数理解析研究所講究録
第353巻 1979年 40-49

マルコフゲームへの学習アルゴリズムの応用

新大 理学部　田中謙輔

§1. マルコフゲーム

ここでは離散時間の二人零和マルコフゲームを次のような一つの組 (S, A, B, p, r, β) で定義する：(1) $S=\{1,2,\cdots,s\}$ でシステムの状態空間と呼ぶ有限集合，(2) $A=\{a^1, a^2, \cdots, a^{m_1}\}$ はプレイヤーIが選択できる純戦略の集りでプレイヤーIの行動空間と呼ぶ有限集合，(3) $B=\{b^1, b^2, \cdots, b^{m_2}\}$ はプレイヤーIIが選択できる純戦略の集りでプレイヤーIIの行動空間と呼ぶ有限集合，(4) $p(l'|l,a,b)$ は $l\in S$, $a\in A$, $b\in B$ によって定まる，次の段階で状態 l' がおこる推移確率，(5) $r(l,a,b)$ は $S\times A\times B$ の上で定義されているプレイヤーIの利得関数と呼ぶ有限な実数値関数，(6) β は割引因子と呼ぶ $0\leqq\beta<1$ なる実数．

このマルコフゲームでは二人のプレイヤーが最初0段階でシステムの状態 $l_0\in S$ を観測し，互に独立に戦略 $a_0\in A$, $b_0\in B$

を選択する．この結果としてプレイヤーIは利得 $r(l_0, a_0, b_0)$，プレイヤーIIは利得 $-r(l_0, a_0, b_0)$ を受ける．次にシステムの状態 l_0 は推移確率 $p(l_1|l_0, a_0, b_0)$ にしたがって新しい状態 l_1 に移る．次に1段階で2人のプレイヤーはシステムの状態 l_1 を観測し，互に独立に戦略 a_1, b_1 を選択する．この結果としてプレイヤーIは利得 $\beta r(l_1, a_1, b_1)$，プレイヤーIIは利得 $-\beta r(l_1, a_1, b_1)$ を受ける．次にシステムの状態 l_1 は推移確率 $p(l_2|l_1, a_1, b_1)$ にしたがって新しい状態 l_2 に移る．以下同様にして2人のプレイヤーはゲームを無限に続けることにする．このとき各プレイヤーが政策 π，σ を用いたとき，プレイヤーIの全期待割引利得は

$$\psi(l_0, \pi, \sigma) = E^{\pi,\sigma}\left[\sum_{n=0}^{\infty} \beta^n r(l_n, a_n, b_n)\right]$$

となる．ただし $\pi = (f_0, f_1, \cdots)$，$\sigma = (g_0, g_1, \cdots)$ で f_n，g_n は n 段階で各プレイヤーが戦略を選択する写像とする．このときプレイヤーIは $\psi(l_0, \pi, \sigma)$ をできるだけ大きくするように，プレイヤーIIはできるだけ小さくするように互に政策 π，σ を選択する基準で考えることにする．

このようなマルコフゲームに対して，ベクトル空間からベクトル空間への縮小写像に関する不動点定理から式：

$$V_l^* = \max_P \min_Q \{r(l, P, Q) + \beta \sum_{l'=1}^{S} V_{l'}^* p(l'|l, P, Q)\} \qquad (1.1)$$
$$= r(l, P^*(l), Q^*(l)) + \beta \sum_{l'=1}^{S} V_{l'}^* p(l'|l, P^*(l), Q^*(l)),\ l = 1, 2, \cdots, S,$$

ただし $r(\ell, P, q) = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} r(\ell, a^i, b^j) p_i q_j$,

$p(\ell'|\ell, P, q) = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} p(\ell'|\ell, a^i, b^j) p_i q_j$,

が成立する. このとき $V^* = (V_1^*, V_2^*, \cdots, V_S^*)$ はマルコフゲームのゲーム値で, $P^*(\ell)$, $q^*(\ell)$ がそれぞれシステムの状態 ℓ におけるプレイヤーI, プレイヤーIIの最適混合戦略となることが知られている. よって2人のプレイヤーの最適政策は定常政策 $\pi^* = (P^*, P^*, \cdots)$, $\sigma^* = (q^*, q^*, \cdots)$ となることがわかる. さらにゲーム値 V^* は縮小写像の原理より唯一つ存在し次のように逐次的に求めることができる:

$V^{(0)}$: 任意に与えられるS次元ベクトル

$$V_\ell^{(n+1)} = \max_P \min_q \{ r(\ell, P, q) + \beta \sum_{\ell'=1}^{S} V_{\ell'}^{(n)} p(\ell'|\ell, P, q) \}$$

$\ell = 1, 2, \cdots, S$.

このとき不等式:

$$\| V^{(n)} - V^* \| \leq \beta^n \| V^{(0)} - V^* \| \qquad (1.2)$$

が成立する. ただし $V^{(n)} = (V_1^{(n)}, V_2^{(n)}, \cdots, V_S^{(n)})$ で, ノルムは各要素の絶対値の最大値を表している.

§2. dummy game の正規化

マルコフゲームについての説明から dummy game と呼ばれる, システムの各状態 ℓ で利得行列 $\{ V_\ell^*(a,b) ; a \in A, b \in B \}$ をもつ2人零和ゲームの最適混合戦略の組, すなわち, 2人のプ

プレイヤーの任意の混合戦略 P, Q に対して

$$V_\ell^*(P^*(\ell), Q) \geq V_\ell^* \geq V_\ell^*(P, Q^*(\ell))$$

が成立する $(P^*(\ell), Q^*(\ell))$ を求めることになる。ただし

$$V_\ell^*(a,b) = r(\ell,a,b) + \beta \sum_{\ell'=1}^{S} V_{\ell'}^* p(\ell'|\ell,a,b),$$

$$V_\ell^*(P,Q) = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} V_\ell^*(a^i,b^j) p_i q_j.$$

ここではこのようなゲームを正規化するために利得行列を，$\delta>0$ に対して，$\{V_\ell^*(a,b) - \frac{\delta}{2}(p_a - q_b);\ a\in A, b\in B\}$ と修正する．このときプレイヤーIの期待利得は

$$V_{\ell,\delta}^*(P,Q) = V_\ell^*(P,Q) - \frac{\delta}{2}(\|P\|^2 - \|Q\|^2)$$

で与えられる。ただし $P=(p_1,p_2,\cdots,p_{m_1})$, $Q=(q_1,q_2,\cdots,q_{m_2})$ で，$\|\cdot\|$ はユークリッドノルムを表している。したがってプレイヤーIIの期待利得は $-V_{\ell,\delta}^*(P,Q)$ で与えられる。

次に，各プレイヤーが利用できる混合戦略 P, Q はそれぞれ $S_\varepsilon^{m_1}$, $S_\varepsilon^{m_2}$ の中にあると仮定する。ただし

$$S_\varepsilon^m = \left\{X=(x_1,x_2,\cdots,x_m);\ x_i \geq \varepsilon, (i=1,2,\cdots,m), \sum_{i=1}^{m} x_i = 1, \left(0\leq\varepsilon\leq\frac{1}{m}\right)\right\}.$$

このように制限された2人零和ゲームでは，$\delta>0$ に対して $V_{\ell,\delta}(P,Q)$ は厳密に凸となる．よって任意の $P\in S_\varepsilon^{m_1}$, $Q\in S_\varepsilon^{m_2}$ に対して

$$V_{\ell,\delta}^*(P^*(\ell,\varepsilon,\delta), Q) \geq V_{\ell,\delta}^*(P^*(\ell,\varepsilon,\delta), Q^*(\ell,\varepsilon,\delta)) \geq V_{\ell,\delta}^*(P, Q^*(\ell,\varepsilon,\delta)) \tag{2.1}$$

となる唯一の最適混合戦略の組 $(P^*(\ell,\varepsilon,\delta), Q^*(\ell,\varepsilon,\delta))$ が存在する．

このような最適混合戦略の組について次のような2つの補助定理が重要となる.

補助定理1.

数列 $\{\varepsilon(n)\}, \{\delta(n)\}$ が次のような条件をみたしているとする:

$$\varepsilon(n) \in (0, \hat{\varepsilon}) \quad , \quad \delta(n) > 0 \quad , \quad \lim_{n\to\infty} \varepsilon(n) = \lim_{n\to\infty} \delta(n) = 0,$$

$$\lim_{n\to\infty} \frac{\varepsilon(n)}{\delta(n)} = \mu \in [0, \infty).$$

このとき $(P^*(\ell, \varepsilon(n), \delta(n)), q^*(\ell, \varepsilon(n), \delta(n))$ は最初のゲームの鞍部点 $(P^*(\ell), q^*(\ell))$ に收束する.

補助定理2.

任意の利得行列に対して, 次のような条件をみたす $\delta' \in (0, \infty)$ と定数 K_1, K_2, K_3 が存在する: $\forall \varepsilon_1, \varepsilon_2 \in [0, \hat{\varepsilon}], \forall \delta_1, \delta_2 \in (0, \delta')$,

$$\| P^*(\ell, \varepsilon_1, \delta_1) - P^*(\ell, \varepsilon_2, \delta_2) \| + \| q^*(\ell, \varepsilon_1, \delta_1) - q^*(\ell, \varepsilon_2, \delta_2) \|$$
$$\leq K_1 |\varepsilon_1 - \varepsilon_2| + K_2 |\delta_1 - \delta_2| + K_3 \left| \frac{\varepsilon_1}{\delta_1} - \frac{\varepsilon_2}{\delta_2} \right| . \qquad (2.2)$$

§3. 学習アルゴリズム

ここでは各プレイヤー(生徒)がマルコフゲームにおける利得関数, システムの状態を変化させる推移確率を知っていないとする. このとき dummy game を通して外3者(教師)が各プレイヤーに最適戦略 $P^*(\ell), q^*(\ell)$ をそれぞれ学習させる方法を次のように考えることにする.

まず外3者は初期ベクトル $V^{(0)} = (V_1^{(0)}, V_2^{(0)}, \cdots, V_s^{(0)})$ を定め,

各プレイヤーは次の段階の各システムの状態で用いる初期混合戦略 $P^{(0)}(\ell)$, $q^{(0)}(\ell)$, $\ell=1,2,3,\cdots,S$, を定める。今n段階で予想着はゲーム値の近似ベクトル $V^{(n)}=(V_1^{(n)}, V_2^{(n)}, \cdots, V_S^{(n)})$ を定めており，各プレイヤーは次の段階の各状態 ℓ で用いる混合戦略 $P^{(n)}(\ell)$, $q^{(n)}(\ell)$ をそれぞれ定めているとする。このとき $(n+1)$ 段階では，各状態 ℓ で各プレイヤーは $P^{(n)}(\ell)$, $q^{(n)}(\ell)$ でゲームをプレイする。この結果各プレイヤーが用いた純戦略をそれぞれ $x_{n+1}(\ell)$, $y_{n+1}(\ell)$ とするとき，次の着は各プレイヤーに

$$V_\ell^{(n+1)}(x_{n+1}(\ell), y_{n+1}(\ell)) = r(\ell, x_{n+1}(\ell), y_{n+1}(\ell)) + \beta \sum_{\ell'=1}^{S} V_{\ell'}^{(n)} p(\ell'|\ell, x_{n+1}(\ell), y_{n+1}(\ell))$$

の値を教える。これを用いて各プレイヤーは次の段階で用いる混合戦略を次のように作ることにする。プレイヤーIはシステムの各状態 ℓ で

$$P^{(n+1)}(\ell) = \Pi_{S_{\varepsilon(n+1)}^{m_1}}\left[P^{(n)}(\ell) + \gamma(n+1) A^{(\ell)}(x_{n+1}(\ell), y_{n+1}(\ell))\right] \qquad (3.1)$$

と作る。ただし $A^{(\ell)}(a^i, b^j)$ の k 要素 $A_k^{(\ell)}(a^i, b^j)$ は

$$A_k^{(\ell)}(a^i, b^j) = \begin{cases} \dfrac{V_\ell^{(n+1)}(a^i, b^j)}{p_i^{(n)}(\ell)} - \delta(n+1) & , k = a^i \\ -\dfrac{1}{m_1 - 1}\left(\dfrac{V_\ell^{(n+1)}(a^i, b^j)}{p_i^{(n)}(\ell)} - \delta(n+1)\right) & , k \neq a^i, \end{cases}$$

で，記号 $\Pi_S[X]$ は有界閉集合 S への X の射影作用素を表している。

同様にして，プレイヤーIIはシステムの各状態 ℓ で

$$q_\ell^{(n+1)} = \Pi_{S_{\varepsilon(n+1)}^{m_2}}[q_\ell^{(n)} - \gamma(n+1)B^{(\ell)}(a_{n+1}(\ell), y_{n+1}(\ell))] \quad (3.2)$$

と作る．ただし $B^{(\ell)}(a^i, b^j)$ の k 要素 $B_k^{(\ell)}(a^i, b^j)$ は

$$B_k^{(\ell)}(a^i, b^j) = \begin{cases} \dfrac{V_\ell^{(n+1)}(a^i, b^j)}{q_j^{(n)}(\ell)} + \delta(n+1) & , k = b^j \\ -\dfrac{1}{m_2 - 1}\left(\dfrac{V_\ell^{(n+1)}(a^i, b^j)}{q_j^{(n)}(\ell)} + \delta(n+1)\right) & , k \neq b^j \end{cases}$$

で表している．

また第3者は次の段階で用いるゲーム値の近似ベクトル $V^{(n+1)} = (V_1^{(n+1)}, V_2^{(n+1)}, \cdots, V_S^{(n+1)})$ を次のように作る．

$$V_\ell^{(n+1)} = \max_P \min_q \{r(\ell, P, q) + \beta \sum_{\ell'=1}^{S} V_{\ell'}^{(n)} p(\ell' | \ell, P, q)\}.$$

このようにして $(n+1)$ 段階での第3者の作業と各プレイヤーの学習は終る．以下の各段階で同様にして各プレイヤーの学習は進行する．

このとき次の定理1，定理2によって各プレイヤーは学習の目的を達成することが保証されている．

定理1.

数列 $\{\varepsilon(n)\}, \{\delta(n)\}, \{\gamma(n)\}$ が次のような条件をみたしている:

(a) $\gamma(n) > 0$, $\delta(n) > 0$, $\varepsilon(n) \in (0, \hat{\varepsilon})$, $n = 1, 2, \cdots\cdots$,

$\delta(n) \to 0$, $n \to \infty$, ただし $\hat{\varepsilon} = \min\left(\frac{1}{m_1}, \frac{1}{m_2}\right)$,

(b) $\lim_{n\to\infty} \varepsilon(n)/\delta(n) = \mu < \infty$, (c) $\sum_{n=1}^{\infty} \gamma(n)\delta(n) = \infty$,

(d) $\sum_{n=1}^{\infty}\gamma^2(n)\delta^2(n)<\infty$, (e) $\sum_{n=1}^{\infty}\gamma^2(n)/\varepsilon(n-1)<\infty$,

(f) $\sum_{n=1}^{\infty}|\varepsilon(n)-\varepsilon(n-1)|<\infty$, (g) $\sum_{n=1}^{\infty}|\delta(n)-\delta(n-1)|<\infty$,

(h) $\sum_{n=1}^{\infty}\left|\frac{\varepsilon(n)}{\delta(n)}-\frac{\varepsilon(n-1)}{\delta(n-1)}\right|<\infty$.

このとき，任意の初期混合戦略の組 $(P^{(0)}(\ell), q^{(0)}(\ell)) \in S^{m_1}_{\varepsilon(0)} \times S^{m_2}_{\varepsilon(0)}$ に対して，列 $(P^{(n)}(\ell), q^{(n)}(\ell))$ は $n\to\infty$ のとき $(P^*(\ell), q^*(\ell))$ に確率1で收束する．

証明. 補助定理2を用いて，$\forall n \geq n_0$ に対して

$$E[d(n+1)\,|\,P^{(n)}(\ell), q^{(n)}(\ell), \ell=1,2,\cdots,s] \leq (1-L_1\gamma(n+1)\delta(n+1))d(n) +$$
$$+K_1|\varepsilon(n+1)-\varepsilon(n)| + K_2|\delta(n+1)-\delta(n)| + K_3\left|\frac{\varepsilon(n+1)}{\delta(n+1)}-\frac{\varepsilon(n)}{\delta(n)}\right| +$$
$$+K_4\beta^{n+1} + K_5\frac{\gamma^2(n+1)}{\varepsilon(n)} + K_6\gamma^2(n+1)\delta^2(n+1)$$

が成立するような n_0 と定数 $K_1, K_2, K_3, K_4, K_5, K_6$ が存在する．ただし

$$d(n+1)=\sum_{\ell=1}^{s}\left(\frac{m_1-1}{m_1}\|P^{(n+1)}(\ell)-P^{(n+1)*}(\ell)\|^2+\frac{m_2-1}{m_2}\|q^{(n+1)}(\ell)-q^{(n+1)*}(\ell)\|^2\right).$$

ここで $D(n)=d(n)+\sum_{k=n+1}^{\infty}\beta(k)$ とおく．ただし

$$\beta(n+1)=K_1|\varepsilon(n+1)-\varepsilon(n)|+K_2|\delta(n+1)-\delta(n)|+K_3\left|\frac{\varepsilon(n+1)}{\delta(n+1)}-\frac{\varepsilon(n)}{\delta(n)}\right|+$$
$$+K_4\beta^{n+1}+K_5\gamma^2(n+1)/\varepsilon(n)+K_6\gamma^2(n+1)\delta^2(n+1).$$

このとき $\{D(n)\}$ は semi-martingale となり，semi-martingale に関する收束定理と $\sum_{n=0}^{\infty}\gamma(n+1)\delta(n+1)E[d(n)]<\infty$ が成立することよりこの定理の証明は完成できる．

次に2乗平均收束を示すために次の補助定理が重要である.

補助定理3.

非負数列$\{u(n)\}$が次のような条件をみたしているとする:

$\forall n \geq n_0$

(a) $u(n) \leqq u(n-1)(1-\lambda(n)) + \theta(n)$,

(b) $\lambda(n) \in (0, 1]$, $\sum_{n=1}^{\infty} \lambda(n) = \infty$, $\lim_{n\to\infty} \frac{\theta(n)}{\lambda(n)} = 0$.

このとき $\lim_{n\to\infty} u(n) = 0$ が成立する.

定理2.

定理1における条件(d)~(h)が次のような条件に置き換えられている:

(d') $\gamma(n+1)\delta(n+1) \longrightarrow 0$ as $n\to\infty$,

(e') $\frac{\gamma(n+1)}{\varepsilon(n)\delta(n+1)} \longrightarrow 0$ as $n\to\infty$,

(f') $\frac{1}{\gamma(n+1)\delta(n+1)} |\varepsilon(n+1) - \varepsilon(n)| \longrightarrow 0$ as $n\to\infty$,

(g') $\frac{1}{\gamma(n+1)\delta(n+1)} |\delta(n+1) - \delta(n)| \longrightarrow 0$ as $n\to\infty$,

(h') $\frac{1}{\gamma(n+1)\delta(n+1)} \left| \frac{\varepsilon(n+1)}{\delta(n+1)} - \frac{\varepsilon(n)}{\delta(n)} \right| \longrightarrow 0$ as $n\to\infty$.

さらに, 次の条件がみたされている:

(i) $\frac{\beta^{n+1}}{\gamma(n+1)\delta(n+1)} \longrightarrow 0$ as $n\to\infty$.

このとき $(p^{(n+1)}(\ell), q^{(n+1)}(\ell))$ は2乗平均の意味で$(p^*(\ell), q^*(\ell))$に收束する.

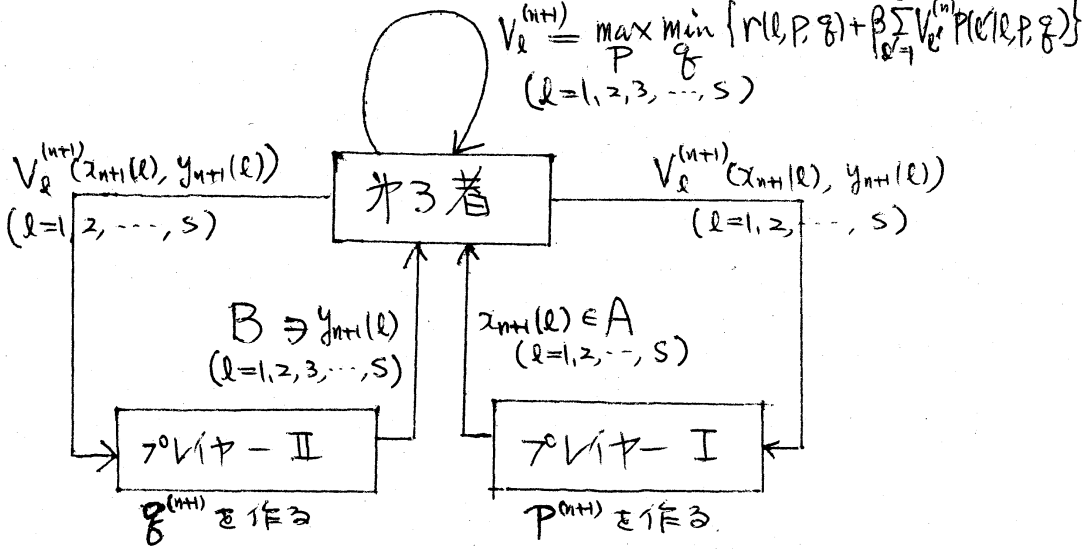
(注意) アルゴリズムにおける数列の例として

$$\gamma(n)\sim\frac{1}{n^{\alpha}}\quad,\quad \varepsilon(n)\sim\frac{1}{n^{\beta}}\quad,\quad \delta(n)\sim\frac{1}{n^{\sigma}}\quad,$$

$$\frac{\varepsilon(n)}{\delta(n)}-\mu\sim\begin{cases}\dfrac{1}{n^{\nu}} & ,\ \beta=\sigma\\[2ex] \dfrac{1}{n^{\beta-\sigma}} & ,\ \beta>\sigma,\end{cases}$$

とおき定理1に対しては，$\beta\geq\sigma>0$ ，$\nu>0$，$\frac{1}{2}<\alpha+\sigma\leq 1$，$2\alpha-\beta>1$，定理2に対しては $\beta\geq\sigma>0$，$\nu>0$，$\alpha+\sigma\leq 1$，$\alpha-\beta-\sigma>0$ 等に選択すればよい。

(注意) 学習システムの図示。



参考文献

1. Tsypkin, Ya. Z., Adaptation and Learning in automatic systems, (1968) (in Russian).
2. Nazin, A. Z. and Poznyak, A. S., Stochastic zero-sum game of two automata, Avtom. Telemekh (1977) (in Russian).
3. Tanaka, k. and Homma, H., On the learning algorithm of 2-person zero-sum game, to appear.