数理解析研究所講究録
1132巻 2000年 15-23



# 確率最適化における再帰式と決定樹表

九大・経済　岩本誠一

## 1 はじめに

動的計画法 [1] は、多変数実数値関数の**同時最適化**を1変数ずつ逐次最適化する方法である。これは、目的関数の**再帰性** recursiveness（**可分性** separability ともいう）と**単調性** monotonicity の下で可能である [6][7][8][18]。このとき、**逐次最適化**の最適解が本来の同時最適解でなければならない。すなわち、逐次最適化よる最適値が本来の最適値に一致し、逐次最適点が同時最適点（全体の少なくとも一部を）を与えることである。動的計画法が最も適用される目的関数は**加法型関数**であり、**確定性の下**である。この状況では再帰性と狭義単調性は満たされている。したがって、ただちに動的計画法を適用することができる。実際、**最適性の原理**より再帰式を導いて、これを解いて本来の多変数同時最適化の最適解を求めることができる。通常、動的計画法とはこのことを指している [1]。

最近、不確実性下における経済動学やファジィ環境下における多段意思決定には**非加法型評価基準**が用いられている [3][5][16]。ここでは加法型と同様なアプローチがとられているが、再帰性を政策（決定関数列）との兼ね合いでとらえなければならない [10][11][12][13]。それに応じて、政策を、(i) 現状態のみ、(ii) 現在までの状態列、(iii) 現在までの状態と決定の交互列、の依存性によって、(i) マルコフ、(ii) 一般、(iii) 原始、の三つ考える必要がある [14][15]。

本報告では、加法型評価（自身の直接の期待値最適ではなく）の閾値確率最適化を（マルコフ政策クラスではなく）一般政策クラスおよび原始政策クラスの中で考える。まず、（1）パラメータを導入した動的計画法、と（2）履歴（本来の状態と決定の交互列）を新状態にした動的計画法、の二つで解く。次に、前述のように同時最適解を与えるという意味で、二つの動的計画法が正しく機能しているかをチェックする簡単な方法として**多段確率決定樹表** multi-stage stochastic decision tree-table による方法が有効であることを示す。これは決定樹と決定表からなり、問題記述から最適解に至るまでが再帰式を解く順に構成されている。

## 2 記号と用語

まず、本論文で用いる記号と用語を述べておこう。

(1) $N \geq 2$ は**段の総数** (total number of stage) を表す正整数

(2) $X = \{s_1, s_2, \ldots, s_p\}$ は有限**状態空間** (state space)

(3) $U = \{a_1, a_2, \ldots, a_k\}$ は有限**決定空間** (action space)

(4) $r_n : X \times U \to R^1$ は**第 $n$ 利得関数** ($n$-th reward function)　$(0 \leq n \leq N-1)$
$r_N : X \to R^1$ は**終端利得関数** (terminal reward function)

(5) $p = \{p(y|x,u)\}$ は**マルコフ推移法則** (Markov transition law)
: $p(y|x,u) \geq 0 \quad \forall (x,u,y) \in X \times U \times X$
$\sum_{y \in X} p(y|x,u) = 1 \quad \forall (x,u) \in X \times U$
; $p(y|x,u)$ は現在状態 $\tilde{X}$ が $x$ で現在の決定 $\tilde{U}$ が $u$ になったとき、次の状態 $\tilde{Y}$ が $y$ になる条件付き確率を表す：$P(\tilde{Y} = y|\tilde{X} = x, \tilde{U} = u) = p(y|x,u)$. ただし ˜ は確率変数を表わす。この確率的推移を $\tilde{Y} \sim p(\cdot|x,u)$ で表現する。

(6) $\pi = \{\pi_0, \pi_1, \ldots, \pi_{N-1}\}$ は**マルコフ政策** (Markov policy)

: $\pi_0 : X \to U, \quad \pi_1 : X \to U, \quad \ldots \quad , \quad \pi_{N-1} : X \to U$
マルコフ政策の全体を $\Pi$ で表わす。

(6)′ $\sigma = \{\sigma_0, \sigma_1, \ldots, \sigma_{N-1}\}$ は**一般政策** (general policy)
: $\sigma_0 : X \to U, \quad \sigma_1 : X \times X \to U, \quad \ldots \quad , \quad \sigma_{N-1} : X \times \cdots \times X \to U$
一般政策の全体を $\Pi_g$ で表わす。

(6)″ $\mu = \{\mu_0, \mu_1, \ldots, \mu_{N-1}\}$ は**原始政策** (primitive policy)
: $\mu_0 : X \to U, \quad \mu_1 : X \times U \times X \to U, \quad \ldots$
$\quad , \quad \mu_{N-1} : X \times U \times X \times U \times \cdots \times U \times X \to U$
原始政策の全体を $\Pi_p$ で表わす。

# 3 閾値確率最適化

問題は、不確実な状況の下で得られる利得の総和があらかじめ定められた（所定の）値 $c$ 以上になる確率を最大にするように行動するには、意思決定者が各段で、それまでの状態に応じてどのように決定を取っていけばよいかである。これは次の閾値確率最大化問題になる [4],[19]：

$$\mathrm{P}_0(x_0) \qquad \begin{array}{ll} \text{Maximize} & P^{\sigma}_{x_0}(r_0 + r_1 + \cdots + r_{N-1} + r_N \geq c) \\ \text{subject to} & \text{(i)} \quad X_{n+1} \sim p(\cdot\,|x_n, u_n) \\ & \text{(ii)} \quad u_n \in U \end{array} \qquad n = 0, 1, \ldots, N-1 \tag{1}$$

ただし $P^{\sigma}_{x_0}$ は、初期状態 $x_0$, マルコフ推移法則 $p$ および 一般政策 $\sigma$ から履歴の直積空間

$$H_N = X \times U \times X \times U \cdots \times U \times X \quad (2N+1)\text{個}$$

上に唯一定まる確率測度である。また、この確率測度による期待値作用素を $E^{\sigma}_{x_0}$ で表わす。

## 3.1 一般政策クラス問題

この小節では、一般政策クラス $\Pi_g$ 上での閾値確率最適化を考える 意思決定者が一般政策 $\sigma(\in \Pi_g)$ を採用すると、最大化問題 (1) の閾値確率は「部分」多重和

$$\begin{aligned} & P^{\sigma}_{x_0}(r_0 + r_1 + \cdots + r_{N-1} + r_N \geq c) \\ = & \sum\sum\cdots\sum_{(x_1, x_2, \ldots, x_N) \in (*)} p(x_1|x_0, u_0) p(x_2|x_1, u_1) \cdots p(x_N|x_{N-1}, u_{N-1}) \end{aligned} \tag{2}$$

で表わされる。ただし、多重和をとる領域 $(*)$ は

$$r_0(x_0, u_0) + r_1(x_1, u_1) + \cdots + r_{N-1}(x_{N-1}, u_{N-1}) + r_N(x_N) \quad \geq \quad c \tag{3}$$

を満たす $(x_1, x_2, \ldots, x_N) \in X \times X \times \cdots \times X$ 全体にわたる多重和である。ここに、式 (2),(3) における決定列 $\{u_0, u_1, \cdots, u_{N-1}\}$ は一般政策 $\sigma = \{\sigma_0, \ldots, \sigma_{N-1}\}$ を通して定まっている：

$$u_0 = \sigma_0(x_0), \; u_1 = \sigma_0(x_0, x_1), \; \ldots, \; u_{N-1} = \sigma_0(x_0, x_1, \ldots, x_{N-1}). \tag{4}$$

さて、われわれの求める最適解は問題 (1) の最大値関数 $v_0 = v_0(x_0)$ および最大値を与える最適政策 $\sigma^*$ である：

$$\begin{aligned} v_0(x_0) &= P^{\sigma^*}_{x_0}(r_0 + \cdots + r_{N-1} + r_N \geq c) \\ &= \operatorname*{Max}_{\sigma \in \Pi_g} P^{\sigma}_{x_0}(r_0 + \cdots + r_{N-1} + r_N \geq c) \quad x_0 \in X. \end{aligned} \tag{5}$$

この問題を**一般政策クラス問題**、または短く**一般問題**と呼ぶ。

一般に、確率変数 $Y$ が $c$ 以上になる確率 $P(Y \geq c)$ は、定義関数

$$\psi(y) := 1_{[c,\infty)}(y) := \begin{cases} 1 & y \geq c \\ 0 & \text{その他} \end{cases} \tag{6}$$

を通した確率変数 $\psi(Y)$ の期待値 $E[\psi(Y)]$ である：

$$P(Y \geq c) = E[\psi(Y)]. \tag{7}$$

したがって、一般問題 (1) の閾値確率は定義関数 $\psi = \psi(\cdot)$ を通した期待値になる：

$$P^{\sigma}_{x_0}(r_0 + \cdots + r_{N-1} + r_N \geq c) \;=\; E^{\sigma}_{x_0}[\psi(r_0 + \cdots + r_{N-1} + r_N)]. \tag{8}$$

すなわち、「部分」多重和は定義関数を通した「全」多重和に等しい：

$$\begin{aligned} &\sum_{(x_1,x_2,\ldots,x_N)\in(*)} \sum\sum\cdots\sum p(x_1|x_0,u_0)p(x_2|x_1,u_1)\cdots p(x_N|x_{N-1},u_{N-1}) \\ &= \sum_{(x_1,x_2,\ldots,x_N)\in X\times X\times\cdots\times X} \sum\sum\cdots\sum \{\psi(r_0(x_0,u_0) + \cdots + r_{N-1}(x_{N-1},u_{N-1}) + r_N(x_N)) \\ &\qquad \times p(x_1|x_0,u_0)p(x_2|x_1,u_1)\cdots p(x_N|x_{N-1},u_{N-1})\}. \end{aligned} \tag{9}$$

## 3.2 拡大マルコフ政策クラス問題

したがって、「閾値確率」最大化問題は次の「期待値」最大化問題になる：

$$\begin{array}{ll} \text{Maximize} & E^{\sigma}_{x_0}[\psi(r_0 + r_1 + \cdots + r_{N-1} + r_N)] \\ \text{subject to} & \text{(i)} \;\; X_{n+1} \sim p(\cdot\,|x_n,u_n) \\ & \text{(ii)} \;\; u_n \in U \end{array} \qquad n = 0,1,\ldots,N-1 \tag{10}$$

この問題を、新たに過去値をパラメータとする問題に埋め込んで考える [2][10][11][12][13][14][17]。まず、第 $n$ 段までの**集積値**確率変数列 $\{\tilde{\Lambda}_n\}$ および それらが取り得る**過去値集合**列 $\{\Lambda_n\}$ をそれぞれ次で定義する：

$$\begin{aligned} \tilde{\Lambda}_0 &\triangleq 0 \\ \tilde{\Lambda}_n &\triangleq r_0(X_0,U_0) + \cdots + r_{n-1}(X_{n-1},U_{n-1}) \quad n = 1,\ldots,N \end{aligned} \tag{11}$$

$$\begin{aligned} \Lambda_0 &\triangleq \{0\} \\ \Lambda_n &\triangleq \{\lambda_n \,|\, \lambda_n = r_0(x_0,u_0) + \cdots + r_{n-1}(x_{n-1},u_{n-1}), \\ &\qquad (x_0,u_0,\ldots,x_{n-1},u_{n-1}) \in X\times U\times\cdots\times X\times U\} \\ &\qquad\qquad n = 1,\ldots,N \end{aligned} \tag{12}$$

このとき、総利得は次になる：

$$r_0 + r_1 + \cdots + r_{N-1} + r_N = \tilde{\Lambda}_N + r_N. \tag{13}$$

式 (11) は漸化式

$$\begin{aligned} \tilde{\Lambda}_0 &= 0 \\ \tilde{\Lambda}_{n+1} &= \tilde{\Lambda}_n + r_n(X_n,U_n) \quad n = 0,\ldots,N-1 \end{aligned} \tag{14}$$

に同値である。また、相隣る過去値集合 $\{\Lambda_{n-1},\ \Lambda_n\}$ 間には次の前向き再帰式が成り立つ：

**補題 3.1** （過去値集合間の再帰式）

$$\Lambda_0 = \{0\}$$
$$\Lambda_n = \{\lambda + r_{n-1}(x,u) \,|\, \lambda \in \Lambda_{n-1},\ (x,u) \in X \times U\} \quad n = 1,2,\ldots,N. \tag{15}$$

さらに、本来の状態空間 $X$ に過去値集合を貼り合せた拡大状態空間列 $\{Y_n\}$ を直積で定義する：

$$Y_n \triangleq X \times \Lambda_n \qquad n = 0,1,\ldots,N. \tag{16}$$

この新状態空間列上のマルコフ政策 $\gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_N\}$ は**マルコフ決定関数列**

$$\gamma_n : Y_n \to U,\ \ (n = 1,\ 2,\ \ldots, N)$$

で定まる。これを**拡大マルコフ政策**といい、その全体を $\tilde{\Pi}$ で表わす。新たに**終端利得関数** $T$ を

$$T(x;\lambda) \triangleq \psi(\lambda + r_N(x)) \quad (x;\lambda) \in Y_N \tag{17}$$

で定義する。さらに、定常マルコフ推移法則 $p = \{p(y|x,u)\}$ および パラメータ・ダイナミックス $\{\lambda_{n+1} = \lambda_n + r_n(x_n, u_n)\}$ によって拡大状態空間列上に定まる**非定常マルコフ推移法則** $q = \{q_n\}$ を

$$q_n((y;\mu) \,|\, (x;\lambda),\, u) \triangleq \begin{cases} p(y|x,\,u) & \lambda + r_{n-1}(x,u) = \mu \text{ のとき} \\ 0 & \text{その他.} \end{cases} \tag{18}$$

で定義する。このとき、拡大マルコフ政策空間上の**終端型評価問題**

$$\begin{array}{lll} & \text{Maximize} & \tilde{E}^{\gamma}_{y_0}[\psi(\tilde{\Lambda}_N + r_N(X_N))] \\ \mathrm{Q}_0(y_0) & \text{subject to} & \text{(i)}\ \ X_{n+1} \sim p(\cdot\,|x_n, u_n) \\ & & \text{(i)}'\ \ \tilde{\Lambda}_{n+1} = \tilde{\Lambda}_n + r_n(X_n, U_n) \quad n = 0,1,\ldots,N-1 \\ & & \text{(ii)}\ \ u_n \in U \end{array} \tag{19}$$

を考える。ただし、 $y_0 = (x_0; 0)$. ここに $\tilde{E}^{\gamma}_{y_0}$ は、初期状態 $y_0$ 、拡大マルコフ政策 $\gamma$ および新マルコフ推移法則 $q$ によって拡大状態空間列上に定まる確率測度 $\tilde{P}^{\gamma}_{y_0}$ に基づく期待値作用素である ([15])。

さて、第 $n$ 段の状態 $y_n = (x_n; \lambda_n)(\in Y_n)$ から始まる部分過程

$$\begin{array}{lll} & \text{Maximize} & \tilde{E}^{\gamma}_{y_n}[\psi(\tilde{\Lambda}_N + r_N(X_N))] \\ \mathrm{Q}_n(y_n) & \text{subject to} & \text{(i)}\ \ X_{m+1} \sim p(\cdot\,|x_m, u_m) \\ & & \text{(i)}'\ \ \tilde{\Lambda}_{m+1} = \tilde{\Lambda}_m + r_m(X_m, U_m) \quad m = n,\ldots,N-1 \\ & & \text{(ii)}\ \ u_m \in U \end{array} \tag{20}$$

の最大値を $u^n(x_n; \lambda_n)$ とする。ただし

$$u^N(x_N; \lambda_N) \triangleq \psi(\lambda_N + r_N(x_N)) \qquad (x_N; \lambda_N) \in Y_N. \tag{21}$$

このとき、次の後向きの再帰式が成り立つ：

**定理 3.1** （拡大マルコフ政策クラス問題の再帰式）

$$\begin{aligned} u^N(x;\lambda) &= \psi(\lambda + r_N(x)) \qquad x \in X,\ \lambda \in \Lambda_N \\ u^n(x;\lambda) &= \operatorname*{Max}_{u \in U} \sum_{y \in X} u^{n+1}(y; \lambda + r_n(x,u)) p(y|x,u) \\ & \qquad\qquad x \in X,\ \lambda \in \Lambda_n,\ \ 0 \le n \le N-1. \end{aligned} \tag{22}$$

さて、式 (22) の最大（値を与える）点を $\gamma_n^*(x;\lambda)$ とすると、拡大マルコフ政策クラス $\tilde{\Pi}$ の中での最適政策 $\gamma^* = \{\gamma_0^*, \gamma_1^*, \ldots, \gamma_{N-1}^*\}$ が得られる ([15, Theorem 4.2])。さらに、$\gamma^*$ から、以下のように一般政策 $\sigma^* = \{\sigma_0^*, \sigma_1^*, \ldots, \sigma_{N-1}^*\}$ を生成する：ただし、$\sigma_n^*(x_0, x_1, \ldots, x_n)$ は

$$\begin{aligned}
&u_0 := \gamma_1^*(x_0; 0), \quad \lambda_1 := 0 + r(x_0, u_0) \\
&u_1 := \gamma_1^*(x_1; \lambda_1), \quad \lambda_2 := \lambda_1 + r(x_1, u_1) \\
&\qquad\vdots \\
&u_{n-1} := \gamma_{n-1}^*(x_{n-1}; \lambda_{n-1}), \quad \lambda_n := \lambda_{n-1} + r(x_{n-1}, u_{n-1}) \\
&\sigma_n^*(x_0, x_1, \ldots, x_n) := \gamma_n^*(x_n; \lambda_n).
\end{aligned} \tag{23}$$

このとき、次が成り立つ：

**定理 3.2** （拡大マルコフクラスと一般クラスの等価性 [15, Theorem 6.1]）
(i) 政策 $\sigma^*$ は一般政策クラス $\Pi_g$ の中での最適である。
(ii) 拡大マルコフ政策クラス $\tilde{\Pi}$ の最大値は一般政策クラス $\Pi_g$ の最大値に等しい：

$$u^0(x_0; 0) = v_0(x_0). \tag{24}$$

## 3.3 原始政策クラス問題

本来の閾値確率制御問題 (1) は一般政策クラス $\Pi_g$ の中での最大化であるが、本節ではこのクラスより広い原始政策クラス $\Pi_p$ の中で最適化しても最適政策が一般政策として得られることを示す。

まず、原始政策クラス $\Pi_p$ 上の閾値確率最大化問題は次になる：

$$\begin{array}{lll}
 & \text{Maximize} & P_{x_0}^{\mu}(r_0 + r_1 + \cdots + r_{N-1} + r_N \geq c) \\
\mathrm{R}_0(x_0) & \text{subject to} & \text{(i)} \quad X_{n+1} \sim p(\cdot\,|x_n, u_n) \qquad n = 0, 1, \ldots, N-1 \\
 & & \text{(ii)} \quad u_n \in U
\end{array} \tag{25}$$

これを**原始政策クラス問題**、または簡単に**原始問題**と呼ぶ。ただし、決定列 $u_0, u_1, \ldots, u_N$ のはそれ以前の決定列にも依存して定まっている：

$$u_0 = \mu_0(x_0),\ u_1 = \mu_0(x_0, u_0, x_1),\ \ldots,\ u_{N-1} = \mu_0(x_0, u_0, x_1, u_1, \ldots, u_{N-2}, x_{N-1}). \tag{26}$$

また原始問題も「期待値」最大化問題：

$$\begin{array}{lll}
 & \text{Maximize} & E_{x_0}^{\mu}[\psi(r_0 + \cdots + r_{N-1} + r_N)] \\
\mathrm{R}_0(x_0) & \text{subject to} & \text{(i)} \quad X_{n+1} \sim p(\cdot\,|x_n, u_n) \qquad n = 0, 1, \ldots, N-1 \\
 & & \text{(ii)} \quad u_n \in U
\end{array} \tag{27}$$

になる。この期待値最大化問題に対して、履歴 $h_n = (x_0, u_0, \ldots, x_{n-1}, u_{n-1}, x_n)(\in H_n)$ から始まる部分問題

$$\begin{array}{lll}
 & \text{Maximize} & E_{h_n}^{\mu}[\psi(r_0 + \cdots + r_{N-1} + r_N)] \\
\mathrm{R}_n(h_n) & \text{subject to} & \text{(i)} \quad X_{n+1} \sim p(\cdot\,|x_n, u_n) \qquad n = 0, 1, \ldots, N-1 \\
 & & \text{(ii)} \quad u_n \in U
\end{array} \tag{28}$$

の原始政策 $\mu = \{\mu_n, \mu_{n+1}, \ldots, \mu_{N-1}\}$ にわたる最大値を $w_n(h_n)$ とする。ただし

$$w_N(h_N) \triangleq \psi(r_0(x_0, u_0) + \cdots + r_{N-1}(x_{N-1}, u_{N-1}) + r_N(x_N)). \tag{29}$$

このとき、後向きの再帰式が成り立つ：

**定理 3.3** （原始政策クラス問題の再帰式）

$$w_n(h) = \underset{u\in U}{\mathrm{Max}} \sum_{y\in X} w_{n+1}(h,u,y)p(y|x,u) \quad h \in H_n, \quad 1 \le n \le N-1 \tag{30}$$

$$w_{N+1}(h) = \psi(r_0(x_0,u_0) + \cdots + r_{N-1}(x_{N-1},u_{N-1}) + r_N(x_N)) \quad h \in H_N. \tag{31}$$

さて、式 (30) の最大点を $\hat{\mu}_n(h)$ とすると、原始政策クラス $\Pi_g$ の中での最適政策 $\hat{\mu} = \{\hat{\mu}_0, \hat{\mu}_1, \ldots, \hat{\mu}_{N-1}\}$ が得られる ([15, Theorem 4.4])。さらに、$\hat{\mu}$ から、以下のように一般政策 $\hat{\sigma} = \{\hat{\sigma}_0, \hat{\sigma}_1, \ldots, \hat{\sigma}_{N-1}\}$ を生成する：ただし

$$\begin{aligned}
\hat{\sigma}_0(x_0) &:= \hat{\mu}_0(x_0) \\
\hat{\sigma}_1(x_0,x_1) &:= \hat{\mu}_1(x_0,u_0,x_1) \quad \text{ただし } u_0 = \hat{\mu}_0(x_0) \\
&\vdots \\
\hat{\sigma}_{N-1}(x_1,x_2,\ldots,x_{N-1}) &:= \hat{\mu}_{N-1}(x_1,u_1,x_2,\ldots,u_{N-2},x_{N-1}) \\
&\quad \text{ただし } u_0 = \hat{\mu}_0(x_0), \quad u_1 = \hat{\mu}_1(x_0,u_0,x_1), \\
&\qquad \ldots, \ u_{N-2} = \hat{\mu}_{N-2}(x_0,u_0,x_1,\ldots,u_{N-3},x_{N-2}).
\end{aligned} \tag{32}$$

このとき、次が成り立つ：

**定理 3.4** （原始クラスと一般クラスの等価性 [15, Lemma 5.1]）
(i) 政策 $\hat{\sigma}$ は一般政策クラス $\Pi_g$ の中での最適である。
(ii) 原始政策クラス $\Pi_p$ 上の最大値は一般政策クラス $\Pi_g$ 上の最大値に等しい：

$$w_0(x_0) = v_0(x_0). \tag{33}$$

# 4 3状態2決定2段問題

この節では、3-2-2（3状態2決定2段）モデルにおいて（加法型）総合評価値が $c = 2.5$ 以上になる閾値確率を最大化する問題を考える：

$$\begin{aligned}
\text{Maximize} \quad & P^{\sigma}_{x_0}(\, r_0(U_0) + r_1(U_1) + r_2(X_2) \ge 2.5\,) \\
\text{subject to} \quad & \text{(i)} \ \ X_{n+1} \sim p(\cdot|x_n,u_n) \quad n = 0,1 \\
& \text{(ii)} \ \ u_0 \in U, \ u_1 \in U
\end{aligned} \tag{34}$$

ただし、次のような Bellman and Zadeh[3, pp.B154] の数値例を用いる：

$r_2(s_1) = 0.3 \quad r_2(s_2) = 1.0 \quad r_2(s_3) = 0.8$

$r_1(a_1) = 1.0 \quad r_1(a_2) = 0.6$

$r_0(a_1) = 0.7 \quad r_0(a_2) = 1.0$

$u_t = a_1$

| $x_t \setminus x_{t+1}$ | $s_1$ | $s_2$ | $s_3$ |
|---|---|---|---|
| $s_1$ | 0.8 | 0.1 | 0.1 |
| $s_2$ | 0.0 | 0.1 | 0.9 |
| $s_3$ | 0.8 | 0.1 | 0.1 |

$u_t = a_2$

| $x_t \setminus x_{t+1}$ | $s_1$ | $s_2$ | $s_3$ |
|---|---|---|---|
| $s_1$ | 0.1 | 0.9 | 0.0 |
| $s_2$ | 0.8 | 0.1 | 0.1 |
| $s_3$ | 0.1 | 0.0 | 0.9 |

## 4.1 再帰式

まず、再帰式 (22) を解いて拡大マルコフ政策クラス $\tilde{\Pi}$ の中で最適値関数列 $\{u^0(x_0;\lambda_0),\ u^1(x_1;\lambda_1),\ u^2(x_1;\lambda_2)\}$ および最適政策 $\gamma^* = \{\gamma_0^*(x_0;\lambda_0),\ \gamma_1^*(x_1;\lambda_1))\}$ が求められる。これをまとめると、表 1 になる。

| | $u^2(x_2;\lambda_2)$ | | | | $u^1(x_1;\lambda_1)$ | $\gamma_1^*(x_1;\lambda_1)$ | $u^0(x_0;0)$ | $\gamma_0^*(x_0;0)$ |
|---|---|---|---|---|---|---|---|---|
| $x_n \diagdown \lambda_n$ | 1.3 | 1.6 | 1.7 | 2.0 | 0.7 | 1.0 | 0 | |
| $s_1$ | 0 | 0 | 0 | 0 | 0.2 | $a_1$ | 0.9 | $a_2$ | 0.99 | $a_2$ |
| $s_2$ | 0 | 1 | 1 | 1 | 1.0 | $a_1$ | 1.0 | $a_1$ | 0.84 | $a_2$ |
| $s_3$ | 0 | 0 | 1 | 1 | 0.2 | $a_1$ | 0.2 | $a_1$ | 0.28 | $a_1$ |

表 1 拡大マルコフ政策クラスの最適解

次に、再帰式 (30) を解くと、原始政策クラス $\Pi_p$ の中で最適値関数列 $\{w^0(x_0),\ w^1(x_0,u_0,x_1),\ w^2(x_0,u_0,x_1,u_1,x_2)\}$ および最適政策 $\hat{\mu} = \{\ \hat{\mu}_0(x_0),\ \ \hat{\mu}_1(x_0,u_0,x_1))\}$ が得られる。これをまとめると、表 2,3,4 になる（表 3,4 省略）。

このとき、$\gamma^*$ から生成される一般政策 $\sigma^*$ は $\hat{\mu}$ から生成される一般政策 $\hat{\sigma}$ に一致し、これはまた次の多段確率決定樹表によっても得られることが分かる。

| $x_0$ | $w_0(x_0)$ | $\hat{\mu}_0(x_0)$ |
|---|---|---|
| $s_1$ | 0.99 | $a_2$ |
| $s_2$ | 0.84 | $a_2$ |
| $s_3$ | 0.28 | $a_1$ |

表 2 $\{w_0(x_0)\ \ \hat{\mu}_0(x_0)\}$

## 4.2 多段確率決定樹表

多段確率決定樹表は、問題のデータを過程の進行状況に応じて配列し、あらゆる可能な経路とその評価値・確率を図示し、各段における最適決定の選択を明示している。この意味では列挙法の解構成を与えている。しかし、最適解に至るまでは動的計画法の再帰式を解く順に構成されている。この樹表ではあらゆる型の評価関数期待値最適化が解かれる。

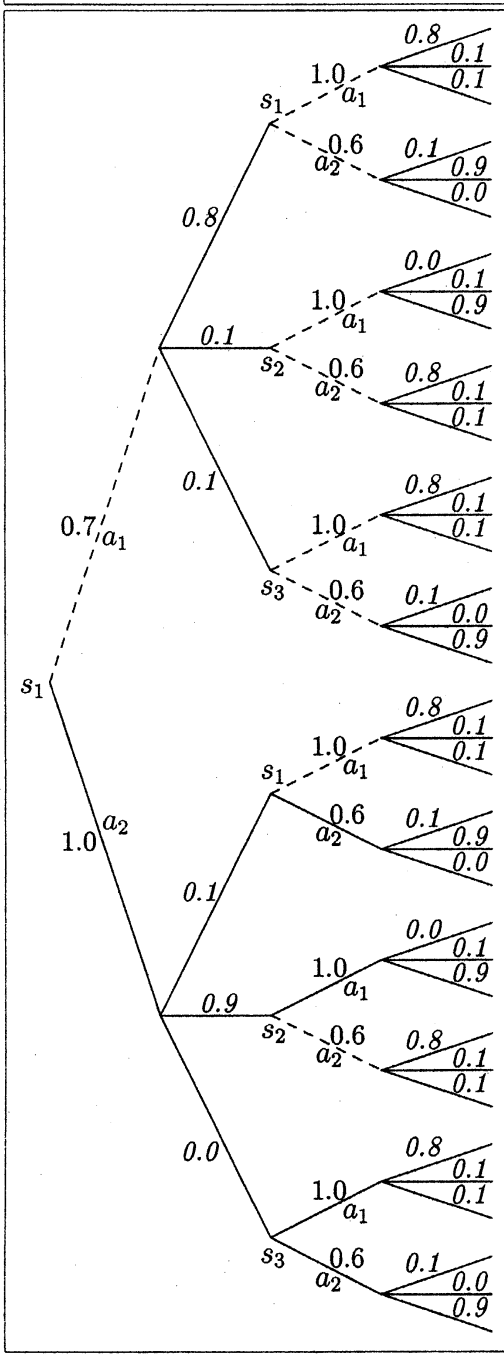
次頁の樹表（図 1）では、3-2-2 型（3 状態 2 決定 2 段）加法モデルに対して次のように簡略化している (数値自体は Bellman and Zadeh (1970) のデータ)：

$$
\begin{aligned}
&\text{履歴} = x_0\ \ r_0(u_0)/u_0\ \ p_0\ \ x_1\ \ r_1(u_1)/u_1\ \ p_1\ \ x_2\ \ r_G(x_2)\\
&\qquad \text{ただし } p_0 = p(x_1 \mid x_0,u_0),\quad p_1 = p(x_2 \mid x_1,u_1)\\
&\text{加法} = \text{加法型評価値} = r_0(u_0) + r_1(u_1) + r_G(x_2)\\
&\text{経路} = \text{経路確率} = p_0p_1\\
&\text{閾確} = \text{閾値確率} = \psi(r_0(u_0) + r_1(u_1) + r_G(x_2))p_0p_1\\
&\qquad \text{ただし } \psi(y) = 1_{[2.5,\infty)}(y)\\
&\text{部分確} = \text{部分確率} = \sum_{x_2}\psi(r_0(u_0) + r_1(u_1) + r_G(x_2))p_0p_1\\
&\text{全確率} = \text{全体確率} = \sum_{x_1}\sum_{x_2}\psi(r_0(u_0) + r_1(u_1) + r_G(x_2))p_0p_1.
\end{aligned}
$$

イタリック体は確率を、ボールド体は上下の確率のうち大きい方を選択したことを表す。特に、履歴の欄では 5 つの数値 $r_0 = r_0(u_0)$, $p_0$, $r_1 = r_1(u_1)$, $p_1$, $r_G = r_G(x_2)$ のみを記している。

$$V^0(s_1) = \underset{\mu}{\mathrm{Max}}\, P^{\mu}_{s_1}(\, r_0(U_0) + r_1(U_1) + r_G(X_2) \geq 2.5\,)$$

図 1：状態 $s_1$ からの 2 段確率決定樹表




| $r_0$ | $p_0$ | 履歴 | $r_1$ | $p_1$ | $r_G$ | 加法 | 経路 | 閾確 | 部分確 | 全確率 |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | $s_1$ 0.3 | 2.0 | *0.64* | 0 | | |
| | | | | | $s_2$ 1.0 | 2.7 | *0.08* | 0.08 | **0.16** | |
| | | | | | $s_3$ 0.8 | 2.5 | *0.08* | 0.08 | | |
| | | | | | $s_1$ 0.3 | 1.6 | *0.08* | 0 | | |
| | | | | | $s_2$ 1.0 | 2.3 | *0.72* | 0 | 0 | |
| | | | | | $s_3$ 0.8 | 2.1 | *0.0* | 0 | | |
| | | | | | $s_1$ 0.3 | 2.0 | *0.0* | 0 | | |
| | | | | | $s_2$ 1.0 | 2.7 | *0.01* | 0.01 | **0.1** | |
| | | | | | $s_3$ 0.8 | 2.5 | *0.09* | 0.09 | | 0.28 |
| | | | | | $s_1$ 0.3 | 1.6 | *0.08* | 0 | | |
| | | | | | $s_2$ 1.0 | 2.3 | *0.01* | 0 | 0 | |
| | | | | | $s_3$ 0.8 | 2.1 | *0.01* | 0 | | |
| | | | | | $s_1$ 0.3 | 2.0 | *0.08* | 0 | | |
| | | | | | $s_2$ 1.0 | 2.7 | *0.01* | 0.01 | **0.02** | |
| | | | | | $s_3$ 0.8 | 2.5 | *0.01* | 0.01 | | |
| | | | | | $s_1$ 0.3 | 1.6 | *0.01* | 0 | | |
| | | | | | $s_2$ 1.0 | 2.3 | *0.0* | 0 | 0 | |
| | | | | | $s_3$ 0.8 | 2.1 | *0.09* | 0 | | |
| | | | | | $s_1$ 0.3 | 2.3 | *0.08* | 0 | | |
| | | | | | $s_2$ 1.0 | 3.0 | *0.01* | 0.01 | 0.02 | |
| | | | | | $s_3$ 0.8 | 2.8 | *0.01* | 0.01 | | |
| | | | | | $s_1$ 0.3 | 1.9 | *0.01* | 0 | | |
| | | | | | $s_2$ 1.0 | 2.6 | *0.09* | 0.09 | **0.09** | |
| | | | | | $s_3$ 0.8 | 2.4 | *0.0* | 0 | | |
| | | | | | $s_1$ 0.3 | 2.3 | *0.0* | 0 | | |
| | | | | | $s_2$ 1.0 | 3.0 | *0.09* | 0.09 | **0.9** | |
| | | | | | $s_3$ 0.8 | 2.8 | *0.81* | 0.81 | | **0.99** |
| | | | | | $s_1$ 0.3 | 1.9 | *0.72* | 0 | | |
| | | | | | $s_2$ 1.0 | 2.6 | *0.09* | 0.09 | 0.09 | |
| | | | | | $s_3$ 0.8 | 2.4 | *0.09* | 0 | | |
| | | | | | $s_1$ 0.3 | 2.3 | *0.0* | 0 | | |
| | | | | | $s_2$ 1.0 | 3.0 | *0.0* | 0.0 | **0** | |
| | | | | | $s_3$ 0.8 | 2.8 | *0.0* | 0.0 | | |
| | | | | | $s_1$ 0.3 | 1.9 | *0.0* | 0 | | |
| | | | | | $s_2$ 1.0 | 2.6 | *0.0* | 0.0 | **0** | |
| | | | | | $s_3$ 0.8 | 2.4 | *0.0* | 0 | | |

ただし $P^{\mu}_{s_1}(\, r_0(U_0) + r_1(U_1) + r_G(X_2) \geq 2.5\,) = E^{\mu}_{s_1}\psi(r_0 + r_1 + r_G)$ に注意.

# References


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

[2] R.E. Bellman and E.D. Denman, Invariant Imbedding, Lect. Notes in Operation Research and Mathematical Systems **52**, Springer-Verlag, Berlin, 1971.

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

[4] M. Bouakiz and Y. Kebir, Target-level criterion in Markov decision processes, Journal of Optimization Theory and Applications **86** (1995), 1-15.

[5] A.O. Esogbue and R.E. Bellman, Fuzzy dynamic programming and its extensions, TIMS/Studies in the Management Sciences **20** (1984), 147-167.

[6] 岩本誠一, 動的計画の理論と応用, 数学 **31** 巻 4 号, 1979 年, 331-348.

[7] 岩本誠一, 動的計画論, 九大出版会, 1987.

[8] 岩本誠一, 動的計画の最近の進歩, 第 2 回ＲＡＭＰシンポジウム論文集, 1990 年, 129-140.

[9] S. Iwamoto, Associative dynamic programs, Journal of Mathematical Analysis and Applications, **201** (1996), 195-211.

[10] S. Iwamoto, Decision-making in fuzzy environment: a survey from stochastic decision process, Ed. L.C. Jain and R.K. Jain, Proceedings of The Second International Conference on Knowledge-based Intelligent Electronics Systems (KES '98), Adelaide, AUSTRALIA, April, 1998, pp.542-546.

[11] S. Iwamoto, Conditional decision processes with recursive reward function, Journal of Mathematical Analysis and Applications, **230** (1999), 193-210.

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

[13] S. Iwamoto and M. Sniedovich, Sequential decision making in fuzzy environment, Journal of Mathematical Analysis and Applications, **222** (1998), 208-224.

[14] S. Iwamoto, K. Tsurusaki and T. Fujita, Conditional decision-making in a fuzzy environment, J. Operations Res. Soc. Japan **42** (1999), 198-218.

[15] S. Iwamoto, T. Ueno and T. Fujita, Controlled Markov chains with utility functions, Proc. of The International Workshop on Markov Processes and Controlled Markov Chains, Changsha, China, August, 1999, under submission.

[16] J. Kacprzyk, Decision-making in a fuzzy environment with fuzzy termination time, Fuzzy Sets and Systems **1** (1978), 169-179.

[17] E.S. Lee, Quasilinearization and Invariant Imbedding, Academic Press, New York, 1968.

[18] M. Sniedovich, Dynamic Programming, Marcel Dekker, Inc. NY, 1992.

[19] C. Wu and Y. Lin, Minimizing risk models in Markov decision processed with policies depending on target values, Journal of Mathematical Analysis and Applications **231** (1999), 47-67.