数理解析研究所講究録
第460巻 1982年 198-213

状態観測の不完全なセミマルコフ決定過程について

長岡工業高専　涌田和芳

§1. 序

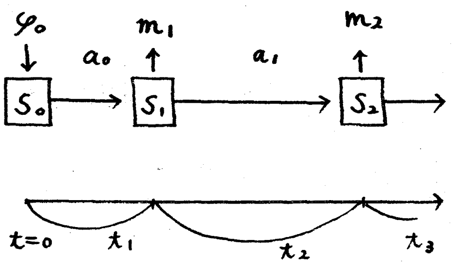
ここでは、状態観測の不完全なセミマルコフ決定過程について述べる(これをSMDP-IIと表わし、通常のセミマルコフ決定過程はSMDP-Iと表わす)。簡単にこの問題をふりかえってみよう。状態観測の不完全なマルコフ決定過程(これをMDP-IIと表わす)は多くの人々により研究された。Dynkin [4] と Shiryaev [9] は数理統計学より生じた部分的観測ランダム列の制御を研究した。Åström [1]は、有限段階のMDP-IIを扱った。また、Bellman [2]でも確率過程の部分的観測について言及している。Sawaragi and Yoshikawa [8]は、これらの研究と Blackwell [3], Strauch [11]の研究との関連を明確にした。Rhenius [6]は、非定常で一般の状態空間をもつ場合を扱った。Sondik [10], Sawaki and Ichikawa [7]は、ε-最適政策を求めるアルゴリズムを、それぞれ、政策改良法と逐次近似法より求めた。Kurano [5]は、平均基準を扱った。一方、MDP-IIでは、各状態での滞在時間はつねに1単位時間であるが、この制約を除いて C.C. White [13]はSMDP-IIについて研究した。そこでは、有限段階で離散時間(整数値上でのみ推移が起る)の場合が扱われた。ここでは、無限段階で(

離散時間を含む）より一般的な推移の場合について述べる．

§2．定義と準備

X，Y はボレル集合とする（ボレル集合とは完備可分距離空間のボレル部分集合をいう）．このとき，X 上のすべての確率測度を P(X) で表わし，X が与えられたときの Y の条件付確率測度のすべてを Q(Y|X) で表わす．また，X 上のすべての有界ボレル可測関数を M(X) で表わす．

SMDP−II は $(S, M, A, p_s, p_t, q, \varphi_0, c, \alpha)$ により決定される．S，M は可算集合で，それぞれ，システムの状態集合，観測信号の集合．A はボレル集合で，行動集合．$p_s \in Q(S|SA)$ は，状態の推移確率．$p_t \in Q(R_+|SAS)$ （$R_+ \equiv [0,+\infty)$）は，状態の滞在時間分布．p_t は，ある σ 有限測度 λ に関して，密度関数 $f(t|s,a,s')$ をもち，f は (t,s,a,s') に関してボレル可測とする．$q \in Q(M|S)$ は観測信号を伝える確率．$\varphi_0 \in \Phi$ （$\Phi \equiv P(S)$）はシステムの初期分布．$c \in M(R_+SA)$ はコスト関数．α （α > 0）は割引因子．



第1区間でのコスト $\int_0^{t_1} c(t, s_0, a_0) e^{-\alpha t} dt$

第2区間でのコスト $e^{-\alpha t_1} \int_0^{t_2} c(t, s_1, a_1) e^{-\alpha t} dt$

・・・

次の条件を全体を通して仮定する．

条件1．

$$\sum_{s'} p_t([0,\delta] \mid s, a, s') \, p_s(s' \mid s, a) \leqq 1 - \varepsilon \quad (\forall s, a)$$

なる $\delta > 0$，$\varepsilon > 0$ が存在する．

政策は $\omega = \{\omega_0, \omega_1, \cdots\}$ と表わされる．ここで，$\omega_n \in Q(A \mid H_n)$，$H_n \equiv \Phi(AR_+M)^n$ $(\forall n)$ である．政策 ω を用いたときの期待合計割引コストは，

$$J_\omega^\alpha(\varphi_0) \equiv E_\omega\left[\sum_{n=0}^{\infty} e^{-\alpha(t_1 + \cdots + t_n)} \int_0^{t_{n+1}} c(t, s_n, a_n) e^{-\alpha t} dt \,\middle|\, \varphi_0 \right]$$

で与えられる．ここで，$E_\omega[\cdot \mid \varphi_0]$ は

$$p_\omega\{\cdot \mid \varphi_0\} \equiv q^p \otimes \bigotimes_{n=0}^{\infty} (\omega_n \otimes p_s \otimes p_t \otimes q)$$

ただし，$q^p(s_0 \mid \varphi_0) \equiv \varphi_0(s_0)$

による条件付期待値である．我々の目的は，すべての政策の中で，$J_\omega^\alpha(\varphi_0)$ を最小にすることである．

$$J_{\omega^*}^\alpha(\varphi_0) \leqq \inf_\omega J_\omega^\alpha(\varphi_0) \quad (\forall \varphi_0)$$

ならば，ω^* は α－最適であるという．

システムの履歴 $h_n \equiv (\varphi_0, a_0, t_1, m_1, \cdots, m_n)$ が観測されたときの状態 s_n の条件付確率を $q_n(\cdot \mid h_n)$ で表わすと，ベイズの公式より

$$g_{n+1}(s_{n+1} \mid h_{n+1})$$

$$= g_{n+1}(s_{n+1} \mid h_n, a_n, t_{n+1}, m_{n+1})$$

$$(3.1) = \frac{\sum_{s_n} f(t_{n+1} \mid s_n, a_n, s_{n+1})\, g(m_{n+1} \mid s_{n+1})\, p_s(s_{n+1} \mid s_n, a_n)\, g_n(s_n \mid h_n)}{\sum_{s_n}\sum_{s_{n+1}} [\qquad \text{〃} \qquad]}$$

ここで, $\varphi_n(s_n) \equiv g_n(s_n \mid h_n)$ $(\forall s_n)$ として, (3.1)をおきかえると $\varphi_{n+1}(s_{n+1}) = u(\varphi_n, a_n, t_{n+1}, m_{n+1})(s_{n+1})$ $(\forall s_{n+1})$ なるボレル可測写像 $u: \Phi A R_+ M \longrightarrow \Phi$ が存在する. この u を繰り返し適用すれば, $h_n \equiv (\varphi_0, a_0, t_1, m_1, \cdots, m_n)$ に対して $b_n \equiv (\varphi_0, a_0, t_1, \varphi_1, \cdots, \varphi_n)$ が定まる. この b_n にもとずく政策をI-政策という. すなわち, I-政策は, $\pi = \{\pi_0, \pi_1, \cdots\}$, $\pi_n \in Q(A \mid B_n)$, $B_n \equiv \Phi(A R_+ \Phi)^n$ $(\forall n)$ と表わされる. このI-政策 π に対する期待合計割引コストは

$$I_\pi^\alpha(\varphi_0) \equiv \bar{E}_\pi\left[\sum_{n=0}^{\infty} e^{-\alpha(t_1+\cdots+t_n)} \int_0^{t_{n+1}} c(t, s_n, a_n)\, e^{-\alpha t} dt \,\Big|\, \varphi_0\right]$$

で与えられる. ここで, $\bar{E}_\pi[\cdot \mid \varphi_0]$ は

$$\bar{P}_\pi\{\cdot \mid \varphi_0\} \equiv g^p \otimes \bigotimes_{n=0}^{\infty} (\pi_n \otimes p_s \otimes p_t \otimes g \otimes u)$$

による条件付期待値である. 政策 ω に対しても同様に, $\bar{P}_\omega$, $\bar{E}_\omega$ を定義する. そうすると, J_ω^α は E_ω のかわりに $\bar{E}_\omega$ でおきかえて表わすことができる.

§3. 主な結果

<u>命題1.</u>

$$J_\omega^\alpha(\varphi_0) = \bar{E}_\omega\left[\sum_{n=0}^{\infty} e^{-\alpha(t_1+\cdots+t_n)}\, \bar{c}_\alpha(\varphi_n, a_n) \,\Big|\, \varphi_0\right]$$

が成りたつ．ただし，

$$C_\alpha(t,s,a) \fallingdotseq \int_0^t c(t,s,a)e^{-\alpha t}dt,$$

$$\bar{C}_\alpha(\varphi,a) \fallingdotseq \sum_s \sum_{s'} \int_0^\infty C_\alpha(t,s,a)\,dp_t(t|s,a,s')\,p_s(s'|s,a)\varphi(s),$$

$$\varphi \in \Phi.$$

証明

$$J_\omega^\alpha(\varphi_0) = \bar{E}_\omega\Big[\sum_{n=0}^\infty e^{-\alpha(t_1+\cdots+t_n)}\int_0^{t_{n+1}} c(t,s_n,a_n)e^{-\alpha t}dt\,\Big|\,\varphi_0\Big]$$

$$= \sum_{n=0}^\infty \bar{E}_\omega\Big[e^{-\alpha(t_1+\cdots+t_n)}\bar{E}_\omega[C_\alpha(t_{n+1},s_n,a_n)\,|\,h_n']\,\Big|\,\varphi_0\Big]$$

ここで，$h_n' \fallingdotseq (\varphi_0, a_0, t_1, m_1, \varphi_1, \cdots, \varphi_n, a_n)$．一方，

$$\bar{E}_\omega[C_\alpha(t_{n+1},s_n,a_n)\,|\,h_n']$$

$$= \sum_{s_n}\sum_{s_{n+1}}\int_0^\infty C_\alpha(t,s_n,a_n)\,dp_t(t|s_n,a_n,s_{n+1})\,p_s(s_{n+1}|s_n,a_n)\varphi_n(s_n)$$

$$= \bar{C}_\alpha(\varphi_n,a_n).$$

ゆえに，命題が成りたつ．

注意 1．この命題は I-政策 π に対しても成りたつ．

定理 3.1．任意に行動の列 $\{a_0, a_1, \cdots\}$ を固定し，a_n は n 番目の推移区間で選択される行動とする．このとき，過程 $\{\varphi_n, t_n; n\in N\}$ は次の性質をもつ．

(i) 過程が φ_n にあるとき，次の推移が φ_{n+1} に起る確率 $\bar{g}$ は φ_n と a_n にだけ依存し，次式で与えられる．

$$\bar{g}(\Gamma|\varphi_n,a_n) = \sum_{s_n}\sum_{s_{n+1}}\sum_{m_{n+1}}\int_{\bar{\Gamma}_m} f(t_{n+1}|s_n,a_n,s_{n+1})\,d\lambda(t_{n+1})$$

$$\times\, g(m_{n+1}|s_{n+1})\,p_s(s_{n+1}|s_n,a_n)\varphi_n(s_n),$$

ただし，Φの任意のボレル部分集合 Γ に対して

$$\overline{\Gamma} = \overline{\Gamma}(\varphi_n, a_n; \Gamma) \equiv \{ (t_{n+1}, m_{n+1}) ;\ u(\varphi_n, a_n, t_{n+1}, m_{n+1}) \in \Gamma \}$$

$$\overline{\Gamma}_m = \overline{\Gamma}_m(\varphi_n, a_n, m_{n+1}; \Gamma) \equiv \{ t_{n+1} ;\ u(\varphi_n, a_n, t_{n+1}, m_{n+1} \in \Gamma \}.$$

(ii) 過程の次の状態が φ_{n+1} であるという条件の下での φ_n での滞在時間分布は，$\varphi_n, a_n, \varphi_{n+1}$ にだけ依存し，次式を満す．

$$\int_{\Phi} \overline{p}(B \mid \varphi_n, a_n, \varphi_{n+1})\, d\,\overline{q}(\varphi_{n+1} \mid \varphi_n, a_n)$$
$$= \sum_{s_n} \sum_{s_{n+1}} p_t(B \mid s_n, a_n, s_{n+1})\, p_s(s_{n+1} \mid s_n, a_n)\, \varphi_n(s_n),$$

ただし，B は R_+ の任意のボレル部分集合．

$\{\varphi_n, t_n ;\ n \in N\}$ は，$\overline{q}$ と $\overline{p}$ により定まるマルコフ再生過程である．

証明

$$p\{ \varphi_{n+1} \in \Gamma,\ t_{n+1} \in B \mid \varphi_0, a_0, t_1, \varphi_1, \cdots, \varphi_n, a_n \}$$
$$= p\{ (t_{n+1}, m_{n+1}) \in \overline{\Gamma},\ t_{n+1} \in B \mid \quad '' \quad \}$$
$$= \sum_{s_n} \sum_{s_{n+1}} p\{ (t_{n+1}, m_{n+1}) \in \overline{\Gamma},\ t_{n+1} \in B \mid s_n, s_{n+1}, '' \}$$
$$\times\ p\{ s_{n+1} \mid s_n, '' \}\, \varphi_n(s_n)$$
$$= \sum_{s_n} \sum_{s_{n+1}} \sum_{m_{n+1}} \int_{\overline{\Gamma}_m \cap B} f(t_{n+1} \mid s_n, a_n, s_{n+1})\, d\lambda(t_{n+1})$$
$$\times\ g(m_{n+1} \mid s_{n+1})\, p_s(s_{n+1} \mid s_n, a_n)\, \varphi_n(s_n).$$

ここで，$B = R_+$ とおけば (i) が得られる．また，

$$p\{ \varphi_{n+1} \in \Gamma,\ t_{n+1} \in B \mid \varphi_n, a_n \}$$
$$= \int_{\Gamma} p\{ t_{n+1} \in B \mid \varphi_n, a_n, \varphi_{n+1} \}\, d\,p\{ \varphi_{n+1} \mid \varphi_n, a_n \}$$

が成りたつ．そこで，

$$\bar{p}(B \mid \varphi_n, a_n, \varphi_{n+1}) \equiv p\{t_{n+1} \in B \mid \varphi_n, a_n, \varphi_{n+1}\}$$

とおくと，この $\bar{p} \in Q(R_+ \mid \Phi A \Phi)$ が次の状態が φ'_{n+1} であるという条件の下での φ_n での滞在時間分布を表わす．そして，(i)の証明より (ii)の式が成りたつ．更に，

$$p\{\{\varphi_{n+1}, t_{n+1}\} \mid \varphi_0, a_0, t_1, \varphi_1, \cdots, \varphi_n, a_n\}$$

$$= p\{\{\varphi_{n+1}, t_{n+1}\} \mid \varphi_n, a_n\}$$

$$= \bar{q} \otimes \bar{p}(\{\varphi_{n+1}, t_{n+1}\} \mid \varphi_n, a_n)$$

が成りたつので，$\{\varphi_n, t_n ; n \in N\}$ は $\bar{q}$ と $\bar{p}$ により定まるマルコフ再生過程である．

<u>定理 3.2.</u> 任意の政策 ω に対して

$$J^{\alpha}_{\pi}(\varphi_0) = J^{\alpha}_{\omega}(\varphi_0) \quad (\forall \varphi_0)$$

なる I－政策 π が存在する．

<u>証明</u> 任意に与えられた政策 ω に対して I－政策 $\pi^{\omega} = \{\pi^{\omega}_0, \pi^{\omega}_1, \cdots\}$ を次のように定義する．

$$\pi^{\omega}_0(\{a_0\} \mid \varphi_0) \equiv \omega_0(\{a_0\} \mid \varphi_0)$$

$$\pi^{\omega}_n(\{a_n\} \mid \varphi_0, a_0, t_1, \varphi_1, \cdots, \varphi_n)$$

$$\equiv \sum_{m_1, \cdots, m_n} \omega_n(\{a_n\} \mid \varphi_0, a_0, t_1, m_1, \cdots, m_n)$$

$$\times \bar{P}_{\omega}\{m_1, \cdots, m_n \mid \varphi_0, a_0, t_1, \varphi_1, \cdots, \varphi_n\},$$

ここで，$b^{h}_{n} \equiv (\varphi_0, a_0, t_1, \varphi_1, \cdots, \varphi_n)$ は $h_n \equiv (\varphi_0, a_0, t_1, m_1, \cdots, m_n)$ に対応する．この π^{ω} の作り方から，ω と π^{ω} は A上に同一の条件付確率

$$\bar{P}_{\pi^\omega}\{\{a_n\} \mid \varphi_0, a_0, t_1, \varphi_1, \cdots, \varphi_n\} = \bar{P}_\omega\{\{a_n\} \mid \quad " \quad\}$$

を与える．一方，定理3.1.より，

$$\bar{P}_{\pi^\omega}\{\{\varphi_{n+1}, t_{n+1}\} \mid \varphi_0, a_0, t_1, \varphi_1, \cdots, \varphi_n, a_n\}$$
$$= \bar{P}_\omega\{\{\quad " \quad\} \mid \quad " \quad\}$$
$$= \bar{q} \otimes \bar{P}(\{\varphi_{n+1}, t_{n+1}\} \mid \varphi_n, a_n).$$

そこで，

$$\bar{P}_{\pi^\omega}\{\{a_0, t_1, \varphi_1, \cdots, \varphi_n, a_n\} \mid \varphi_0\} = \bar{P}_\omega\{\{\quad " \quad\} \mid \varphi_0\}$$

が成りたつ．命題1とその注意より，π^ωが求めるものである．

次に，新しいモデルSMDP-I(Φ, A, $\bar{q}$, $\bar{P}$, $\bar{C}_\alpha$; α)を考える．このSMDP-Iの政策は，I-政策と同じものであり，I-政策πによる期待合計割引コストは

$$I_\pi^\alpha(\varphi_0) \equiv E_\pi\left[\sum_{n=0}^\infty e^{-\alpha(t_1+\cdots+t_n)} \bar{C}_\alpha(\varphi_n, a_n) \mid \varphi_0\right]$$

で与えられる．ここで，$E_\pi[\cdot \mid \varphi_0]$は，

$$P_\pi\{\cdot \mid \varphi_0\} \equiv \bigotimes_{n=0}^\infty (\pi_n \otimes \bar{q} \otimes \bar{P})$$

による条件付期待値である．

<u>定理3.3.</u> 任意のI-政策πに対して，

$$J_\pi^\alpha(\varphi_0) = I_\pi^\alpha(\varphi_0) \quad ({}^\forall\varphi_0)$$

が成りたつ．

<u>証明</u> 定理3.1.より，$(A\Phi R_+)^n A$上の条件付確率$\bar{P}_\pi\{\cdot \mid \varphi_0\}$は，次のように分解される．

$$\bar{P}_\pi\{\cdot\,|\,\varphi_0\} = \bar{P}_\pi\{\{a_0\}|\varphi_0\} \otimes \bar{P}_\pi\{\{\varphi_1, t_1\}\,|\,\varphi_0, a_0\}$$

$$\otimes \cdots \otimes \bar{P}_\pi\{\{a_n\}\,|\,\varphi_0, a_0, t_1, \varphi_1, \cdots, \varphi_n\}$$

$$= \pi_0 \otimes \bar{g} \otimes \bar{P} \otimes \cdots \otimes \pi_n$$

$$= p_\pi\{\cdot\,|\,\varphi_0\}.$$

したがって，注意1と I_π^α の定義より

$$J_\pi^\alpha(\varphi_0) = I_\pi^\alpha(\varphi_0) \quad (\forall \varphi_0)$$

が成りたつ．

以上の2つの定理より，SMDP-II（$S, M, A, p_s, p_t, g, \varphi_0, C, \alpha$）は，SMDP-I（$\Phi, A, \bar{g}, \bar{P}, \bar{C}_\alpha, \alpha$）に同値変形されることがわかる．次の命題は，条件1と定理3.1（ii）より明らかである．

<u>命題2.</u>

$$\int_\Phi \bar{P}([0,\delta]\,|\,\varphi, a, \varphi')\,d\bar{g}(\varphi'|\varphi, a) \leqq 1-\varepsilon \quad (\forall \varphi, a)$$

なる $\delta>0$ と $\varepsilon>0$ が存在する．

$$I^\alpha(\varphi) \equiv \inf_\pi I_\pi^\alpha(\varphi) \quad (\forall \varphi)$$

とおく．

<u>定理3.4.</u> A は可算集合とする．このとき，$I^\alpha(\varphi)$ は，Φ 上のボレル可測関数で次の方程式の一意の解である．

$$(3.2)\qquad I^\alpha(\varphi) = \inf_{a\in A}\Big\{\bar{C}_\alpha(\varphi, a) + \int_\Phi\int_0^\infty e^{-\alpha t} I^\alpha(\varphi') \times d\bar{P}(t\,|\,\varphi, a, \varphi')\,d\bar{g}(\varphi'|\varphi, a)\Big\} \quad (\forall \varphi).$$

この方程式は次のように表わすことができる．

(3.3) $$I^{\alpha}(\varphi)=\inf_{a\in A}\{\bar{c}_{\alpha}(\varphi,a)+\sum_{s}\sum_{s'}\sum_{m'}\int_0^{\infty}e^{-\alpha t}\times I^{\alpha}(u(\varphi,a,t,m'))\,dp_t(t|s,a,s')\,p_s(s'|s,a)\,q(m'|s')\varphi(s)\}$$

(3.2) or (3.3) で, もし各φについて右辺を最小にする行動が存在すれば, そのように定める定常I-政策 f_{α} は α-最適である.

<u>証明</u> MDP-I での Strauch [11] の結果を SMDP-I に適用する. その結果, $I^{\alpha}(\varphi)$は絶対可測であることと (p,ε)-最適I-政策が存在することがわかる. そこで, まず, $I^{\alpha}(\varphi)$が(3.2)の解であることを示す.

$$T_a u(\varphi)\equiv\bar{c}_{\alpha}(\varphi,a)+\int_{\Phi}\int_0^{\infty}e^{-\alpha t}u(\varphi')\times d\bar{p}(t|\varphi,a,\varphi')\,d\bar{q}(\varphi'|\varphi,a)$$

とおく. 命題2より

$$\int_{\Phi}\int_0^{\infty}e^{-\alpha t}\,d\bar{p}(t|\varphi,a,\varphi')\,d\bar{q}(\varphi'|\varphi,a)\leqq 1-\varepsilon+\varepsilon e^{-\alpha\delta}<1 \quad (\forall\varphi,a)$$

が成りたつことに注意する.

$$I_{\pi}^{\alpha}(\varphi_0)=E_{\pi}\left[\sum_{n=0}^{\infty}e^{-\alpha(t_1+\cdots+t_n)}\bar{c}_{\alpha}(\varphi_n,a_n)\,\middle|\,\varphi_0\right]$$
$$\geqq\sum_{a_0\in A}p_{a_0}T_{a_0}I^{\alpha}(\varphi_0),$$

ここで, $p_{a_0}\equiv\pi_0(a_0|\varphi_0)$ $(\forall a_0)$. したがって,

$$I^{\alpha}(\varphi_0)\geqq\inf_{a_0\in A}T_{a_0}I^{\alpha}(\varphi_0).$$

もう一方の不等式を得るために, まず $t=0$ で a_0を選択し, 続いて $p\equiv\bar{q}(\cdot|\varphi_0,a_0)$に関する$(p,\varepsilon)$-最適I-政策 π'を

選択するI－政策をπとする．このとき，

$$I^{\alpha}_{\pi}(\varphi_0) = T_{a_0} I^{\alpha}_{\pi'}(\varphi_0)$$
$$\leqq T_{a_0}(I^{\alpha}+\varepsilon)(\varphi_0)$$
$$\leqq T_{a_0} I^{\alpha}(\varphi_0) + \varepsilon(1-\varepsilon+\varepsilon e^{-\alpha\delta}).$$

したがって，

$$I^{\alpha}(\varphi_0) \leqq \inf_{a_0 \in A} T_a I^{\alpha}(\varphi_0) + \varepsilon(1-\varepsilon+\varepsilon e^{-\alpha\delta}).$$

$\varepsilon > 0$ は任意なので，

$$I^{\alpha}(\varphi_0) \leqq \inf_{a_0 \in A} T_{a_0} I^{\alpha}(\varphi_0).$$

ゆえに，$I^{\alpha}(\varphi)$ は (3.2)の解である．

次に，$I^{\alpha}(\varphi)$ はボレル可測であることを示す．再び Strauch [11] より，(3.2) に対する他の有界な解は存在しないことがわかる．すなわち，$I^{\alpha}(\varphi)$は有界な関数の中での (3.2)の一意な解である．一方，

$$\| \inf_{a \in A} T_a u - \inf_{a \in A} T_a v \| \leqq (1-\varepsilon+\varepsilon e^{-\alpha\delta}) \| u - v \|$$

なので，$\inf_{a \in A} T_a$ は有界ボレル可測関数の中で一意の不動点をもつ．ゆえに $I^{\alpha}(\varphi)$ はボレル可測である．

次に定理の後半を証明する．定常I－政策 f に対して，

$$T_f u(\varphi) \equiv T_{f(\varphi)} u(\varphi) \quad (\forall \varphi)$$

とおく．f_α の定義より

$$I^{\alpha}(\varphi) = T_{f_\alpha} I^{\alpha}(\varphi) \quad (\forall \varphi).$$

一方，$I^{\alpha}_{f_\alpha}(\varphi)$ は，

$$I^{\alpha}_{f_\alpha}(\varphi) = T_{f_\alpha} I^{\alpha}_{f_\alpha}(\varphi) \quad (\forall \varphi)$$

の一意な解である．したがって，

$$I^{\alpha}_{f_\alpha}(\varphi) = I^{\alpha}(\varphi) \quad (\forall \varphi).$$

ゆえに，f_α は α-最適である．

最後に，(3.2)と(3.3)が同値であることを示す．g は (φ, t) に関して有界なボレル可測関数とする．このとき，定理3.1より

$$\int_{\Phi}\int_0^\infty g(\varphi', t)\, d\bar{F}(t \mid \varphi, a, \varphi')\, d\bar{q}(\varphi' \mid \varphi, a)$$
$$= \sum_{s}\sum_{s'}\sum_{m'} \int_0^\infty g(u(\varphi, a, t, m'), t)\, dp_t(t \mid s, a, s')$$
$$\times p_s(s' \mid s, a)\, q(m' \mid s')\, \varphi(s)$$

が成りたつ．すなわち，(3.2)と(3.3)は同値である．

§4. 平均基準

これまでは，割引コスト基準のSMDP-IIについて述べたが，ここでは平均コスト基準のSMDP-IIについて述べる．平均コスト基準のSMDP-IIも§2と同様に $(S, M, A, p_s, p_t, q, c)$ で定義する．割引因子 α はもはや必要なくなる．また，政策 ω を用いたときの期待平均コストは

$$J_\omega(\varphi_0) \equiv \overline{\lim_{n\to\infty}} \frac{E_\omega\left[\sum_{i=0}^{n-1} c'(t_{i+1}, s_i, a_i) \mid \varphi_0\right]}{E_\omega\left[\sum_{i=1}^{n} t_i \mid \varphi_0\right]}$$

で与えられる．ただし，

$$c'(t_{i+1}, s_i, a_i) \equiv \int_0^{t_{i+1}} c(t, s_i, a_i)\, dt$$

とし，以下 c' は有界とする．

$$\overline{J}_{\omega^*}(\varphi_0) \leqq \inf_{\omega} \overline{J}_{\omega}(\varphi_0) \quad (\forall \varphi_0)$$

が成りたつとき，ω^* は平均最適であるという．

$$c(s,a) \doteqdot \sum_{s'} \int_0^\infty c'(t,s,a)\, dp_t(t|s,a,s')\, p_s(s'|s,a)$$

$$\tau(s,a) \doteqdot \sum_{s'} \int_0^\infty t\, dp_t(t|s,a,s')\, p_s(s'|s,a)$$

とおくと，

$$\overline{J}_{\omega}(\varphi_0) = \overline{\lim_{n\to\infty}} \frac{E_\omega\left[\sum_{i=0}^{n-1} c(s_i,a_i) \,\middle|\, \varphi_0\right]}{E_\omega\left[\sum_{i=0}^{n-1} \tau(s_i,a_i) \,\middle|\, \varphi_0\right]}$$

が成りたつ．更に，

$$\overline{c}(\varphi,a) \doteqdot \sum_{s} c(s,a)\, \varphi(s)$$

$$\overline{\tau}(\varphi,a) \doteqdot \sum_{s} \tau(s,a)\, \varphi(s), \quad \varphi \in \Phi,$$

とおく．

<u>定理4.1.</u>

(a) $\overline{E}_\omega\left[c(s_n,a_n) \mid \varphi_0\right] = \overline{E}_\omega\left[\overline{c}(\varphi_n,a_n) \mid \varphi_0\right]$

(b) $\overline{E}_\omega\left[\tau(s_n,a_n) \mid \varphi_0\right] = \overline{E}_\omega\left[\overline{\tau}(\varphi_n,a_n) \mid \varphi_0\right] \quad (\forall \varphi_0)$

が成りたつ．

<u>証明</u>

$$\overline{E}_\omega\left[c(s_n,a_n) \mid \varphi_0\right]$$

$$= \int_{S\Phi A} c(s_n,a_n)\, d\overline{P}_\omega\{(s_n,\varphi_n,a_n) \mid \varphi_0\}$$

$$= \int_{\Phi A} \sum_{s_n} c(s_n,a_n)\, \overline{P}_\omega\{s_n \mid \varphi_0,\varphi_n,a_n\}\, d\overline{P}_\omega\{(\varphi_n,a_n) \mid \varphi_0\}$$

$$= \int_{\Phi A} \bar{c}(\varphi_n, a_n)\, d\bar{P}_\omega\{(\varphi_n, a_n) \mid \varphi_0\}$$
$$= \bar{E}_\omega\left[\bar{c}(\varphi_n, a_n) \mid \varphi_0\right].$$

τに対しても同様に成りたつ.

注意2. これは, I-政策πに対しても成りたつ.

定理4.2. 任意の政策ωに対して, 次の(a), (b)を満すI-政策πが存在する.

(a) $\bar{E}_\pi\left[\bar{c}(\varphi_n, a_n) \mid \varphi_0\right] = \bar{E}_\omega\left[\bar{c}(\varphi_n, a_n) \mid \varphi_0\right]$

(b) $\bar{E}_\pi\left[\bar{\tau}(\varphi_n, a_n) \mid \varphi_0\right] = \bar{E}_\omega\left[\bar{\tau}(\varphi_n, a_n) \mid \varphi_0\right] \quad (^\forall\varphi_0)(^\forall n)$

証明は定理3.2と同様にできる.

平均コスト基準のSMDP-Iも§3の後半と同様に(Φ, A, $\bar{q}$, $\bar{P}$, $\bar{c}$)で定義する. また, I-政策πを用いたときの期待平均コストは,

$$\bar{I}_\pi(\varphi_0) \equiv \overline{\lim_{n\to\infty}} \frac{E_\pi\left[\sum_{i=0}^{n-1} \bar{c}(\varphi_i, a_i) \mid \varphi_0\right]}{E_\pi\left[\sum_{i=1}^{n} t_i \mid \varphi_0\right]}$$
$$= \overline{\lim_{n\to\infty}} \frac{E_\pi\left[\sum_{i=0}^{n-1} \bar{c}(\varphi_i, a_i) \mid \varphi_0\right]}{E_\pi\left[\sum_{i=1}^{n} \bar{\tau}(\varphi_i, a_i) \mid \varphi_0\right]}$$

で与えられる. ここで第2の等式は, $\bar{\tau}$の定義と定理3.1より成りたつことに注意する.

定理4.3. 任意のI-政策πに対して

$$\bar{J}_\pi(\varphi_0) = \bar{I}_\pi(\varphi_0) \quad (^\forall\varphi_0)$$

が成りたつ.

証明は定理3.3と同様にできる.

以上の2つの定理より，平均コスト基準のSMDP-II (S, M, A, p_S, p_M, q, c) はSMDP-I ($\bar{S}$, A, $\bar{q}$, $\bar{p}$, $\bar{c}$) へ同値変形されることがわかる. なお，平均最適なI-政策が存在するための十分条件については，Wakuta [12] が議論している.

参考文献

[1] Åström, K. J.(1965). Optimal control of Markov processes with incomplete state information. J. Math.Anal.Appl. 10, 174-205.

[2] Bellman, R.(1968). New classes of stochastic processes. J. Math. Anal. Appl. 22,602-617.

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

[4] Dynkin, E. B.(1965). Controlled random sequences. Theory Probability Appl. 10, 1- 14.

[5] Kurano, M.(1977). On the existence of an optimal stationary I-policy in non-discounted Markov decision processes with incomplete state information. Bull. Math. Statist. 17, 75-81

[6] Rhenius, D.(1974). Incomplete information in Markovian decision models. Ann. Math. Statist. 2, 1327-1334.

[7] Sawaki, K. and A. Ichikawa.(1978). Optimal control for partially observable Markov decision processes over an infinite horizon. J. Oper. Res. Soc. Japan, 21,1-16.

[8] Sawaragi, Y. and T. Yoshikawa.(1970). Discrete time Markov decision processes with incomplete state observation. Ann. Math. Statist. 41,78-86.

[9] Shiryaev, A.N. Some new results in the theory of controlled random sequences.

[10] Sondik, E.(1978). The optimal control of partially observable Markov processes over the infinite horizon; Discounted costs. Oper. Res. 26, 282-304.

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

[12] Wakuta, K.(1980). Semi-Markov decision processes with incomplete state observation - Average cost criterion-. J. Oper. Res. Soc. Japan. 24, 95-108.

[13] White, C.C.(1976). Procedure for the solutions of a finite horizon, partially observed, semi-Markov optimization problem. Oper. Res. 24, 348-358.