

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

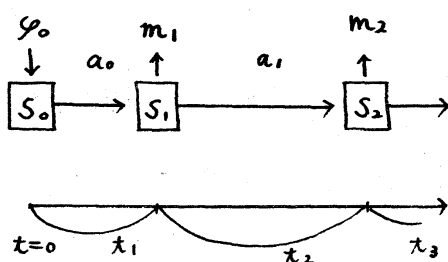
ここでは、状態観測の不完全なセミマルコフ決定過程について述べる（これも SMDP-II と表わし、通常のセミマルコフ決定過程は SMDP-I と表わす）。簡単にこの問題をふりかえってみよう。状態観測の不完全なマルコフ決定過程（これも MDP-II と表わす）は多くの人々により研究された。Dynkin [4] と Shiryaev [9] は数理統計学より生じた部分的観測ランダム列の制御を研究した。Åström [1] は、有限段階の MDP-II を扱った。また、Bellman [2] でも確率過程の部分的観測について言及している。Sawaragi and Yoshikawa [8] は、これらの研究と 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も、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.

SHDP-II は $(S, M, A, p_s, p_t, g, y_0, c, \alpha)$ 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. $g \in Q(M|S)$ は観測信号を伝える確率. $y_0 \in \Omega$ ($\Omega \equiv P(S)$) はシステムの初期分布. $c \in M(R_+SA)$ はコスト関数. α ($\alpha > 0$) は割引因子.



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

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

...

次の条件も全体を通して仮定する.

条件1.

$$\sum_{s'} \phi_{\star}([0, \delta] | s, a, s') \phi_s(s' | s, a) \leq 1 - \varepsilon \quad (\forall s, a)$$

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

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

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

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

$$\phi_{\omega}[\cdot | \varphi_0] \equiv g^P \otimes \bigotimes_{n=0}^{\infty} (\omega_n \otimes \phi_s \otimes \phi_{\star} \otimes g)$$

ただし, $g^P(s_0 | \varphi_0) \equiv \varphi_0(s_0)$

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.

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

ならば, ω^* は α -最適であるという.

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

$$\begin{aligned}
 & g_{n+1}(S_{n+1} | h_{n+1}) \\
 &= g_{n+1}(S_{n+1} | h_n, a_n, t_{n+1}, m_{n+1}) \\
 (3.1) &= \frac{\sum_{s_n} f(t_{n+1} | s_n, a_n, S_{n+1}) g(m_{n+1} | S_{n+1}) \phi_S(S_{n+1} | s_n, a_n) g_n(s_n | h_n)}{\sum_{s_n} \sum_{s_{n+1}} [\quad \quad \quad]}
 \end{aligned}$$

ここで, $g_n(s_n) \equiv g_n(s_n | h_n)$ ($\forall s_n$) とし, (3.1) をおきかえ
 ると $g_{n+1}(S_{n+1}) = u(g_n, a_n, t_{n+1}, m_{n+1})(S_{n+1})$ ($\forall S_{n+1}$) なるボレル
 可測写像 $u: \mathfrak{R}AR + M \rightarrow \mathfrak{R}$ が存在する. この u を繰り返し適
 用すれば, $h_n \equiv (y_0, a_0, t_1, m_1, \dots, m_n)$ に対して $b_n \equiv (y_0, a_0, t_1,$
 $y_1, \dots, y_n)$ が定まる. この b_n にもとづく政策を I -政策とい
 う. すなわち, I -政策は, $\pi = \{\pi_0, \pi_1, \dots\}$, $\pi_n \in Q(A | B_n)$,
 $B_n \equiv \mathfrak{R}(AR + \mathfrak{R})^n$ ($\forall n$) 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

$$I_\pi^\alpha(y_0) \equiv \bar{E}_\pi \left[\sum_{n=0}^{\infty} e^{-\alpha(t_1 + \dots + t_n)} \int_0^{t_{n+1}} c(t, s_n, a_n) e^{-\alpha t} dt \mid y_0 \right]$$

で与えられる. ここで, $\bar{E}_\pi[\cdot | y_0]$ は

$$\bar{P}_\pi(\cdot | y_0) \equiv g^p \otimes_{n=0}^{\infty} (\pi_n \otimes \phi_s \otimes \phi_t \otimes g \otimes u)$$

による条件付期待値である. 政策 ω に対して同様に, $\bar{P}_\omega$,
 $\bar{E}_\omega$ 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

命題 1.

$$J_\omega^\alpha(y_0) = \bar{E}_\omega \left[\sum_{n=0}^{\infty} e^{-\alpha(t_1 + \dots + t_n)} \bar{c}_\alpha(y_n, a_n) \mid y_0 \right]$$

が成り立つ。ただし,

$$C(\tau, s, a) \equiv \int_0^\tau c(\tau, s, a) e^{-a\tau} d\tau,$$

$$\bar{C}_\alpha(\varphi, a) \equiv \sum_s \sum_{s'} \int_0^\infty C_\alpha(\tau, s, a) d\phi_\tau(\tau | s, a, s') \phi_s(s' | s, a) \varphi(s),$$

$$\varphi \in \Xi.$$

証明

$$\begin{aligned} J_\omega^\alpha(\varphi_0) &= \bar{E}_\omega \left[\sum_{n=0}^\infty e^{-\alpha(\tau_1 + \dots + \tau_n)} \int_0^{\tau_{n+1}} c(\tau, s_n, a_n) e^{-\alpha\tau} d\tau | \varphi_0 \right] \\ &= \sum_{n=0}^\infty \bar{E}_\omega \left[e^{-\alpha(\tau_1 + \dots + \tau_n)} \bar{E}_\omega [C_\alpha(\tau_{n+1}, s_n, a_n) | h_n'] | \varphi_0 \right] \end{aligned}$$

ここで, $h_n' \equiv (\varphi_0, a_0, \tau_1, m_1, \varphi_1, \dots, \varphi_n, a_n)$. $-\bar{\sigma}$,

$$\begin{aligned} &\bar{E}_\omega [C_\alpha(\tau_{n+1}, s_n, a_n) | h_n'] \\ &= \sum_{s_n} \sum_{s_{n+1}} \int_0^\infty C_\alpha(\tau, s_n, a_n) d\phi_\tau(\tau | s_n, a_n, s_{n+1}) \phi_s(s_{n+1} | s_n, a_n) \varphi_n(s_n) \\ &= \bar{C}_\alpha(\varphi_n, a_n). \end{aligned}$$

ゆえに, 命題が成り立つ。

注意 1. この命題は π -政策 π に対しても成り立つ。

定理 3.1. 任意に行動の列 $\{a_0, a_1, \dots\}$ 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 $\{\varphi_n, \tau_n; n \in N\}$ 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 だけに依存し, 次式で与えられる。

$$\begin{aligned} \bar{g}(\tau | \varphi_n, a_n) &= \sum_{s_n} \sum_{s_{n+1}} \sum_{m_{n+1}} \int_{\bar{T}_m} f(\tau_{n+1} | s_n, a_n, s_{n+1}) d\lambda(\tau_{n+1}) \\ &\quad \times g(m_{n+1} | s_{n+1}) \phi_s(s_{n+1} | s_n, a_n) \varphi_n(s_n), \end{aligned}$$

ただし, π の任意のボレル部分集合 Γ に対して

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

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

(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 $\bar{p}$ は, $\varphi_n, a_n, \varphi_{n+1}$ にだけ依存し, 次式を満たす.

$$\begin{aligned} & \int_{\bar{\Gamma}} \bar{p}(B | \varphi_n, a_n, \varphi_{n+1}) d\bar{q}(\varphi_{n+1} | \varphi_n, a_n) \\ &= \sum_{s_n} \sum_{s_{n+1}} \phi_{\bar{p}}(B | s_n, a_n, s_{n+1}) \phi_s(s_{n+1} | s_n, a_n) \varphi_n(s_n), \end{aligned}$$

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, $\bar{q}$ と $\bar{p}$ により定まるマルコフ再生過程である.

証明

$$\begin{aligned} & P\{\varphi_{n+1} \in \Gamma, t_{n+1} \in B | \varphi_0, a_0, t_1, \varphi_1, \dots, \varphi_n, a_n\} \\ &= P\{(t_{n+1}, m_{n+1}) \in \bar{\Gamma}, t_{n+1} \in B | \quad " \quad \} \\ &= \sum_{s_n} \sum_{s_{n+1}} P\{(t_{n+1}, m_{n+1}) \in \bar{\Gamma}, t_{n+1} \in B | s_n, s_{n+1}, " \quad \} \\ & \quad \times P\{s_{n+1} | s_n, " \quad \} \varphi_n(s_n) \\ &= \sum_{s_n} \sum_{s_{n+1}} \sum_{m_{n+1}} \int_{\bar{\Gamma}_m \cap B} f(t_{n+1} | s_n, a_n, s_{n+1}) d\lambda(t_{n+1}) \\ & \quad \times q(m_{n+1} | s_{n+1}) \phi_s(s_{n+1} | s_n, a_n) \varphi_n(s_n). \end{aligned}$$

ここで, $B = R_+$ とおけば (i) が得られる. また,

$$\begin{aligned} & P\{\varphi_{n+1} \in \Gamma, t_{n+1} \in B | \varphi_n, a_n\} \\ &= \int_{\Gamma} P\{t_{n+1} \in B | \varphi_n, a_n, \varphi_{n+1}\} d\phi\{\varphi_{n+1} | \varphi_n, a_n\} \end{aligned}$$

が成り立つ. そこで,

$$\bar{p}(B | y_n, a_n, y_{n+1}) \equiv p\{t_{n+1} \in B | y_n, a_n, y_{n+1}\}$$

とおくと、この $\bar{p} \in Q(R_+ | \mathfrak{A}^{\mathfrak{A}})$ が次の状態が y_{n+1}' であるという条件の下での y_n での滞在時間分布を表わす。そして、(i)の証明より (ii)の式が成り立つ。更に、

$$\begin{aligned} & p\{\{y_{n+1}, t_{n+1}\} | y_0, a_0, t_1, y_1, \dots, y_n, a_n\} \\ &= p\{\{y_{n+1}, t_{n+1}\} | y_n, a_n\} \\ &= \bar{g} \otimes \bar{p}(\{y_{n+1}, t_{n+1}\} | y_n, a_n) \end{aligned}$$

が成り立つので、 $\{y_n, t_n; n \in N\}$ は $\bar{g}$ と $\bar{p}$ により定まるマルコフ再生過程である。

定理 3.2. 任意の政策 ω に対して

$$J_{\pi}^{\omega}(y_0) = J_{\omega}^{\omega}(y_0) \quad (\forall y_0)$$

なる I-政策 π が存在する。

証明 任意に与えられた政策 ω に対して I-政策 $\pi^{\omega} = \{\pi_0^{\omega}, \pi_1^{\omega}, \dots\}$ を次のように定義する。

$$\begin{aligned} \pi_0^{\omega}(\{a_0\} | y_0) &\equiv \omega_0(\{a_0\} | y_0) \\ \pi_n^{\omega}(\{a_n\} | y_0, a_0, t_1, y_1, \dots, y_n) \\ &\equiv \sum_{m_1, \dots, m_n} \omega_n(\{a_n\} | y_0, a_0, t_1, m_1, \dots, m_n) \\ &\quad \times \bar{p}_n\{m_1, \dots, m_n | y_0, a_0, t_1, y_1, \dots, y_n\}, \end{aligned}$$

ここで、 $b_n^h \equiv (y_0, a_0, t_1, y_1, \dots, y_n)$ は $h_n \equiv (y_0, a_0, t_1, m_1, \dots, m_n)$ に対応する。この π^{ω} の作り方から、 ω と π^{ω} は A 上同一の条件付確率

$$\overline{P}_{\pi^{\omega}}\{\{a_n\} | y_0, a_0, x_1, y_1, \dots, y_n\} = \overline{P}_{\omega}\{\{a_n\} | \text{ " } \}$$

も与える. 一方, 定理 3.1. より,

$$\begin{aligned} \overline{P}_{\pi^{\omega}}\{\{y_{n+1}, x_{n+1}\} | y_0, a_0, x_1, y_1, \dots, y_n, a_n\} \\ = \overline{P}_{\omega}\{\{ \text{ " } \} | \text{ " } \} \\ = \overline{g} \otimes \overline{P}(\{y_{n+1}, x_{n+1}\} | y_n, a_n). \end{aligned}$$

そこで,

$$\overline{P}_{\pi^{\omega}}\{\{a_0, x_1, y_1, \dots, y_n, a_n\} | y_0\} = \overline{P}_{\omega}\{\{ \text{ " } \} | y_0\}$$

が成り立つ. 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 $(\Phi, A, \overline{g}, \overline{P}, \overline{C}_{\alpha}, \alpha)$ 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

$$I_{\pi}^{\alpha}(y_0) \equiv E_{\pi} \left[\sum_{n=0}^{\infty} e^{-\alpha(x_1 + \dots + x_n)} \overline{C}_{\alpha}(y_n, a_n) | y_0 \right]$$

で与えられる. ここで, $E_{\pi}[\cdot | y_0]$ は,

$$P_{\pi}\{\cdot | y_0\} \equiv \bigotimes_{n=0}^{\infty} (\pi_n \otimes \overline{g} \otimes \overline{P})$$

による条件付期待値である.

定理 3.3. 任意の I-政策 π に対して,

$$J_{\pi}^{\alpha}(y_0) = I_{\pi}^{\alpha}(y_0) \quad (\forall y_0)$$

が成り立つ.

証明 定理 3.1. より, $(A \oplus R_+)^n A$ 上の条件付確率 $\overline{P}_{\pi}\{\cdot | y_0\}$

は, 次のように分解される.

$$\begin{aligned}
\bar{P}_\pi \{ \cdot | y_0 \} &= \bar{P}_\pi \{ \{a_0\} | y_0 \} \otimes \bar{P}_\pi \{ \{y_1, t_1\} | y_0, a_0 \} \\
&\quad \otimes \cdots \otimes \bar{P}_\pi \{ \{a_n\} | y_0, a_0, t_1, y_1, \dots, y_n \} \\
&= \pi_0 \otimes \bar{g} \otimes \bar{p} \otimes \cdots \otimes \pi_n \\
&= \bar{p}_\pi \{ \cdot | y_0 \}.
\end{aligned}$$

したがって、注意1と I_π^α の定義より

$$J_\pi^\alpha(y_0) = I_\pi^\alpha(y_0) \quad (\forall y_0)$$

が成り立つ。

以上の2つの定理より、SMDP-II $(S, M, A, \bar{p}_s, \bar{p}_*, \bar{g}, y_0, C, \alpha)$ は、SMDP-I $(\bar{\pi}, A, \bar{g}, \bar{p}, \bar{c}_\alpha, \alpha)$ 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。

命題2.

$$\int_{\bar{\pi}} \bar{p}([0, \delta] | y, a, y') d\bar{g}(y' | y, a) \leq 1 - \varepsilon \quad (\forall y, a)$$

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

$$I^\alpha(y) \equiv \inf_{\pi \in \Pi} I_\pi^\alpha(y) \quad (\forall y)$$

とおく。

定理3.4. A は可算集合とする。このとき、 $I^\alpha(y)$ は、 $\bar{\pi}$ 上のボレル可測関数で次の方程式の一意的解である。

$$\begin{aligned}
(3.2) \quad I^\alpha(y) &= \inf_{a \in A} \left\{ \bar{c}_\alpha(y, a) + \int_{\bar{\pi}} \int_0^\infty e^{-\alpha t} I^\alpha(y') \right. \\
&\quad \left. \times d\bar{p}(t | y, a, y') d\bar{g}(y' | y, a) \right\} \quad (\forall y).
\end{aligned}$$

この方程式は次のように表わすことができる。

$$(3.3) \quad I^\alpha(\varphi) = \inf_{a \in A} \left\{ \bar{c}_\alpha(\varphi, a) + \sum_s \sum_{s'} \sum_{m'} \int_0^\infty e^{-\alpha t} \right. \\ \left. \times I^\alpha(u(\varphi, a, t, m')) d\bar{\phi}_t(t|s, a, s') \bar{\phi}_s(s'|s, a) \bar{q}(m'|s') \bar{q}(s) \right\}$$

(3.2) の (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.

証明 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 (ϕ, 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.

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

とおく. 命題 2 より

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

が成り立つことに注意する.

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

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

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

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 $\phi \equiv \bar{q}(\cdot | \varphi_0, a_0)$ に属する (ϕ, ε) -最適 I-政策 π' を

選択する I -政策 を π とする. このとき,

$$\begin{aligned} I_{\pi}^*(\varphi_0) &= T_{a_0} I_{\pi}^*(\varphi_0) \\ &\leq T_{a_0} (I^{\alpha} + \varepsilon)(\varphi_0) \\ &\leq T_{a_0} I^{\alpha}(\varphi_0) + \varepsilon(1 - \varepsilon + \varepsilon e^{-\alpha\delta}). \end{aligned}$$

したがって,

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

$\varepsilon > 0$ は任意なので,

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

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,

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

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, $I_{f_{\alpha}}^{\alpha}(\varphi)$ は,

$$I_{f_n}^{\alpha}(\varphi) = T_{f_n} I_{f_n}^{\alpha}(\varphi) \quad (\forall \varphi)$$

の一意的な解である。したがって、

$$I_{f_n}^{\alpha}(\varphi) = I^{\alpha}(\varphi) \quad (\forall \varphi).$$

ゆえに、 f_n 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

$$\begin{aligned} & \int_{\mathbb{R}} \int_0^{\infty} g(\varphi', t) d\bar{P}(t | \varphi, a, \varphi') d\bar{g}(\varphi' | \varphi, a) \\ &= \sum_s \sum_{s'} \sum_{m'} \int_0^{\infty} g(u(\varphi, a, t, m'), t) d\phi_t(t | s, a, s') \\ & \quad \times \phi_s(s' | s, a) g(m' | s') \varphi(s) \end{aligned}$$

が成り立つ。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 $(S, M, A, \phi_s, \phi_t, g, c)$ で定義する。割引因子 α はもはや必要なくなる。また、政策 w を用いたときの期待平均コストは

$$\bar{J}_w(\varphi_0) \equiv \lim_{n \rightarrow \infty} \frac{E_w \left[\sum_{i=0}^{n-1} c'(t_{i+1}, s_i, a_i) | \varphi_0 \right]}{E_w \left[\sum_{i=0}^{n-1} 1 | \varphi_0 \right]}$$

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.

$$\bar{J}_{\omega^*}(\varphi_0) \leq \liminf_{\omega} \bar{J}_{\omega}(\varphi_0) \quad (\forall \varphi_0)$$

が成り立つとき, ω^* は平均最適であるという.

$$c(s, a) \equiv \sum_{s'} \int_0^{\infty} c'(t, s, a) d\phi_*(t | s, a, s') \phi_s(s' | s, a)$$

$$\tau(s, a) \equiv \sum_{s'} \int_0^{\infty} t d\phi_*(t | s, a, s') \phi_s(s' | s, a)$$

とおくと,

$$\bar{J}_{\omega}(\varphi_0) = \overline{\lim}_{n \rightarrow \infty} \frac{E_{\omega} \left[\sum_{i=0}^{n-1} c(s_i, a_i) \mid \varphi_0 \right]}{E_{\omega} \left[\sum_{i=0}^{n-1} \tau(s_i, a_i) \mid \varphi_0 \right]}$$

が成り立つ. 更に,

$$\bar{c}(\varphi, a) \equiv \sum_s c(s, a) \varphi(s)$$

$$\bar{\tau}(\varphi, a) \equiv \sum_s \tau(s, a) \varphi(s), \quad \varphi \in \Phi,$$

とおく.

定理 4.1.

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

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

が成り立つ.

証明

$$\begin{aligned} & \bar{E}_{\omega} [c(s_n, a_n) \mid \varphi_0] \\ &= \int_{S \times A} c(s_n, a_n) d\bar{\Phi}_{\omega} \{ (s_n, \varphi_n, a_n) \mid \varphi_0 \} \\ &= \int_{\Phi \times A} \sum_{s_n} c(s_n, a_n) \bar{\Phi}_{\omega} \{ s_n \mid \varphi_0, \varphi_n, a_n \} d\bar{\Phi}_{\omega} \{ (\varphi_n, a_n) \mid \varphi_0 \} \end{aligned}$$

$$\begin{aligned}
 &= \int_{\mathbb{F}_A} \bar{c}(y_n, a_n) d\bar{p}_\omega \{ (y_n, a_n) | y_0 \} \\
 &= \bar{E}_\omega [\bar{c}(y_n, a_n) | y_0].
 \end{aligned}$$

これに対して同様に成り立つ。

注意 2. これは、I-政策 π に対して成り立つ。

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

$$(a) \quad \bar{E}_\pi [\bar{c}(y_n, a_n) | y_0] = \bar{E}_\omega [\bar{c}(y_n, a_n) | y_0]$$

$$(b) \quad \bar{E}_\pi [\bar{c}(y_n, a_n) | y_0] = \bar{E}_\omega [\bar{c}(y_n, a_n) | y_0] \quad (\forall y_0)(\forall n)$$

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 $(\mathbb{A}, \mathbb{B}, \bar{p}, \bar{c})$ 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、

$$\begin{aligned}
 \bar{I}_\pi(y_0) &\equiv \lim_{n \rightarrow \infty} \frac{E_\pi \left[\sum_{i=0}^{n-1} \bar{c}(y_i, a_i) | y_0 \right]}{E_\pi \left[\sum_{i=1}^n \tau_i | y_0 \right]} \\
 &= \lim_{n \rightarrow \infty} \frac{E_\pi \left[\sum_{i=0}^{n-1} \bar{c}(y_i, a_i) | y_0 \right]}{E_\pi \left[\sum_{i=1}^n \bar{c}(y_i, a_i) | y_0 \right]}
 \end{aligned}$$

で与えられる。ここで示す等式は、 $\bar{c}$ の定義と定理 3.1 より成り立つことに注意する。

定理 4.3. 任意の I-政策 π に対して

$$\bar{J}_\pi(y_0) = \bar{I}_\pi(y_0) \quad (\forall y_0)$$

が成り立つ。

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 $(S, M, A, \phi_s, \phi_A, g, c)$ は SMDP-I $(\bar{S}, \bar{A}, \bar{g}, \bar{P}, \bar{c})$ へ同値変形とれることがわかる. 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, Wakata [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.