数理解析研究所講究録
第1939巻 2015年 189-195

相互依存型マルコフ決定過程 − 結合型評価 −

九州工業大学・大学院工学研究院　藤田 敏治
Toshiharu Fujita
Graduate School of Engineering, Kyushu Institute of Technology

1 はじめに

相互依存型決定過程 [6, 8] とは，複数の決定過程から構成され、その利得関数を通して互いに再帰的に依存する関係を持つものである。これは、nonserial な状態推移 [3, 15] をもつ決定過程の一種と考えることもでき、ある種の枝分かれ構造をもつ状態推移上に、特殊な決定過程構造を構成した問題と解釈することも可能である。ここで導入する依存構造は，各決定過程問題の利得関数が他の決定過程問題群の最適値の関数として定まるものであり，その際、他の決定過程の初期状態は，もとの決定過程における各期の状態と決定により確定的に定まる．このような依存構造が再帰的に与えられ，また、各決定過程における終端状態に対しては，通常の終端利得関数が割り当てられる．

この種の相互依存構造により、多角形から凸多面体を構成する問題がうまく扱えることは [7] で示している。相互依存型決定過程を用いることで、ある種の非常に複雑な再帰的構造をもつ問題が、比較的容易に表現可能となるのである。

2 定式化

初期状態 $x_0 \in X_i \backslash T_i$ に対する次の m 個の決定過程問題を定める $(i = 1, 2, \ldots, m)$.

$$\begin{aligned} \mathrm{P}_i(x_0) \quad & \text{Maximize } E^{\sigma}\left[r_i(x_0, u_0) \circ_i r_i(x_1, u_1) \circ_i \cdots \circ_i r_i(x_{N-1}, u_{N-1}) \circ_i k_i(x_N)\right] \\ & \text{subject to} \quad x_{n+1} \sim p_i(\cdot\,|x_n, u_n) \quad n = 0, 1, \ldots, N-1 \\ & \qquad \sigma = (\sigma_0, \sigma_1, \ldots, \sigma_{N-1}) \in \Sigma_i \\ & \qquad (\ N = N(x_0, u_0, x_1, u_1, \ldots) = \max\{n : x_n \notin T_i\} + 1\ ), \end{aligned}$$

ただし

(1) X_i は有限状態空間をあらわし，$T_i \subset X_i$ は終了状態集合とする。また、$x_n\ (\in X_i)$ は時刻 n における状態をあらわし、$x_N \in T_i$ となる時刻 N で推移は終了する。

(2) U_i は有限決定空間をあらわし、$u_n \in U_i$ は時刻 n における決定をあらわす。以下、集合 S に対し

$$2^S = \{A : \text{a set} \,|\, A \subset S\}$$

とおく。ここで、写像 $U_i : X \setminus T_i \to 2^{U_i} \setminus \{\phi\}$ は各非終了状態に対し実行可能な決定全体を与えるものとする。すなわち、$U_i(x)$ は状態 $x \in X_i \backslash T_i$ に対し選択可能な決定全体をあらわす。

(3) $r_i : G_r(U_i) \to D_i$ は利得関数をあらわし、$k_i : X_i \to D_i$ は終端利得関数をあらわす。ただし、$D_i \subset \mathbf{R}$ で

$$G_r(U_i) = \{(x, u) \,|\, u \in U_i(x),\ x \in X_i \setminus T_i\}$$

である。

(4) $\circ_i : D_i \times D_i \to D_i$ は結合律を満たす2項演算子をあらわし、単位元 $\tilde{\lambda}_i \in D_i$ の存在を仮定する。

(5) $p_i(\cdot\,|\,x,u)$, $(x,u) \in G_r(U_i)$ はマルコフ推移法則をあらわす。すなわち、ある時刻において状態 x に対し決定 u を選んだ際，次の時刻で状態 $y \in X_i$ へ確率 $p_i(y\,|\,x,u)$ で推移する。この状態推移を $y \sim p_i(\cdot\,|\,x,u)$ とあらわす。

(6) $\sigma_n : X_i^{n+1} \to U_i, n = 0,1,\ldots,N-1$ は時刻 n における一般決定関数をあらわし、各期の決定 $u_n = \sigma_n(x_0,x_1,\ldots,x_n) \in U_i(x_n)$ はその時刻までの状態列に依存して定まる．また一般決定関数列 $\sigma = (\sigma_0,\sigma_1,\ldots,\sigma_{N-1})$ を一般政策と呼び，一般政策全体を Σ_i で表す．

各決定過程において，履歴 $\{u_0,x_1,u_1,x_2,\ldots,u_{N-1},x_N\}$ は，一般政策 $\sigma = (\sigma_0,\sigma_1,\ldots,\sigma_{N-1}) \in \Sigma_i$ および初期状態 $x_0 \in X_i \backslash T_i$ により次のように確率的に生成される：

$$\begin{array}{lll}
\sigma_0(x_0) = u_0 & \to & p_i(\cdot\,|x_0,u_0) \sim x_1 \to \\
\sigma_1(x_0,x_1) = u_1 & \to & p_i(\cdot\,|x_1,u_1) \sim x_2 \to \\
\quad\vdots & & \quad\vdots \\
\sigma_{N-1}(x_0,\ldots,x_{N-1}) = u_{N-1} & \to & p_i(\cdot\,|x_{N-1},u_{N-1}) \sim x_N.
\end{array}$$

そして、目的関数である期待値は明示的には次のようにあらわされる：

$$\begin{aligned}
& E^\sigma\left[r_i(x_0,u_0) \circ_i r_i(x_1,u_1) \circ_i \cdots \circ_i r_i(x_{N-1},u_{N-1}) \circ_i k_i(x_N)\right] \\
= & \sum_{(x_1,x_2,\ldots,x_N)\in X\times\cdots\times X} \sum \cdots \sum \Big\{\left[r_i(x_0,u_0) \circ_i r_i(x_1,u_1) \circ_i \cdots \circ_i r_i(x_{N-1},u_{N-1}) \circ_i k_i(x_N)\right] \\
& \qquad \times p_i(x_1|x_0,u_0)p_i(x_2|x_1,u_1)\cdots p_i(x_N|x_{N-1},u_{N-1})\Big\}
\end{aligned}$$

ここで，状態空間 X_i と X_j の間には確定的推移：

$$f_{ij} : X_i \times U_i \to X_j$$

を与える。すなわち、第 i 決定過程の状態 $x \in X_i$ とその時の決定 $u \in U_i(x)$ に対し、第 j $(j \neq i)$ 決定過程の各初期状態が $f_{ij}(x,u) \in X_j$ と定まるのである。さらに、問題 $\mathrm{P}_i(x_0)$ の最適値を $V_i(x_0)$ $(i = 1,2,\ldots,m)$ で表す：

$$V_i(x_0) \;=\; \begin{cases} k_i(x_0), & x_0 \in T_i \\ \max\limits_{\sigma\in\Sigma_i} E^\sigma\Big[r_i(x_0,u_0) \circ_i r_i(x_1,u_1) \circ_i \cdots \circ_i r_i(x_{N-1},u_{N-1}) \circ_i k_i(x_N)\Big], & x_0 \notin T_i. \end{cases}$$

このとき、問題 $\mathrm{P}_i(x_0)$ の利得関数 $r_i(x,u)$ は，$V_j(\cdot)$ $(j \neq i)$ に依存して関数

$$R_i : D_1 \times D_2 \times \cdots \times D_{i-1} \times D_{i+1} \times \cdots \times D_m \;\to\; D_i$$

により次のように定まるものとする：

$$\begin{aligned}
r_i(x,u) \;=\; R_i\Big(& V_1(f_{i1}(x,u)),\, V_2(f_{i2}(x,u)),\ldots,\, V_{i-1}(f_{i(i-1)}(x,u)), \\
& V_{i+1}(f_{i(i+1)}(x,u)),\ldots,\, V_m(f_{im}(x,u))\Big), \qquad i = 1,2,\ldots,m.
\end{aligned}$$

ただし，再帰的に生じる過程を通じて生成される全状態列に対し有限段での終了を仮定する．なお、解くべき問題は第1決定過程で与えられるものとし、$\bar{x}_0 \in X_1 \backslash T_1$ に対する $\mathrm{P}_1(\bar{x}_0)$ とする．

3 再帰式

各決定過程 $\mathrm{P}_i(x_0)$ $(i=1,2,\ldots,m)$ に対し，パラメータ $\lambda \in D_i$ を導入した次の埋め込み問題を考え，その最適値関数を W_i とおく：

$$W_i(x_0,\lambda) \;=\; \begin{cases} \lambda \circ_i k_i(x_0), & x_0 \in T_i,\ \lambda \in D_i \\ \max\limits_{\sigma\in\Sigma_i} E^{\sigma}\Big[\lambda \circ_i r_i(x_0,u_0) \circ_i \cdots \circ_i r_i(x_{N-1},u_{N-1}) \circ_i k_i(x_N)\Big], & x_0 \notin T_i,\ \lambda \in D_i. \end{cases}$$

このとき，各 $i=1,2,\ldots,m$ に対して次の再帰式が成り立つ [10].

$$\begin{aligned} W_i(x,\lambda) &= \lambda \circ_i k_i(x) \qquad\qquad x \in T_i, \lambda \in D_i, \\ W_i(x,\lambda) &= \max_{u\in U_i(x)} \Big[\sum_{y\in X_i} W_i\big(y, \lambda \circ_i r_i(x,u)\big) p_i(y \,|\, x,u) \Big] \quad x \notin T_i, \lambda \in D_i. \end{aligned}$$

ここで，

$$V_i(x) \;=\; W_i(x,\tilde{\lambda}_i)$$

であるので

$$\begin{aligned} r_i(x,u) \;=\; R_i\Big(& W_1(f_{i1}(x,u),\tilde{\lambda}_1),\ W_2(f_{i2}(x,u),\tilde{\lambda}_2),\ldots,\ W_{i-1}(f_{i(i-1)}(x,u),\tilde{\lambda}_{i-1}), \\ & W_{i+1}(f_{i(i+1)}(x,u),\tilde{\lambda}_{i+1}),\ldots,\ W_m(f_{im}(x,u),\tilde{\lambda}_m)\Big), \qquad i=1,2,\ldots,m \end{aligned}$$

が成り立つ．これと再帰式を組み合わせることにより，相互依存型決定過程問題 $\mathrm{P}_1(\overline{x}_0)$ の最適値 $W_1(\overline{x}_0,\tilde{\lambda}_1)$ を再帰的手法により求めることができる．

さらに，再帰式の計算と同時に

$$\pi^{i*}(x,\lambda) \;\in\; \operatorname*{argmax}_{u\in U_i(x)} \Big[\sum_{y\in X_i} W_i\big(y, \lambda \circ_i r_i(x,u)\big) p_i(y \,|\, x,u) \Big] \quad x \notin T_i, \lambda \in D_i.$$

により定まる決定関数 $\pi^{1*},\pi^{2*},\ldots,\pi^{m*}$ を用いて，相互依存型決定過程問題 $\mathrm{P}_1(\overline{x}_0)$ を構成する各過程に対する最適一般政策 $\sigma^{1*},\sigma^{2*},\ldots,\sigma^{m*}$ を以下のように構築することができる．

まず、$x_0^1=\overline{x}_0$，$\lambda_0^1=\tilde{\lambda}_1$ とおき、次のように定める。

$$\sigma_0^{1*}(x_0^1) \;=\; \pi^{1*}(x_0^1,\lambda_0^1)$$

そして 各 $n=1,2,\ldots$ に対し、マルコフ推移法則

$$p_1(\cdot \,|\, x_{n-1}^1, u_{n-1}^1) \quad \text{ただし} \quad u_{n-1}^1 \;=\; \sigma_{n-1}^{1*}(x_0^1,x_1^1,\ldots,x_{n-1}^1)$$

にしたがって正の確率で生じる非終了状態 $x_n^1 \notin T_1$ ごとに

$$\lambda_n^1 \;=\; \lambda_{n-1}^1 \circ_1 r_1(x_{n-1}^1,u_{n-1}^1),$$

$$\sigma_n^{1*}(x_0^1,x_1^1,\ldots,x_n^1) \;=\; \pi^{1*}(x_n^1,\lambda_n^1)$$

と定める。この手続きを、初期状態 $x_0^1=\overline{x}_0$ から正の確率で生じるすべての履歴 $(x_0^1,u_0^1,x_1^1,u_1^1,\ldots,x_N^1)$ に対し実行する。

同様に各 $i=2,3,\ldots,m$ および $l=0,1,\ldots$ に対し

$$x_0^i \;=\; f_{1i}(x_l^1,u_l^1), \quad \lambda_0^i \;=\; \tilde{\lambda}_i$$

$$\sigma_0^{i*}(x_0^i) \;=\; \pi^{i*}(x_0^i, \lambda_0^i)$$

とおき、第1過程 ($i=1$ のとき) に対して行った手続きをすべての第 i 過程へ実行する。この手続きを、終了状態に達するまで再帰的に実行する。

このとき

$$\sigma^{i*} \;=\; \{\sigma_0^{i*}, \sigma_1^{i*}, \ldots\}, \quad i = 1, 2, \ldots, m$$

が、所与の問題の最適一般政策群を与える。

4 ファジィ環境下における相互依存型マルコフ決定過程

ファジィ環境下における意思決定は、Bellman と Zadeh [2] により提唱された。ここでは、結合型評価の応用としてファジィ環境下における相互依存型マルコフ決定過程を扱う。ファジィ環境下では、決定がファジィゴールとファジィ制約をあらわすファジィ集合の交わりとしてあらわされる。そして、ファジィ集合の交わりは、メンバーシップ関数で置き換えた場合、それらの最小値としてあらわされる。ゆえに、意思決定過程問題として定式化した場合、本質的には最小型評価の最大化問題となるのである。したがって、ファジィ環境下での相互依存型決定過程問題は、次で与えられる ($i = 1, 2, \ldots, m$) :

$$\begin{aligned} \mathrm{P}_i(x_0) \qquad & \text{Maximize } E^{\sigma}\left[r_i(x_0, u_0) \wedge r_i(x_1, u_1) \wedge \cdots \wedge r_i(x_{N-1}, u_{N-1}) \wedge k_i(x_N)\right] \\ & \text{subject to} \quad x_{n+1} \sim p_i(\,\cdot\,|x_n, u_n) \quad n = 0, 1, \ldots, N-1 \\ & \qquad \sigma = (\sigma_0, \sigma_1, \ldots, \sigma_{N-1}) \in \Sigma_i \\ & \qquad (\; N = N(x_0, u_0, x_1, u_1, \ldots) = \max\{n : x_n \notin T_i\} + 1 \;), \end{aligned}$$

ただし

$$a \wedge b = \min(a, b) \quad a, b \in \mathbf{R}$$

であり、2項演算子 $\wedge$ が結合律をみたすことは容易にわかる。また、演算子 $\wedge$ に対する単位元は十分大きな値を選べばよく、特にファジィ環境下においては、利得関数すなわちメンバーシップ関数が $[0,1]$ にその値をとることから $\tilde{\lambda}_i = 1$ と選べばよい。よって、前節の結果より、各 $i = 1, 2, \ldots, m$ に対して次の再帰式が成り立つ :

$$\begin{aligned} W_i(x, \lambda) &= \lambda \wedge k_i(x) && x \in T_i,\ \lambda \in [0,1], \\ W_i(x, \lambda) &= \max_{u \in U_i(x)} \Big[\sum_{y \in X_i} W_i\big(y, \lambda \wedge r_i(x,u)\big) p_i(y \,|\, x, u) \Big] && x \notin T_i,\ \lambda \in [0,1]. \end{aligned}$$

たとえば $m = 3$ で、各過程の利得関数が

$$r_i(x, u) \;=\; \bigvee_{j=1(j \neq i)}^{3} W_j(f_{ij}(x, u)) \quad i = 1, 2, 3$$

で与えられるとき、相互依存型再帰式 :

$$\begin{aligned} W_1(x, \lambda) &= \lambda \wedge k_1(x) \qquad x \in T_1,\ \lambda \in [0,1], \\ W_2(x, \lambda) &= \lambda \wedge k_2(x) \qquad x \in T_2,\ \lambda \in [0,1], \\ W_3(x, \lambda) &= \lambda \wedge k_3(x) \qquad x \in T_3,\ \lambda \in [0,1], \end{aligned}$$

$$W_1(x,\lambda) = \max_{u\in U_1(x)}\Big[\sum_{y\in X_1} W_1\Big(y, \lambda\wedge\big(W_2(f_{12}(x,u))\vee W_3(f_{13}(x,u))\big)\Big)p_1(y\,|\,x,u)\Big]$$
$$x\notin T_1,\ \lambda\in[0,1],$$

$$W_2(x,\lambda) = \max_{u\in U_2(x)}\Big[\sum_{y\in X_2} W_2\Big(y, \lambda\wedge\big(W_1(f_{21}(x,u))\vee W_3(f_{23}(x,u))\big)\Big)p_2(y\,|\,x,u)\Big]$$
$$x\notin T_2,\ \lambda\in[0,1],$$

$$W_3(x,\lambda) = \max_{u\in U_3(x)}\Big[\sum_{y\in X_3} W_3\Big(y, \lambda\wedge\big(W_1(f_{31}(x,u))\vee W_2(f_{32}(x,u))\big)\Big)p_3(y\,|\,x,u)\Big]$$
$$x\notin T_3,\ \lambda\in[0,1].$$

を用いて、解を求めることができる。

5 決定過程間の推移が確率的な場合

ここまでは、相互依存型決定過程を構成する各決定過程間の推移が確定的な問題を扱ってきた。本節では、決定過程間における状態の推移が確率的な場合についても、若干の拡張で対応可能であることを述べる。

5.1 各決定過程において初期状態が確率的に生じる場合

第 i 決定過程の状態 $x\in X_i$ とその時の決定 $u\in U_i(x)$ に対し、第 j 決定過程の初期状態 $y\in X_j$ が確率的に生じる場合を考える（図 5.1）。この推移が

$$y\sim p_{ij}(\,\cdot\,|\,x,u)\quad\Big(\sum_{y\in X_j} p_{ij}(\,y\,|\,x,u)=1\Big)$$

とあらわされるとする。

このとき、第 j 決定過程の状態空間 X_j にダミーの状態 $\hat{y}$ を加えて

$$f_{ij}(x,u)=\hat{y}$$

とし、さらに、決定空間 U_j にダミーの決定 $\hat{v}$ を加え

$$U_j(\hat{y})=\{\hat{v}\},\qquad p_j(y\,|\,\hat{y},\hat{v})=p_{ij}(y\,|\,x,u)\ \ (y\in X_j,\ y\neq\hat{y})$$

とおく。すなわち、まずダミー状態 $\hat{y}$ へ確定的に推移し、その後、確率推移を 1 段階追加するのである。あとは

$$r_j(\hat{y},\hat{v})=\tilde{\lambda}_j,\qquad p_j(\hat{y}\,|\,y,v)=0\ \ (y\in X_j,\ v\in U_j(y))$$

とおけば、所与の問題と同値な問題が本論文の枠組みで構成可能となる。

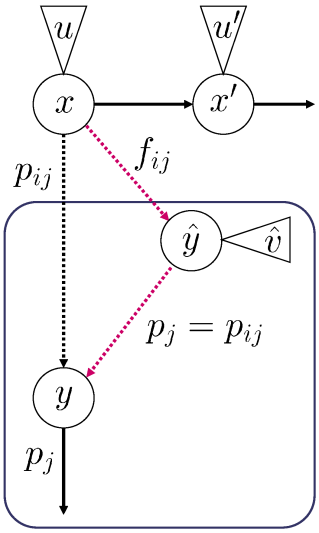

図 5.1: 確率的推移（その 1）

5.2 決定過程の各初期状態へ確率的に推移する場合

第 i 決定過程の状態 $x \in X_i$ とその時の決定 $u \in U_i(x)$ に対し、第 j $(j \neq i)$ 決定過程の各初期状態

$$y_1 \in X_1,\ y_2 \in X_2, \ldots, y_{i-1} \in X_{i-1},$$

$$y_{i+1} \in X_{i+1}, \ldots, y_m \in X_m$$

へ確率的に推移する場合を考える（図 5.2）。推移確率は条件付き確率 q_i により

$$q_i(\, y_j \,|\, x, u) \quad \left(j \neq i,\ \sum_{j \neq i} q_i(y_j \,|\, x, u) = 1 \right)$$

で与えられるとする。このときは、評価方法に合わせて利得関数をうまく定義すればよい。たとえば、第 i 決定過程の利得関数が、各第 j 決定過程 $(j \neq i)$ の最適値の期待値で与えられるときは、R_i を

$$R_i : X_i \times U_i \times D_1 \times D_2 \times \cdots \times D_{i-1} \times D_{i+1} \times \cdots \times D_m \ \rightarrow \ D_i$$

と拡張し

$$R_i\big(x, u, V_1(y_1), V_2(y_2), \ldots, V_{i-1}(y_{i-1}), V_{i+1}(y_{i+1}), \ldots V_m(y_m)\big) \ = \ \sum_{j \neq i} V_j(y_j) q(y_j \,|\, x, u)$$

と定義する。そして、利得関数を

$$r_i(x, u) \ = \ R_i\big(x, u, V_1(y_1), V_2(y_2), \ldots, V_{i-1}(y_{i-1}), V_{i+1}(y_{i+1}), \ldots V_m(y_m)\big)$$

とおけばよい。

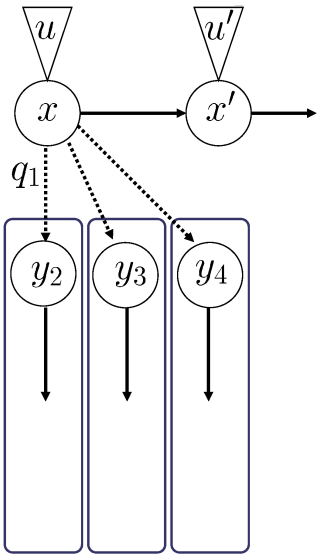

図 5.2: 確率的推移（その 2）

謝辞

本研究は科研費 23654038 の助成を受けたものである。

参考文献

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

[2] R.E. Bellman and L.A. Zadeh, Decision-making in a fuzzy environment, Management Science, 17, B141-B164, 1970.

[3] U. Bertelé and F. Brioschi, Nonserial Dynamic Programming, Academic Press, New York, 1972.

[4] T. Fujita, Re-examination of Markov policies for additive decision process, Bulletin of Informatics Cybernetics, 29, No.1, 51-65, 1997.

[5] T. Fujita, On policy classes in dynamic programming theory, Proceedings of the 9th Bellman Continuum International Workshop on Uncertain System and Soft Computing, Series of Information & Management Sciences 2, 39-43, 2002.

[6] T. Fujita, Associative Criteria in Mutually Dependent Markov Decision Processes, Proceedings of IIAI International Conference on Advanced Applied Informatics (IIAI AAI 2014), 147-150, 2014.

[7] 藤田敏治, 長友健太郎, 折り紙ユニットを用いた凸多面体の構成 ー相互依存型決定過程によるアプローチー, 京都大学数理解析研究所講究録, 1912, 17-25, 2014.

[8] T. Fujita and A. Kira, Mutually Dependent Markov Decision Processes, Journal of Advanced Computational Intelligence and Intelligent Informatics, 18(6), 2014, 992-968.

[9] T. Fujita, and K. Tsurusaki, Stochastic optimization of multiplicative functions with negative value, Journal of Operations Research Society of Japan, 41(3), 351-373, 1998.

[10] S. Iwamoto: Associative dynamic programs. Journal of Mathematical Analysis and Applications, 201, 195-211,1996.

[11] S. Iwamoto and T. Fujita, Stochastic decision-making in a fuzzy environment, Journal of Operations Research Society of Japan, 38, 467–482, 1995.

[12] S. Iwamoto, T. Ueno and T. Fujita, Controlled Markov Chains with Utility Functions, Eds. H. Zhenting, J. A. Filar and A. Chen, Markov Processes and Controlled Markov Chains, Chap.8, 135-148, Kluwer, 2002.

[13] R. A. Howard, Dynamic Programming and Markov Processes, MITPress, Cambridge, Mass., 1960.

[14] A. Kira, T. Ueno and T. Fujita Threshold probability of non-terminal type in finite horizon Markov decision processes, Journal of Mathematical Analysis and Applications, 386, 461-472, 2012

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