数理解析研究所講究録
128巻 1971年 25-39



# マルコフ型決定過程における最適政策

九大 理学部 古川長太

## 第1章 最適政策の存在と構成

### §1. 準備

Borel set : ある complete separable metric space の Borel subset

$X$, $Y$ : Borel set

$\mathcal{B}(X)$ : $X$ における Borel field

$\mathcal{B}(Y)$ : $Y$ における Borel field

$P(X)$ : $X$ 上の確率測度の全体

$F(x)$ : $X$ 上で定義された集合値関数で

$$F(x) \subset Y \quad \forall x \in X$$

をみたす。

$Q(\{F(x)\} \mid X)$ : つぎの (i), (ii), (iii) をみたす条件付確率測度 $q$ の全体

(i) 各 $x \in X$ に対し, $q(\cdot \mid x)$ は $\mathcal{B}(Y)$ 上の確率測度

(ii) $q(F(x) \mid x) = 1 \quad \forall x \in X$

(iii) 各 $B \in \mathcal{B}(Y)$ に対し, $q(B \mid \cdot)$ は $Y$ 上の $\mathcal{B}(X)$ 可測関数

$M(Z)$ : Borel set $Z$ 上の有界 Borel 可測関数の全体

$A$ を $n$ 次元ユークリッド空間 $R^n$ の compact subset, $R^n$ の座標を $\{x_1, x_2, \cdots, x_n\}$ とする。

$$A_{x_1} \equiv \{y : y \in A, \ (x_1, y) = \max_{z \in A} (x_1, z)\}$$

以下帰納的に

$$A_{x_1 x_2 \cdots x_i} \equiv \{y : y \in A_{x_1 x_2 \cdots x_{i-1}}, \ (x_i, y) = \max_{z \in A_{x_1 \cdots x_{i-1}}} (x_i, z)\}$$

$e(A) \equiv A_{x_1 x_2 \cdots x_n}$ は1点として定まる。 $e(A)$ を, $A$ の lexicographic maximum と云う。

## §2. Selector の定理

$X$ : Borel set

$Y$ : compact metric space

$2^Y$ : $Y$ の closed subset の全体

$\mathcal{B}(2^Y)$ : $2^Y$ における Hausdorff metric $h$ で生成された Borel field

補題 2.1

$F(x) \in 2^Y \ \forall x \in X$, かつ $F(\cdot) \in \mathcal{B}(X)/\mathcal{B}(2^Y)$ ならば

$$\{(x, F(x)) ; x \in X\} \in \mathcal{B}(X) \otimes \mathcal{B}(Y)$$

補題 2.2

$g \in Q(\{F(x)\} \mid X)$ なる $g$ と，$g(\Gamma_x \mid x) > 0 \quad \forall x \in X$

かつ $\Gamma_x \subset F(x) \quad \forall x \in X$ なる $\Gamma \in \mathcal{B}(X) \otimes \mathcal{B}(Y)$ に対して，

$$(x, f(x)) \in \Gamma \qquad \forall x \in X$$

なる $\mathcal{B}(X)$ 可測な $f$ が存在する。

補題 2.3

$F(x) \in 2^Y \quad \forall x \in X$ かつ $F(\cdot) \in \mathcal{B}(X) / \mathcal{B}(2^Y)$ とする。

このとき，任意の $g \in Q(\{F(x)\} \mid X)$，任意の $u \in M(\{(x, F(x)) ; x \in X\})$，任意の $\varepsilon > 0$ に対して

(i) $gu \geq gu$ かつ $f(x) \in F(x) \quad \forall x \in X$

なる $\mathcal{B}(X)$ 可測関数 $f$ が存在する。

(ii) $g(\{y \in F(x_0) ; u(x_0, y) \leq u(x_0, f(x_0)) + \varepsilon\} \mid x_0) = 1 \quad \forall x_0 \in X$

かつ $f(x) \in F(x) \quad \forall x \in X$

なる $\mathcal{B}(X)$ 可測関数 $f$ が存在する。

## §3. $(p, \varepsilon)$-optimal stationary policy と optimality equation

マルコフ型決定過程の記号と定義

S : state space (ある Borel set)

A : ある compact metric space

$A(s)$ : state $s$ における feasible action の空間で,

$A(s) \in 2^A \quad \forall s \in S$ および $A(\cdot) \in \mathcal{B}(S)/\mathcal{B}(2^A)$

をみたすもの とする。

$D \equiv \{(s, A(s)) ; s \in S\}$

補題 2.1 により $D \in \mathcal{B}(S) \otimes \mathcal{B}(A)$ となる。

$q$ : 推移確率測度で $q \in Q(S \mid D)$ とする。

$r$ : 利得関数で $r \in M(SAS)$ とする。

$\beta$ : 割引率で $0 \leq \beta < 1$ とする。

$H_n \equiv DD \cdots DS$ ($n$ factors)

$\pi = \{\pi_1, \pi_2, \pi_3, \cdots\}$ : 政策(policy)で

$\pi_n \in Q(\{A(s)\} \mid H_n) \quad n=1,2,\cdots$ とする。

$I(\pi) = \sum_{n=1}^{\infty} \beta^{n-1} \pi_1 q \pi_2 q \cdots \pi_n q r$ : policy $\pi$ による総利得の期待値

<u>補題 3.1</u> 任意の $p \in P(S)$, 任意の $\varepsilon > 0$ に対して $(p,\varepsilon)$-optimal policy が存在する。

<u>補題 3.2</u> 任意の $p \in P(S)$, 任意の $\pi$, 任意の $\varepsilon > 0$ に対して, $\pi$ を $(p,\varepsilon)$-dominate する Markov policy が存在する。

<u>定理 3.1</u> 任意の $p \in P(S)$, 任意の $\varepsilon > 0$ に対して $(p,\varepsilon)$-optimal Markov policy が存在する。

Operator $L_f$, Operator $L_a$ をつぎのように定義する.

$f(s) \in A(s) \quad \forall s \in S$ なる $\mathcal{B}(S)$ 可測関数 $f$ と,

$u \in M(S)$ に対して

$$L_f u(s) = \int_S [r(s, f(s), t) + \beta u(t)] \, dq(t \mid s, f(s)) \qquad \forall s \in S$$

$a \in A(s)$, $u \in M(S)$ に対して

$$L_a u(s) = \int_S [r(s, a, t) + \beta u(t)] \, dq(t \mid s, a)$$

Optimality equation

$u \in M(S)$ が

$$u(s) = \sup_{a \in A(s)} L_a u(s) \qquad \forall s \in S$$

をみたすとき, $u$ は optimality equation をみたすと云う。

定理 3.2

(a) 任意の $p \in P(S)$, 任意の $\varepsilon > 0$ に対して $(p, \varepsilon)$-optimal stationary policy が存在する.

(b) そして $\varepsilon$-optimal policy が存在すれば, $\varepsilon/(1-\beta)$-optimal stationary policy が存在する ($\varepsilon \geqq 0$)。

(c) $u \in M(S)$ に対して

$$L_a u(s) \leq u(s) \quad \forall a \in A(s), \ \forall s \in S$$

ならば

$$I(\pi) \leq u \qquad \forall \pi$$

が成立する。

(d) $\pi$ が最適であることと，$I(\pi)$ が optimality equation をみたすこととは同値である。

## §4. Optimal stationary policy

補題 4.1

$A$ : $R^m$ の compact subset

$S$ : Borel set

$u(s,a)$ : $SA \to R^m$ なる mapping で有界かつ

各 $a$ に対し，$s$ につき $\mathcal{B}(S)$ 可測

各 $s$ に対し，$a$ につき連続 とする。

$X^m$ : $R^m$ の compact subset の全体

このとき，$U(s) \equiv u(s,A) \equiv \{x : x = u(s,a),\ a \in A\}$ は $S \to X^m$ なる mapping で $U(\cdot) \in \mathcal{B}(S)/\mathcal{B}(X^m)$ となる。

仮定 (I)

$A$ は $R^m$ の compact subset で，$A(s) \in 2^A\ \forall s \in S$ かつ $A(\cdot) \in \mathcal{B}(S)/\mathcal{B}(2^A)$

補題 4.2

(I) を仮定し，$u(s,a)$ について 補題 4.1 と同じ仮定をおく。

このとき $\tilde{U}(s) \equiv u(s, A(s)) \equiv \{x;\ x = u(s,a),\ a \in A(s)\}$ は

$S \to X^m$ なる mapping で $\tilde{U}(\cdot) \in \mathcal{B}(S)/\mathcal{B}(X^m)$ となる。

補題 4.3

$G(s) : S \to X^n$ への mapping で 有界 かつ $G(\cdot) \in \mathcal{B}(S)/\mathcal{B}(X^n)$

とすると，$e(G(s))$ は $S \to R^n$ への $\mathcal{B}(S)$ 可測 関数である。

補題 4.4 （可測 陰関数の 補題）

$A$ : $R^n$ の compact subset

$u$ については 補題 4.1 と同じ仮定をおく。

$U(s) \equiv u(s, A)$

$\phi(s)$ : $S \to R^m$ への $\mathcal{B}(S)$ 可測 map で かつ，

$$\phi(s) \in U(s) \qquad \forall s \in S$$

とする。 このとき

$$\phi(s) = u(s, f(s)) \qquad \forall s \in S$$

をみたす $S \to A$ への $\mathcal{B}(S)$ 可測関数 $f$ が存在する。

補題 4.5 （可測 陰関数の 補題）

(I) を仮定する。 $u$ について 補題 4.1 と同じ仮定をおく。

$\tilde{U}(s) \equiv u(s, A(s))$

$\phi(s)$ : $S \to R^m$ への $\mathcal{B}(S)$ 可測 map で かつ

$$\phi(s) \in \tilde{U}(s) \qquad \forall s \in S$$

とする。 このとき

$$f(s) \in A(s) \qquad \forall s \in S$$

$$\phi(s) = u(s, f(s)) \qquad \forall s \in S$$

をみたす $\mathcal{B}(S)$ 可測関数 $f$ が存在する。

定理 4.1 (I) を仮定し, $u$ については $m=1$ として補題 4.1 と同じ仮定をおく。 このとき

$$f(s) \in A(s) \qquad \forall s \in S,$$

$$u(s, f(s)) = \max_{a \in A(s)} u(s, a) \qquad \forall s \in S$$

をみたす $\mathcal{B}(S)$ 可測関数 $f$ が存在する。

仮定 (II) $r \in M(SA)$, $r$ は各 $s$ に対し, $a \in A$ につき連続とする。

仮定 (III) 各 $s$ に対し, いかなる有界可測関数 $w(\cdot)$ に対しても

$$\int_S w(\cdot)\, dq(\cdot \mid s, a)$$ が $a \in A$ につき連続とする。

Operator $T$ をつぎのように定義する:

$w \in M(S)$ に対し

$$Tw(s) = \max_{a \in A(s)} \left[ \int_S \{ r(s, a) + \beta w(\cdot) \}\, dq(\cdot \mid s, a) \right]$$

定理 4.2　仮定 (I), (II), (III) をおく。このとき, optimal stationary policy が存在する。

## §5. Policy improvement

$N$ : measurable space $(S, \mathcal{B}(S))$ 上の Markov kernel

$E$ : measurable space $(S, \mathcal{B}(S))$ 上の Markov identity kernel

$S$ 上の実数値 $\mathcal{B}(S)$ 可測関数 $f$ に対して

$$h = \lim_{n\to\infty} (E + \beta N + \beta^2 N^2 + \cdots + \beta^n N^n) f$$

が well-defined かつ有限のとき, $f$ は $\beta$-charge であると云い, $h$ を $f$ の $\beta$-Potential と云う。

$G_\beta \rightleftharpoons \sum_{n=0}^{\infty} (\beta N)^n$ : $\beta$-Potential kernel

補題 5.1　任意の $h \in M(S)$ に対して, $f = (E - \beta N)h$ とおくと, $h$ は $f$ の $\beta$-Potential で, $h = G_\beta f$ となる。

補題 5.2　$f_1, f_2$ はともに $\mathcal{B}(S)$ 可測で $f_i(\omega) \in A(\omega)$ $\forall \omega \in S$ $(i=1,2)$ とする。このとき

$L_{f_2} I(f_1^\infty) \geq I(f_1^\infty)$ ならば $I(f_2^\infty) \geq I(f_1^\infty)$

$L_{f_2} I(f_1^\infty) \leq I(f_1^\infty)$ ならば $I(f_2^\infty) \leq I(f_1^\infty)$

となる。

定理 5.1　(I), (II), (III) を仮定する。

(i)　$\mathcal{B}(S)$ 可測な $f_0$ : $f_0(s) \in A(s)$　$\forall s \in S$　を任意に与える。

(ii)　各 $n \geq 0$ について，つぎの手順を行う。

$$L_{f_{n+1}} I(f_n^\infty) = T I(f_n^\infty)$$

なる $\mathcal{B}(S)$ 可測な $f_{n+1}$　( $f_{n+1}(s) \in A(s)$　$\forall s \in S$ ) を求める。

このとき

(a)　$I(f_0^\infty) \leq I(f_1^\infty) \leq \cdots \leq I(f_n^\infty) \leq I(f_{n+1}^\infty) \leq \cdots \uparrow \sup_\pi I(\pi)$

(b)　もしある $n_0$ で　$f_{n_0} = f_{n_0+1}$　となれば，$f_{n_0}^\infty$ は optimal policy である。

# 第2章　確率オートマトンとの関連

Probabilistic automaton　$\sigma = (\Sigma, S, Q)$

$\Sigma$ : input alphabet ( finite set)

$S$ : state space　(denumerable set)

$Q$ : $\Sigma \to Q(S|S)$　なる map

$\Sigma^* = W(\Sigma)$ : $\Sigma$ 上の words の全体

$Q$ は $\Sigma^* \to Q(S|S)$ へ拡張される

$x = \sigma_1 \sigma_2 \cdots \sigma_n$ に対し $Q(x) \equiv Q(\sigma_1) Q(\sigma_2) \cdots Q(\sigma_n)$ と
おく。

Prop. 1　第1章で特に S, A finite (Aは一定) の場合
を考えると，これはつぎのような利得をもつ Mealy type の
確率オートマトン $(\mathfrak{A}(\Sigma, S, Q), \gamma)$ で表わされる。

$\Sigma$ : $S \to A$ なる map の全体 (finite)

$\gamma$ : $S\Sigma \to R^1$ なる利得関数で，$\gamma(s, \sigma) = \gamma(s, \sigma(s))$
をみたす。

Def. 利得をもつ確率オートマトン

確率オートマトン $\mathfrak{A} = (\Sigma, S, Q, F)$

ただし　$Q : \Sigma \to Q(S|S)$

$F$ : $F \subset S$, final state の集合

$g, f$ ともに利得関数でつぎの性質をもつとする。

$$E^{\sigma_1 \sigma_2 \cdots \sigma_m \cdots \sigma_n} g = E^{\sigma_1 \cdots \sigma_m \cdots \sigma_n} g(E^{\sigma_1 \cdots \sigma_m} g, s_{m+1}, \cdots, s_{n+1})$$

$$\forall (\sigma_1 \cdots \sigma_n) \in \Sigma^*$$

$f$ は $(\sigma_1 \cdots \sigma_n)$ に対し $s_{n+1}$ のみの関数 :

$$E^{\sigma_1 \cdots \sigma_n} f(s_{n+1})$$

empty word $\lambda$ に対して

$$E^{\lambda} g = 0, \qquad E^{\lambda} f = f(s_1)$$

以上の仮定のもとで，$\Pi = (\sigma, g, k)$ を利得のある確率オートマトンと云う。

Def. 利得をもつ確率オートマトン $\Pi$ における最適化問題.

$$E^x(g+k) \longrightarrow \mathrm{Max} \quad \text{in} \quad x \in \Sigma^*$$

subject to

$$P^x(F) = 1.$$

(註) $P^x(F) = 1$ は $P^x_{s_0}(F) = 1 \quad \forall_{s_0} \in S$ の意味.

Def. 非逐次的確率オートマトン

$\Sigma$ : input alphabet (finite set)

$S$ : state space (denumerable)

$L \subset \Sigma^*$ : language

$Q^* : L \to Q(S|S)$

$f$ : output function

このとき $\mathfrak{M} = (\Sigma, S, L, Q^*, f)$ を非逐次的確率オートマトンと云う。

Def. 非逐次的確率オートマトン $\mathfrak{M}$ における最適化問題.

$$E^x f \longrightarrow \mathrm{Max} \quad \text{in} \quad x \in L$$

Def. $\Pi = (\mathcal{A}(\Sigma, S, Q, F), g, k)$ が $\mathcal{M} = (\Sigma, S, L, Q^*, f)$ を represent するとは,

(i) $L = \{x \in \Sigma^* \mid p^x(F) = 1\}$

(ii) $E^x f = E^x(g + k) \qquad \forall x \in L$

Prop. 2 利得をもつ確率オートマトン $\Pi$ を任意に与えると, 自然な方法で $\Pi$ は一つの非逐次的確率オートマトン $\mathcal{M}$ を represent する。

Def. $\Pi$ が つぎの性質をもつとき, monotone であると云う:
任意の $m$ に対し,

$$E^{\sigma_1 \cdots \sigma_m} g \leq E^{\sigma_1' \cdots \sigma_m'} g$$

ならば, $E^{\sigma_1 \cdots \sigma_m \sigma_{m+1}} g \leq E^{\sigma_1' \cdots \sigma_m' \sigma_{m+1}} \quad \forall \sigma_{m+1} \in \Sigma$

が成立する。

Prop. 3 $\Pi$ が monotone なら, つぎのことが成立する。

$$E^{\sigma_1 \cdots \sigma_m} g \leq E^{\sigma_1' \cdots \sigma_m'} g$$

ならば, $E^{\sigma_1 \cdots \sigma_m \sigma_{m+1} \cdots \sigma_n} g \leq E^{\sigma_1' \cdots \sigma_m' \sigma_{m+1} \cdots \sigma_n} g$

$$\forall (\sigma_{m+1} \cdots \sigma_n) \in \Sigma^*, \ \forall n \geq 1$$

Def. $L \subset \Sigma^*$ に対して, $L$ によって induce された $\Sigma^*$ 上の relation $R_L$ をつぎのように定義する:

$$x R_L y \overset{\text{def}}{\Leftrightarrow} v \in \Sigma^* \text{ に対し } xv \in L \rightleftarrows yv \in L$$

Prop. 4　$R_L$ は right congruent, equivalence relation である。

ただし right congruent とは

$$x \equiv y \quad \Rightarrow \quad xw \equiv yw \quad \forall w \in \Sigma^*$$

Def. equi-transition relation

$x = (\sigma_1, \sigma_2, \cdots, \sigma_n) \in \Sigma^*$ に対し, $Q(x) \equiv Q(\sigma_1)Q(\sigma_2)\cdots Q(\sigma_n)$

$x, y \in \Sigma^*$ に対し, (ただし $lg\, x$ と $lg\, y$ は一致しなくてよい),

$Q(x) = Q(y)$ のとき $x$ と $y$ は equi-transition という。

Prop. 5　equi-transition relation は right congruent な equivalence relation で, 任意の FCS に対し

$$L \equiv \{ x \in \Sigma^* \mid P^x(F) = 1 \}$$

とおくと, equi-transition relation は $R_L$ の細分である。

定理 1

$L \subset \Sigma^*$,

$\sim$ を $\Sigma^*$ 上の right congruent equivalence relation で denumerable rank とする。このとき,

$$L = \{ x \in \Sigma^* \mid P^x(F) = 1 \}$$

かつ　$\sim$ が $\mathfrak{A}$ の equi-transition relation

となるような有限確率オートマトン $\mathfrak{A} = (\Sigma, S, Q, F)$ が存在するための必要十分条件は, $\sim$ が $R_L$ の細分となることである。

定理2　$\mathcal{M} = (\Sigma, S, L, Q^*, f)$ を非逐次的確率オートマトン，$\sim$ を $\Sigma^*$ 上の denumerable rank, equivalence relation とする。このとき $\Pi$ が $\mathcal{M}$ を represent し，$\sim$ が $\mathcal{O}$ 上の equi-transition relation となるような monotone $\Pi = (\mathcal{O}, g, k)$ が存在するための必要十分条件はつぎの (i)～(iii) が成立することである。

(i)　$\sim$ が $R_L$ を細分する。

(ii)　$x \sim y$ なら (a) または (b) のどちらか一方が成立

(a)　$E^{xv} f \leq E^{yv} f$　　$\forall v$ such that $xv \in L$

(b)　$E^{xv} f \geq E^{yv} f$　　$\forall v$ such that $xv \in L$

(iii)　$x \sim y$, $x \in L$, $y \in L$, $E^x f = E^y f$ ならば

$E^{xv} f = E^{yv} f$　　$\forall v$ such that $xv \in L$

定理3　monotone な，利得をもつ確率オートマトン $\Pi$ における最適化問題に対しては，"最適性の原理" が成立する。