数理解析研究所講究録
28巻 1967年 87-104

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，云いかえると 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 が問題にされない点に在る。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云っ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。

ここでは，これらの点について，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

$\mathcal{A}$; X における Borel 集合族

$\mathcal{B}$; Y における Borel 集合族

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

$g(y|x)$; conditional probability distribution (各 x について，$g(\cdot|x)$ は Y 上の prob. dist., 各 $B \in \mathcal{B}$ について $g(B|\cdot)$ は X 上の $\mathcal{A}$ 可測関数)

$Q(Y|X)$; $g(y|x)$ の class

Def. 2.2

$M(X)$:
- X 上の実数値有界可測関数の class (D-case)
- X 上の non-positive 実数値有界可測関数の class (N-case)
- X 上の non-negative 実数値有界可測関数の class (P-case)

(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

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

$q u$; $q u(x) \equiv \int_Y u(x,y)\, dq(y|x)$ for $q \in Q(Y|X)$, $u \in M(XY)$

Def. 2.5 以上の諸定義を拡張して

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

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

$M(X_1 X_2 \cdots X_n)$

Def. 2.6

$p \in P(X)$ 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

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

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

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

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

discount factor vector

Def. 2.8

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

Def. 2.9

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

$\pi_n \in Q(A|H_n)$, $n=1,2,\cdots$ として

$\pi \equiv (\pi_1, \pi_2, \cdots)$; policy

Def. 2.10 $e_\pi \equiv \pi_1 q \pi_2 q \cdots \in Q(ASAS\cdots|S)$

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

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

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

Def. 2.11

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

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

Markov policy を $(f_1, f_2, \cdots)$ と書く。

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

§3. 基本定理

Def. 3.1 $u, v \in M(X)$ に対して

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

Lem. 3.1 (Blackwell and Nardzewski [ ])

$g \in Q(Y|X)$, $S \in \mathcal{A} \times \mathcal{B}$, $g(S_x|x) > 0$ for all $x \in X$

$\Rightarrow$ $(x, g(x)) \in S$ for all $x \in X$ なる $\mathcal{A}$ 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) $fu \geq gu$

(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 ; $\sum_{j=1}^{\infty} \beta_1 \beta_2 \cdots \beta_j \equiv L$: 收束, $\gamma \in M(SAS)$

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

P-case ; $\beta_i = 1$ for $i = 1, 2, \cdots$, $\gamma \in M(SAS)$, かつ
すべての 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 $i = 1, 2, \cdots$

D-n-case ; D-case で D-h case 以外

Lem 3.2

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

(P-case) $I_n(\pi) \updownarrow I(\pi)$ 收束

(N-case) $I_n(\pi) \downarrow I(\pi)$ ($-\infty$ に発散のこともある)

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

Def. 4.2 (Furukawa)

T_{n_j} ; $T_{n_j}u(s) = \int [r(s, f_n(s), t) + \beta_j u(t)]\, dq(t \mid s, f_n(s))$

$\pi = (f_1, f_2, \cdots)$ に対応して

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

T_{nj} を (f_n, β_j) に対応する operator と云い，$(f_n, \beta_j) \rightsquigarrow T_{nj}$ と書く。

$\sqcup_j$ を $(\pi=(f_1, f_2, \cdots), \beta_j)$ に対応する operator と云い，

$(\pi, \beta_j) \rightsquigarrow \sqcup_j$ と書く。

Theorem 4.2 (Furukawa)

(a) T_{nj} は単調，即ち $u \le v \Rightarrow T_{nj}u \le T_{nj}v$

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

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

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

Def. 4.3

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

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

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

Theorem 4.2 (d)

$\pi=(f_1, f_2, \cdots)$ に対して

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

Def. 4.4 (Blackwell [ ])

π を Markov policy として

f が π-generated ; f は $S \to A$ なる可測 mapping

S の partition $\{S_n\}$ が存在して

$f = f_n$ on S_n for each n.

policy $\hat{\pi}$ が π-generated ; $\hat{\pi} = (g_1, g_2, \cdots)$ 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 $\pi = (f_1, f_2, \cdots)$, 各 $\varepsilon > 0$, 各 β_j に対して $\hat{f}_j \in F(\pi)$ が
存在し, $(\hat{f}_j, \beta_j) \rightsquigarrow \hat{T}_{jj}$ 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
$(\hat{f}, \beta_j) \rightsquigarrow \hat{T}_j$ とすれば"

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

Def. 4.6 (Furukawa)

$(\pi, \beta_j) \rightsquigarrow U_j$ として, $\lim_{n\to\infty} U_j U_{j+1} \cdots U_n u \equiv u^*_{j-1}$ とおき
これを $(\pi, j\to\beta)$ に対応する limit point と云う。記号で
$(\pi, j\to\beta) \rightsquigarrow u^*_{j-1}$ 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 ${}^{j}\beta$ $(j \geq 0)$ に対して ε-optimal policy (${}^{j}\beta$ に依存する) が有れば, β に対する $(1+L)\varepsilon$-optimal Markov policy が存在する。 $(\varepsilon \geq 0)$

(d) $(f \equiv a, \beta_j)$ に対応する operator を T_{aj} とする。各 $\varepsilon > 0$, 各 ${}^{j}\beta$ に対して ε-optimal policy (${}^{j}\beta$ に依存する) が有れば Markov policy $\hat{\pi}$ が存在して $(\hat{\pi}, {}^{j}\beta) \rightsquigarrow \hat{u}^*_j$ については $\hat{u}^*_j$ は可測関数で, かつ

$$\hat{u}^*_{j-1} = \sup_{a \in A} T_{aj} \hat{u}^*_j \qquad \text{for each } j$$

が成立する。

(e) 各 j につき, ${}^{j}\pi^*$ が ${}^{j}\beta$ に対して optimal となるための必十条件は

$$I({}^{j-1}\pi^*, {}^{j-1}\beta) = \sup_{a \in A} T_{aj} I({}^{j}\pi^*, {}^{j}\beta) \qquad \text{for each } j$$

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。

$\hat{\pi} \equiv (g_1, g_2, \cdots) \in G(\pi)$, $(g_j, \beta_j) \rightsquigarrow \hat{T}_j$ とする。

Theorem 4.2 (d) より $\hat{T}_{mn} I({}^{n}\hat{\pi}, {}^{n}\beta) = I({}^{n-1}\hat{\pi}, {}^{n-1}\beta)$ for each n

$\therefore \hat{T}_{n-1,n-1} \hat{T}_{nn} I({}^{n}\hat{\pi}, {}^{n}\beta) = \hat{T}_{n-1,n-1} I({}^{n-1}\hat{\pi}, {}^{n-1}\beta)$

$= I({}^{n-2}\hat{\pi}, {}^{n-2}\beta)$.

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

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

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

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

一方 任意の $u \in M(S)$ にたして

$\|\hat{T}_{11}\hat{T}_{22}\cdots\hat{T}_{mm}u_m - \hat{T}_{11}\hat{T}_{22}\cdots\hat{T}_{mm}u\| \leq \beta_1\beta_2\cdots\beta_m\|u_m - u\|$

$\leq \beta_1\beta_2\cdots\beta_m\left(L\|r\| + \|u\|\right).$

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

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

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

$\hat{T}_{jj}u \leq \cup_j u \quad$ for each j.

$\therefore \hat{T}_{j-1,j-1}(\hat{T}_{jj}u) \leq \hat{T}_{j-1,j-1}(\cup_j u) \leq \cup_{j-1}\cup_j u.$

$\therefore \hat{T}_{11}\hat{T}_{22}\cdots\hat{T}_{jj}u \leq \cup_1\cup_2\cdots\cup_j u \quad$ for $\forall j$

故に (1) より

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

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

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

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

$\hat{T}_{jj}u \geq \cup_j u - \varepsilon' \quad \cdots\cdots (2)$

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

$(\hat{f}_{j-1}, \beta_{j-1}) \rightsquigarrow \hat{T}_{j-1,j}$ とすると

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

(2) により

$\hat{T}_{j-1,j-1}(\hat{T}_{jj}u) \geq \hat{T}_{j-1,j-1}(\cup_j u - \varepsilon') \geq \cup_{j-1}(\cup_j u - \varepsilon') - \varepsilon'$

$= \cup_{j-1}\cup_j u - \beta_{j-1}\varepsilon' - \varepsilon'.$

$\therefore \hat{T}_{11}\hat{T}_{22}\cdots\hat{T}_{jj}u \geq \cup_1\cup_2\cdots\cup_j u - \varepsilon'\left(1 + \sum_{m=1}^{j-1}\beta_1\beta_2\cdots\beta_m\right)$

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

(c) 各 β $(\beta \geq 0)$ にたして ε-optimal policy があるとする。

π^*_β を β にたする ε-optimal policy とする。

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

$\therefore I(\pi^{*j-1}, {}^{j}\beta) = \pi_{j-1,1} g[r + \beta_j I({}^{1}\pi^{*j-1}, {}^{j}\beta)]$

$\leq \pi_{j-1,1} g[r + \beta_j \{I(\pi^{*j}, {}^{j}\beta) + \varepsilon\}]$

$\leq f_j g[r + \beta_j \{I(\pi^{*j}, {}^{j}\beta) + \varepsilon\}]$ (Th. 3.1 による)

$= T_{jj} I(\pi^{*j}, {}^{j}\beta) + \beta_j \varepsilon$.

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

(3)は各 j について成立するから

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

なる degenerate f_{j-1} が存在。

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

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

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

$\hat{\pi} \equiv (f_1, f_2, \cdots)$ とおくに 上式より

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

→ π^{*0} は β にたする ε-op だから

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

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

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

(d) 各 ε, 各 ${}^{j}\beta$ にたして ε-optimal policy があると仮定する。

従って、各 n にたして、${}^{j}\beta$ にたする $\frac{1}{n(1+L)}$-optimal policy が存在する。

Th. 4.4 (c) により ${}^{j}\beta$ にたする $\frac{1}{n}$-optimal Markov policy が存在する。これを π^{jn} とする。

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

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

Th. 4.4 (a) により

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

$\therefore I(\pi^{jn}, {}^{j}\beta) \leq \hat{u}^{*}_{j}$ for $\forall n$ ..... (4)

一方 π^{j_m} の定義から

$$I(\pi^{j_m}, {}^{j}\beta) \geq I(\pi, {}^{j}\beta) - \frac{1}{m} \quad \text{for } \forall \pi \quad \cdots\cdots (5)$$

(4), (5) より

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

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

Th. 4.4 (b) により、各 m, j にたいして $\tilde{\pi}^{m_j} \in G(\hat{\pi})$ がなたして

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

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

$$\therefore T_{a_j} \hat{u}^*_j \leq T_{a_j}\left[I(\tilde{\pi}^{m_j}, {}^{j}\beta) + \frac{1}{m} \right]$$

$$= I((a, \tilde{\pi}^{m_j}), {}^{j-1}\beta) + \frac{\beta_j}{m}$$

$$\leq \hat{u}^*_{j-1} + \frac{\beta_j}{m}. \qquad ((6) \text{による})$$

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

一方 $\sup_a T_{a_j} \hat{u}^*_j \geq \bigcup_j \hat{u}^*_j = \hat{u}^*_{j-1}$.

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

(e) π^* があって 各 ${}^{j}\pi^*$ が ${}^{j}\beta$ にたいして optimal と仮定する。

Th. 4.4 (c) により π^* は Markov policy としてよい。

$$\pi^* \rightleftharpoons (f^*_1, f^*_2, \cdots\cdots)$$

Th. 4.2 (d) により

$$I({}^{j-1}\pi^*, {}^{j-1}\beta)_{s_0} = T_{f^*_j(s_0), j}\, I({}^{j}\pi^*, {}^{j}\beta)_{s_0} \quad \text{for } \forall j$$

$$\therefore I({}^{j-1}\pi^*, {}^{j-1}\beta)_{s_0} \leq \sup_a T_{a_j} I({}^{j}\pi^*, {}^{j}\beta)_{s_0} \quad \text{for } \forall j$$

これは各 s_0 で成立するから

$$I({}^{j-1}\pi^*, {}^{j-1}\beta) \leq \sup_a T_{a_j} I({}^{j}\pi^*, {}^{j}\beta) \quad \cdots\cdots (7)$$

一方

$$I({}^{j-1}\pi^*, {}^{j-1}\beta) \geq I((a, {}^{j}\pi^*), {}^{j-1}\beta)$$

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

$$\therefore I({}^{j-1}\pi^*, {}^{j-1}\beta) \geq \sup_a T_{a_j} I({}^{j}\pi^*, {}^{j}\beta) \quad \text{for } \forall j \quad \cdots\cdots (8)$$

(7), (8) より

$$I({}^{j-1}\pi^*, {}^{j-1}\beta) = \sup_a T_{a_j} I({}^{j}\pi^*, {}^{j}\beta) \quad \text{for } \forall j \quad \cdots (9)$$

(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 $\hat{\pi} = (\hat{f}_1, \hat{f}_2, \cdots)$ が存在するから

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

故に 特に,

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

$(\hat{f}_j, \beta_j) \rightsquigarrow \hat{T}_{\beta_j}$ とすると (9) の仮定の下では

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

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

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

$$\therefore \hat{T}_{\beta_1} \hat{T}_{\beta_2} \cdots \hat{T}_{\beta_j} I(\beta^j \pi^*, \beta^j \beta) \leq I(\pi^*, \beta) \quad \text{for } \forall \beta_j$$

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

(10), (11) より

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

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

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 β_j につき, $\beta^j \pi^*$ が $\beta^j \beta$ に対して optimal なることも全く同じにて証明出来る。

[Theorem 4.4 の証明終り]

Corollary 各 β_j $(j \geq 0)$ に対して optimal policy があれば, β に対する optimal Markov policy が存在する。

Def. 4.7 (Blackwell)

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

$r(s, a, \cdot) = r(s, b, \cdot)$ かつ

$q(\cdot \mid s, a) = q(\cdot \mid s, b)$ のとき, a と b は

sにおいて equivalent と云う。

essentially countable by π ;

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

essentially finite by π;

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

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 $f_1(s), f_2(s), \cdots, f_n(s)$ の中の少くとも一つが s において a と equivalent になるとき, A は ess. finite by π であると云う。

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 $\beta < 1$ でなければならぬ。

Def. 5.1

T_n ; (f_n, β) に対応する operator

$\sqcup$; (π, β) に対応する operator

Lem. 5.1 $\sqcup$ は Banach space における contraction mapping (たゞし，D-h case のみ)

Def. 5.2 $\sqcup$ の fixed point を u^* とする。(たゞし，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 $\Leftrightarrow$ $I(\pi^*) = \sup_a T_a I(\pi^*)$

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 が少くとも一つ有ることを仮定する。

(註) 一般に N-case では (p,ε)-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 が有れば 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 が少くとも一つ有れば，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，次の式を充たす semi-Markov policy τ 及び Markov policy σ が存在する。;

$$I(\tau)\geq I(\pi)-\varepsilon, \qquad pI(\sigma)\geq pI(\pi)-\varepsilon$$

Theorem 6.2 (P)

各 $p\in P(S)$，各 $\varepsilon>0$ に対して，(p,ε)-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)

$v^* \equiv \sup_\pi I(\pi)$ とおくと,

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

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 に関し一様に最小の non-negative solution である。

Theorem 6.4 (D, N)

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

Theorem 6.5 (D, N)

σ, τ を policy として, π を次の様に定義する。

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

ただし

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

$$u_n(s_1, a_1, \cdots, a_{n-1}, s_n) = \sum_{j=n}^{\infty} \beta^{j-1} \sigma_n q \cdots \sigma_j q r,$$

$$v_n(s_1, a_1, \cdots, a_{n-1}, s_n) = \sum_{j=n}^{\infty} \beta^{j-1} \tau_n q \cdots \tau_j q r$$

$$\Rightarrow \qquad I(\pi) \geqq \max\left(I(\sigma), I(\tau) \right)$$

Theorem 6.6 (D, N)

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

$\pi = (h_1, h_2, \cdots)$ を次の様に定める。

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

$$\Rightarrow \quad I(\pi) \geqq \max\left(I(\sigma), I(\tau)\right)$$

Theorem 6.7 (D, N)

$f^{(\infty)}, g^{(\infty)}$ を stationary policy として,

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

$$\Rightarrow \quad I(h^{(\infty)}) \geqq \max\left(I(f^{(\infty)}), I(g^{(\infty)})\right)$$

§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, これを列挙する。

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