

マルコフ型決定過程 (II)

九大理 古川長太

§ 1. 序

Wald 以来の sequential analysis と Bellman の Dynamic Programming の中の discrete parameter の部分を, まとめて Markovian Decision Process として最初に定式化したものは Blackwell である。

sequential analysis と所謂 D. P. との相異は, 前者においては stopping rule と stop したときの decision のとり方, ないかえると sequential decision rule の追究が問題であるのに反し, 後者においては殆んど stopping rule が問題にされない臭に在る。而るに前者は吸収壁をもつマルコフ過程とみなされることから, 結局両者は同じ範ちゅうに属する問題として定式化されることが分る。

この様な一般的な定式化では, 従来研究されて来た様な個々の問題に対する解を, 又はその近似解を求めるとなつた具体的な結果を期待出来ないのは当然であるが, 反面それらの

問題に共通の本質的な部分を浮きぼりにし、解明することが出来る。

こゝでは、これらの事について、主として Blackwell [], Strauch [] の結果と、併せて Furukawa の拡張した部分をつけ加えて報告する。

§2. 諸定義

Def. 2.1

X, Y ; ある complete separable metric space の Borel subset

$\mathcal{O}$; X における Borel 集合族

$\mathcal{O}$; Y における Borel 集合族

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

$\mathcal{P}(Y|X)$; conditional probability distribution (各 x について,

$\mathcal{P}(\cdot|x)$ は Y 上の prob. dist., 各 $B \in \mathcal{O}$ について

$\mathcal{P}(B|\cdot)$ は X 上の $\mathcal{O}$ 可測関数)

$\mathcal{Q}(Y|X)$; $\mathcal{P}(Y|X)$ の class

Def. 2.2

$M(X) : \begin{cases} X \text{ 上の 実数値有界可測関数の class (D-case)} \\ X \text{ 上の non-positive 実数値有界可測関数の class (N-case)} \\ X \text{ 上の non-negative 実数値有界可測関数の class (P-case)} \end{cases}$

(D, N, P の定義は後述)

Def. 2.3 $M(XY)$; Def. 2.2 において定義空間を直積空間 XY としたもの

Def. 2.4

$$p u; \quad p u \equiv \int_X u(x) dp(x) \quad \text{for } p \in P(X), \quad u \in M(X)$$

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

Def. 2.5 以上の諸定義を拡張して

$$P(x_1, x_2, \dots, x_n), \quad P(x_1, x_2, \dots),$$

$$Q(x_{n+1}|x_1, x_2, \dots, x_n), \quad Q(x_{n+1}, x_{n+2}|x_1, x_2, \dots, x_n),$$

$$M(x_1, x_2, \dots, x_n)$$

Def. 2.6

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

$$g \in Q(Y|X) \text{ なる } g \text{ が degenerate} \Leftrightarrow g(\cdot|x) \text{ が } x \text{ について degenerate}$$

Def. 2.7

S ; state space (non-empty Borel set)

A ; action space (non-empty Borel set)

$S \ni s$; state

$A \ni a$; action

$Q(S|SA) \ni g$; transition probability distribution

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

$$\beta \equiv (\beta_1, \beta_2, \beta_3, \dots), \quad (0 \leq \beta_i \leq 1, i=1, 2, \dots);$$

discount factor vector

Def. 2.8

(S, A, g, r, β) ; discrete Dynamic Programming の構成要素

Def. 2.9

$$H_n \equiv SASA \dots SAS \quad (2n-1 \text{ factors})$$

$$\pi_n \in Q(A|H_n), \quad n=1, 2, \dots \in \mathcal{T}$$

$$\pi \equiv (\pi_1, \pi_2, \dots); \text{ policy}$$

Def. 2.10 $c_\pi \equiv \pi_1 g \pi_2 g \dots \in Q(ASAS \dots | S)$

$$I_n(\pi, \beta, v) \equiv c_\pi \left[\sum_{j=1}^n \beta_1 \beta_2 \dots \beta_{j-1} r(s_j, a_j, s_{j+1}) + \beta_1 \beta_2 \dots \beta_n v \right], \quad (v \in M(S))$$

$$I_n(\pi, \beta) \equiv I_n(\pi, \beta, 0)$$

$$I(\pi, \beta) \equiv c_\pi \sum_{j=1}^{\infty} \beta_1 \beta_2 \dots \beta_{j-1} r(s_j, a_j, s_{j+1}); \text{ expected total reward}$$

Def. 2.11

random Markov policy; π において $\pi_n \in Q(A|S), n=1, 2, \dots$

Markov policy; random Markov policy π において 各 π_n が degenerate

Markov policy f $(f_1, f_2, \dots) \in \mathcal{F}$.

stationary policy; $(f, f, f, \dots)$

§ 3. 基本定理

Def. 3.1 $u, v \in M(X)$ に対して

$$u \geq v \iff u(v) \geq v(x) \text{ for all } x \in X$$

Lem. 3.1 (Blackwell and Nardzewski [1])

$g \in Q(Y|X)$, $S \in \sigma \times \beta$, $g(S_x|x) > 0$ for all $x \in X$

$\Rightarrow (x, g(x)) \in S$ for all $x \in X$ なる σ 可測 $g(x)$ が存在する。

Theorem 3.1

$\forall g \in Q(Y|X)$, $\forall u \in M(XY)$, $\forall \varepsilon > 0$,

$\exists$ degenerate $f \in Q(Y|X)$

; (i) $f u \geq g u$

(ii) $g(\{y; u(x_0, y) \geq u(x_0, f(x_0)) + \varepsilon\} | x_0) = 0$
for all $x_0 \in X$

Def. 3.2

D-case ; $\sum_{j=1}^{\infty} \beta_1 \beta_2 \cdots \beta_j \equiv L$: 収束, $r \in M(SAS)$

N-case ; $\beta_i = 1$ for $i=1, 2, \dots$, $r \in M(SAS)$

P-case ; $\beta_i = 1$ for $i=1, 2, \dots$, $r \in M(SAS)$, かつ
 π の policy に対して exp. total reward が有界

D-case を更に、次の二つの場合に分ける。

D-h-case ; $\beta_i = b$ for $i=1, 2, \dots$

D-n-case ; D-case と D-h-case を外

Lem 3.2

(D-case) $I_n(\pi, \beta, v) \rightarrow I(\pi, \beta)$ for $\forall v \in M(S)$ 収束

(P-case) $I_n(\pi) \uparrow I(\pi)$ 収束

(N-case) $I_n(\pi) \downarrow I(\pi)$ ($-\infty$ に発散することもある)

Def. 3.3

π^* が (p, ε) -optimal ; $P\{I(\pi^*, \beta) \geq I(\pi, \beta) - \varepsilon\} = 1$ for $\forall \pi$

π^* が ε -optimal ; $I(\pi^*, \beta) \geq I(\pi, \beta) - \varepsilon$ for $\forall \pi$

π^* が optimal ; $I(\pi^*, \beta) \geq I(\pi, \beta)$ for $\forall \pi$

ただし N-case では上の定義に, $I(\pi^*, \beta) > -\infty$ を加える。

§4. D-case

Lem. 4.1 (Blackwell []))

各 $\beta \in P(S)$, 各 $\varepsilon > 0$ に対して (p, ε) -optimal policy が存在。

Def. 4.1 π^* が π を (p, ε) -dominate する ;

$$P\{I(\pi^*, \beta) \geq I(\pi, \beta) - \varepsilon\} = 1$$

Lem. 4.2 (Blackwell []))

各 $\beta \in P(S)$, 各 $\varepsilon > 0$, 各 π に対して, π を (p, ε) -dominate する Markov policy が存在する。

Theorem 4.1 (Blackwell []))

各 $\beta \in P(S)$, 各 $\varepsilon > 0$ に対して (p, ε) -optimal Markov policy が存在する。

Def. 4.2 (Furukawa)

$$T_{n_j} ; T_{n_j} u(s) = \int [\gamma(s, f_n(s), t) + \beta_j u(t)] d\beta_j(t | s, f_n(s))$$

$$\pi = (f_1, f_2, \dots) \text{ に対して}$$

$$\sqcup_j ; \sqcup_j u(s) = \sup_n T_{n_j} u(s)$$

$T_{nj} \in (f_n, \beta_j)$ に対する operator と云い. $(f_n, \beta_j) \rightsquigarrow T_{nj}$ と書く.

$U_j \in (\pi = (f_1, f_2, \dots), \beta_j)$ に対する operator と云い,
 $(\pi, \beta_j) \rightsquigarrow U_j$ と書く.

Theorem 4.2 (Furukawa)

(a) T_{nj} は 単調, 即ち $u \leq v \Rightarrow T_{nj} u \leq T_{nj} v$

(b) c が定数なら $T_{nj}(u+c) = T_{nj} u + \beta_j c$

(c) Markov $\pi = (f_1, f_2, \dots)$ に対しては

$$T_{11} T_{22} \dots T_{nn} v = I_n(\pi, \beta, v)$$

Def. 4.3

$\pi = (\pi_1, \pi_2, \dots)$ に対して ${}^n \pi \equiv (\pi_{n+1}, \pi_{n+2}, \dots)$

従って $\pi = (f_1, f_2, \dots)$ に対して ${}^n \pi = (f_{n+1}, f_{n+2}, \dots)$

$\beta = (\beta_1, \beta_2, \dots)$ に対して ${}^n \beta \equiv (\beta_{n+1}, \beta_{n+2}, \dots)$

Theorem 4.2 (d)

$\pi = (f_1, f_2, \dots)$ に対して

$$T_{nn} I({}^n \pi, {}^n \beta) = I({}^{n-1} \pi, {}^{n-1} \beta) \quad \text{for } n=1, 2, \dots$$

Def. 4.4 (Blackwell [1])

π は Markov policy とし

f が π -generated; f は $S \rightarrow A$ なる 可測 mapping
 S の partition $\{S_n\}$ が存在して

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

policy $\hat{\pi}$ が π -generated ; $\hat{\pi} = (g_1, g_2, \dots)$ は Markov.
各 g_n が π -generated

Def. 4.5 $F(\pi)$; π -generated function の class

$G(\pi)$; π -generated policy の class

Theorem 4.3 (Furukawa)

(a) 各 $\pi = (f_1, f_2, \dots)$, 各 $\varepsilon > 0$, 各 β_j に対して $\hat{f}_j \in F(\pi)$ が

存在し, $(\hat{f}_j, \beta_j) \leadsto \hat{T}_{jj}$ とすると

$$\hat{T}_{jj} u \geq U_j u - \varepsilon \quad \text{for } \forall u$$

(b) π を任意の Markov policy とする。各 $\hat{f} \in F(\pi)$ に対して

$(\hat{f}, \beta_j) \leadsto \hat{T}_j$ とすれば

$$\hat{T}_j u \leq U_j u \quad \text{for } \forall j, \forall u$$

Def. 4.6 (Furukawa)

$(\pi, \beta_j) \leadsto U_j$ とし, $\lim_{n \rightarrow \infty} U_j U_{j+1} \dots U_n u \equiv u_{j-1}^*$ とおき

これを $(\pi, j-1\beta)$ に対応する limit point と置く。記号で

$(\pi, j-1\beta) \leadsto u_{j-1}^*$ と書く。特に, $u_0^* = u^*$ と書く。

Theorem 4.4 (Furukawa)

(a) π を任意の Markov policy とすると

$$I(\hat{\pi}, \beta) \leq u^* \quad \text{for } \forall \hat{\pi} \in G(\pi)$$

(b) π を任意の Markov policy とすると

$$\forall \varepsilon > 0, \exists \hat{\pi} \in G(\pi); I(\hat{\pi}, \beta) \geq u^* - \varepsilon$$

- (c) 各 $\delta\beta$ ($j \geq 0$) に対して ϵ -optimal policy ($\delta\beta$ に依存する) が有れば, β に対する $(1+\epsilon)\epsilon$ -optimal Markov policy が存在する. ($\epsilon \geq 0$)
- (d) $(f \equiv a, \beta_j)$ に対する operator を Ta_j とする. 各 $\epsilon > 0$, 各 $\delta\beta$ に対して ϵ -optimal policy ($\delta\beta$ に依存する) が有れば Markov policy $\hat{\pi}$ が存在して $(\hat{\pi}, \delta\beta) \leadsto \hat{u}_j^*$ については $\hat{u}_j^*$ は 可測関数で,かつ
- $$\hat{u}_{j-1}^* = \sup_{a \in A} Ta_j \hat{u}_j^* \quad \text{for each } j$$
- が成立する.

- (e) 各 j につき, $\delta\pi^*$ が $\delta\beta$ に対して optimal なるための必要条件は

$$I(\delta\pi^*, \delta\beta) = \sup_{a \in A} Ta_j I(\delta\pi^*, \delta\beta) \quad \text{for each } j$$

である. (この条件式を, π^* に関する optimality equation と云う)

[証明]

- (a) π を任意の Markov とする.

$$\hat{\pi} \equiv (g_1, g_2, \dots) \in G(\pi), \quad (g_j, \beta_j) \leadsto \hat{T}_j \text{ とする.}$$

$$\text{Theorem 4.2 (a) より} \quad \hat{T}_{nn} I(n\hat{\pi}, n\beta) = I(n+1\hat{\pi}, n+1\beta) \quad \text{for each } n$$

$$\begin{aligned} \therefore \hat{T}_{n-1, n-1} \hat{T}_{nn} I(n\hat{\pi}, n\beta) &= \hat{T}_{n-1, n-1} I(n+1\hat{\pi}, n+1\beta) \\ &= I(n-2\hat{\pi}, n-2\beta). \end{aligned}$$

$$\therefore \hat{T}_{11} \hat{T}_{22} \dots \hat{T}_{nn} I(n\hat{\pi}, n\beta) = I(\hat{\pi}, \beta)$$

$$u_n \equiv I(n\hat{\pi}, n\beta).$$

$$\therefore \hat{T}_{11} \hat{T}_{22} \cdots \hat{T}_{nn} u_n = I(\hat{\pi}, \beta).$$

$$\|u\| \equiv \sup_S |u(s)|$$

一方、任意の $u \in M(S)$ に対して

$$\begin{aligned} \|\hat{T}_{11} \hat{T}_{22} \cdots \hat{T}_{nn} u_n - \hat{T}_{11} \hat{T}_{22} \cdots \hat{T}_{nn} u\| &\leq \beta_1 \beta_2 \cdots \beta_n \|u_n - u\| \\ &\leq \beta_1 \beta_2 \cdots \beta_n (L \|r\| + \|u\|). \end{aligned}$$

$$\therefore \|I(\hat{\pi}, \beta) - \hat{T}_{11} \hat{T}_{22} \cdots \hat{T}_{nn} u\| \leq \beta_1 \beta_2 \cdots \beta_n (L \|r\| + \|u\|)$$

$$\therefore \hat{T}_{11} \hat{T}_{22} \cdots \hat{T}_{nn} u \rightarrow I(\hat{\pi}, \beta) \quad \dots (1)$$

各 g_i は $g_i \in F(\pi)$ から Th. 4.3 (b) より

$$\hat{T}_{ji} u \leq U_j u \quad \text{for each } j.$$

$$\therefore \hat{T}_{j-1, j-1}(\hat{T}_{ji} u) \leq \hat{T}_{j-1, j-1}(U_j u) \leq U_{j-1} U_j u.$$

$$\therefore \hat{T}_{11} \hat{T}_{22} \cdots \hat{T}_{ji} u \leq U_1 U_2 \cdots U_j u \quad \text{for } \forall j$$

故に (1) より

$$I(\hat{\pi}, \beta) \leq u^*$$

$$(b) \quad \varepsilon' \equiv \varepsilon / (1+L)$$

Th. 4.3 (a) により β_j, u に depend して $\hat{f}_j \in F(\pi)$ が存在し

$$(\hat{f}_j, \beta_j) \rightsquigarrow \hat{T}_{ji} \text{ となる}$$

$$\hat{T}_{ji} u \geq U_j u - \varepsilon' \quad \dots (2)$$

$\beta_{j-1}, U_j u - \varepsilon'$ に depend して $\hat{f}_{j-1} \in F(\pi)$ が存在し

$$(\hat{f}_{j-1}, \beta_{j-1}) \rightsquigarrow \hat{T}_{j-1, j-1} \text{ となる}$$

$$\hat{T}_{j-1, j-1}(U_j u - \varepsilon') \geq U_{j-1}(U_j u - \varepsilon') - \varepsilon'$$

(2) により

$$\begin{aligned} \hat{T}_{j-1, j-1}(\hat{T}_{ji} u) &\geq \hat{T}_{j-1, j-1}(U_j u - \varepsilon') \geq U_{j-1}(U_j u - \varepsilon') - \varepsilon' \\ &= U_{j-1} U_j u - \beta_{j-1} \varepsilon' - \varepsilon'. \end{aligned}$$

$$\therefore \hat{T}_{11} \hat{T}_{22} \cdots \hat{T}_{ji} u \geq U_1 U_2 \cdots U_j u - \varepsilon' (1 + \sum_{m=1}^{j-1} \beta_1 \beta_2 \cdots \beta_m)$$

$$\therefore I(\hat{\pi}, \beta) \geq u^* - \varepsilon' (1+L) = u^* - \varepsilon.$$

(c) 各 $\delta\beta$ ($\delta \geq 0$) に対して ε -optimal policy が存在する。

$\pi^{\delta\beta} \in \delta\beta$ に 対応する ε -optimal policy となる。

$$\pi^{*j-1} \equiv (\pi_{j-1,1}, \pi_{j-1,2}, \dots)$$

$$\begin{aligned} \therefore I(\pi^{*j-1}, \beta) &= \pi_{j-1,1} g[r + \beta_j I(\pi^{*j-1}, \beta)] \\ &\leq \pi_{j-1,1} g[r + \beta_j \{I(\pi^{*j}, \beta) + \varepsilon\}] \\ &\leq f_j g[r + \beta_j \{I(\pi^{*j}, \beta) + \varepsilon\}] \quad (\text{Th. 3.1 に } \beta) \\ &= T_{jj} I(\pi^{*j}, \beta) + \beta_j \varepsilon. \end{aligned}$$

$$\therefore T_{jj} I(\pi^{*j}, \beta) \geq I(\pi^{*j-1}, \beta) - \beta_j \varepsilon \quad \dots (3)$$

(3) は各 j について成立するから

$$T_{j-1,j-1} I(\pi^{*j-1}, \beta) \geq I(\pi^{*j-2}, \beta) - \beta_{j-1} \varepsilon$$

今 β degenerate f_{j-1} を用いる。

$$\begin{aligned} \therefore T_{j-1,j-1} T_{jj} I(\pi^{*j}, \beta) &\geq T_{j-1,j-1} I(\pi^{*j-1}, \beta) - \beta_{j-1} \beta_j \varepsilon \\ &\geq I(\pi^{*j-2}, \beta) - \beta_{j-1} \varepsilon - \beta_{j-1} \beta_j \varepsilon \end{aligned}$$

$$\therefore T_{11} T_{22} \dots T_{jj} I(\pi^{*j}, \beta) \geq I(\pi^{*0}, \beta) - \varepsilon \left(\sum_{n=1}^j \beta_1 \beta_2 \dots \beta_n \right)$$

$$\hat{\pi} \equiv (f_1, f_2, \dots) \text{ に対し } \varepsilon \text{ 上式より}$$

$$I(\hat{\pi}, \beta) \geq I(\pi^{*0}, \beta) - \varepsilon L$$

$\rightarrow \pi^{*0}$ は β に対する ε -optimal から

$$I(\pi^{*0}, \beta) \geq I(\pi, \beta) - \varepsilon \quad \text{for } \forall \pi$$

$$\therefore I(\hat{\pi}, \beta) \geq I(\pi, \beta) - \varepsilon(1+L) \quad \text{for } \forall \pi$$

従って $\hat{\pi}$ は β に対する $(1+L)\varepsilon$ -optimal Markov policy である。

(d) 若し ε , 若し β に対して ε -optimal policy があることを仮定する。

従って, 若し n に対して, β に対して $\frac{1}{n(1+L)}$ -optimal policy がある。

Th. 4.4 (c) により β に対して $\frac{1}{n}$ -optimal Markov policy がある。これを π^{*n} とする。

$\hat{\pi} \in \pi^{*n} \in G(\hat{\pi})$ for $\forall j, \forall n$ 若し Markov policy とする。

$$(\hat{\pi}, \beta) \rightsquigarrow \hat{u}_j^* \text{ とする。}$$

Th. 4.4 (a) により

$$I(\pi, \beta) \leq \hat{u}_j^* \quad \text{for } \forall \pi \in G(\hat{\pi})$$

$$\therefore I(\pi^{*n}, \beta) \leq \hat{u}_j^* \quad \text{for } \forall n \quad \dots (4)$$

→ π^{δ^n} の定義から

$$I(\pi^{\delta^n}, \delta\beta) \geq I(\pi, \delta\beta) - \frac{1}{n} \quad \text{for } \forall \pi \quad \dots (5)$$

(4), (5) より

$$\hat{u}_j^* \geq I(\pi, \delta\beta) - \frac{1}{n} \quad \text{for } \forall \pi$$

$$\therefore \hat{u}_j^* \geq I(\pi, \delta\beta) \quad \text{for } \forall \pi, \forall j \quad \dots (6)$$

Th. 4.4 (b) により、各 n, j に対して $\tilde{\pi}^{n,j} \in G(\hat{\pi})$ が存在して

$$I(\tilde{\pi}^{n,j}, \delta\beta) \geq \hat{u}_j^* - 1/n$$

$$\therefore \hat{u}_j^* \leq I(\tilde{\pi}^{n,j}, \delta\beta) + \frac{1}{n}$$

$$\begin{aligned} \therefore T_{a_j} \hat{u}_j^* &\leq T_{a_j} [I(\tilde{\pi}^{n,j}, \delta\beta) + \frac{1}{n}] \\ &= I((a, \tilde{\pi}^{n,j}), \delta\beta) + \frac{\beta_j}{n} \\ &\leq \hat{u}_{j-1}^* + \frac{\beta_j}{n}. \quad ((6) \text{ による}) \end{aligned}$$

$$\therefore \sup_{a_j} T_{a_j} \hat{u}_j^* \leq \hat{u}_{j-1}^*.$$

$$\rightarrow \sup_{a_j} T_{a_j} \hat{u}_j^* \geq \bigcup_j \hat{u}_j^* = \hat{u}_{j-1}^*.$$

$$\therefore \sup_{a_j} T_{a_j} \hat{u}_j^* = \hat{u}_{j-1}^* \quad \text{for } \forall j.$$

(e) π^* があって、各 $\delta\pi^*$ が $\delta\beta$ に対して optimal と仮定する。

Th. 4.4 (c) により π^* は Markov policy とする。

$$\pi^* \equiv (f_1^*, f_2^*, \dots)$$

Th. 4.2 (d) により

$$I(\delta\pi^*, \delta\beta)_{s_0} = T_{f_j^*(s_0), j} I(\delta\pi^*, \delta\beta)_{s_0} \quad \text{for } \forall j$$

$$\therefore I(\delta\pi^*, \delta\beta)_{s_0} \leq \sup_{a_j} T_{a_j} I(\delta\pi^*, \delta\beta)_{s_0} \quad \text{for } \forall j$$

これは各 s_0 に対して成立するから

$$I(\delta\pi^*, \delta\beta) \leq \sup_{a_j} T_{a_j} I(\delta\pi^*, \delta\beta) \quad \dots (7)$$

→

$$I(\delta\pi^*, \delta\beta) \geq I((a, \delta\pi^*), \delta\beta)$$

$$= T_{a_j} I(\delta\pi^*, \delta\beta) \quad \text{for } \forall j$$

$$\therefore I(\delta\pi^*, \delta\beta) \geq \sup_{a_j} T_{a_j} I(\delta\pi^*, \delta\beta) \quad \text{for } \forall j \quad \dots (8)$$

(7), (8) より

$$I(\delta\pi^*, \delta\beta) = \sup_{a_j} T_{a_j} I(\delta\pi^*, \delta\beta) \quad \text{for } \forall j \quad \dots (9)$$

(9) は即ち optimality equation である。

次に逆の証明。

(9) を仮定する。

Th. 4.1 により (ρ, ϵ) -optimal Markov $\hat{\pi} = (\hat{f}_1, \hat{f}_2, \dots)$ が存在するから

$$P\{I(\hat{\pi}, \beta) \geq I(\pi, \beta) - \epsilon\} = 1 \quad \text{for } \forall \pi$$

次に今度は、

$$I(\hat{\pi}, \beta)_{s_0} \geq I(\pi, \beta)_{s_0} - \epsilon \quad \text{for } \forall \pi \quad \dots (10)$$

$(\hat{f}_i, \beta_i) \rightsquigarrow \hat{T}_{\beta_i}$ とおくと (9) の仮定の下では

$$\hat{T}_{\beta_i} I(i\pi^*, i\beta) \leq I(i-1\pi^*, i-1\beta) \quad \text{for } \forall i$$

$$\therefore \hat{T}_{\beta_{i-1}} \hat{T}_{\beta_i} I(i\pi^*, i\beta) \leq \hat{T}_{\beta_{i-1}} I(i-1\pi^*, i-1\beta)$$

$$\leq I(i-2\pi^*, i-2\beta) \quad \text{for } \forall i \quad ((9) \text{ による})$$

$$\therefore \hat{T}_{\beta_1} \hat{T}_{\beta_2} \dots \hat{T}_{\beta_i} I(i\pi^*, i\beta) \leq I(\pi^*, \beta) \quad \text{for } \forall i$$

$$\therefore I(\hat{\pi}, \beta) \leq I(\pi^*, \beta) \quad \dots (11)$$

(10), (11) より

$$I(\pi, \beta)_{s_0} \leq I(\hat{\pi}, \beta)_{s_0} + \epsilon \leq I(\pi^*, \beta)_{s_0} + \epsilon \quad \text{for } \forall \pi$$

$$\therefore I(\pi, \beta)_{s_0} \leq I(\pi^*, \beta)_{s_0} \quad \text{for } \forall \pi$$

これは各 s_0 で成立するから

$$I(\pi, \beta) \leq I(\pi^*, \beta) \quad \text{for } \forall \pi$$

故に、 π^* は β に対する optimal policy.

各 i につき、 $i\pi^*$ が $i\beta$ に対して optimal 存在をこの全系列にて証明出来る。

[Theorem 4.4 の証明 終り]

Corollary 各 $i\beta$ ($i \geq 0$) に対して optimal policy が"ある", β に対する optimal Markov policy が存在する。

Def. 4.7 (Blackwell)

equivalent action ; $a, b \in A,$

$$r(s, a, \cdot) = r(s, b, \cdot) \quad \text{かつ}$$

$$g(\cdot | s, a) = g(\cdot | s, b) \quad \text{のとき, } a \sim b \text{ といふ}$$

S_1 において equivalent となる。

essentially countable by π ;

$\pi = (f_1, f_2, \dots)$, 各 (s, a) に対して, $f_n(s)$ と a とが S_1 において equivalent になる n が存在するとき, A は ess. count. by π であるとなる。

essentially finite by π ;

$\pi = (f_1, f_2, \dots)$,

S の partition $\{S_n\}$ が存在して, 各 n について $s \in S_n$ なる各 (s, a) に対して $f_1(s), f_2(s), \dots, f_n(s)$ の中の少なくとも一つが S において a と equivalent になるとき, A は ess. finite by π であるとなる。

Theorem 4.5 (Furukawa)

- (a) A が ess. count. by some π なら, 各 $\varepsilon > 0$ に対して ε -optimal Markov policy が存在する。
- (b) A が ess. finite by some π なら, optimal Markov policy が存在する。

§ 5. D-h case 及び N-case

§ 4 の諸定理において, homogeneous discount factor を代入すると, 更に詳しい結果が得られる。

この場合, D-h case では, discount factor を改めて β と

おく。Def. 3.2 の D-case の定義より, 当然 $\beta < 1$ になければならぬ。

Def. 5.1

T_n ; (f_n, β) に対応する operator

U ; (π, β) に対応する operator

Lem. 5.1 U は Banach space における contraction mapping
($k \geq 1$, D-h case のみ)

Def. 5.2 U の fixed point $\in U^*$ とする。($k \geq 1$, D-h case のみ)

Theorem 5.1 (D-h), (Blackwell)

(a) 各 $p \in P(S)$, 各 $\varepsilon > 0$ に対して, (p, ε) -Optimal stationary policy が存在する。

(b) ε -Optimal policy があれば, $\varepsilon/(1-\beta)$ -Optimal stationary policy が存在する。 ($\varepsilon \geq 0$)

(c) π^* が optimal $\Leftrightarrow I(\pi^*) = \sup_{\pi} T_{\alpha} I(\pi^*)$

Theorem 5.2 (D-h), (Blackwell)

(a) A が ess. count. by some π なら, 各 $\varepsilon > 0$ に対して,

ε -Optimal stationary policy が存在する。

(b) A が ess. finite by some π なら, optimal stationary policy が存在する。

Def. 5.3 π^* が 強義 (p, ε) -optimal ; $p\{I(\pi^*) \geq \sup_{\pi} I(\pi) - \varepsilon\} = 1$

Theorem 5.3 (D-h, N) (Strauch)

各 p , 各 $\varepsilon > 0$ に対して, 強義 (p, ε) -Optimal Markov policy

が存在する。ただし, N -case では収束する expected total reward をもつ policy が少なくとも一つあることを仮定する。

(註) 一般に N -case では (p, θ) -optimal stationary policy が存在するかどうか明らかにされていない。

Theorem 5.4 (N) (Strauch)

Optimal policy があれば optimal stationary policy が存在する。

Theorem 5.5 (N) (Strauch)

A が ess. finite by some π , かつ, 収束する expected total reward をもつ policy が少なくとも一つあれば, optimal stationary policy が存在する。

§ 6. Additional results.

Theorem 6.1 (P)

各 π , 各 $\varepsilon > 0$, 各 $p \in P(S)$ に対して, 次の式を満たす semi-Markov policy τ 及び Markov policy σ が存在する. ;

$$I(\tau) \geq I(\pi) - \varepsilon, \quad p I(\sigma) \geq p I(\pi) - \varepsilon$$

Theorem 6.2 (P)

各 $p \in P(S)$, 各 $\varepsilon > 0$ に対して, (p, θ) -optimal semi-Markov policy が存在する。

(註) Th. 5.4 が P -case でも成立するかどうかは明らかでない。

Theorem 6.3 (D, P, N)

$$v^* \equiv \sup_{\pi} I(\pi) \quad \text{とおく,}$$

$$v^*(s) = \sup_a T_a v^*(s) \quad \text{for all } s \in S.$$

かつ, D-case では v^* は上式の unique bounded solution であり, P-case では v^* は上式の, s に 対し 一様に 最小の non-negative solution である。

Theorem 6.4 (D, N)

$$I(f, \pi) \geq I(\pi) \Rightarrow I(f^{(\infty)}) \geq I(\pi)$$

Theorem 6.5 (D, N)

σ, τ は policy として, π は 次の 様に 定義 する。

$$\pi_n = \begin{cases} \sigma_n & \text{if } (s_1, a_1, \dots, a_{n-1}, s_n) \in B_n \\ \tau_n & \text{if } (s_1, a_1, \dots, a_{n-1}, s_n) \in B_n^c \end{cases}$$

ただし,

$$B_n = \{(s_1, a_1, \dots, a_{n-1}, s_n) \mid u_n > v_n\},$$

$$u_n(s_1, a_1, \dots, a_{n-1}, s_n) = \sum_{j=n}^{\infty} \beta^{j-n} \sigma_n g \dots \sigma_j g \gamma,$$

$$v_n(s_1, a_1, \dots, a_{n-1}, s_n) = \sum_{j=n}^{\infty} \beta^{j-n} \tau_n g \dots \tau_j g \gamma$$

$$\Rightarrow I(\pi) \geq \max(I(\sigma), I(\tau))$$

Theorem 6.6 (D, N)

$\sigma = (f_1, f_2, \dots)$, $\tau = (g_1, g_2, \dots)$ は Markov policy とする。

$\pi = (h_1, h_2, \dots)$ は 次の 様に 定める。

$$h_n = \begin{cases} f_n & \text{if } I(n^{-1}\sigma) > I(n^{-1}\tau) \\ g_n & \text{if otherwise} \end{cases}$$

$$\Rightarrow I(\pi) \geq \max(I(\sigma), I(\tau))$$

Theorem 6.7 (D, N)

$f^{(\infty)}, g^{(\infty)}$ is stationary policy τ 17,

$$h = \begin{cases} f & \text{if } I(f^{(\infty)}) \geq I(g^{(\infty)}) \\ g & \text{if otherwise} \end{cases}$$

$$\Rightarrow I(h^{(\infty)}) \geq \max(I(f^{(\infty)}), I(g^{(\infty)}))$$

§ 7. Remarks.

マルコフ型決定過程として定式化される問題としては、discrete time D.P. problem の他に、Bayesian 的な統計の諸問題がある。以下、これを利用する。

- 1) Sequential Analysis
- 2) Bayesian Adaptive Control
- 3) Replacement Problem
- 4) Taxicab Problem
- 5) Allocation Problem
- 6) Inventory Problem
- 7) Smoothing Problem