty),\ \{q_n\}_1^\infty)$$

where

$$\begin{array}{ccc} & S_n = S = \{1, 2, ..., N\} & \\ A_n = A, & A_n(i) = \{1,2,...,K_i\} & i = 1, 2, ..., N \\ r_n = r, & \beta_n = \beta, & q_n = q. \end{array}$$

The Markov bidecision process **B** represents the optimization problem

$$\begin{array}{ll} \text{Opt} & E^\nu[r_1 + \beta_1 r_2 + \beta_1\beta_2 r_3 + \cdots + \beta_1\beta_2\cdots\beta_n r_{n+1} + \cdots] \\ \text{s.t.} & \text{(i)} \quad s_{n+1} \sim q_n(\cdot|s_n, a_n) \\ & \text{(ii)} \quad a_n \in A_n(s_n) \quad n = 1, 2, ... \end{array}$$

where

$$\beta_n = \beta(s_n, a_n), \qquad r_n = r(s_n, a_n).$$

We assume that there exist m, M such that

$$-1 < m \le \beta(i,k) \le M < 1 \qquad 1 \le k \le K_i,\ 1 \le i \le N. \tag{9}$$

Let $U(i)$ be the maximum expected value from an initial state i and $u(i)$ be the minimum. Then we have the following system of two alternate-recursive equations:

Theorem 3 *(Bicursive Equation)*

$$\begin{aligned} U(i) \ = \ & \bigvee_{\beta(i,k)\le 0} \left(r(i,k) + \beta(i,k)\sum_{j=1}^{N} u(j)q(j|i,k)\right) \\ \vee \ & \bigvee_{\beta(i,k)>0} \left(r(i,k) + \beta(i,k)\sum_{j=1}^{N} U(j)q(j|i,k)\right) \qquad (10) \\ & \qquad\qquad i = 1, 2, ..., N \end{aligned}$$

$$\begin{aligned} u(i) \ = \ & \bigwedge_{\beta(i,k)\le 0} \left(r(i,k) + \beta(i,k)\sum_{j=1}^{N} U(j)q(j|i,k)\right) \\ \wedge \ & \bigwedge_{\beta(i,k)>0} \left(r(i,k) + \beta(i,k)\sum_{j=1}^{N} u(j)q(j|i,k)\right). \qquad (11) \end{aligned}$$

We call this system *optimality equation* in Markov bidecision process.

Let $W = (U, u)'$ be any $2N$-dimensional real column-vector with $U' = (U_1, ..., U_N)$ and $u' = (u_1, ..., u_N)$, where $'$ is tranpose. We take the maximum norm $\|W\| = \|U\| \vee \|u\|$, where

$$\|U\| = \bigvee_{i=1}^{N} |U_i| , \quad \|u\| = \bigvee_{i=1}^{N} |u_i|.$$

Let TW be the $2N$-dimensional real column-vector composed of the right-hand sides of (10),(11).

Theorem 4 *The operator $T : R^{2N} \to R^{2N}$ is a contraction mapping with the contraction coefficient $\beta < 1$, where*

$$\beta = \mathrm{Max}_{s,a;-}|\beta(s,a)| \vee \mathrm{Max}_{s,a;+}\beta(s,a).$$

Here $s, a; -$ and $s, a; +$ denote $a \in A(s)$, $\beta(s,a) \leqq 0$, $s \in S$ and $a \in A(s)$, $\beta(s,a) > 0$, $s \in S$, respectively. Therefore the bicursive equation (10),(11) has a unique solution in R^{2N}.

Theorem 5 *(Successive Approximation) Take any W_0 in R^{2N}. If we generate the sequence $\{W_n\}$ by*

$$W_{n+1} = TW_n \qquad n \geqq 0$$

then $\{W_n\}$ converges to the unique solution W of (10),(11), that is,

$$TW = W.$$

5 Bi-policy Iteration Algorithm

We have the following algorithm for finding the unique solution.
Bi-policy Iteration Algorithm (BIP)
step 1 (initial selection) Let $n = 0$. Take any pair of feasible selections (f_0, F_0).
step 2 (value determination) Calculate $X^n = X(f_n, F_n) =$
$(X_1(f_n, F_n), \ldots, X_N(f_n, F_n))$ and $x^n = x(f_n, F_n) = (x_1(f_n, F_n), \ldots, x_N(f_n, F_n))$ satisfying the system of $2N$ linear equations :

$$X_i^n = \begin{cases} \beta(i, F_n(i)) \sum_{j=1}^{N} q_{ij}^{F_n(i)} x_j^n + r_i^{F_n(i)} & \text{for } \beta(i, F_n(i)) \leq 0 \\ \beta(i, F_n(i)) \sum_{j=1}^{N} q_{ij}^{F_n(i)} X_j^n + r_i^{F_n(i)} & \text{for } \beta(i, F_n(i))) > 0 \end{cases}$$

and $i = 1, 2, ..., N$

$$x_i^n = \begin{cases} \beta(i, f_n(i)) \sum_{j=1}^{N} q_{ij}^{f_n(i)} X_j^n + r_i^{f_n(i)} & \text{for } \beta(i, f_n(i)) \leq 0 \\ \beta(i, f_n(i)) \sum_{j=1}^{N} q_{ij}^{f_n(i)} x_j^n + r_i^{f_n(i)} & \text{for } \beta(i, f_n(i))) > 0. \end{cases}$$

step 3 (optimality test) If (X^n, x^n) satisfies

$$X_i^n = \bigvee_{\beta(i,k) \leq 0} (\beta(i,k) \sum_{j=1}^{N} q_{ij}^k x_j^n + r_i^k) \vee \bigvee_{\beta(i,k) > 0} (\beta(i,k) \sum_{j=1}^{N} q_{ij}^k X_j^n + r_i^k)$$

$$i = 1, 2, ..., N$$

$$x_i^n = \bigwedge_{\beta(i,k)\leq 0} (\beta(i,k) \sum_{j=1}^{N} q_{ij}^k X_j^n + r_i^k) \wedge \bigwedge_{\beta(i,k)>0} (\beta(i,k) \sum_{j=1}^{N} q_{ij}^k x_j^n + r_i^k),$$

then go to step 6. Otherwise, go to step 4.

step 4 (selection improvement) Choose a pair of feasible selections (F_{n+1}, f_{n+1}) satisfying

$$\bigvee_{\beta(i,k)\leq 0} (\beta(i,k) \sum_{j=1}^{N} q_{ij}^k x_j^n + r_i^k) \vee \bigvee_{\beta(i,k)>0} (\beta(i,k) \sum_{j=1}^{N} q_{ij}^k X_j^n + r_i^k)$$

$$= \begin{cases} \beta(i, F_{n+1}(i)) \sum_{j=1}^{N} q_{ij}^{F_{n+1}(i)} x_j^n + r_i^{F_{n+1}(i)} & \text{for } \beta(i, F_{n+1}(i)) \leq 0 \\ \beta(i, F_{n+1}(i)) \sum_{j=1}^{N} q_{ij}^{F_{n+1}(i)} X_j^n + r_i^{F_{n+1}(i)} & \text{for } \beta(i, F_{n+1}(i)) > 0, \end{cases}$$

and

$$i = 1, 2, ..., N$$

$$\bigwedge_{\beta(i,k)\leq 0} (\beta(i,k) \sum_{j=1}^{N} q_{ij}^k X_j^n + r_i^k) \wedge \bigwedge_{\beta(i,k)>0} (\beta(i,k) \sum_{j=1}^{N} q_{ij}^k x_j^n + r_i^k)$$

$$= \begin{cases} \beta(i, f_{n+1}(i)) \sum_{j=1}^{N} q_{ij}^{f_{n+1}(i)} X_j^n + r_i^{f_{n+1}(i)} & \text{for } \beta(i, f_{n+1}(i)) \leq 0 \\ \beta(i, f_{n+1}(i)) \sum_{j=1}^{N} q_{ij}^{f_{n+1}(i)} x_j^n + r_i^{f_{n+1}(i)} & \text{for } \beta(i, f_{n+1}(i)) > 0. \end{cases}$$

step 5 (next step) Let $n = n + 1$. Go to step 2.

step 6 (optimal solution) The pair of selections (F_n, f_n) is optimal and (X^n, x^n) is the desired solution.

References

[1] R. Bellman, *Dynamic Programming*, Princeton Univ. Press, NJ, 1957.

[2] D. Blackwell, Discounted dynamic programming, Ann. Math. Stat. **36**(1965), 226-235.

[3] E.V. Denardo, Contraction mappings in the theory underlying dynamic programming, SIAM Review **9**(1968), 165-177.

[4] E.V. Denardo, *Dynamic Programming : Models and Applications*,Prentice-Hall, N.J., 1982.

[5] F. Furukawa and S. Iwamoto, Markovian decision processes with recursive reward functions, Bull. Math. Statist. **15**(1973), 79-91.

[6] F. Furukawa and S. Iwamoto, Dynamic programming on recursive reward systems, Bull. Math. Statist. **17**(1976), 103-126.

[7] R.A. Howard, *Dynamic Programming and Markov Processes*, John Wiley and Sons, New York, 1960.

[8] S. Iwamoto, Finite-horizon Markov games with recursive payoff systems, Mem. Fac. Sci. Kyushu Univ. Ser. A **29**(1975), 123-147.

[9] S. Iwamoto, The second principle of optimality, Bull. Math. Statist. **17**(1977), 104-114.

[10] S. Iwamoto, Sequential minimaximization under dynamic programming structure, J. Math. Anal. Appl. **108**(1985), 267-282.

[11] S. Iwamoto, From dynamic programming to bynamic programming, to appear in J. Math. Anal. Appl..

[12] S. Iwamoto, On minimax linear equations, in preparation.

[13] L.G. Mitten, Composition principles for sysnthesis of optimal multi-stage processes, Operations Res. **12**(1964), 610-619.

[14] G.L. Nemhauser, *Introduction to Dynamic Programming*, Wiley, New York, 1966.

[15] N.L. Stokey and R.E. Lucas,Jr.(with E.C. Prescott), *Recursive Methods in Economic Dynamics*, Harvard Univ. Press, Cambridge, MA, 1989.