数理解析研究所講究録 1306 巻 2003 年 210-216

動的計画論における決定ツリーと政策

九工大・工　藤田敏治
Kyushu Institute of Technology　Toshiharu Fujita

1　はじめに

我々は，動的計画法 ([1]) で種々の評価関数をもつ問題を扱い ([2], [3], [4], [5])，その解析のために政策クラスの概念を広げてきた ([6], [7])．そして，問題構造の正確な把握および厳密な理論的解析のためには，まず問題における実行可能解の正確な記述が必要と感じている．

ここでは，動的計画問題における実行可能解を直感的な決定ツリーとして捉え，分類し，それらの政策 (Policy) による表現について考える．実際，実行可能解の集合は，一般にある条件を満たす状態・決定ツリーの集まりと考えられる．この定式化のもとに，問題の性質に応じた解析を行い，問題を扱うにあたって必要十分な政策クラスを見出す必要がある．

以下では，状態・決定ツリーおよび政策について述べた後，確定システム上での決定過程問題に対し，実行可能集合として必要となる政策クラスについて考える．この研究の最終目的は，様々なタイプの動的計画問題を統一的に表現すること，そして汎用的な解法の枠組みを構築することである．

なお，$N \geq 1$ でシステムの終端時刻をあらわし，$X = \{s_1, \cdots, s_m\}$ は状態集合を，$U = \{a_1, \cdots, a_l\}$ は決定集合をそれぞれあらわすものとする．

2　状態・決定ツリー

2.1　状態・決定ツリーとは

状態・決定ツリーとは，システムにおいて状態と決定の続く様子をツリー上に表現したものである．もっとも単純な例として，たとえば確定システム上での

$$s_1 \text{ — } a_2 \text{ — } s_3 \text{ — } a_1 \text{ — } s_3 \tag{1}$$

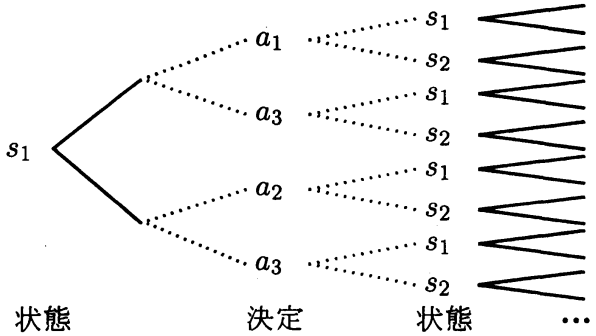
は，「初期状態 s_1 が与えられた際，まず決定 a_2 をとる．その後，状態 s_3 に移るが，ここでは決定 a_1 をとる．そして最終状態 s_3 で終了」を表す．

状態と決定について，一般には次のように考えられる．

(1) 状態および決定は確率変数で与えられる（確定的な場合はこの特殊ケース）．

(2) 決定は自由に選べる複数の選択肢を持ち得る．

従って，状態・決定ツリーを図示する際には 2 通りの分岐が発生する．ひとつは，確率分布に応じた推移を表す分岐であり，もうひとつは選択を表す分岐である．実線で選択，点線で確率的推移をあらわした場合，たとえば次のような状態・決定ツリーが考えられる．



例えば，最初の分岐および2番目の分岐は「初期状態 s_1 に対し，確率的に選ばれる決定 $\{a_1, a_3\}$ をとるか，もしくは確率的に選ばれる決定 $\{a_1, a_2\}$ をとる」ことを意味し，3番目の（一番上の）分岐は「第1期において状態 s_1 に対し決定 a_1 をとった場合，第2期に状態 s_1, s_2 に確率的に推移する」ことを意味する．

以後，簡略化のため，状態・決定ツリーの型を表す際に以下の表記を用いる．

$$X_1 \text{ — } (U_1) \text{ — } X_2 \text{ — } (U_2) \text{ — } \cdots \text{ — } (U_N) \text{ — } X_{N+1} \tag{2}$$

ここで，X_n と U_n がそれぞれ状態と決定を代表する．大文字はそれが確率変数であることをあらわし，カッコにより決定が選択可能であることをあらわす．なお，状態あるいは決定が確定的な場合を特に区別する際には，確定的なものを小文字によりあらわす．次の例は，状態は確率的だが，決定は確定的であることを表す．

$$X_1 \text{ — } (u_1) \text{ — } X_2 \text{ — } (u_2) \text{ — } X_3$$

また，状態・決定ツリー (1) の型は

$$x_1 \text{ — } u_1 \text{ — } x_2 \text{ — } u_2 \text{ — } x_3$$

と表される．

2.2 状態・決定ツリーの分類

確率システム上での最も一般的な状態・決定ツリーの型は (2) であった．この型で表される決定ツリー全体を T_S^* とおく．また，確定システム上では

$$x_1 \text{ — } (U_1) \text{ — } x_2 \text{ — } (U_2) \text{ — } \cdots \text{ — } (U_N) \text{ — } x_{N+1}$$

で表されるツリー全体を考え，T_D^* とおく．なお，決定を確定的な場合に限定した状態・決定ツリーの全体を，それぞれ T_S, T_D とする．すなわち，下付の S, D によりシステムが確率的か確定的かを表し，$*$ のあるなしで決定が確率的か，確定的かを表す．

3 政策

3.1 定義

各期において決定を与える関数は決定関数と呼ばれ，その決定関数の列が政策である．決定が何に依存して定まるかにより，次の3種類の政策が定義される．([6], [7])

マルコフ政策

現時刻の状態のみに依存し決定を定める決定関数の列を意味し，$\pi = \{\pi_1, \pi_2, \ldots, \pi_N\}$：

$$\pi_n : X \to U, \quad n = 1, 2, \ldots, N$$

と表される．すなわち，時刻 n の状態を x_n とするとき，マルコフ政策 π による時刻 n の決定 u_n は

$$u_n = \pi_n(x_n), \quad n = 1, 2, \ldots, N$$

と与えられる.

一般政策
現時刻までの状態すべてに依存し決定を定める決定関数の列を意味し，$\sigma = \{\sigma_1, \sigma_2, \ldots, \sigma_N\}$：

$$\sigma : X^n \to U, \quad n = 1, 2, \ldots, N$$

と表される．すなわち，時刻 n までの状態を $x_1, x_2, \ldots, x_n$ とするとき，一般政策 σ による時刻 n の決定 u_n は

$$u_n = \sigma_n(x_1, x_2, \ldots, x_n), \quad n = 1, 2, \ldots, N$$

と与えられる.

原始政策
履歴 (現時刻までの状態と決定の交互列) に依存し決定を定める決定関数からなる列であり，$\gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_N\}$：

$$\gamma_n : (X \times U)^{n-1} \times X \to U, \quad n = 1, 2, \ldots, N$$

と表される．すなわち，時刻 n までの履歴を $x_1, u_1, x_2, u_2, \ldots, u_{n-1}, x_n$ とするとき，一般政策 γ による時刻 n の決定 u_n は

$$u_n = \gamma_n(x_1, u_1, x_2, u_2, \ldots, u_{n-1}, x_n), \quad n = 1, 2, \ldots, N$$

と与えられる.

また，上記の3政策はいずれも決定関数が U への写像となっているが，U のべき集合 2^U への写像として扱う場合もある．ただし 2^U を想定した場合，集合として与えられる決定の意味は「集合の中から任意にひとつの決定を取ることができる」と解釈するものとする．以上，計 $3 \times 2 = 6$ 通りの最適政策のクラスが考えられるのである.

以後，単に政策，あるいはマルコフ政策，一般政策，原始政策と表現した場合には，それを構成する決定関数の写像先は U であるものとし，集合 2^U への写像を想定する際には "集合値" を付けて表現することとする．たとえば "集合値一般政策" などと呼ぶ.

3.2 状態・決定ツリーの表現

一般に，T_S や T_D あるいは T_D^* に属す状態・決定ツリーは，単一の集合値政策で表現できる．ただし T_S^* に限っては，単一の政策で表現することは必ずしもできない．この場合，複数の原始政策による表現となる．また，T_S, T_D, T_D^* に属す状態・決定ツリーも複数の政策による表現は可能である.

この小節では，政策や集合値政策によって状態・決定ツリーがどのように表現されるかをいくつかの例で示す.

例 1 (確定システム)
第2節の状態・決定ツリー (1) は，マルコフ政策 ($\to U$) によって以下のように表される.

$$\{\pi_1(s_1) = a_2, \quad \pi_2(s_3) = a_1\}$$

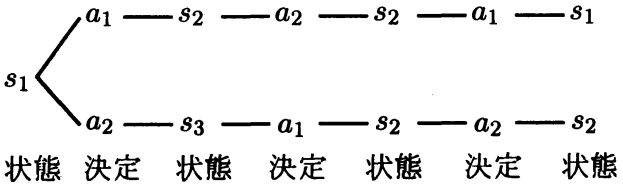
また，集合値マルコフ政策 ($\to 2^U$):

$$\{\pi_1(s_1) = \{a_2\}, \quad \pi_2(s_3) = \{a_1\}\}$$

で表されているとも解釈できる. □

なお，例 1 におけるツリー (1) の表現において，$\pi_1(s_2), \pi_2(s_1)$ 等は関与しないのでその値は任意である．以後の例も含めて，対象とする状態・決定ツリーの表現において任意の決定関数は省略する．

例 2 (確定システム)



この状態・決定ツリーは集合値一般政策 ($\to 2^U$) により表される.

$$\left\{\begin{array}{l} \sigma_1(s_1) = \{a_1, a_2\}, \\ \sigma_2(s_1, s_2) = \{a_2\}, \quad \sigma_2(s_1, s_3) = \{a_1\} \\ \sigma_3(s_1, s_2, s_2) = \{a_1\}, \quad \sigma_3(s_1, s_3, s_2) = \{a_2\} \end{array}\right\}$$

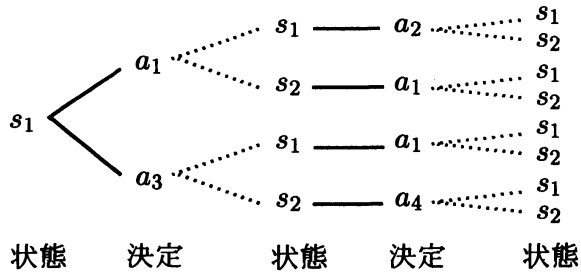
また，2 つのマルコフ政策 ($\to U$):

$$\{\pi_1(s_1) = a_1, \quad \pi_2(s_2) = a_2, \quad \pi_3(s_2) = a_1\}$$

$$\{\pi_1'(s_1) = a_2, \quad \pi_2'(s_3) = a_1, \quad \pi_3'(s_2) = a_2\}$$

により表されると解釈することもできる. □

例 3 (確率システム)



この状態・決定ツリーは集合値原始政策 ($\to 2^U$):

$$\left\{\begin{array}{l} \gamma_1(s_1) = \{a_1, a_3\}, \\ \gamma_2(s_1, a_1, s_1) = \{a_2\}, \quad \gamma_2(s_1, a_1, s_2) = \{a_1\} \\ \gamma_2(s_1, a_3, s_1) = \{a_1\}, \quad \gamma_2(s_1, a_3, s_2) = \{a_4\} \end{array}\right\}$$

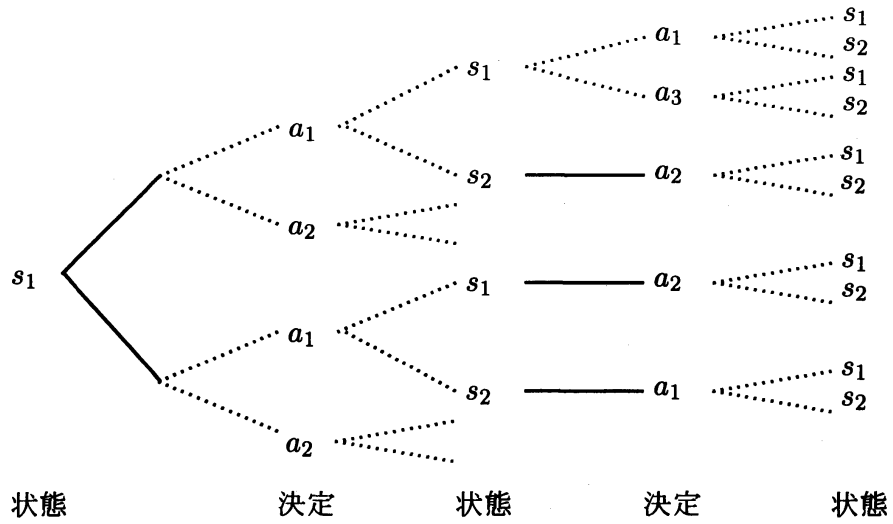
または，2 つの一般政策 ($\to U$):

$$\{\sigma_1(s_1) = a_1, \quad \sigma_2(s_1, s_1) = a_2, \quad \sigma_2(s_1, s_2) = a_1\}$$

$$\{\sigma_1'(s_1) = a_3, \quad \sigma_2'(s_1, s_1) = a_1, \quad \sigma_2'(s_1, s_2) = a_4\}$$

(一部省略) により表される. □

例 4 (確率システム)



この状態・決定ツリーは単一の政策で表すことはできない.
次の 2 つの原始政策 ($\to U$):

$$\left\{\begin{array}{l}\gamma_1(s_1)=(a_1,a_2)\\ \gamma_2(s_1,a_1,s_1)=(a_1,a_3),\quad \gamma_2(s_1,a_1,s_2)=a_2\\ \gamma_2(s_1,a_2,s_1)=a_2,\quad \gamma_2(s_1,a_2,s_2)=a_1\end{array}\right\}$$

$$\left\{\begin{array}{l}\gamma_1(s_1)=(a_1,a_2)\\ \gamma_2(s_1,a_1,s_1)=a_2,\quad \gamma_2(s_1,a_1,s_2)=a_1\\ \gamma_2(s_1,a_2,s_1)=a_1,\quad \gamma_2(s_1,a_2,s_2)=a_2\end{array}\right\}$$

により表される. □

4 多段決定過程問題

4.1 定式化

確定的推移法則 $f: X\times U\to X$ が与えられ, 確定的に決定をとる場合の問題を考える. なお, 確定的推移法則 f は, 現時刻の状態が x, 決定が u であるとき, 状態が次の時刻で $f(x,u)$ へ推移することをあらわす. この f のもと, 考えうる状態・決定ツリー全体は T_D である. 従って, 初期状態 x_1 を与えた場合, 確定システム上での結合型評価最大化問題 ([4]) は次のように表される.

$$\begin{array}{l}\mathrm{Max}\ r_1(x_1,u_1)\circ r_2(x_2,u_2)\circ\cdots\circ r_N(x_N,u_N)\circ r_G(x_{N+1})\\ \quad\text{s.t. (i)}\ \ x_{n+1}=f(x_n,u_n)\quad n=1,2,\ldots N\\ \qquad\text{(ii)}\ \ (x_1,u_1,x_2,u_2,\ldots,x_N,u_N,x_{N+1})\ll t\\ \qquad\text{(iii)}\ \ t\in T_D\end{array}\qquad(3)$$

なお,

$r_n: X\times U\to\mathbf{R}$ 時刻 n における利得
$r_G: X\to\mathbf{R}$ 終端利得

であり，$\circ$ は結合演算子（$(x \circ y) \circ z = x \circ (y \circ z)$ をみたす）をあらわす．また，(ii) は，決定ツリー t から任意に一つのパスをとることをあらわすものとする．ここでいうパスとは，x_1 をツリーの根とした際，根からひとつの葉（最終状態）までの鎖のことである．

ここで一つ問題になるのは，状態・決定ツリー $t \in T_D$ に対し，目的関数の値が一意に定まらない場合がありうる点である．ひとつの解決策は，状態・決定ツリーを

$$x_1 \text{ — } u_1 \text{ — } x_2 \text{ — } u_2 \text{ — } \cdots \text{ — } u_N \text{ — } x_{N+1}$$

の型に限定することである．もうひとつは，目的関数の値を一意に与えない状態・決定ツリーを除外することである．

前者の場合，最適解（最適状態・決定ツリー）は一般に複数のツリーで与えられる．一方，後者では，極大の意味で単一の状態・決定ツリーとして与えられる．言い換えると，すべての最適状態・決定ツリーをその部分ツリーとする最適状態・決定ツリーが存在する．

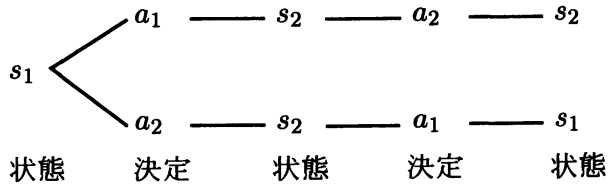
なお，状態・決定ツリーを表現する政策に関しても同様の注意が必要で，状態・決定ツリーの代わりに政策を考えている場合，目的関数の値を一意に与えない政策は除外することとする．

4.2 政策クラス

決定過程問題 (3) の最適状態・決定ツリーは，どの政策クラスによって表現可能であろうか．それは，目的関数の型に応じて決まる．

まず，目的関数が加法型 ($\circ = +$) の場合について考える．この場合，[2] の結果より最適政策はマルコフ政策のクラス内に存在することが示されている．したがって，問題 (3) の実行可能解はマルコフ政策で表現されるものに限定してよい．実際に解く際には，マルコフ政策あるいは集合値マルコフ政策のいずれかに関する最大化問題として解けばよい．

次に一般の場合だが，実際，次のような状態・決定ツリーが考えられる．



この場合，もはや単一の集合値マルコフ政策では表現することはできない．この種の状態・決定ツリーを単一政策で表現するには，集合値原始政策を用いる必要がある．

$$\left\{ \begin{array}{l} \gamma_1(s_1) = \{a_1, a_2\}, \\ \gamma_2(s_1, a_1, s_2) = \{a_2\}, \quad \gamma_2(s_1, a_2, s_2) = \{a_1\} \end{array} \right\}$$

なお，複数の政策による表現を考えた場合には，マルコフ政策で十分である．したがって，一般に問題 (3) においては，マルコフ政策クラスあるいは集合値原始政策クラスでの最大化を考える必要がある．

5 まとめ

動的計画問題における実行可能解を状態・決定ツリーとして捉え，その政策による表現について考えた．政策については，決定の依存先に関する３通りの分類と，決定関数の写像先に関する２通

りの分類から，6 通りを定義した．一般に結合型評価をもつ決定過程問題を扱う場合，最適政策をマルコフ政策では表現できない場合が起こる．よって，政策クラスとしてより広いクラスを想定しなければならないのである．

今回提案した動的計画法の枠組みは，様々な問題，特にこれまでに扱ったことのないような型の問題を動的計画法により正確に解析するために有効と考える．

References

[1] R.E. Bellman, *Dynamic Programming*, NJ: Princeton Univ. Press (1957).

[2] T. Fujita, Re-examination of Markov Policies for Additive Decision Process, Bull. Infor. Cyber., **29**(1997), pp.51-65

[3] T. Fujita and K. Tsurusaki, Stochastic optimization of multiplicative functions with negative value, J. Oper. Res. Soc. Japan, **41**(1998), 351-373.

[4] S. Iwamoto, Associative dynamic programs, J. Math. Anal. Appl., **201**(1996), No.1, 195-211.

[5] S. Iwamoto and T. Fujita, Stochastic decision-making in a fuzzy environment, J. Oper. Res. Soc. Japan, **38**(1995), 467-482.

[6] S. Iwamoto, K. Tsurusaki and T. Fujita, On Markov Policies for Minimax Decision Processes, J. Math. Anal. Appl., **253**(2001), 58-78.

[7] S. Iwamoto, T. Ueno and T. Fujita, Controlled Markov Chains with Utility Functions, Markov Process and Controlled Markov Chains, Chap. 8, Kluwer (2002)