学習プロセスと後悔

大阿久　博
（武蔵野大学政治経済学部教授）

一　学習理論

従来のゲーム理論において中心となってきた解概念はナッシュ均衡であり、ナッシュ均衡が一般的に存在することはナッシュ自身によって証明された（Nash[8.9]）。しかし、均衡が（一般には複数）存在することと、それらを具体的に求めること、さらにそれらの中のどの均衡が実現するか見極めることは、決して同じことではない。様々な実験を通して、こうしたことが明らかにされてきた。つまりゲーム理論においては、「均衡が具体的に求められないという非決定性の問題」と「均衡を計算して求めるのは原理的に可能であるが、実際には膨大な時間がかかって困難であるという非決定性の問題」という異なる問題があるのである。（川越[11]第三章、p.79）。後者の問題が生じたとき、人間（プレイヤー）はどのように均衡に達するのかを研究するのが**学習理論**である。

こうした学習理論においては、プレイヤーは従来のゲーム理論が想定してきたような合理性は持っておらず、必然的に「限定合理的」になる。[1]

学習理論では、主に二つのタイプの学習プロセスが考えられている。一つは「信念学習」、もう一つは「強化学習」と呼ばれるものである。前者は、相手プレイヤーが過去に選択した戦略の頻度から、次に相手がとる戦略を予想し、その予想に対して最適な戦略を選択するという学習行動であり、後者は、相手のプレイヤーの戦略の選択頻度ではなく、過去に選んだ自分自身の戦略がどの程度の利得を上げてきたかを考え、より利得の高かったものをより高い頻度で選択するという学習行動である。信念学習はゲーム理論においても「仮想プレー（fictitious play）」などで議論されてきた。一方、強化学習は心理学などで研究されてきた分野である。

以下では、「強化学習」の考え方、その中でも特に「後悔しない」戦略を選択するといった行動様式（regret matching モデル）について既存研究を概観する。[2]

二．後悔（regret）

1．例

「後悔しない戦略をとる」という状況を定式化するに当たり、例として次の表1のような 2×2ゲームを考えよう（The Soda Game（Young[13]））。このゲームでは、二人のプレイヤーはそれぞれLとRの二つの行動を持ち、利得の代わりに財（Coke, Sprite, Seven-up, Pepsi）が表示されている。

表１　The Soda Game

		Player2	
		L	R
Player1	L	Coke, Coke	Sprite, Seven-up
	R	Seven-up, Sprite	Pepsi, Pepsi

表２　Soda Game が11回繰り返されたときのプレイヤー１の利得

	１期	２期	３期	４期	５期	６期	７期	８期	９期	10期	11期	12期
Player1	L	R	R	R	L	L	L	L	R	R	R	?
Player2	L	L	R	L	R	L	R	L	R	L	L	?
１の利得	1	0	1	0	0	1	0	1	1	0	0	?

表３

	L	R
L	1, 0	0, 1
R	0, 1	1, 0

表４

	L	R
L	1, 1	0, 0
R	0, 0	1, 1

表５　Soda Game：常にＲを選択

	１期	２期	３期	４期	５期	６期	７期	８期	９期	10期	11期	12期
Player1	R	R	R	R	R	R	R	R	R	R	R	?
Player2	L	L	R	L	R	L	R	L	R	L	L	?
１の利得	0	0	1	0	1	0	1	0	1	0	0	?

表６　Soda Game：常にＬを選択

	１期	２期	３期	４期	５期	６期	７期	８期	９期	10期	11期	12期
Player1	L	L	L	L	L	L	L	L	L	L	L	?
Player2	L	L	R	L	R	L	R	L	R	L	L	?
１の利得	1	1	0	1	0	1	0	1	0	1	1	?

各々のプレイヤーはこれらの財から利得を得るが、自分自身の利得についてはわかるが、相手が得る利得についてはわからないものとする。このような相手の得る利得がわからない（利得の分布さえ分からないかもしれない）ゲームを**不確実性ゲーム**（uncertain game）と呼ぶことにする。

以下、このゲームを繰り返し行い、十一回の結果が表2のようになったとしよう。一番下の数字がプレイヤー1の各期の利得である。割引を考えなければ、プレイヤー1の利得は十一期合計で5になっている。

プレイヤー1の立場で考えたとき、十二回目はLとRのどちらを選んだらよいか。このゲームでは相手の利得がわからないので、そもそも自分（プレイヤー1）がどんなゲームに参加しているのかがわからない。協調ゲーム（表3）であるかもしれないし、コイン合わせ（matching pennies）ゲーム（表4）であるかもしれない（もちろんそれら以外かもしれない）。

ここで、プレイヤー1がすべての期で同じ行動を選択しなければならないことを条件として、このゲームをやり直せる機会が与えられたとしよう。むろんやり直しが可能になったとき、相手のプレイヤー2がどのように対応してくるかはわからないが、ここでは非常に単純に、相手は行動を変更してこないと考えよう。

このとき、プレイヤー1が「常にRを選択」（このような行動様式をR^*と記す）に変更すると、利得は表5のようになる。このときにはプレイヤー1の利得は4に減少する。

また、プレイヤー1が「常にLを選択」（L^*とする）に変更すると、プレイヤー1の利得は7に増加する（表6）。

したがって、

・R^*を取らなかったことについては**後悔しない**が、

・L^*を取らなかったことには後悔する

であろう。よって後悔を減らすため、将来はLの選択にもっと大きなウェイトを置くべきと考えるだろう。

以下ではより一般的に、一人のプレイヤーが「自然（Nature）」を相手にプレーする場合を考える。ここで自然とは、相手のプレイヤーと考えてもよいし、自分が置かれている状態と考えてもよいが、いずれにせよ上述のように、意思決定を行うプレイヤーとのinteractiveな状況は考えない。自然の選ぶ行動（あるいは状態）は観察できるとする。もちろん意思決定プレイヤーの利得は、自身の意思決定と自然の選ぶ行動に依存する。

2. Hannan, Fudenberg and Levine, Hart and Mas-Colell モデル

有限個の要素からなる状態の集合をΩとする。「自然」はゲームの行われる各期においてΩから一つの状態を選ぶ。また、この自然とゲームを行う一人のプレイヤーを想定し、彼女／彼は行動の集合Aから一つを選ぶとする。このゲームが繰り返し行われ、各n期において、自然が状態$\omega_n \in \Omega$を選び、プレイヤーの行動が$a_n \in A$であったとき、プレイヤーは利得$u(\omega_n, a_n)$を得る。また、$h_n = (\omega_1, \ldots, \omega_n; a_1, \ldots, a_n)$で長さnの履歴を表す。便宜上、履歴がない初期状態を空履歴h_0と表し、この空履歴と有限個の履歴すべてからなる集合をHと表記する。

行動の集合A上の確率分布を$\Delta(A)$とする。プレイヤーの取る戦略はHから$\Delta(A)$への関数$f: H \to \Delta(A)$で与えられる。

ゲームは次のように進行する。まず、第1期に関数fが最初の行動（これをa_1と書く）が従う分布を決定する（$f(h_0) \in \Delta(A)$）。次の第2期では、実現した状態ω_1とa_1に基づいて、fは第2期の行動a_2が従う分布を決定する。第3期以降も同様である。つまりn+1期の行動a_nの従う確率分布は$f(h_n) = f(\omega_1,\ldots,\omega_n;a_1,\ldots,a_n) \in \Delta(A)$である。また、無限の状態の列$\omega = (\omega_1, \omega_2, \ldots, \omega_n, \ldots)$に対して、戦略$f$と$\Omega$は無限の行動の列$A^N$上の確率分布を導く（この確率分布を$(\Omega, f)$と書くことにする。ただし$N$は自然数の集合）。

さて、今まで取ってきた戦略を後悔するかどうかは、その戦略を取らずに「他の戦略」を取ったときに利得が増加するかどうかにかかっている。「他の戦略」として、まず一番単純に過去の履歴に関係なく、ある戦略$a \in A$をずっと取り続ける戦略a^*を考えよう。状態の列$\omega=(\omega_1, \omega_2, \ldots, \omega_n)$に対して、t期の行動を$a_t (t = 1,2,\ldots,n)$とすると、t期における$a^*$との利得の差は、$u(\omega_t, a_t) - u(\omega_t, a)$であり、第1期からn期までの利得の差の平均は

$$L_n(f, a^*) = \frac{\sum_{t=1}^{n} [u(\omega_t, a_t) - u(\omega_t, a)]}{n}$$

である。ここでゲームの回数を大きくしていったとき（$n \to \infty$）、$\liminf_n L_n(f,a^*) \geq 0$であるならば、$a^*$を取らなかったことを後悔しないだろう。Hannan[3]によって、次のことが証明されている。

定理1（Hannan[3]）

任意の$a \in$と任意の状態の列$\omega = (\omega_1, \omega_2, \ldots, \omega_n, \ldots)$に対して、$a^*$と比べ後悔しない戦略$f$が存在する。

Hannan[3] の結果は、同一の行動 a がずっと取られたときという極めてナイーブな戦略と比較した場合を扱っている。この結果は、

（1）期の列を互いに素な部分集合に分割し、ある部分集合上では特定の行動 $a \in A$ をとる戦略、

（2）二つの異なる行動 a, b に対して、a がプレーされていた場合にはそれをすべて b で置き換えるような戦略

に対して後悔しない戦略 f を求めるという二つの方向に拡張された。（1）については Fudenberg and Levine [2]、（2）については Hart and Mas-Colell[4] によって分析されている。Fudenberg and Levine[2] では、まず期の集合（自然数の集合 N と同一視できる）の分割 $B_1, B_2, \ldots B_k$ を考える。そして B_j については任意の行動を a で置き換え、B_j 以外のところでは行動を変更しないという戦略を $a^* \mid B_j$ と記すことにする。このとき Fudenberg and Levine[2] の結果は次のように書ける（Lehrer[6]）。

定理2（Fudenberg and Levine[2]）

任意の期の集合の分割 $B_1, B_2, \ldots B_k$ について、任意の状態の列 $\omega = (\omega_1, \omega_2, \ldots, \omega_n, \ldots)$ と任意の $a \in A$, j（ただし $1 \leq j \leq k$）に対して、$a^* \mid B_j$ と比較して後悔しない戦略 f が存在する。

Hart and Mas-Colell[4] では、任意の行動のペア (a, b) を考え、ある期 t の行動が a であった場合それを b で置き換え、a 以外の場合には置き換えない戦略を考える。これは関数

$g_{a,b} : H \times A \to A - \{a\}$、

$g_{a,b} = (h_{t-1}, a) = b$

$g_{a,b} = (h_{t-1}, c) = c : c \neq a$

で表せる。このとき Hart and Mas-Colell[4]（Theorem A）の結果は、次のようになる。

定理3（Hart and Mas-Colell[4]）

任意の状態の列 $\omega = (\omega_1, \omega_2, ..., \omega_n, ...)$ に対して、$g_{a,b}$ と比較して後悔しない戦略 f が存在する。

3. Lehrer モデル

さらに Lehrer[6] では、「置換スキーム」と「アクティブ関数」という二つの関数を導入することで、Fudenberg and Levine[2] と Hart and Mas-Colell[4] の分析の一部をスペシャルケースとして含む、より一般的な学習プロセスが提案されている。

「置換スキーム」は過去の履歴において採用してきたある行動を、別の行動で置き換える関数

$$g : H \times A \to A$$

である。行動の置換には様々なケースが考えられる。前述の Hart and Mas-Colell[4] は、任意の二つの行動のペアを考え、その間で置換が起きるという特別の場合を扱っていると見なせる。

さらに Lehrer[6] では、プレイヤーが特定の期において戦略のパフォーマンスを重視するケースを想定する。例えばプレイヤーが、休暇日や週末には高い利得が得られるようにしたいと望むような場合である。そのために、こうした「特別な期（日）」を特定化する関数、アクティブ関数を次のように定義する。

$$I : H \times A \to \{0, 1\}$$

このアクティブ関数により、$I(h_{n-1}, a_n) = 1$ であるなら、行動 a_n の取られた第 n 期はアクティブであるという

ことになる。履歴が $h_{n-1} \in H$ であるとき、n 期までに関数 I がアクティブになった、つまり関数の値が 1 になった回数を

$$I^n(h_{n-1}, a_n) = \sum_{t=1}^{n} I(h_{t-1}, a_t)$$

とする。これは履歴にしたがって n 期までに何回アクティブ期があったかを表している。ここで二つの関数のペア $(g, I): H \times A \to A \times \{0, 1\}$ により、ある期がアクティブであるかどうか、またその期にどのような行動の置換が行われるかが定められることになる。

次に戦略 f と、置換スキーム g によって置き換えられた戦略が、アクティブな期を通して平均的にどちらが良いか、あるいは、戦略 f を取ることで後悔しないかを考えるよう。したがって以下ではプレイヤーは f と (g, I) を比較すると考えることができる。

いまある期 t において行動 a_t が取られるとする。このときの利得は $u(\omega_t, a_t)$ である。これに対して、置換スキーム g で行動を置き換えた場合、t 期の行動は $g(h_{t-1}, a_t)$ であるから、t 期の利得は $u(\omega_t, g(h_{t-1}, a_t))$ となる。したがってプレイヤーにとって、t 期において g で行動を変えた場合との利得の差は、$u(\omega_t, a_t) - u(\omega_t, g(h_{t-1}, a_t))$ となる。この差が生まれるのは t 期がアクティブ関数 I によってアクティブになる場合である。

以上から、第 n 期までの利得の差の平均は、

$$L_n(f, g) = \frac{\sum_{t=1}^{n} I(h_{t-1}, a_t)[u(\omega_t, a_t) - u(\omega_t, g(h_{t-1}, a_t))]}{I^n(h_{n-1}, a_n)}$$

となる。ゲームの回数を$n\to\infty$と考えたとき、アクティブ関数がアクティブになる回数が無限になるようなほとんどすべての行動の列に対して、$\liminf_n L_n(f,g)\geq 0$となる場合、戦略fはペア(g, I)と**同等以上に良い**と言えるだろう。これは、アクティブな期を通じた利得の平均がfに従った行動をとる方がgによって変更するときに得られる利得より高い、つまり戦略fを取ったことを後悔しないということである。

ここですべての(g, I)からなる集合Rを考える。[3] Lehrer[6]よって次が示されている。

定理4（Lehrer[6]）

任意の状態の列$\omega=(\omega_1, \omega_2, \ldots, \omega_n, \ldots)$に対して、ほとんどすべての$(g, I)\in R$と比べ後悔しない戦略$f$が存在する。

これら一連の「後悔なし(no regret)」の定理はBlackwell[1]のapproachability定理によって証明されている。特にLehrer[6]では、オリジナルのBlackwell定理を無限次元に拡張したapproachability定理（Lehrer[5]）が使われている。[4]

三．有限記憶

次にN人による標準形ゲームを考えよう。各プレイヤーは有限個の要素から成る行動集合A_iを持つものとする。$A=\Pi_{i\in N}A_i$とし、一般に各プレイヤーの行動の組は$a=(a_i, a_{-i})\in A$と表され、a_iはプレイヤーiの行動、

a_{-i}はプレイヤーi以外のプレイヤーの行動の組を示す。今まで同様$u_i: A \to \mathbf{R}$（実数体）をプレイヤーiの利得関数とする。行動の組が(a_i, a_{-i})のとき、プレイヤーiの利得は$u_i(a_i, a_{-i})$である。

このときすべての$i \in \mathrm{N}$と$a_i, b_i \in \mathrm{A}$に対して、

$$\sum_{a_{-i}} u_i(b_i, a_{-i}) q(a_i, a_{-i}) \leq \sum_{a_{-i}} u_i(a_i, a_{-i}) q(a_i, a_{-i})$$

を満たすA上の分布$\mathrm{q}: \mathrm{A} \to \Delta(\mathrm{A})$を**相関均衡**という。

このゲームが各期t(t=1, 2,…)毎に繰り返し行なわれる場合を考える。t期の行動の組を(a^t_i, a^t_{-i})とすると、Hart and Mas-Colell[4]では、上述のように行動a^t_iをb_iで置き換える場合を想定していた。ここでの置換スキームは、

$$g_{a^t_i,\, b_i}(a^t_i, a^t_{-i}) = b$$
$$g_{a^t_i,\, b_i}(a^k_i, a^t_{-i}) = a^k_i: \; a^k_i \neq a^t_i$$

と書ける。

また、t期までに(a_i, a_{-i})組が生じる確率

$$\mu(a_i, a_{-i}) = \frac{1}{t} \left| \{ 1 \leq k \leq t : (a^t_i, a^t_{-i}) = (a_i, a_{-i}) \} \right|$$

を**経験分布**（empilical distribution）と呼ぶ。

Hart and Mas-Colell[4]では、$t \to \infty$のときこの経験分布が（ほとんど確実に）相関均衡の集合に収束することが明らかにされた。また、経験分布が相関均衡の集合へ収束することが、後悔しない戦略が存在することと同値になることも示されている。しかしながら、経験分布が実際にどの「点」に収束するのかといった収

束の細かい特徴についてはほとんど知られていない（Hart and Mas-Colell[4] Section 2）。

この Hart and Mas-Colell[4] のモデルをはじめ、いままで考えてきたモデルではプレイヤーは履歴について完全に記憶していた。つまり、過去に行われた各期でのゲームの利得を完全に知っており、それをもとに戦略を変更するかどうか検討するという行動様式を考えてきたのである。

しかし過去の自分の行動や利得をすべて記憶しているという状況は、現実的ではない場合が多いであろう。その点に注目し、限られた過去の記憶しか持たないプレイヤーを分析したのが Saran and Serrano [10] である。こうすることで、Hart and Mas-Colell[4] よりも、経験分布についてのシャープな結果が得られる。以下それを紹介する。

Saran and Serrano [10] では、プレイヤーが戦略を考えるに当たり初期時点からすべての期の状況を考慮に入れるのではなく、直近の m 期のみ考慮する、言い換えると、直近の m 期の記憶しか持たない場合を想定した。プレイヤー i の（t 期から遡って）直近 m 期の平均利得は、

$$L_i^m(a_i^t, a_{-i}^t) = \frac{1}{m}\sum_{k=t-m+1}^{t} u_i(a_i^t, a_{-i}^t)$$

となる。

次に、直近 m 期間の行動で、t 期の行動 a_i^t と同じ行動を取った期をすべて b_i で置き換える場合を考えよう。ただし a_i^t 以外の行動を取っている場合は変更しないものとする。したがって直近 m 期間の置換スキームは、Hart and Mas-Colell[4] と同じものになる。この置換スキームを使った m 期の平均利得は、

$$L_i^m(b_i, a_{-i}^t) = \frac{1}{m}\sum_{k=t-m+1}^{t} u_i(g_{a_i b_i}, a_{-i}^t)$$

と書ける。直近の m 期間における行動a_i^tをb_iに変換したときとしないときの平均利得の差が正であった場合には、t + 1 期には正の確率で行動b_iを取り、また、そのままa_i^tを取り続ける確率も正であるとする。つまり、$\hat{L}(b_i, a_i^t) = L_i^m(b_i, a_{-i}^t) - L_i^m(a_i^t, a_{-i}^t)$とおき、$\hat{L}(b_i, a_i^t) > 0$の場合に確率$q(\hat{L}(b_i, a_i^t)) > 0$で行動$b_i$への変更が起きる。ただし、$\sum_{b_i \neq a_i} q(\hat{L}(b_i, a_i^t)) < 1$と仮定する。

このような状況で、直近 m 期の行動の組の履歴を一つの**状態**と考える。したがって状態の集合はA^mである。確率分布 q が与えられたとき、こうした有限記憶によって推移するプロセスは、A^m上の非周期的マルコフ・プロセス$M(q)$になる。

任意の行動の組$(a_i, a_{-i}) \in A$に対して、プレイヤーiの同等以上の反応（same or better reply）からなる集合を

$$R_i(a_i, a_{-i}) = \{b_i \in A_i \mid b_i = a_i \text{ あるいは } u_i(b_i, a_{-i}) > u_i(a_i, a_{-i})\}$$

とする。さらに(a_i, a_{-i})に対する同等以上の反応対応（same or better reply correspondence）を

$$R_G(a_i, a_{-i}) = \prod_{i \in N} R_i(a_i, a_{-i})$$

とする。

A の部分集合$\hat{A}$で、すべての$a \in \hat{A}$に対して、$R_G(a) \subseteq \hat{A}$なる場合、$\hat{A}$を同等以上の反応に関して閉じている集合（Closed Under Same-Or-Better Reply set）という（以下 CUSOBR 集合）。

$\hat{A}$をCUSOBR集合とする。R_Gの定義から任意の$a \in \hat{A}$について$a \in R_G(a)$である。したがって$a \in \bigcup_{a \in \hat{A}} R_G(a)$である。また逆に，任意の$a \in \bigcup_{a \in \hat{A}} R_G(a)$に対して，ある$a' \in \hat{A}$が存在し，$a \in R_G(a')$である。CUSOBR集合の定義から$a \in R_G(a') \subseteq \hat{A}$である。よって，$\hat{A}$がCUSOBR集合の場合は

$$\widetilde{R_G}(\hat{A}) = \bigcup_{a \in \hat{A}} \left(\prod_{i \in N} R_i(a_i, a_{-i}) \right)$$

の不動点になっている（つまり$\widetilde{R_G}(\hat{A}) = \hat{A}$）。

また，Aの部分集合$\hat{A}$が直積集合で，かつすべての$a \in \hat{A}$に対して，$R_G(a) \subseteq \hat{A}$なる場合，$\hat{A}$を同等以上の反応に関して閉じた行動の組の直積集合（Product set of action profiles that is Closed Under Same-Or-Better Reply set）という（以下PCUSOBR集合）。CUSOBR集合の場合と同様に，$\hat{A}$がPCUSOBR集合の場合は

$$\widehat{R_G}(\hat{A}) = \prod_{i \in N} \left(\bigcup_{a \in \hat{A}} R_i(a_i, a_{-i}) \right)$$

の不動点になっている（つまり$\widehat{R_G}(\hat{A}) = \hat{A}$）。

また，真部分集合となるCUSOBR集合を持たないCUSOBR集合（or PCUSOBR集合）を最小CUSOBR集合（or 最小PCUSOBR集合）という[5]。

これらの集合を「グラフ」を使って考える。まずゲームの行動の組を頂点とする。いまA_iは有限であるから頂点も有限個になる。各頂点と他の頂点を結ぶ辺を考える（こうした頂点と辺からなる構造を**グラフ**という）。

いま次のようにグラフの辺に「向き」を付ける（辺に向きの付いたグラフを有向グラフという）。頂点$(a_1, a_2, \ldots, a_N)$と$(a'_1, a'_2, \ldots, a'_N)$について、$(a_1, a_2, \ldots, a_N) \neq (a'_1, a'_2, \ldots, a'_N)$かつ$(a'_1, a'_2, \ldots, a'_N) \in R_G(a_1, a_2, \ldots, a_N)$であるときに$(a_1, a_2, \ldots, a_N)$から$(a'_1, a'_2, \ldots, a'_N)$へ向かって矢印を付ける。例えば、次の利得表7で表される戦略形ゲームを考える（Saran and Serrano[10] Sec.3）。

表7　例（Saran and Serrano[10]）

	L	C	R
T	0, 0	0, 0	1, 1
U	1, 0	0, −1	0, 0
M	0, 1	2, 0	0, 0
D	2, 0	0, 1	0, 0

このゲームでは、各$R_1(i = 1, 2)$は次のようになる。

$R_1(T, L) = \{T, U, D\}$; $R_2(T, L) = \{L, R\}$

$R_1(T, C) = \{T, M\}$; $R_2(T, C) = \{C, R\}$

$R_1(T, R) = \{T\}$; $R_2(T, R) = \{R\}$

$R_1(U, L) = \{U, D\}$; $R_2(U, L) = \{L\}$

$R_1(U, C) = \{U, M\}$; $R_2(U, C) = \{C\}$

$R_1(U, R) = \{T, U\}$; $R_2(U, R) = \{R\}$

$R_1(M, L) = \{U, M, D\}$; $R_2(M, L) = \{L\}$

$R_1(M, C) = \{M\}$; $R_2(M, C) = \{L, C\}$

$R_1(M, R) = \{T, M\}$; $R_2(M, R) = \{L, R\}$

$R_1(D, L) = \{D\}$; $R_2(D, L) = \{L, C\}$

図1　同等反応グラフ

$R_1(D, C) = \{M, D\}$ ； $R_2(D, C) = \{C\}$

$R_1(D, R) = \{T, D\}$ ； $R_2(D, R) = \{C, R\}$

これから次の R_G 群が得られる。

$R_G(T, L) = \{(T, L), (T, R), (U, L), (U, R), (D, L), (D, R)\}$

$R_G(T, C) = \{(T, C), (T, R), (M, C), (M, R)\}$

$R_G(T, R) = \{(T, R)\}$

$R_G(U, L) = \{(U, L), (D, L)\}$

$R_G(U, C) = \{(U, C), (M, C)\}$

$R_G(U, R) = \{(T, R), (U, R)\}$

$R_G(M, L) = \{(U, L), (M, L), (D, L)\}$

$R_G(M, C) = \{(M, L), (M, C)\}$

$R_G(M, R) = \{(T, L), (T, R), (M, L), (M, R)\}$

$R_G(D, L) = \{(D, L), (D, C)\}$

$R_G(D, C) = \{(M, C), (D, C)\}$

$R_G(D, R) = \{(T, C), (T, R), (D, C), (D, R)\}$

以上から有向グラフをかくと図1のようになる（ここでは自分自身に向いている辺は省いている。以下では同等反応グラフと呼ぶ）。

同等反応グラフで自分以外へ出ていく辺がないものをシンクと呼ぶ。このグラフでは頂点 (T, R) がシンクである。自分以外へは向かわないということは、誰も行動を変更するのものがいない、つまりシンクは純粋戦略ナッシュ均衡になる。また、頂点 (a^l, b^l) (ここで $a^l = T, U, M, D$; $b^l = L, C, R$) から (a^{l+1}, b^{l+1}) へ向けて繋がる辺があるとき (a^l, b^l) は (a^{l+1}, b^{l+1}) へ隣接するといい、頂点 $(a^1, b^1), (a^2, b^2), \ldots, (a^L, b^L)$ が隣接した頂点の列のとき同等反応パスという。図1のグラフから最小 CUSOBR 集合は {(T, R)} と {(U, L), (M, L), (M, C), (D, L), (D, C)} であることがわかり[6]、また最小 P CUSOBR 集合は {(T, R)} だけである。

ある状態の集合 $\hat{H}$ があり、もし上記の学習プロセスがいったん $\hat{H}$ のどれかに達したら、$\hat{H}$ から抜け出ることがなく、かつそうした性質を満たす真部分集合を含まない場合、$\hat{H}$ を**再帰クラス** (recurrent class) という。

命題1 (Saran and Serrano[10] Prop.4.1)

任意の状態の集合 $\hat{H} \subseteq A^m$ に対して、$\hat{H}$ のある状態でプレーされる行動の組すべてを $A(\hat{H})$ で表すとする。$\hat{A}$ がゲームの最小 PCUSOBR 集合であるとき、$A(\hat{H}) \subseteq \hat{A}$ なるマルコフ・プロセス $M(q)$ の再帰クラス $\hat{H}$ が存在する。

また、ゲームが**同等反応の下で弱非周期的**であるとは、いかなる頂点からも少なくとも一つのシンク（ナッシュ均衡）へ向かう同等反応パスがある場合をいう。したがって、もしゲームが同等反応の下で弱非周期的であるなら、すべての CUSOBR 集合は純粋戦略ナッシュ均衡を含む。

これらから次が得られる。

系1（Saran and Serrano[10] Cor.4.2）

同等反応の下で弱非周期的なゲームを考える。$A(\hat{H}) \subseteq A$ が純粋戦略ナッシュ均衡であることと、$\hat{H}$ がマルコフ・プロセス $M(q)$ の再帰クラスになることとは同値である。

この系1によって、Hart and Mas-Colell[4] ―記憶が有限でない場合―では不明であった点が明らかになる。つまり、期ごとの行動の組 a がほとんど確実に有限時間で純粋戦略ナッシュ均衡に達するということである。(7)

四．結びに代えて

本稿では、学習理論の中でも特に心理学で用いられることの多い「強化学習」、特に「後悔の度合いに応じて行動を選ぶ戦略」を概観した。この分野の理論研究も進んで来てはいるが、実験で得られた結果をどこまで説明できるかという点に関しては、経済学で使われてきた「信念学習」との比較で、どちらが一般に優れているかといった結論は得られていないようである。

強化・信念学習双方のさらなる研究が望まれるが、いかにシンプルで実験データに適合したモデルを作れるかが問題となる。「後悔しない」行動を取る確率が高くなるという考え方はわかりやすいが、どの程度実用性があるかは今後十分研究・検討されなければならない。本稿後半で、Saran and Serrano[10] による有限の記憶への拡張のケースを見たが、これ以外にも指数関数的に記憶を割り引いて行く方法などが考えられている（Marden,Arslan and Shamma[7]）。

参考文献

[1] Blackwell,D.,“A vector valued analog of mini-max theorem”, Pacsfic Jpurnal of Mathematics, 6, 1956, 1--8.

[2] Fudenberg,D. and Levine,D., “Conditional Universal Consistency”, Games and Economic Behavior, 29, 1999, 104--130.

[3] Hannan,J., “Approximation to Bayes Risk in Reapeated Plays”, in Contribution to the Theory of Games, 3, 97--139. Princeton Univ. Press, 1957

[4] Hart,S. and Mas-Colell,A., “A Simple Adaptive Procedure Leading to Correslated Equilibrium”, Econometrica 68, 2000, 1127--1150.

[5] Lehrer, E., “Approachability in infinite dimension space”, International Journal of Game Theory 31, 2002, 253--268.

[6] Lehrer, E., “A wide range no-regret theorem”, Games and economic behavior 42, 2003, 101--115.

[7] Marden,J.R.,Arslan,G. and Shamma,J.S., “Regret based dynamics: Convergence in weakly acyclic games”, AAMAS 2007, Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems, 2007.

[8] Nash,J.F., “Equilibrium Points in N-Person Games”, Proceedings of the National Academy of Science of the United States of America 36, 1950, 48--49.

[9] Nash,J.F., “Non-Cooperative Games”, Annals of Mathematics} 54, 1951, 286--295.

[10] Saran, R. and Serrano, Roberto., “Regret matching with finite memory”, Working Paper Institute Madriedo de Estudios Aranzados (IMDEA) Ciencias Sociales, 2010.

[11] 川越敏司 『行動ゲーム理論入門』 NTT出版、二〇一〇年。

[12] van Damme, E., Stability and Perfection of Nash Equilibria, Springer-Verlag,Berlin, 1987.

[13] Young, H.P., {\it Strategic Learning and its Limits}, Oxford Press, 2004.

注

（1）八十年代くらいまで、ゲーム理論での研究はプレイヤーの合理性を仮定し、ナッシュ均衡をいかに絞り込むかに焦点が置かれていた（ナッシュ均衡の「精緻化」と呼ばれる）。こうした問題に関する一つの到達点が van Damme [12] である。

（2）以下では数学的に厳密な議論は行っていない。また、regret matching モデルについて網羅するものでもない。

（3）ここで集合Rは可測で、μをR上の確率測度とする。

（4）Blackwell の approachability 定理については Yong [13] 第 4 章に簡潔な解説がある。

（5）もし最小 CUSOBR が直積集合であるなら、それは最小 PCUSOBR 集合であり、さらに純粋戦略ナッシュ均衡は一点からなる最小 CUSOBR 集合でもあり、最小 PCUSOBR 集合でもある。また、すべての最小 PCUSOBR 集合はどれかの最小 CUSOBR 集合に含まれる。

（6）グラフから頂点$\equiv(T, R)$は 1 点のみで同等以上の反応に関して「閉じて」おり、$\{(U, L),(M, L),(M, C),(D, L),(D, C)\}$も「閉じて」いることがわかる。

（7）この系では「同等反応の下で弱非周期的なゲーム」という条件がつけられているが、Saran and Serrano [10] では、さらに、各プレイヤーが直近 m 期の中からランダムに s 期抜き取り、後悔の度合いを考慮するという場合も分析されている。この場合$\frac{s}{m}$が十分小さければ、$A(\vec{H})$が最小 PCUSOBR 集合であることと、$\vec{H}$がマルコフ・プロセス$M(q)$の再帰クラスになることとは同値であることが示されている（Saran and Serrano [10] Cor.4.3）。