数理解析研究所講究録 1252 巻 2002 年 132-138

動的計画論における政策クラスについて

九工大・工　藤田敏治 (Toshiharu Fujita)

1 はじめに

動的計画法は R.Bellman により創出され ([1])，現在までに幅広い研究・応用がなされている．離散・連続，確定・確率等を問わず，多方面で利用可能な強力なツールである．また，Bellman と Zadeh による [2] 以降，ファジィ環境下においても様々な研究がなされている．

我々は，[2] で扱われていた確率システム上での問題に対し，その再帰式の不整合な点を指摘し，埋め込み法を用いて新たに再帰式を導いた ([4])．[2] では，同時最適化（もとの問題）と逐次最適化（再帰式による解法）とで異なる解が生じていたのである．また，このファジィ環境下での問題においては，最適政策が必ずしもマルコフ政策の中に存在するわけではないことも分かった．

それ以降，我々は政策クラスの概念を一般政策，原始政策へとひろげてきた．そして，動的計画法を用いて種々の評価関数をもつ問題を扱ってきたが，同時最適化と逐次最適化は同値でなければならないという観点，および最適化においては決定関数列としての政策を基本とすべきであるという観点を重視して解析を行っている ([3], [5], [6])．その中で，より厳密な理論展開を行うためには，まず政策について整理しておく必要性がでてきた．

そこで本論分では，政策をその構成要素である決定関数の型に応じて分類し，6種類のクラスとして定義する．各クラスに属する政策が表現可能な決定ツリーを例により示し，一般的な有限段決定過程問題に対して解の構成方法について述べる．

2 多段決定過程問題

ここで扱う問題について定義する．以下の記号を用いる：

$N \geq 2$	**終端時刻**
$X = \{s_1, s_2, \cdots, s_m\}$	**状態集合**
$U = \{a_1, a_2, \cdots, a_l\}$	**決定集合**
$x_n \in X$	時刻 n における**状態** $(n = 1, 2, \ldots, N+1)$
$u_n \in U$	時刻 n における**決定** $(n = 1, 2, \ldots, N)$
$r_n : X \times U \to \mathbf{R}$	時刻 n における**利得**
$r_G : X \to \mathbf{R}$	**終端利得**
$\circ : \mathbf{R} \times \mathbf{R} \to \mathbf{R}$	**結合演算子** $(x \circ y) \circ z = x \circ (y \circ z)$

演算子 $\circ$ は各時刻において得られる利得を結びつけるもので，足し算 $(+)$ や掛け算 $(\times)$ あるいは最小演算子 $(\wedge)$ 等を一般化したものである．それぞれに応じて，加法型評価，乗法型評価，最小型評価等をもつ問題を表現する．

確定システム上での問題

状態が確定的に推移するシステム上での多段決定過程問題を考える．ここで，確定的推移法則 f とは，現時刻の状態が $x \in X$，決定が $u \in U$ であるとき，状態が次の時刻で $f(x, u) \in X$ へ確定的に推移することをあらわすものとする．この f のもと，初期状態 x_1 を与えた場合，確定システム

上での多段決定過程問題は次のように表される.

$$\begin{aligned}&\text{Maximize } g(r_1(x_1,u_1)\circ r_2(x_2,u_2)\circ\cdots\circ r_N(x_N,u_N)\circ r_G(x_{N+1}))\\&\text{subject to } (\text{i})_\text{n}\quad x_{n+1}=f(x_n,u_n)\quad n=1,2,\ldots,N\\&\qquad\quad (\text{ii})_\text{n}\quad \mu=\{\mu_1,\mu_2,...,\mu_N\}\ :\ \text{政策}\end{aligned}$$

ただし, $g:\mathbf{R}\to\mathbf{R}$ とする.

確率システム上での問題

次に, 状態が確率的に推移するシステム上での多段決定過程問題を考える. ここで, マルコフ推移法則 p とは, 現時刻の状態が $x\in X$, 決定が $u\in U$ であるとき, 次の時刻で状態 $y\in X$ へ確率 $p(y|x,u)$ で推移することをあらわすものとする. この推移を記号で $y\sim p(\cdot|x,u)$ と表す. このとき, 与えられた初期状態 x_1 に対し, 確率システム上での問題は次のように表される.

$$\begin{aligned}&\text{Maximize } E^{\mu}_{x_1}[g(r_1(x_1,u_1)\circ r_2(x_2,u_2)\circ\cdots\circ r_N(x_N,u_N)\circ r_G(x_{N+1}))]\\&\text{subject to } (\text{i})_\text{n}\quad x_{n+1}\sim p(\cdot|\,x_n,u_n)\quad n=1,2,\ldots,N\\&\qquad\quad (\text{ii})_\text{n}\quad \mu=\{\mu_1,\mu_2,...,\mu_N\}\ :\ \text{政策}\end{aligned}$$

ここでの $E^{\mu}_{x_1}$ は条件付き確率 $p(x_{n+1}|\,x_n,u_n)$, 政策 μ 及び初期状態 $x_1\in X$ に依存して定まる $X\times U\times X\times U\times\cdots\times U\times X$ 上の期待値を表す.

より一般には, 確定および確率システム上で, それぞれ次の目的関数を考えることができる.

$$h(x_1,u_1,x_2,u_2,\ldots,x_N,u_N,x_{N+1})$$

$$E[h(x_1,u_1,x_2,u_2,\ldots,x_N,u_N,x_{N+1})]$$

また, 政策に関しては, 次節で詳しく述べる.

3 原始・一般・マルコフ政策

各期においてとり得べき決定を与えるものが決定関数であり, その決定関数の列が政策である. 決定を何に依存して定めるかに応じて, 3 通りの分類が考えられる.

原始政策

履歴に依存して決定を定める決定関数からなる列である. ここで履歴とは, 現時刻までのすべての状態と決定の交互列を意味する. すなわち原始政策は $\gamma=\{\gamma_1,\gamma_2,\ldots,\gamma_N\}$:

$$\gamma_1:X\to U$$

$$\gamma_2:X\times U\times X\to U$$

$$\gamma_3:X\times U\times X\times U\times X\to U$$

$$\gamma_N : X \times U \times X \times \cdots \times U \times X \to U$$

と表される．以後，原始政策全体を Γ であらわす．また，Γ_n と表記した場合には n 期以降のみを考えた場合の原始政策 $\gamma = \{\gamma_n, \gamma_{n+1}, \ldots, \gamma_N\}$:

$$\gamma_n : X \to U$$

$$\gamma_{n+1} : X \times U \times X \to U$$

$$\vdots$$

$$\gamma_N : X \times U \times X \times \cdots \times U \times X \to U$$

の全体を表すものとする．

一般政策

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

$$\sigma_1 : X \to U$$

$$\sigma_2 : X \times X \to U$$

$$\sigma_3 : X \times X \times X \to U$$

$$\vdots$$

$$\sigma_N : X \times X \times \cdots \times X \to U$$

と表される．

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，$\pi = \{\pi_1, \pi_2, \ldots, \pi_N\}$:

$$\pi_1 : X \to U$$

$$\pi_2 : X \to U$$

$$\vdots$$

$$\pi_N : X \to U$$

と表される．

また，上記の3政策はいずれも決定関数が決定集合 U への写像となっているが，U のべき集合 2^U への写像としても定義できる．ただし 2^U を想定した場合，集合として与えられ決定の意味は「その中のいずれの決定を取ることもできる」と解釈するものとする．またこの場合，初期状態を与えても，1つの政策に対し目的関数値が一意に定まらないことがある．よって，べき集合 2^U への写像として決定関数を考える場合には，政策全体に関する最大化（または最小化）は，目的関数の値を一意に定めない政策は除外して考えるものとする．

以上，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通りの分類が考えられ，一般には計 $3 \times 2 = 6$ 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"）．

4 政策による決定ツリーの表現

各政策による決定ツリー（状態とその状態に対する決定の列をツリー上にあらわしたもの）の表現例を挙げる．

例 1 (確定システム)
もっとも単純な決定ツリーであり，いずれの政策クラスによっても同様に表現可能である．

$$s_1 \xrightarrow{a_2} s_2 \xrightarrow{a_1} s_2$$

マルコフ政策（集合値）

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

一般政策（集合値）

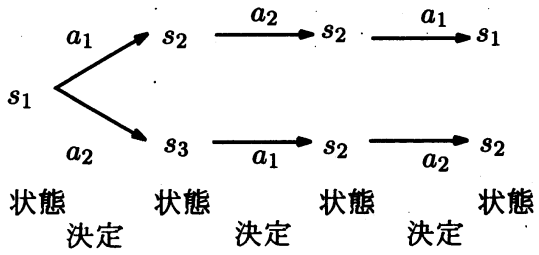
$$\{\ \sigma_1(s_1) = a_2,\ \ \sigma_2(s_1, s_3) = a_1\ \} \qquad \Big(\{\ \sigma_1(s_1) = \{a_2\},\ \ \sigma_2(s_1, s_3) = \{a_1\}\ \}\Big)$$

原始政策（集合値）

$$\{\ \gamma_1(s_1) = a_2,\ \ \gamma_2(s_1, a_2, s_3) = a_1\ \} \qquad \Big(\{\ \gamma_1(s_1) = \{a_2\},\ \ \gamma_2(s_1, a_2, s_3) = \{a_1\}\ \}\Big)$$

□

例 2 (確定システム)
以下の決定ツリーは単一のマルコフ政策では表現できないが，集合値一般政策のクラスで考えればひとつの政策として表現可能である．なお，集合値マルコフ政策では表現不可である．



マルコフ政策

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

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

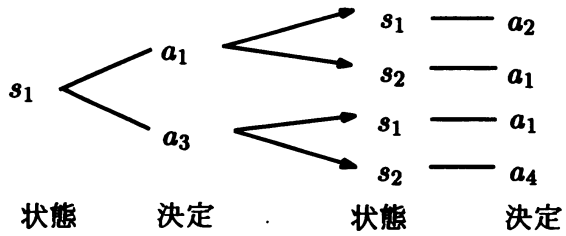
集合値一般政策

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

ただし，決定ツリーの表現に無関係な $\sigma_2(s_1, s_1)$ 等は任意で構わないため，ここでは省略している（以下同様）．

□

例3 (確率システム)
以下の決定ツリーは単一の一般政策では表現できないが，集合値原始政策のクラスで考えればひとつの政策として表現可能である．



一般政策

$$\{\sigma_1(s_1) = a_1,\ \sigma_2(s_1, s_1) = a_2,\ \sigma_2(s_1, s_2) = a_1\}$$
$$\{\sigma_1(s_1) = a_3,\ \sigma_2(s_1, s_1) = a_1,\ \sigma_2(s_1, s_2) = a_4\}$$

集合値原始政策

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

□

ここで挙げた決定ツリーの例は，人為的なものではなく2節の問題において実際に生じるものである．正確には，加法型評価のみを考えている場合には起こらないが，より一般の評価関数を考えた場合に生じる．

5 再帰式と最適政策の導出

前節の例からもわかるように，政策のマルコフ性がはっきりと仮定できない場合には，より広い政策クラスのもとでの定式化がなされるべきである．そうでなければ，真の最適政策を見落とすことになりかねないばかりか，誤った再帰式を導いてしまうことにもなりかねない．

実際，ここで考えている問題に対しては，一般政策または集合値原始政策のもとでの定式化がなされるべきである．そして，パラメータの追加あるいは状態の拡大等により，部分問題を構成して再帰式を導く．その再帰式を解くことで得られる最適政策は，パラメータつきマルコフ政策あるいは，拡大状態空間上のマルコフ政策とみなされるが，最終的には，その政策からもとの問題の最適政策を導く．

以下に，解法の概略を述べる．確定システムは確率システムの特殊な場合とみなせるので，ここでは確率システム上での問題について考える．また，政策クラスは集合値原始政策クラスとする．

$$\text{Maximize } E^{\gamma}_{x_1}[g(r_1(x_1, u_1) \circ \cdots \circ r_N(x_N, u_N) \circ r_G(x_{N+1}))]$$
$$\text{subject to } (\text{i})_n \quad x_{n+1} \sim p(\cdot \mid x_n, u_n) \quad n = 1, 2, \ldots, N$$
$$(\text{ii})_n \quad \gamma = \{\gamma_1, \gamma_2, ..., \gamma_N\} \in \Gamma$$

ただし，3節でも述べたように，最大化は目的関数の値を一意に定める政策のみに関するもので

なお，集合値を取る決定関数からなる政策を考える場合には，次の2つの点に注意すべきである．政策クラスにおける同値類，そして極大の概念についてである．ここで言う同値類とは，まったく同じ決定ツリーを構成する政策を同一視する概念であり，極大とは，評価関数の値を等しく定めるすべての政策を含む政策に対する概念である．これらを考慮することにより，最適決定ツリーと政策が同値類の意味で1対1に対応し，(もし存在すれば) その極大元によってすべての最適決定ツリーが表現可能となる．

5.1 再帰式

パラメータ λ を加えた次の問題を考える．

埋め込み問題

$$\begin{aligned}&\text{Maximize } E^{\gamma}_{x_1}[g(\lambda \circ r_1(x_1,u_1)\circ r_2(x_2,u_2)\circ\cdots\circ r_N(x_N,u_N)\circ r_G(x_{N+1}))]\\ &\text{subject to (i)}_n \quad x_{n+1}\sim p(\cdot|\, x_n,u_n)\quad n=1,2,\ldots,N\\ &\qquad\text{(ii)}_n \quad \gamma=\{\gamma_1,\gamma_2,\ldots,\gamma_N\}\in\Gamma\end{aligned}$$

この問題において，λ に演算子 $\circ$ の単位元を代入すれば，元の問題と同値になることは明らかである．すなわち，これは，元の問題を埋め込んだ問題とみなせる．この埋め込み問題に対し再帰式を求めるべく，n 期以降に問題を限定した次の部分問題群を考え，その最適値関数を v^n であらわす:

部分問題群

$$\begin{aligned}v^N(x_{N+1},\lambda) &= g(\lambda\circ r_G(x_{N+1})),\qquad x_{N+1}\in X\\ v^n(x_n,\lambda) &= \text{Max}\ \{\ E^{\gamma}_{x_n}[g(\lambda\circ r_n(x_n,u_n)\circ\cdots\circ r_G(x_{N+1}))]\ |\\ &\qquad x_i\sim p(\cdot\,|x_{i-1},u_{i-1}),\ \gamma\in\Gamma_n,\ \ i=n,\ldots,N\},\qquad x_n\in X\end{aligned}$$

このとき，次の再帰式が成り立つ．

再帰式

$$\begin{aligned}v^{N+1}(x,\lambda) &= g(\lambda\circ r_G(x)),\qquad x\in X\\ v^n(x,\lambda) &= \max_{u\in U}\sum_{y\in X} v^{n+1}(y,\lambda\circ r_n(x,u))p(y\,|\,x,u),\ x\in X\qquad n=1,2,\ldots,N\end{aligned}$$

5.2 最適政策

各再帰式の計算においてその最大値 $(v^n(x,\lambda))$ を与える**決定の集合**を

$$\pi^*_n(x,\lambda),\quad n=1,2,\ldots,N$$

とおく．このとき，元の問題に対する最適集合値原始政策 $\gamma^*=\{\gamma^*_1,\gamma^*_2,\ldots,\gamma^*_N\}$ は次のように構成できる．

$$\gamma^*_1(x_1)=\pi^*_1(x_1,\lambda_1)$$

$$\lambda_1=\hat{\lambda}\quad(\hat{\lambda}\text{は}\circ\text{の単位元})$$

$$\gamma_2^*(x_1,u_1,x_2)=\pi_2^*(x_2,\lambda_2),$$
$$\lambda_2=\lambda_1\circ r_1(x_1,u_1),\ u_1\in\gamma_1^*(x_1)$$

$$\gamma_3^*(x_1,u_1,x_2,u_2,x_3)=\pi_3^*(x_3,\lambda_3),$$
$$\lambda_3=\lambda_2\circ r_2(x_2,u_2),\ u_1\in\gamma_1^*(x_1),\ u_2\in\gamma_2^*(x_1,u_1,x_2)$$
$$\vdots$$

$$\gamma_N^*(x_1,u_1,x_2,\ldots,u_{N-1},x_N)=\pi_N^*(x_N,\lambda_N),$$
$$\lambda_N=\lambda_{N-1}\circ r_{N-1}(x_{N-1},u_{N-1}),$$
$$u_1\in\gamma_1^*(x_1),\ u_2\in\gamma_2^*(x_1,u_1,x_2),\ldots,$$
$$u^{N-1}\in\gamma_{N-1}^*(x_1,u_1,x_2,\ldots,u^{N-2},x^{N-1})$$

6 まとめ

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通りの政策クラスを定義した．そして，それぞれのクラスの表現力の違いを例により示した．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．ただし，一般政策あるいは集合値原始政策のクラスを用いれば，ここで想定している問題に対しては，すべての決定ツリーが表現可能である．原始政策の必要性は，確率的に決定を取るという状況で（それがもし考えられるならば）起こる．

一般に同時最適化の観点からは，1点をとる決定関数のほうが考えやすいが，一方，逐次最適化の観点からは，集合値を取る決定関数のほうが扱いやすいように思われる．

References

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

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

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

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

[6] S. Iwamoto, T. Ueno and T. Fujita, Controlled Markov Chains with Utility Functions, Proc. of Intl Workshop on Markov Process and Controlled Markov Chains; Changsha, China, 2000.