数理解析研究所講究録
611巻 1987年 42-53

コンパクトな行動空間をもつ部分的観測セミマルコフ決定過程

長岡高専　涌田和芳（Kazuyoshi Wakuta）

§1. 序

部分的観測マルコフ決定過程は多くの人により研究されてきた。その一般化として、White [5]は部分的観測セミマルコフ決定過程（POSMDP）を導入した。そこでは、基礎過程として有限段階で状態が離散的に推移するセミマルコフ過程が用いられた。ここではWakuta [2,3]に引き続き、無限段階で状態が連続的に推移する場合を考えるが、状態空間は有限集合で、行動空間はコンパクト距離空間とする。

§2. POSMDP

まず記号について説明する。一般に、XとYを空でないボレル集合とするとき、X上の確率測度の全体を$P(X)$、XからYへの推移確率の全体を$Q(Y|X)$と表わす。また、任意の$p\in P(X)$と$g\in Q(Y|X)$に対して、pgは$X\times Y$で定義される通

常の確率測度とする。qが退化しているとき、すなわち、すべての$x\in X$に対して、$q(\{x\}|x)=1$なるXからYへのボレル可測写像ψが存在するとき、ψ_qのかわりにψとも書く。X上の有界ボレル可測関数の全体を$M(X)$、有界連続関数の全体を$C(X)$と表わす。

POSMDPは次のもので定められる：Sは有限集合で、システムの状態集合；Mは有限集合で、可能な信号の集合；Aはボレル集合で、行動集合；$q\in Q(S\times R_+|S\times A)$はシステムの運動法則、ここで$R_+=[0,\infty)$、$q$は

$$(2.1)\quad q(\{j\}\times[0,t]|i,a)=\int_0^t k(j,s|i,a)\,ds,\qquad i,j\in S,\ a\in A,\ t\geq 0$$

なる確率密度関数kをもつと仮定する。ただし、dsはルベーグ測度で、$k(j,\cdot|i,\cdot)$, $i,j\in S$,は(s,a)の非負ボレル可測関数とする；$r\in M(S\times A)$は、利得関数；αは正数で、割引因子。

有限時間区間で無限回の推移が起こらないように、次の仮定をおく。

Condition (1). (cf. Ross [1]) すべての$(i,a)\in S\times A$に対して、

$$(2.2)\quad q(S\times[0,\delta]|i,a)\leq 1-\varepsilon$$

なる$\varepsilon>0$と$\delta>0$が存在する。

簡単のために、$\mathcal{P} = P(S)$とおく。すなわち、$\mathcal{P}$はS上のすべての確率測度の集合を表わす。$\varphi \in \mathcal{P}$のi番目の成分を$\varphi(i)$とすると、$\varphi(i), i \in S$, はφについて連続なので、$p(i|\varphi) = \varphi(i), i \in S, \varphi \in \mathcal{P}$, として$p \in Q(S|\mathcal{P})$を定義することができる。行動を選択するための政策は、$\omega = \{\omega_0, \omega_1, \cdots\}$, $\omega_n \in Q(A|H_n), n \geq 0$, である。ここで、$H_n = \mathcal{P} \times (A \times R_+ \times M)^n$はn番目の行動が選択されるときの観測可能な履歴の集合であり、$\varphi \in H_0 = \mathcal{P}$はシステムの状態の初期情報を表わす。$i_n$, m_n, a_n, t_nは、それぞれ、n番目の状態、観測信号、行動、滞在時間を表わす。$T_n = t_1 + \cdots + t_n$ ($T_0 = 0$) とおくと、T_nは、丁n番目の決定時刻を表わす。政策ωとp, g, gに対して、$Q(S \times (A \times S \times R_+ \times M)^\infty | \mathcal{P})$の要素$p_\omega = p\omega_0 g_g \omega_1 g g \cdots$が定義できる。政策$\omega$を用いたときの期待合計割引利得を

$$(2.3) \quad J_\omega(\varphi) = E_\omega\left[\sum_{n=0}^{\infty} e^{-\alpha T_n} r(i_n, a_n) \,|\, \varphi\right], \quad \varphi \in \mathcal{P},$$

と定義する。ここで、$E_\omega[\cdot|\varphi]$は$p_\omega\{\cdot|\varphi\}$による条件付期待値を表わす。我々の最適化問題は、すべての政策の中でJ_ωを最大にすることである。すべての政策ωとすべての$\varphi \in \mathcal{P}$に対して、$J_{\omega^*}(\varphi) \geq J_\omega(\varphi)$が成り立つとき、$\omega^*$は最適であるという。

Remark 1. 利得関数 r は、次に推移する状態と滞在時間に依存するように一般化できる。

§3. 同値なSMDPの構成

POSMDPと同値な通常の（完全に観測可能な）セミマルコフ決定過程（SMDP）を構成する。

政策を任意に固定し、$h_n=(\varphi, a_0, t_1, m_1, \cdots, a_{n-1}, t_n, m_n)$ が与えられたときの i_n の条件付分布を ψ_n とする。そして、ψ_n に対して、$\varphi_n(i)=\psi_n\{i_n=i \mid h_n\}$, $i\in S$, として $\varphi_n\in\Pi$ を対応させる。ベイズの公式より、$\{\varphi_n\}_{n=0,1,2,\cdots}$ の再帰式

$$\varphi_{n+1}(j) \tag{3.1}$$
$$=\frac{\sum_{i\in S} k(j,t\mid i,a)\,g(m\mid j)\,\varphi_n(i)}{\sum_{i\in S}\sum_{j\in S} k(j,t\mid i,a)\,g(m\mid j)\,\varphi_n(i)},\quad j\in S,$$

を得る。ここで、φ_0 は初期情報である。そこで、写像 $u: D\to\Pi$ を

$$u(\varphi,a,t,m)(j)=\frac{\sum_{i\in S} k(j,t\mid i,a)\,g(m\mid j)\,\varphi(i)}{\sum_{i\in S}\sum_{j\in S} k(j,t\mid i,a)\,g(m\mid j)\,\varphi(i)},\quad j\in S, \tag{3.2}$$

として定義する。ここで、$D=\{(\varphi,a,t,m)\mid \sum_{i\in S}\sum_{j\in S} k(j,t\mid i,a)\,g(m\mid j)\varphi(i)>0\}$。$u$ はボレル可測である。D^c は確率ゼ

ロの集合なので、D^c上に亜の任意の値を与えて、$\hat{u}$の定義域を全空間$\bar{亜}\times A\times R_+\times M$に拡張する。この拡張された写像$\hat{u}$は再びボレル可測である。$\hat{u}$を繰返し用いることにより、観測可能な履歴$h_n=(\varphi, a_0, t_1, m_1, \cdots, a_{n-1}, t_n, m_n)$はボレル可測に$b_n=(\varphi, a_0, \varphi_1, t_1, \cdots, a_{n-1}, \varphi_n, t_n)\in B_n=\bar{亜}\times(A\times\bar{亜}\times R_+)^n$へ変換される。$B_n$を$n$番目の決定時刻までの可能な情報の集合と呼び、この可能な情報を通してのみ観測可能な履歴に依存する政策を情報政策（I-政策）と呼ぶ。I-政策は$\pi=\{\pi_0,\pi_1,\cdots\}$, $\pi_n\in Q(A|B_n)$, $n\geq 0$, である。すべての$b_n\in B_n$に対して、$\pi_n(f(\varphi_n)|b_n)=1$, $n\geq 0$, なるボレル可測写像$f:\bar{亜}\to A$が存在するとき、I-政策πは定常であるといい、f^∞と表わす。任意のI-政策πに対して、$\omega_n^\pi(\cdot|h_n)=\pi_n(\cdot|b_n^{h_n})$, $n\geq 0$, として政策$\omega^\pi=\{\omega_0^\pi, \omega_1^\pi, \cdots\}$を定義する。ここで、$b_n^{h_n}$は$h_n\in H_n$に対応する$B_n$の要素である。このとき、$\pi$と$\omega^\pi$はAに同じ条件付確率を割り当てるので、すべてのI-政策は政策とみなすことができる。I-政策πと$\varphi, q, g, \hat{u}$に対して$Q(S\times(A\times S\times R_+\times M\times\bar{亜})^\infty|\bar{亜})$の要素$\bar{\mathrm{P}}_\pi=\mathrm{P}\pi_0 g q \hat{u}\pi_1 g q \hat{u}\cdots$が定義できる。I-政策$\pi$を用いたときの期待合計割引利得を

$$(3.3)\quad J_\pi(\varphi)=\bar{E}_\pi\Big[\sum_{n=0}^{\infty} e^{-\alpha T_n} r(i_n,a_n)\,\Big|\,\varphi\Big],\quad \varphi\in\bar{亜},$$

と定義する。ここで、$\bar{E}_\pi[\cdot|\varphi]$は$\bar{\mathrm{P}}_\pi\{\cdot|\varphi\}$による条件付期待

値を表わす。政策 ω に対して、$Q(S\times(A\times S\times R_+\times M\times\Xi)^\infty|\Xi)$ の要素 $\bar{p}_\omega = p_{\omega_0}\bar{g}\hat{u}\,\omega_1\bar{g}\hat{u}$ を定義し、$\bar{p}_\omega\{\cdot|\varphi\}$ による条件付期待値を $\bar{E}_\omega[\cdot|\varphi]$ と表わすと、J_ω の定義において、E_ω を $\bar{E}_\omega$ でおきかえることができる。

$\bar{g}\in Q(\Xi\times R_+|\Xi\times A)$ を

$$(3.4)\quad \bar{g}(B|\varphi,a)=\sum_{i\in S}\sum_{j\in S}\sum_{m\in M} g(\{j\}\times B_m(\varphi,a)|i,a)\,g(m|j)\,\varphi(i),$$
$$B\in\mathcal{B}(\Xi\times R_+),\ \varphi\in\Xi,\ a\in A,$$

と定義する。ここで、

$$(3.5)\quad B_m(\varphi,a)=\{t\,|\,(t,m)\in B(\varphi,a)\},$$

$$(3.6)\quad B(\varphi,a)=\{(t,m)\,|\,(\hat{u}(\varphi,a,t,m),t)\in B\}.$$

また、$\bar{r}\in M(\Xi\times A)$ を

$$(3.7)\quad \bar{r}(\varphi,a)=\sum_{i\in S} r(i,a)\varphi(i),\ \varphi\in\Xi,\ a\in A,$$

と定義する。POSMDP(S,M,A,g,g,r,α) は、以上のものから定義されるSMDP$(\Xi,A,\bar{g},\bar{r},\alpha)$ に同値に変換されることを示す。ここで、Ξ は状態の集合；A は行動の集合；$\bar{g}$ は運動法則；$\bar{r}$ は利得関数；α は割引因子。SMDPに対する政策はI-政策と同じなので、$\pi=\{\pi_0,\pi_1,\cdots\}$ と表わす。I-政策 π と $\bar{g}$ に対して、$Q((A\times\Xi\times R_+)^\infty|\Xi)$ の要素 $p_\pi=\pi_0\bar{g}\pi_1\bar{g}\cdots$

が定義できる。I-政策 π を用いたときの期待合計割引利得を

$$(3.8)\quad I_\pi(\varphi) = E_\pi\Big[\sum_{n=0}^{\infty} e^{-\alpha T_n}\, r(\varphi_n, a_n)\,\Big|\,\varphi\Big],\quad \varphi\in\Phi,$$

と定義する。ここで、$E_\pi[\cdot|\varphi]$ は $p_\pi\{\cdot|\varphi\}$ による条件付期待値を表わす。

Remark 2. Condition(1)が満たされるとすると、

$$(3.9)\quad q(\Phi\times[0,\delta]\,|\,\varphi, a) \le 1-\varepsilon,\quad \varphi\in\Phi,\ a\in A,$$

が成り立つ。

Lemma 3.1. 任意の政策 ω と任意のI-政策 π に対して、

$$(3.10)\quad J_\omega(\varphi) = E_\omega\Big[\sum_{n=0}^{\infty} e^{-\alpha T_n}\, r(\varphi_n, a_n)\,\Big|\,\varphi\Big],\ \varphi\in\Phi;$$

$$(3.11)\quad J_\pi(\varphi) = E_\pi\Big[\sum_{n=0}^{\infty} e^{-\alpha T_n}\, r(\varphi_n, a_n)\,\Big|\,\varphi\Big],\ \varphi\in\Phi,$$

が成り立つ。

Lemma 3.2. 任意の政策 ω と任意のI-政策 π に対して、

$$(3.12)\quad q \in p_\omega(\varphi_{n+1}, t_{n+1})\,|\,(\varphi, a_0, \varphi_1, t_1, \cdots, a_{n-1}, \varphi_n, t_n, a_n),\ n\ge 0;$$

$$(3.13)\quad q \in p_\pi(\varphi_{n+1}, t_{n+1})\,|\,(\varphi, a_0, \varphi_1, t_1, \cdots, a_{n-1}, \varphi_n, t_n, a_n),\ n\ge 0,$$

が成り立つ。ここで、$q \in p_{\omega\, X_2|X_1}$ ($q \in p_{\pi\, X_2|X_1}$) は、q は p_ω (p_π) の下で X_1 が与えられたときの X_2 の条件付分布であることを意味する。

Lemma 3.3. 各政策 ω に対して、$J_{\pi^\omega}(\varphi) = J_\omega(\varphi)$, $\varphi \in \overline{S}$, なる I-政策 π^ω が存在する。すなわち、I-政策で十分である。

以上の補題から次の定理を得る (cf. Wakuta [2,3])。

Theorem 3.4. 任意の I-政策 π に対して、$J_\pi(\varphi) = I_\pi(\varphi)$, $\varphi \in \overline{S}$, が成りたつ。この意味で、POSMDP $(S, M, A, q, g, r, \alpha)$ は SMDP $(\overline{S}, A, \overline{q}, \overline{r}, \alpha)$ と同値である。

§4. 最適定常 I-政策の存在

次の条件をおく。

Condition (2). (i) 各 (i, j) に対して、$k(j, \cdot | i, \cdot)$ は $R_+ \times A$ 上連続で、

$$(4.1) \quad k(j, t | i, a) \leq M_{ij}(t), \quad t \geq 0, \ a \in A,$$

$$(4.2) \quad \int_0^\infty e^{-\alpha t} M_{ij}(t)\, dt < \infty,$$

なる可測関数 $M_{ij}(t)$ が存在する;

(ii) A はコンパクト距離空間である;

(iii) $r(i, \cdot)$, $i \in S$, は A 上の有界連続関数である。

いわゆる最適方程式は、

$$(4.3) \quad v(\varphi) = \max_{a \in A} \left\{ \overline{r}(\varphi, a) + \int_{\overline{S}} \int_0^\infty e^{-\alpha t} v(\varphi')\, d\overline{q}((\varphi', t) | \varphi, a) \right\}, \quad \varphi \in \overline{S},$$

と表わされる。

$\bar{r}$の定義から、この方程式は

$$(4.4)\quad v(\varphi)=\max_{a\in A}\Big\{\bar{r}(\varphi,a)+\sum_{m\in M}\int_0^\infty e^{-\alpha t}v(\hat{u}(\varphi,a,t,m))\times\Big[\sum_{i\in S}\sum_{j\in S}k(j,t|i,a)g(m|j)\varphi(i)\Big]dt\Big\},\quad \varphi\in\Phi,$$

と表わすことができる。(4.3)あるいは(4.4)式の右辺のカッコの項を$v(\varphi,a)$と表わし、

$$(4.5)\quad T_a v(\varphi)=v(\varphi,a),\ \varphi\in\Phi,\ a\in A;$$

$$(4.6)\quad Tv(\varphi)=\max_{a\in A}T_a v(\varphi),\ \varphi\in\Phi;$$

$$(4.7)\quad T_f v(\varphi)=v(\varphi,f(\varphi)),\ \varphi\in\Phi,$$

として、オペレーターT_a, T, T_fを導入する。ここで、fはΦからAへのボレル可測写像である。

Lemma 4.1. Conditions (1)と(2)が成り立つとする。このとき、(i)(4.6)で定義されるオペレーターTは$C(\Phi)$上の縮小写像である;(ii)(4.7)で定義されるオペレーターT_fは$M(\Phi)$上の縮小写像である。

(証明) (i)を証明する。(ii)の証明は同様である。$m\in M$を固定する。$u_m(\varphi,a,t)=u(\varphi,a,t,m)$, $\nu_m(\varphi,a,t)=\sum_{i\in S}\sum_{j\in S}k(j,t|i,a)g(m|j)\varphi(i)$, $D_m=\{(\varphi,a,t)\mid \nu_m(\varphi,a,t)>0\}$とおく。Condition (2)より$u_m$は$D_m$上連続で、$\nu_m$は$\Phi\times A\times R_+$上連続である。$(\varphi,a,t)$を$D_m$の境界点とし、$(\varphi_n,a_n,t_n)\in D_m$,

$(\varphi_n, a_n, t_n) \to (\varphi, a, t)$ $(n \to \infty)$ とする。このとき、$\nu_m(\varphi_n, a_n, t_n) \to \nu_m(\varphi, a, t) = 0$。$v$ は有界で、$(\varphi, a, t) \notin D_m$ なら、$\nu_m(\varphi, a, t) = 0$ なので、(4.4) の被積分関数は D_m の境界で連続で、したがって、全空間 $\Xi \times A \times R_+$ 上で連続である。Condition (2) より、$v(\varphi, a)$ は (φ, a) について連続であることが容易に示される。A はコンパクトなので、$Tv(\varphi)$ は φ について連続である。したがって、T は $C(\Xi)$ 上のオペレーターである。(3.9) より、任意の $v, v' \in C(\Xi)$ に対して、

$$\|Tv - Tv'\| \leq (1 - \varepsilon + \varepsilon e^{-\alpha\delta}) \|v - v'\| \tag{4.8}$$

が成り立ち、T は縮小写像である（$\|\cdot\|$ はスーップノルムを表わす）。以上で証明された。

Theorem 4.2. Conditions (1) と (2) が成り立つとする。このとき、I-政策 π は、利得 $I(x)$ が最適方程式 (4.3) あるいは (4.4) を満たすときに限り最適である。

証明は Wakuta [4] を参照のこと。

Theorem 4.3. Conditions (1) と (2) が成り立つとする。このとき、各 φ に対して、T の不動点 v^* をもつ (4.3) 式あるいは (4.4) 式の右辺を最大にする行動を選択する最適定常 I-政策 f が存在する。すなわち、f は次のようなものである：

$$(4.9)\quad \bar{r}(\varphi, f(\varphi)) + \sum_{m\in M} \int_0^\infty e^{-\alpha t} v^*(\hat{u}(\varphi, f(\varphi), t, m)) \times \Big[\sum_{i\in S} \sum_{j\in S} k(j, t \mid i, f(\varphi))\, g(m \mid j)\, \varphi(i) \Big] dt$$

$$= \max_{a\in A} \Big\{ \bar{r}(\varphi, a) + \sum_{m\in M} \int_0^\infty e^{-\alpha t} v^*(\hat{u}(\varphi, a, t, m)) \times \Big[\sum_{i\in S} \sum_{j\in S} k(j, t \mid i, a)\, g(m \mid j)\, \varphi(i) \Big] dt \Big\}$$

証明は、Theorem 4.2 と良く知られた Dubins and Savage の selection theorem から直ちに行なわれる。

References

[1] Ross, S. (1970). Applied Probability Models with Optimization Applications. Holden-Day.

[2] Wakuta, K. (1981). Semi-Markov decision processes with incomplete state observation – Average cost criterion. J. Oper. Res. Soc. Japan, 24, 95–108.

[3] _____. (1982). Semi-Markov decision processes with incomplete state observation – Discounted cost criterion. J. Oper. Res. Soc. Japan, 25, 351–361.

[4] _____. Arbitrary state semi-Markov decision processes with unbounded rewards. To appear in Optimization.

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