数理解析研究所講究録
第67巻 1969年 1-12

非定常な推移確率分布をもつ マルコフ決定過程

九州大学 理学部 古川長太

§1 序

マルコフ決定過程の殆んどすべての論文に見られるように，そこでは，推移確率分布が定常であることに基礎を置く。この推移確率分布の定常性は，割引率の有る場合では，最適政策に関するいくつかのエレガントな性質を導く本質的な素因となっている。例えば，Blackwell [2] は，一般に (p,ε) 最適政策の存在と，特に action が可付番個の場合は ε 最適政策，action が有限個の場合は最適政策の存在を証明し，かつ，これらがいづれも定常政策として得られることを示した。また，Strauch [3] は，Blackwell より強い意味での (p,ε) 最適定常政策の存在を証明した。

この報告では，推移確率分布の定常性をゆるめ，非定常な推移確率分布をもつものとして，いくつかの意味での最適政策の存在と，存在すれば，その最適政策の性質について考察

することにしよう。

§2 諸定義 (I)

定義 2.1

X, Y ; ある separable metric space の Borel subset

$\mathcal{B}$; X の subset の Borel 集合族

$\mathcal{A}$; Y の subset の Borel 集合族

$P(X)$; X の上の probability distribution の class

$g(y|x)$; conditional probability distribution

(各 x につき, $g(\cdot|x)$ は Y 上の prob. distribution

各 $B \in \mathcal{A}$ につき, $g(B|\cdot)$ は X 上の $\mathcal{B}$ 可測関数)

$Q(Y|X)$; $g(y|x)$ の class

$M(X)$; X 上の有界実数値可測関数の class

定義 2.2

定義 2.1 を拡張して

$P(X_1 X_2 \cdots X_n)$, $P(X_1 X_2 \cdots\cdots)$,

$Q(X_{n+1}|X_1 X_2 \cdots X_n)$, $Q(X_{n+1} X_{n+2}|X_1 X_2 \cdots X_n)$

$M(X_1 X_2 \cdots X_n)$ 等を定義する。

定義 2.3

pu ; $pu \equiv \int_X u(x)\, dp(x)$ for $p \in P(X)$, $u \in M(X)$

gu ; $gu \equiv \int_Y u(x,y)\, dg(y|x)$ for $g \in Q(Y|X)$, $u \in M(XY)$

定義 2.4

$p \in P(X)$ なる p が degenerate $\Leftrightarrow$ $p\{x\} = 1$ for some $x \in X$

$q \in Q(Y|X)$ なる q が degenerate $\Leftrightarrow$ $q(\cdot|x)$ が各 x について degenerate

補題 2.1 (Blackwell [2]) 各 $q \in Q(Y|X)$, 各 $u \in M(XY)$ に対して

$$fu \geqq qu \qquad \text{for all } x \in X$$

なるような degenerate $f \in Q(Y|X)$ が存在する.

§3 諸定義 (II)

定義 3.1

S; state space (non-empty Borel set)

A; action space (non-empty Borel set)

$S \ni s$; state

$A \ni a$; action

$Q(S|SA) \ni q_i$; 時刻 i における state から時 $i+1$ における state への transition prob. distribution

$M(SAS) \ni r$; reward function

$0 < \beta < 1$; discount factor

定義 3.2

$H_n \equiv SAS \cdots AS$ ($2n-1$ factors)

$\pi_n \in Q(A|H_n)$, $(n=1.2,\cdots)$ として

$\pi \rightleftharpoons \{\pi_1, \pi_2, \pi_3, \cdots\}$; strategy

定義 3.3

random Markov strategy ; $\pi=\{\pi_1, \pi_3, \cdots\}$ において $\pi_n \in Q(A|S)$ $(n=1.2,\cdots)$

Markov strategy ; random Markov strategy $\pi=\{\pi_1,\pi_2,\cdots\}$ において, 各 π_n が degenerate

Markov strategy を $\{f_1, f_2, \cdots\}$ 等と書く。

定義 3.4

$e_\pi \rightleftharpoons \pi_1 g_1 \pi_2 g_2 \cdots \in Q(X|S)$ だが $X = ASAS\cdots$

$I_n(\pi, g, v) \rightleftharpoons e_\pi\left[\sum_{j=1}^{n} \beta^{j-1} r(s_j, a_j, s_{j+1}) + \beta^n v\right]$

$I_n(\pi, g) \rightleftharpoons I_n(\pi, g, 0)$

$I(\pi, g) \rightleftharpoons e_\pi \sum_{j=1}^{\infty} \beta^{j-1} r(s_j, a_j, s_{j+1})$

Q^* ; $g_n \in Q(S|SA)$ for $n=1.2,\cdots$ なる $\{g_1, g_2, \cdots\}$ の class

補題 3.1 $I_n(\pi, g, v) \longrightarrow I(\pi, g)$ $(n\to\infty)$ for $v \in M(S)$, for $g \in Q^*$

定義 3.5

π^* が (p, ε, g)-optimal ; $p\{I(\pi^*, g) \geqq I(\pi, g) - \varepsilon\} = 1$ for all π

π^* が (ε, g)-optimal ; $I(\pi^*, g) \geqq I(\pi, g) - \varepsilon$ for all π

π^* が g-optimal ; $I(\pi^*, g) \geqq I(\pi, g)$ for all π

§4 最適政策

補題 4.1 各 $p \in P(S)$, 各 $\varepsilon > 0$, 各 $g \in Q^*$ に対して, (p, ε, g)-optimal Markov strategy が存在する。

(証明)

Blackwell [2] の Theorem 2 において $\{g, g, \cdots\}$ の代りに $\{g_1, g_2, g_3, \cdots\}$ とおけば, 直ちに証明される。

定義 4.1

T_{nj} ; $T_{nj}u(s) = \int [r(s, f_n(s), t) + \beta u(t)] \, dg_j(t \mid s, f_n(s))$

T_{nj} を (f_n, g_j) に対応する operator と呼び, $(f_n, g_j) \rightsquigarrow T_{nj}$ と書く。

$\pi = \{f_1, f_2, \cdots\}$ に対して

$\sqcup_j$; $\sqcup_j u(s) = \sup_n T_{nj} u(s)$ ただし $(f_n, g_j) \rightsquigarrow T_{nj}$

$\sqcup_j$ を (π, g_j) に対応する operator と呼び, $(\pi, g_j) \rightsquigarrow T_j$ と書く。

定理 4.1 (a) T_{nj} は単調

(b) $T_{nj}(u + c) = T_{nj}u + \beta c$ ただし c は定数

(c) Markov $\pi = \{f_1, f_2, \cdots\}$ に対して, $(f_n, g_n) \rightsquigarrow T_{nn}$ とすれば $T_{11}T_{22}\cdots T_{nn}v = I_n(\pi, g, v)$ for each n

(d) Markov $\pi = \{f_1, f_2, \cdots\}$ に対して $(f_n, g_n) \rightsquigarrow T_{nn}$ とすれば

$$T_{nn} I({}^n\pi, {}^n g) = I({}^{n-1}\pi, {}^{n-1}g) \quad \text{for each } n$$

ただし ${}^n\pi = \{\pi_{n+1}, \pi_{n+2}, \cdots\}$, ${}^n g = \{g_{n+1}, g_{n+2}, \cdots\}$

定義 4.2

π を Markov strategy として,

f が π-generated ; f は $S \to A$ の可測関数で, S の分割 $\{S_n\}$ が存在して

$$f = f_n \quad \text{on } S_n \quad \text{for each } n.$$

$F(\pi)$; π-generated function の class

Markov $\pi' = \{g_1, g_2, \cdots\}$ が π-generated ; 各 g_n が $g_n \in F(\pi)$

$G(\pi)$; π-generated strategy の class

定理 4.2 (a) π を任意の Markov strategy とする。各 $g_j \in Q(S|SA)$, 各 $\hat{f} \in F(\pi)$ に対して

$$\hat{T}_j u \leqq U_j u$$

ただし $(\hat{f}, g_j) \rightsquigarrow \hat{T}_j$ 。

(b) 各 Markov π, 各 $\varepsilon > 0$, 各 $g_j \in Q(S|SA)$ に対して $\hat{f}_j \in F(\pi)$ が存在し, $(\hat{f}_j, g_j) \rightsquigarrow \hat{T}_{jj}$ とすれば

$$\hat{T}_{jj} u \geqq U_j u - \varepsilon$$

定義 4.3

$(\pi, g_j) \rightsquigarrow U_j$ として, $\lim_{m\to\infty} U_j U_{j+1} \cdots U_m u \equiv u^*_{j-1}$ とおき, これを $(\pi; {}^{j-1}g_j)$ に対応する limit point という。記号で $(\pi, {}^{j-1}g) \rightsquigarrow u^*_{j-1}$ と書く。 特に $u^*_0 = u^*$ と書く。

定理 4.3 (a) π を任意の Markov strategy とすると

$$I(\hat{\pi}, g) \leqq u^* \quad \text{for } \hat{\pi} \in G(\pi)$$

ただし $(\pi, g) \rightsquigarrow u^*$。

(b) 各 Markov π, 各 $\varepsilon>0$, 各 $g_j \in Q(S|SA)$ に対して $\hat{\pi} \in G(\pi)$ が存在して $I(\hat{\pi}, g) \geqq u^* - \varepsilon$。ただし $(\pi, g) \rightsquigarrow u^*$。

(c) もし, 各 $j \geqq 0$ に対して $(\varepsilon, {}^j g)$-optimal strategy が存れば, $(\varepsilon/(1-\beta), g)$-optimal Markov strategy が存在する。$(\varepsilon \geqq 0)$

(d) $(f \equiv a, g_j) \rightsquigarrow T_{aj}$とする。そしも, 各 $\varepsilon>0$, 各 $j \geqq 0$ に対して $(\varepsilon, {}^j g)$-optimal strategy が存在すれば, Markov strategy $\hat{\pi}$ が存在して $(\hat{\pi}, {}^j g) \rightsquigarrow \hat{u}_j^*$ については, $\hat{u}_j^*$ は可測関数でかつ,

$$\hat{u}_{j-1}^* = \sup_{a \in A} T_{aj} \hat{u}_j^* \quad \text{for } j=1,2,\cdots$$

が成立する。

(e) 各 $j \geqq 0$ につき, ${}^j\pi^*$ が ${}^j g$-optimal $\Longleftrightarrow$

$$I({}^{j-1}\pi^*, {}^{j-1}g) = \sup_{a \in A} T_{aj} I({}^j\pi^*, {}^j g) \quad \text{for } j=1,2,\cdots$$

<u>定理 4.3 (c) の系</u> 各 $j \geqq 0$ につき ${}^j g$-optimal strategy が存在すれば g-optimal Markov strategy が存在する。

<u>定義 4.4</u>

action a と action b が (s, g_j) において同等;

$$T_{aj} u(s) = T_{bj} u(s) \quad \text{for all } u \in M(S)$$

action a と action b が (s, g) において同等;

$$T_{aj} u(s) = T_{bj} u(s) \quad \text{for all } u \in M(S),$$
$$\text{for all } g_j \text{ in } \{g_1, g_2, \cdots\}$$

π を Markov strategy として,

A が essentially countable by π ; 各 (s,a) に対して $f_n(s)$ と a が $(s,\mathcal{B})$ において同等となるような n が存在する。

A が essentially finite by π ; S の分割 $\{S_n\}$ が存在して各 n について, $s\in S_n$ なる各 (s,a) に対して $f_1(s), f_2(s), \cdots, f_n(s)$ の内の少くとも一つと a が $(s,\mathcal{B})$ において同等となる。

補題 4.2 A が essentially finite by π ならば, 各 $g_j\in Q(S|SA)$, 各 $u\in M(S)$ に対して $\hat{f}_j\in F(\pi)$ が存在して, $\hat{T}_{\hat{f}_j}u = \sqcup_j u$ が成立つ。ただし $(\hat{f}_j, g_j)\rightsquigarrow \hat{T}_{\hat{f}_j}$, $(\pi, g_j)\rightsquigarrow \sqcup_j$。

定理 4.4 (a) A が essentially countable by π ならば, 各 $\varepsilon>0$, 各 $g\in Q^*$ に対して (ε, g)-optimal Markov strategy が存在する。

(b) A が essentially finite by π ならば, 各 $g\in Q^*$ に対して, g-optimal strategy が存在する。

§5 l-stationary strategy.

定義 5.1

g が l-stationary ; $g_j\in Q(S|SA)$ $(j=1,2,\cdots,l)$ なる

$g_1, g_2, \cdots, g_\ell$ により, $g=\{\bar{g}, \bar{g}, \bar{g}, \cdots\}$ と
書けること。ただし $\bar{g}=\{g_1, g_2, \cdots, g_\ell\}$。

π が ℓ-stationary ; degenerate measurable $f_j : S \to A$;
$(j=1,2,\cdots,\ell)$ により $\pi=\{\bar{f}, \bar{f}, \bar{f}, \cdots\}$ と
書けること。ただし $\bar{f}=\{f_1, f_2, \cdots, f_\ell\}$。

定義 5.2

$\bar{g}=\{g_1, g_2, \cdots, g_\ell\}$, $(\pi, g_j) \rightsquigarrow U_j$ として
$\bar{U}_\ell \equiv U_1 U_2 \cdots U_\ell$; $(\pi, \bar{g})$ に対応する operator
記号で $(\pi, \bar{g}) \rightsquigarrow \bar{U}_\ell$ と書く。

定理 5.1 $(\pi, \bar{g}) \rightsquigarrow \bar{U}_\ell$ とすると, $\bar{U}_\ell$ は contraction coefficient β^ℓ をもつ contraction mapping である。

定理 5.2 (a) π を Markov strategy, $g=\bar{g}^{(\infty)}$ を ℓ-stationary とする。$(\pi, \bar{g}) \rightsquigarrow \bar{U}_\ell$ とする。$\bar{U}_\ell$ の fixed point を $\bar{u}_\ell^*$ とすれば $I(\hat{\pi}, \bar{g}^{(\infty)}) \leqq \bar{u}_\ell^*$ for all $\hat{\pi} \in G(\pi)$。

(b) 各 Markov π, 各 ℓ-stationary $g=\bar{g}^{(\infty)}$ に対して, $\bar{u}_\ell^*$ を (a) の様に定義すれば, 各 $\varepsilon>0$ に対して ℓ-stationary strategy $\bar{f}^{(\infty)} \in G(\pi)$ が存在して, $I(\bar{f}^{(\infty)}, \bar{g}^{(\infty)}) \geqq \bar{u}_\ell^* - \varepsilon$ が成立つ。

(c) 各 $p \in P(S)$, 各 $\varepsilon>0$, 各 ℓ-stationary $g=\bar{g}^{(\infty)}$ に対して $(p, \varepsilon, \bar{g}^{(\infty)})$-optimal ℓ-stationary strategy が存在する。

(d) 各 $\varepsilon \geqq 0$ に対して, 各 $j=1,2,\cdots,\ell$ に対して $(\varepsilon, (g_j, g_{j+1}, \cdots, g_\ell, \bar{g}^{(\infty)}))$-optimal strategy が存在すれば,

$(\varepsilon/(1-\beta),\ \bar{g}^{(\infty)})$-optimal l-stationary strategy が存在する。

(e) g を l-stationary とする。このとき

各 $j=1,2,\cdots,l$ に対して ${}^{j}\pi^*$ が ${}^{j}g$-optimal

$$\Leftrightarrow\quad I({}^{j-1}\pi^*, {}^{j-1}g) = \sup_{a\in A} T_{aj} I({}^{j}\pi^*, {}^{j}g) \quad (j=1,2,\cdots,l)$$

§6 Strong optimality.

定義 6.1

π^* が strong (p,ε,g)-optimal ; $p\{I(\pi^*,g) \geqq \sup_{\pi} I(\pi,g)-\varepsilon\}=1$

定理 6.1 各 $p\in P(S)$, 各 $g\in Q^*$, 各 π に対して,

$$pI(\tilde{\pi},g) \geqq pI(\pi,g)$$

なるような Markov strategy $\tilde{\pi}$ が存在する。

定理 6.2 各 $p\in P(S)$, 各 $\varepsilon>0$, 各 $g\in Q^*$ に対して, (p,ε,g)-optimal strategy が存在する。

定理 6.3 Markov strategy の任意の sequence $\{\pi^j, j=1,2,\cdots\}$, 各 $\varepsilon>0$, 各 $g\in Q^*$ に対して

$$I(\hat{\pi},g) \geqq \sup_{j} I(\pi^j,g) - \varepsilon$$

なる Markov $\hat{\pi}$ が存在する。

定理 6.4 各 $p\in P(S)$, 各 $\varepsilon>0$, 各 $g\in Q^*$ に対して, strong (p,ε,g)-optimal Markov strategy が存在する。

注意 Strauch [3] は特に g が stationary の場合に、この section の定理 6.4 に相当する結果を、"conservation" の概念を導入することにより与えている。而もその証明方法は、conservation に関するいくつかの補題を用意して、可成り遠廻りなものである。我々の場合には、g が stationary でないため、conservation の補題が成立たず、従って Strauch の方法では成功しないことが分る。処し乍ら、我々の定理 4.3 (a), (b) を用いることにより、もっと直接的に定理 6.4 を証明することに成功した。

附記

この報告では、補題、定理の証明を記載しなかったが、それらの証明については、参考文献 [4] に詳細に記されている。

参考文献

[1] R. Bellman, Dynamic Programming, Princeton Univ. Press, (1957).

[2] D. Blackwell, Discounted dynamic programming, Ann. Math. Stat. 36 (1965), 226-235.

[3] R.E. Strauch, Negative dynamic programming, Ann. Math. Stat. 37 (1966), 871-890.

[4] N. Furukawa, A Markov decision process with non-stationary

transition laws, Bulletin of Mathematical Statistics, 13 (1968), 41–52.