数理解析研究所講究録
第358巻 1979年 99-115

On adaptive policies in Markovian Decision Processes with uncertain transition matrices

千葉大 教育　蔵野 正美

1. はじめに

推移確率行列が未知の場合のマルコフ決定過程の研究は色々な見地（たとえばゲーム論的，最尤法，ベイズ的方法）から行われている。ここでは learning policy（学習政策）を定義して，discount case および average case において，learning policy と optimal policy との関係を調べる。§.3 では discount case におけるベイズ的方法が議論され，learning policy が定義される。そして，任意の $\varepsilon > 0$ に対して，ε-optimal な learning policy が存在することを証明する。また optimal policy は必ずしも learning policy でない例が与えられる。§.4 では，average case において，optimal な learning policy が存在することを示す。

証明はいずれも構成的である。Sweat [10] は，a game model involving repeated play of a matrix game with unknown entries

において，"learning"と"optimal"との関係を研究した。この小論の手法は，いわばSweatの方法のMarkov modelへの適応と拡張といってよい。

2. Definitions and Notations

$S=\{1,2,\cdots,L\}$: state space　$A=\{1,2,\cdots,k\}$: action space

r: $S\times A$の上に定義されたreturn function

仮定　$0\leqq r(s,a)\leqq R<\infty$

parameter space of K unknown stochastic matrices を

$\Theta=\{\theta \mid \theta=(\theta^a;a\in A),\ \theta^a_{ss'}\geqq 0,\ \sum_{s'\in S}\theta^a_{ss'}=1 \text{ for } s,s'\in S, a\in A\}$

で表わす。可測空間$(\Theta, \mathcal{F}_\Theta)$の上の確率分布の全体を$\Xi$で表わす。history $w=(x_0,a_0,x_1,a_1,x_2,\cdots)$の全体を$\Omega\equiv S\times(A\times S)^\infty$で表わし，$\mathcal{F}$を$\Omega$の上の通常の$\sigma$-fieldとする。$w\in\Omega$に対して　$X_t(w)=x_t$　はt期のstateを，$\Delta_t(w)=a_t$　はt期のactionを，$H_t(w)=(x_0,a_0,\cdots,x_t)$　はt期までのhistoryを表わす。但し，$w=(x_0,a_0,\cdots)$。

次にPolicyを定義しよう。$\pi=(\pi_0,\pi_1,\cdots)$がPolicyとは$t=0,1,2,\cdots$において，$\pi_t(\cdot/\xi, H_t(w))$は$\forall w\in\Omega$，$\forall\xi\in\Xi$に対して　Aの上の確率分布であるときをいう。Policyの全体をΠで表わす。$\forall\pi\in\Pi$に対して，$\pi_\xi=(\pi_{0,\xi},\pi_{1,\xi},\cdots)$を$t=0,1,\cdots$に対して　$\pi_{t\xi}(\cdot/H_t(w))=\pi_t(\cdot/\xi,H_t(w))$と定義する。$\pi_\xi$の全体を　$\Pi_\xi\equiv\{\pi_\xi;\pi\in\Pi\}$とする。

$\pi \in \Pi$, $\theta \in \Theta$, $\xi \in \Xi$ が与えられたときの system の推移法則は次の条件に従うものとする。

$$P(\Delta_t = a / X_0, \Delta_0, \cdots, X_t = s) = \pi_{\xi,t}(a / X_0, \Delta_0, \cdots, X_t = s)$$

$$P(X_{t+1} = s' / X_0, \cdots, X_t = s, \Delta_t = a) = \theta^a_{ss'}$$

従って，通常の方法で初期値 $x \in S$ に対して $(\Omega, \mathcal{F})$ の上の確率分布 $P^x_{\theta,\pi_\xi}(\cdot)$ が定義される。また $\pi \in \Pi$, $\xi \in \Xi$ が与えられたときの $(\Omega, \mathcal{F})$ の上の確率分布 $P^x_{\xi,\pi}(\cdot)$ は次で定義する。 $P^x_{\xi,\pi}(D) = \int_\Theta \xi(d\theta) P^x_{\theta,\pi_\xi}(D)$ for $D \in \mathcal{F}$

3. Adaptive and learning policies in the discounted case.

β を discount factor $0 \le \beta < 1$ として，$\pi \in \Pi$, $\xi \in \Xi$, $x \in S$ に対して，無限期間の期待利得，最適期待利得を次のように定める。

$$v_\beta[\pi](x,\xi) \equiv E^x_{\xi,\pi}\left[\sum_{t=0}^{\infty} \beta^t r(X_t, \Delta_t)\right]$$

$$v_\beta(x,\xi) \equiv \sup_{\pi \in \Pi} v_\beta[\pi](x,\xi)$$

定義1

π^*; optimal $\Leftrightarrow$ $v_\beta(x,\xi) = v_\beta[\pi^*](x,\xi)$ for $x \in S$, $\xi \in \Xi$.

$\forall \varepsilon > 0$, π^*; ε-optimal $\Leftrightarrow$ $v_\beta[\pi^*](x,\xi) \ge v_\beta(x,\xi) - \varepsilon$ for $x \in S$, $\xi \in \Xi$.

$a \in A$, $s, s' \in S$ に対して Ξ の上の operator $T^a_{ss'}$ を次で定義する。

$$T^a_{ss'}\xi(D) = \int_D \theta^a_{ss'} d\xi(\theta) \Big/ \int_\Theta \theta^a_{ss'} d\xi(\theta) \quad \text{for } D \in \mathcal{F}_\Theta \quad (\text{Bayes の定理})$$

次の定理は chap 3 of Martin [7] につかわれたのと同じ方法で証明される。

定理1 ([7])

$\{v_\beta(x,\xi)\}$ は次の関数方程式を満足する。

(1) $$v_\beta(x,\xi) = \max_{a\in A}\{r(x,a) + \beta\sum_{x'\in S}\bar{\theta}^a_{xx'}(\xi)\, v_\beta(x', T^a_{xx'}(\xi))\}$$

for $x\in S$ and $\xi\in\Xi$ 但し, $\bar{\theta}^a_{xx'}(\xi) = \int_\Theta \theta^a_{xx'}\, d\xi$.

各 $x\in S$, $\xi\in\Xi$ に対して $g^*(x,\xi)$ を(1)の右辺を最大にするactionの1つを表わすとする。そのとき, $\pi^* = (\pi^*_0, \pi^*_1, \cdots)$ を $t=0,1,2,\cdots$ に対して $\pi^*_t(\{g^*(x_t, \xi_{H_t})\}/\xi, H_t) = 1$ で定義すると, π^* は optimal policy である。但し ξ_{H_t} は, history H_t が与えられたときの事後分布である。

$\pi\in\Pi$ が与えられたとき, $\xi\in\Xi$ と n 期までの history $H_n(w)$ に対して policy $\pi(\xi, H_n) = (\pi(\xi, H_n)_0, \pi(\xi, H_n)_1, \cdots)$ を

$\pi(\xi, H_n)_t(\cdot/\xi', H_t(w')) = \pi(\cdot/\xi, H_n, H_t(w'))$ $(t=0,1,2,\cdots)$

for $w'\in\Omega$, $\xi'\in\Xi$ で定義する。

$\Theta^+ \equiv \{\theta / \theta^a_{ss'} > 0$ for $s,s'\in S$ and $a\in A\}$ として, learning policy を次に定義しよう。

定義2

$\delta > 0$ に対して, π が δ-learning policy である

$\Leftrightarrow {}^\forall\xi\in\Xi$, ${}^\forall\theta\in\Theta^+\cap \mathrm{supp}\,\xi$ に対して次が成立する: ${}^\forall\varepsilon > 0$ に対して次の(i),(ii)を満足する $A_\varepsilon\in\mathcal{F}$ が存在する

(i) $P^{x}_{\theta,\pi_\xi}(A_\xi) = 1$ for $x \in S$ and

(ii) $\forall w \in A_\varepsilon$ に対して正の整数 $n_\varepsilon(w)$ が存在して，$\forall n \geq n_\varepsilon(w)$ に対して θ が parameter の実際の値のとき，policy $\pi(\xi, H_n(w))$ は $(\delta+\varepsilon)$-optimal となる。

ie $v_\beta[\pi(\xi, H_n(w))](x, I_\theta) \geq v_\beta(x, I_\theta) - (\varepsilon+\delta)$ for $\forall x \in S$, $\forall n \geq n_\varepsilon(w)$。ここで $I_\theta \in \Xi$ は θ で退化している確率分布を表わす。

δ-learning policy を使うと parameter の実際の値に対して，system の operation は確率 1 で最終的に δ-最適となる。π が任意の $\delta > 0$ に対して δ-learning ならば π を <u>learning policy</u> という。

<u>定理2</u>

$\forall \varepsilon > 0$ に対して $\varepsilon/(1-\beta)$-optimal な learning policy が存在する。

以下，定理2の証明のために準備をする。

$s \in S$, $a \in A$ に対して

$$u_t(s,a)(w) = 1 \quad \text{if } X_t(w) = s \text{ and } \Delta_t(w) = a$$
$$= 0 \quad \text{otherwise}$$

$$W \equiv \{ w \mid \sum_{t=0}^{\infty} u_t(s,a) = +\infty \text{ for } s \in S, a \in A \}$$

そのとき次の lemma を得る。

Lemma 1

$\pi=(\pi_0,\pi_1,\cdots)$ を次を満たす policy とする。すなわち，次の (i),(ii) を満足する正の実数列 $\{\varepsilon_k\}_0^\infty$ が存在する：

(i) $\pi_t(a/\xi, H_k(w)) \geq \varepsilon_k$ for $a\in A$, $\xi\in\Xi$, $w\in\Omega$

(ii) $\sum_0^\infty \varepsilon_k = +\infty$

$\Rightarrow$ $P^\pi_{\theta,\pi_\xi}(W)=1$ for $\theta\in\Theta^+$, $x\in S$, $\xi\in\Xi$

<略証>

任意の $(s,a)\in S\times A$ を固定する。簡単のために $P^\pi_{\theta,\pi_\xi}(\cdot)$ を $P(\cdot)$ で表わす。 $B_t \equiv \{U_t(s,a)=1\}$ $(t=1,2,\cdots)$

任意の n, N $(N>n)$ に対して

$$P(\bigcap_{t=n}^N B_t^c) \leq P(B_n^c)\, e^{-\sum_{\ell=n+1}^N P(B_\ell / B_n^c\cap\cdots\cap B_{\ell-1}^c)}$$

仮定(i)より $P(B_\ell^c / B_n^c\cap\cdots\cap B_{\ell-1}^c) \geq [\min_{s,s',a}\theta^a_{ss'}]\varepsilon_\ell$

ゆえに $P(\bigcap_{t=n}^N B_t^c) \leq P(B_n^c)\, e^{-[\min_{s,s',a}\theta^a_{ss'}]\sum_{\ell=n+1}^N \varepsilon_\ell}$

仮定(ii)より $P(\lim_{t\to\infty} B_t^c)=0$

すなわち $P(\overline{\lim_{t\to\infty}} B_t)=1$ Q.E.D.

Lemma 2 ([8], [9])

$\xi\in\Xi$, $\theta\in\Theta^+\cap \mathrm{supp}\,\xi$, $x\in S$, $\pi\in\Pi$

そのとき

$$P^\pi_{\theta,\pi_\xi}(W)=1 \Rightarrow \xi_{H_n(w)} \to I_\theta \text{ weakly in } \Xi$$

as $n\to\infty$ with P^π_{θ,π_ξ}-probability 1

Lemma 3

$\xi_n \to \xi$ weakly in Ξ as $n \to \infty$

$\Rightarrow V_\beta(x, \xi_n) \to V_\beta(x, \xi)$ for $x \in S$

<略証>

関数列 $\{U^\ell(x,\xi),\ \ell=0,1,2,\cdots\}$ を逐次的に定義する。

$$U^\ell(x,\xi) = \max_{a\in A}\{r(x,a) + \beta \sum_{x'\in S} \theta^a_{xx'}(\xi)\, U^{\ell-1}(x', T^a_{xx'}(\xi))\}$$

$$U^0(x,\xi) = 0$$

そのとき Helley-Bray theorem と induction によって,

$\xi_n \to \xi$ weakly in Ξ as $n\to\infty$ のとき,

$U^\ell(x,\xi_n) \to U^\ell(x,\xi)$ $(\ell = 1,2,\cdots)$ が成り立つ。

従って, $^\forall\varepsilon > 0$ と各 ℓ $(\ell=1,2,\cdots)$ に対して正の整数 N_ℓ が存在して $|U^\ell(x,\xi_n) - U^\ell(x,\xi)| < \frac{\varepsilon}{3}$ for $n \geq N_\ell$ が成り立つ。

また $U^\ell(x,\xi) \to V_\beta(x,\xi)$ uniformly in (x,ξ) as $\ell\to\infty$ が成り立つことはしめされるので, 従って, ある正の整数 M が存在して, $\ell \geq M$ に対して

$|V_\beta(x,\xi_n) - U^\ell(x,\xi_n)| < \frac{\varepsilon}{3}$,

$|V_\beta(x,\xi) - U^\ell(x,\xi)| < \frac{\varepsilon}{3}$ が成り立つ。

結局, $\ell \geq M$ に対して

$|V_\beta(x,\xi_n) - V_\beta(x,\xi)| < \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon$ for $n \geq N_\ell$

を得る。 QED.

<定理の略証>

$\forall \varepsilon > 0$, $\delta_\varepsilon \equiv \varepsilon(1-\beta)/R(L-1)$ とする。

policy ${}^\varepsilon\pi = ({}^\varepsilon\pi_0, {}^\varepsilon\pi_1, \cdots)$ を次のように定義する。

$\xi \in \Xi$, $\omega \in \Omega$, $t = 0, 1, 2, \cdots$ に対して

$${}^\varepsilon\pi_t(a/\xi, H_t(\omega)) = \begin{cases} 1-(K-1)\delta_\varepsilon & \text{if } a = g^*(X_t(\omega), \xi_{H_t(\omega)}) \\ \delta_\varepsilon & \text{otherwise} \end{cases}$$

但し, g^*は定理1で定義された。

与えられた$\varepsilon > 0$に対して, $\{\varepsilon_t\}_0^\infty$を次をみたす数列とする。

(a) $\varepsilon_t \leq \varepsilon$ for $t = 0, 1, 2, \cdots$

(b) $\varepsilon_t \downarrow 0$ as $t \to \infty$ and (c) $\sum_0^\infty \varepsilon_t = +\infty$

そのとき, ${}^\varepsilon\pi^* = ({}^\varepsilon\pi_0^*, {}^\varepsilon\pi_1^*, \cdots)$ を $t = 0, 1, 2, \cdots$ に対して

$${}^\varepsilon\pi_t^*(a/\xi, H_t(\omega)) = {}^{\varepsilon_t}\pi_t(a/\xi, H_t(\omega)) \quad \text{for } \omega \in \Omega, \xi \in \Xi$$

で定義する。今, ${}^\varepsilon\pi^*$は$\varepsilon/(1-\beta)$-optimalでかつlearning policyとなることを示そう。$M(S \times \Xi)$を$S \times \Xi$の上に定義された有界なBair関数の全体とする。$g \in M(S \times \Xi)$によってn期でtruncateされる過程を考えて

$$v_\beta^n[\pi, g(\cdot,\cdot)](x, \xi) = E_{\xi,\pi}^x\left[\sum_0^n \beta^t r(X_t, \Delta_t) + \beta^{n+1} g(X_{n+1}, \xi_{H_{n+1}})\right].$$

そのとき, 任意のn ($n = 1, 2, \cdots$)に対して

$$v_\beta^n[{}^\varepsilon\pi^*, v_\beta(\cdot,\cdot) - \varepsilon/(1-\beta)](x, \xi) \leq v_\beta(x, \xi) - \frac{\varepsilon}{1-\beta} \quad \text{for } \forall x \in S, \forall \xi \in \Xi$$

が示される。また

$n \to \infty$ のとき $v_\beta^n[{}^\varepsilon\pi^*, v_\beta(\cdot,\cdot) - \varepsilon/(1-\beta)](x, \xi) \to v_\beta[{}^\varepsilon\pi^*](x, \xi)$

であるから，$\varepsilon\pi^*$ は $\varepsilon/(1-\beta)$-optimal となる。

次に証明は複雑になるが，lemma 1 ～ 3 および truncation method によって $\varepsilon\pi^*$ が learning policy であることが示される。 Q.E.D.

例 （optimal な learning policy は必ずしも存在しない）

$S = A = \{1, 2\}$, $r(1, a) = 1$ for all $a \in A$,

$r(2, a) = 0$ for all $a \in A$

$\overline{\Theta} \equiv \{\theta \mid \theta'_{21} = \theta'_{22} = \frac{1}{2} \quad \theta^a_{11} = \theta^a_{12} = \frac{1}{2}$ for all $a \in A\}$

$\Xi^* = \{\xi \in \Xi \mid \mathrm{supp}\, \xi = \overline{\Theta}\}$

定理 1 から $\forall \xi \in \Xi^*$ に対して，次式を得る。

$$v_\beta(2, \xi) = \max\left\{\frac{\beta}{2}(v_\beta(1, T'_{21}\xi) + v_\beta(2, T'_{22}\xi)),\right.$$
$$\left.\beta(\overline{\theta}^2_{21}(\xi)\, v_\beta(1, T^2_{21}\xi) + (1 - \theta^2_{21}(\xi)) v_\beta(2, T^2_{21}\xi))\right\}$$

$\theta^{*2}_{21} = \frac{1}{9}$ なる $\theta^* \in \overline{\Theta}$ を選ぶ。そのとき

$\exists \{\xi_n\}_1^\infty$ such that

(a) $\xi_n \to I_{\theta^*}$ weakly in Ξ as $n \to \infty$

(b) $\xi_n \in \Xi^*$ for $n = 1, 2, 3, \cdots$

lemma 3 と Helly-Bray theorem から

$$v_\beta(2, \xi_n) \to v_\beta(2, I_{\theta^*}) = \max\left\{\frac{1}{2}\beta(v_\beta(1, I_{\theta^*}) + v_\beta(2, I_{\theta^*}))\right.$$
$$\left., \beta(\frac{1}{9} v_\beta(1, I_{\theta^*}) + \frac{8}{9} v_\beta(2, I_{\theta^*}))\right\}$$

$v_\beta(1, I_{\theta^*}) > v_\beta(2, I_{\theta^*})$ から

$$v_\beta(2, \xi^*) = \frac{1}{2}\beta(v_\beta(1, T'_{21}\xi^*) + v_\beta(2, T'_{22}\xi^*))$$

なる $\xi^* \in \Xi^*$ が存在する。その結果 $P^2_{\xi^*,\pi^*}(\Delta_t = 1) = 1$ for $t = 0, 1, \cdots$ となる。すなわち，a prior 分布が ξ^* のとき optimal policy は常に action 1 のみを選ぶことになる。ところが，たとえば $\theta^2_{21} = \frac{8}{9}$ なる $\theta \in \Theta$ が実際の値のときは，$\pi^*(\xi^*, H_n)$ は決して optimal にならない。

Corollary

$P_{\theta,\pi^*_\xi}(W) = 1$ for $\xi \in \Xi$, $\theta \in \Theta^+$, $x \in S$ なる optimal policy π^* が存在するならば，π^* はまた learning policy でもある。

4. Non-Bayesian approaches to MDP's with uncertain transition matrices.

この節では未知の推移確率 (parameter) に対する事前分布が存在しなく，かつ無限期間での一期あたりの平均期待利得を最大にする問題をとり扱い，最尤法をもちいて optimal な learning policy を構成する。事前分布が存在する場合は Rose, J.S. [9] によって研究された。$t = 0, 1, 2, \cdots$ に対して $\gamma_t(\cdot / H_t(w))$ は A の上の確率分布であるとき $\gamma = (\gamma_0, \gamma_1, \gamma_2, \cdots)$ をこの節では policy と呼ぶ。policy の全体を Γ で表わす。さらに，θ が真のとき初期値 $x \in S$ で policy γ をつかったときの $(\Omega, \mathcal{F})$ の上の確率分布を $P^x_{\theta,\gamma}(\cdot)$ で表わす。

$x \in S$, $\gamma \in \Gamma$, $\theta \in \Theta$ に対して T 期までの総期待利得を

$\psi_T[\gamma](x,\theta) \equiv E^x_{\theta,\gamma}\left[\sum_0^{T-1} r(X_t,\Delta_t)\right]$ として，無限期間における平均期待利得と最適平均期待利得を次のように定義する：

$$\psi[\gamma](x,\theta) \equiv \lim_{T\to\infty} \frac{1}{T}\psi_T[\gamma](x,\theta)$$

$$\psi(x,\theta) = \sup_{\gamma\in\Gamma} \psi[\gamma](x,\theta)$$

定義3

γ^* ; optimal in the average case

$\Leftrightarrow$ $\psi(x,\theta) = \psi[\gamma^*](x,\theta)$ for $x\in S$, $\theta\in\Theta^+$

$\gamma = (\gamma_0,\gamma_1,\cdots)$; Markov

$\Leftrightarrow$ $\gamma_t(\cdot\,/\,X_0,\cdots,X_t) = \gamma_t(\cdot\,/\,X_t)$ $\quad t=0,1,2,\cdots$

$\gamma = (\gamma_0,\gamma_1,\cdots)$; stationary

$\Leftrightarrow$ $\gamma_t(\cdot\,/\,X_0,\Delta_0,\cdots,X_t) = \gamma_0(\cdot\,/\,X_t)$ $\quad t=0,1,2,\cdots$

γ が stationary のとき，γ を $\gamma_0^{(\infty)}$ と記す。また，deterministic stationary policy を $f^\infty = (f,f,\cdots)$ で表わす。

Lemma 1 ([2])

$\theta\in\Theta^+$ が既知とする。そのとき，optimal deterministic policy $f_\theta^{(\infty)}$ が存在する。また，$\psi[f_\theta^{(\infty)}](x,\theta)$ は初期値 x に無関係である。

そこで，$\theta\in\Theta^+$ に対して，$g_\theta \equiv \psi[f_\theta^{(\infty)}](x,\theta)$ とする。任意の stationary policy $\gamma^{(\infty)} = (\gamma,\gamma,\cdots)$ と $\forall\theta\in\Theta$ に対して $L\times L$ matrix $\theta^\gamma = (\theta^\gamma_{ss'})$ と limiting matrix ${}^*\theta^\gamma$ を次のように定義する。

$$\theta^{r}_{ss'} = \sum_{a\in A} \theta^{a}_{ss'} r(a/s) \qquad {}^{*}\theta^{r} \equiv \lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n} \theta^{r(t)}$$

そのとき次の lemma が成り立つことが知られている。

Lemma 2 ([I], [II])

$$\theta \in \Theta^{+},\ r^{(\infty)} = (r, r, \cdots), \quad H^{r} \equiv [I - \theta^{r} + {}^{*}\theta^{r}]^{-1} - {}^{*}\theta^{r}$$

$$g^{r} = (g^{r}_{1}, \cdots, g^{r}_{L})' = H^{r} l(r)$$

$\Rightarrow$

(i) $\psi[r^{(\infty)}](s,\theta) = \sum_{s'\in S} {}^{*}\theta^{r}_{ss'} l_{s'}(r)$ and $\sum_{s'\in S} {}^{*}\theta^{r}_{ss'} g^{r}_{s'} = 0$

(ii) $l_{s}(r) + \sum_{s'\in S} \theta^{r}_{ss'} g^{r} = \psi[r^{(\infty)}](s,\theta) + g^{r}_{s}$

(iii) $\psi_{T}[r^{(\infty)}](s,\theta) = T\psi[r^{(\infty)}](s,\theta) + g^{r}_{s} - \sum_{s'\in S} \theta^{r(T)}_{ss'} g^{r}_{s'}$

但し, L-vector $l(r) = (l_{1}(r), \cdots, l_{L}(r))'$ は

$l_{s}(r) = \sum_{a\in A} l(s,a) r(a/s)$ $s \in S$ で定義される。

次に ${}^{\forall}\varepsilon > 0$ $\left(0 < \varepsilon < \frac{1}{L}\right)$, ${}^{\forall}\theta \in \Theta^{+}$ に対して, stationary policy ${}^{\varepsilon}r^{(\infty)}_{\theta}$ を次のように定義する。

$${}^{\varepsilon}r_{\theta}(a/s) = \begin{cases} 1 - (L-1)\varepsilon & \text{if } a = f_{\theta}(s) \\ \varepsilon & \text{otherwise} \end{cases}$$

そのとき, 次の lemma を得る。

Lemma 3

$$\theta(n) \in \Theta^{+} \to \theta \in \Theta^{+} \text{ as } n\to\infty$$

$$\Rightarrow \psi[{}^{\varepsilon}r^{(\infty)}_{\theta(n)}](x,\theta) \to g_{\theta} \text{ as } n\to\infty \text{ and } \varepsilon\to 0$$

<略証>

記号を簡単にするために $f_{n} = f_{\theta(n)}$, ${}^{\varepsilon}r_{n} = {}^{\varepsilon}r_{\theta(n)}$ とおく。

Denard & Miller [3] より，各 n $(n=1,2,\cdots)$ に対して，次の(1),(2)を満足する L－vector $y^n=(y_1^n,\cdots,y_L^n)$ が存在する。

(1) $r(s,f_n(s))+\sum_{s'\in S}\theta_{ss'}^{f_n(s)}(n)\,y_{s'}^n = g_{\theta(n)}+y_s^n$

(2) $r(s,a)+\sum_{s'\in S}\theta_{ss'}^{a}(n)\,y_{s'}^n \leqq g_{\theta(n)}+y_s^n \quad \text{for } {}^\forall a\in A$

そこで，$\gamma_{ss'}^n(f_\theta)=\theta_{ss'}^{f_\theta(s)}(n)-\theta_{ss'}^{f_\theta(s)}$ と定義すれば，(2)から次を得る。

(3) $r(s,f_\theta(s))+\sum_{s'\in S}\theta_{ss'}^{f_\theta(s)}y_{s'}^n+\sum_{s'\in S}\gamma_{ss'}^n(f_\theta)y_{s'}^n \leqq g_{\theta(n)}+y_s^n$

$\{{}^*\theta_{xs}^{f_\theta(x)}\}$ を $\{\theta_{xs}^{f_\theta(x)}\}$ の limiting matrix とすると，lemma 2 の(i)と(3)より次を得る。

(4) $g_\theta \leqq g_{\theta(n)}-\sum_{s,s'}{}^*\theta_{xs}^{f_\theta(x)}\gamma_{ss'}^n(f_\theta)y_{s'}^n$

一方，${}^\varepsilon r_n$ の定義から

$$\theta_{ss'}^{f_n(s)}=\theta_{ss'}^{{}^\varepsilon r_n}-\varepsilon\sum_{a\in A}(\theta_{ss'}^a-\theta_{ss'}^{f_n(s)})$$

$$r(s,f_n(s))=r_s({}^\varepsilon r_n)-\varepsilon\sum_{a\in A}(r(s,a)-r(s,f_n(s)))$$

但し，$r_s({}^\varepsilon r_n)=\sum_{a\in A}r(s,a)\,{}^\varepsilon r_n(a)$

これを使って計算すると(1)から次を得る。

(5) $\psi[{}^\varepsilon r_n^{(\infty)}](x,\theta)=g_{\theta(n)}+\varepsilon\sum_{s\in S}{}^*\theta_{xs}^{{}^\varepsilon r_n}Z_s^n$

但し，$Z_s^n=\sum_{s',a}(\theta_{ss'}^a-\theta_{ss'}^{f_n(s)})y_{s'}^n+\sum_a(r(s,a)-r(s,f_n(s)))$
$+\sum_{s'}\gamma_{ss'}^n(f_n)y_s^n$

(4)と(5)より次を得る。

$$g_\theta-\psi[{}^\varepsilon r_n^{(\infty)}](x,\theta)\leqq l_n\|y^n\|+\varepsilon(2R+2\|y^n\|+l_n\|y^n\|)$$

但し，$l_n=\max_{x,x'\in S,a\in A}|\theta_{xx'}^a(n)-\theta_{xx'}^a|$，$\|y^n\|=\left(\sum_{k=1}^L(y_k^n)^2\right)^{\frac12}$

$\{\gamma^m\}$ は有界な vector であるから，$n\to\infty$, $\varepsilon\to\omega$ のとき

$g_0 - \psi[{}^{\varepsilon}\gamma_n^{(\omega)}](x,\theta) \to 0$ となる。 QED.

$s, s' \in S$, $a \in A$ に対して $u_t(s,a,s')$ を次のように定義する。

$u_t(s,a,s') = 1$ if $X_t(\omega)=s$, $\Delta_t(\omega)=a$, $X_{t+1}(\omega)=s'$

$u_t(s,a,s') = 0$ otherwise

$\{N_\ell\}_0^\infty$ を $N_0=0$ なる正の整数の数列とする。

$s, s' \in S$, $a\in A$, $k=1,2,\cdots$ に対して，$\hat{\theta}^a_{ss'}(k)$ を次のように定義する。

$\hat{\theta}^a_{ss'}(k) = \sum_1^{L_k} u_\ell(s,a,s') \Big/ \sum_1^{L_k} u_\ell(s,a)$ 但し，$L_k = \sum_0^k N_\ell$。

matrix $\hat{\theta}(k) = \{\hat{\theta}^a_{ss'}(k)\}$ は L_k 期でのθの最尤推定量である。

$\{\lambda_\ell\}_0^\infty$ を $0<\lambda_\ell<1$, $\lambda_0=1$ なる正の実数列として，$\forall \hat{\theta}_0 \in \textcircled{H}^+$ に対して

${}^*\hat{\theta}_k \equiv \lambda_k \hat{\theta}_0 + (1-\lambda_k)\hat{\theta}(k)$, $k=1,2,\cdots$ とする。

次に，$\{\varepsilon_\ell\}_0^\infty$ を $0<\varepsilon_\ell<\frac{1}{L}$ なる正の実数列として

policy $\gamma^* = (\gamma_0^*, \gamma_1^*, \cdots)$ を次のように定義する：

$\gamma_t^*(\cdot / H_t(\omega)) = {}^{\varepsilon_k}\gamma_{*\hat{\theta}_k}(\cdot / X_t(\omega))$ if $L_k \le t < L_k + N_k = (L_{k+1})$

そのとき，policy γ^* はある条件のもとで optimal & learning policy であることが次の定理で示される。

<u>定理3</u>

(i) $\sum_1^\infty \varepsilon_\ell N_\ell = +\infty$

(ii) $N_\ell \uparrow +\infty$, $\lambda_\ell \downarrow 0$, $\varepsilon_\ell \downarrow 0$

$\Rightarrow$ $\forall \hat{\theta}_0 \in \Theta^+$ に対して γ^* は optimal である。

さらに, $\psi[\gamma^*](x,\theta) = g_\theta$ for all $\theta \in \Theta^+$

〈略証〉

§3 の lemma 1 と大数の強法則から, $k \to \infty$ のとき

${}^*\hat{\theta}_k \to \theta$ with P^x_{θ,γ^*} − probability 1 である。今, 簡単のために $\gamma_k \equiv {}^{\varepsilon_k}\gamma_{{}^*\hat{\theta}_k}$ とする。そのとき

$\eta({}^*\hat{\theta}_k) \equiv \psi[\gamma_k^{(\infty)}](x,\theta) - g_\theta$ とすると, lemma 3 より,

$k \to \infty$ のとき $\eta({}^*\hat{\theta}_k) \to 0$ with P_{θ,γ^*} − probability 1。

lemma 2 から $k = 0, 1, \cdots$ に対して L−vector $y^k = (y^k_1, \cdots, y^k_L)$

が存在して, 次式が成り立つ。

$$\sum_{t=L_k}^{L_{k+1}-1} E^{x'}_{\theta,\gamma_k^{(\infty)}}[r(X_t,\Delta_t) / X_{L_k} = x] = [g_\theta - \eta({}^*\hat{\theta}_k)]N_{k+1} + y^k_{x'} - \sum_{x \in S} \theta^{\gamma_k}_{x x'} y_x$$

ゆえに次を得る。

$$\frac{1}{L_{k+1}} \sum_{t=0}^{L_{k+1}-1} E^{x'}_{\theta,\gamma^*}[r(X_t,\Delta_t)] \geq g_\theta - \frac{1}{L_{k+1}} \sum_{\ell=0}^{k} E^{x}_{\theta,\gamma^*}[\eta({}^*\hat{\theta}_\ell)]N_{\ell+1} - 2\frac{1}{L_{k+1}} \sum_{\ell=0}^{k} \|y^\ell\|.$$

従って, bounded convergence theorem, Toeplity Lemma [6] 等をつかって,

$$\psi[\gamma^*](x,\theta) = \lim_{k\to\infty} \frac{1}{L_{k+1}} \sum_{t=0}^{L_{k+1}-1} E^{x}_{\theta,\gamma^*}[r(X_t,\Delta_t)] \geq g_\theta \quad \text{for all } \theta \in \Theta^+$$

一方, $\psi[\gamma](x,\theta) \leq g_\theta$ for $\gamma \in \Gamma$ より

$\Psi[\tau^*](x,\theta) = g_\theta$ for all $\theta \in \Theta^+$ が成り立つ。

Q E D .

REFERENCE

[1] Blackwell, D. (1962)

Discrete Dynamic programming.

Ann. Math. Statist. 33, 719-726.

[2] Derman, Cyrus (1962)

On sequential decision and Markov chains.

Magi. Sci. 9, 16-24.

[3] Denardo, E. V. and Miller, B. L. (1968)

An optimality condition for discrete dynamic programming with no discounting.

Ann. Math. Statist. 39, 1220-1227.

[4] Howard, Ronald A. (1960)

Dynamic programming and Markov processes.

the M. I. T. press.

[5] M. Kurano (1972)

Discrete-time Markovian decision processes with an unknown parameter, -average return criterion-

Journal of Op. Reser. of Japan. 15 no. 2 67–76.

[6] Loeve, M. (1960)

Probability Theory, Second Edition.

D. Van Nostrand Co., Inc., New Jersey.

[7] Martin, J. J. (1967)

Bayesian Decision problem and Markov chains.

Wiley, New York.

[8] Rose, J. S. (1971)

Markovian Decision Processes under uncertainty.

PH. D. dissertation, Northwestern U.

[9] Rose, J. S. (1975)

Markov Decision Processes under uncertainty – average return criterion. Unpublished report.

[10] C. W. Sweat (1968)

Adaptive competitive decision in repeated play of a matrix game with uncertain entries.

Naval R. L. Q. 15, 425–448.

[11] Veinott, A. F. (1966)

On finding optimal policies in discrete dynamic programming with no discounting.

Ann. Math. Stat., 37, 1284–1294.