# 非定常多腕バンディットゲームでの社会的学習エージェントの相転移的な振る舞い

Phase transitive behavior of social learning collectives in restless multi-armed bandit game

守真太郎 *1　　中山一昭 *2
Shintaro Mori　　Kazuaki Nakayama

*1北里大学理学部物理学科　　*2信州大学理学部数理科学科
Faculty of Science, Kitasato University　　Faculty of Science, Shinshu University

オンラインメディアで観察されるエコー・チェンバー現象が社会的学習の効率を低下させるという報告がある。この現象の単純なモデルとして、非定常多腕バンディットゲームでの社会的学習エージェントシステムを導入した。バンディットはリターンが 1,0 の二値の状態をとるとする。エージェントの状態をエージェントが記憶しているバンディットの状態 $\sigma \in \{0,1\}$ で表すとする。各ターンでエージェントがランダムに選択され、$\sigma = 1$ ならそのまま、$\sigma = 0$ なら確率 $1-r$ で独自学習（成功確率 $q_I$）か、確率 $r$ で社会的学習 (成功確率は他のエージェントがリターン 1 のバンディットを知っていれば $q_O$、知らなければ 0 とする) を行ってリターンが 1 のバンディットを発見しようとする混合戦略を採用する。エージェント数は $N$ とし、各ターン後、リターン 1 のバンディットは確率 $q_c/N$ でリターン 0 のバンディットに変化し、リターン 1 のバンディットが新たに追加される。全エージェントの $r$ が同一のとき、$r$ が小さいときはリターン 1 のバンディットを記憶するエージェント数とリターン 0 を記憶するエージェント数が均衡し、期待リターンは $r$ の増加関数となる。$r$ が大きくなるとエージェントの期待リターンはピークアウトし、$r > r_{Pareto}$ では期待リターンは減少する。さらに $r \simeq 1$ では $r = 0$ の独自学習のみのエージェントの期待リターンを下回る。このとき、リターン 0 のバンディット情報がエージェント間に循環し続け、リターン 1 のバンディットを見つけられないエコー・チェンバー状態となっている。この系を期待リターンを利得表とするゲームと考え、ESS ナッシュ解 $r_{Nash}$ を見出した。$N$ が大きくなると $r_{Nash}$ での期待リターンは $r_{Pareto}$ での期待リターンを大きく下回るようになる。また、エージェントが $r$ を期待リターンが増加する方向に更新したとき、$r_{Nash}$ に収束する（のまわりでゆらぐ）ことを数値的に示す。

## 1. 社会的学習とエコーチェンバー

社会的学習 (social learning) とは、他個体の選択・行動を真似することで行う学習であり、トライ&エラーで行う独自学習 (individual learning) と比べてコストが低く、獲得できる情報の精度もそれほど悪くないというメリットがある。社会的学習の有効性は文明・文化、さらにはヒトの適応能力の高さを説明すると考えられている。一方、A.Rogers は、社会的学習を行う個体の比率の増加は社会的学習の適合度 (fitness) を減少させ、均衡状態では独自学習と社会的学習の適合度が同じになり、社会的学習は独自学習に対して環境への適合度を改善しないことを示した。この結果は上述の社会的学習の有効性を否定するものであり、Rogers パラドックスと呼ばれている。一方、[Rendell 2010] は非定常多腕バンディット (restless multi-armed bandit, 以下 rMAB と呼ぶ) でのエージェントの学習アルゴリズムを競うトーナメントを行い、バンディット情報の獲得において社会的学習のみを用いるアルゴリズムのパフォーマンスがよいという一般的な結論を得た。ここでリターンの異なる多数のスロットマシンのことを多碗バンディットと呼び、非定常はリターンがランダムに変化することを意味する。多腕バンディットにおいて有限回の試行でリターンを最大にする問題をバンディット問題と呼び、情報科学での最適化問題でよく扱われる対象である。リターンが時間変化する非定常の場合、バンディット問題は最適解が知られていず、そのため、[Rendell 2010] では学習アルゴリズムのトーナメントの舞台として選ばれたのであった。

一方、社会的学習の有効性をデータで検証したのが MIT の A.Pentland のグループである [Pentland 2014]。eToro というオンライン証券会社では、ユーザーは自分のポートフォリオを公開して、参照したユーザーからリターンを得ることができるサービスを展開している。このポートフォリオの参照は社会的学習であり、ポートフォリオを参照する度合いと収益率の関係を調べたところ、参照の度合いが高くなると収益率は増加するが、ある点でピークアウトし、収益率は減少に転じることを見出した。コピー比率の高く、収益率の低い投資家集団では、古い情報がぐるぐる巡り、新しい有効な情報が入ってこないエコーチェンバー状態 (Echo chamber) が見られることも報告している。

本稿では、リターンが 0,1 の 2 値の rMAB で、社会的学習と独自学習の混合戦略を採用したエージェント系の示す統計的な振る舞いを解析し、エコーチェンバー状態とパフォーマンスの低下を説明する [Mori 2016]。また、ゲーム論の枠組みで最適な戦略を議論し、ESS ナッシュ解は独自学習よりもパフォーマンスが高いことを示し、Rogers パラドックスを混合戦略の枠組みで解決する [Nakayama 2017]。また、エージェントが期待リターンを改善する方向に社会的学習の比率 $r$ を変化させたとき、ESS ナッシュ均衡状態に収束する (のまわりでゆらぐ) ことを数値的に示す。

## 2. rMAB ゲーム

リターンが 0,1 の 2 値の rMAB ゲームを考える。バンディットは多数あり、そのうち、リターン 1 のバンディット数は 1 とし、リターン 0 のバンディット数は多数とする。エージェントはバンディットを 1 つだけ記憶でき、独自学習でリターン 1 のバンディットを発見する確率は $q_I \in (0,1]$、社会的学習でリターン 1 のバンディットを発見する確率は、リターン 1 のバンディットを記憶するエージェントが 1 以上なら $q_O \in (0,1]$、0 なら 0 とする。リターン 1 のバンディットを記憶したエージェント数を $N_1$ とし、社会的学習の確率を $r$、独自学習の確率を $1-r$ とすると、エージェントがリターン 1 のバンディットを

連絡先: shintaro.mori@gmail.com*1

発見する確率は

$$(1-r)\cdot q_I + r\cdot q_O\cdot(1-\delta_{N_1,0})$$

と書くことができる。エージェント数を $N$、エージェントを $n=1,2,\cdots,N$ でラベルし、エージェント $n$ の社会的学習の確率を $r_n\in[0,1]$ で表すとする。また、エージェント $n$ がリターン 1 のレバーを記憶しているかいないかを 2 値の変数 $\sigma_n\in\{0,1\}$ で記述する。

ゲームのルールは以下の通りである。毎ターン、エージェント $n\in\{1,\cdots,N\}$ がランダムに選ばれ、$\sigma_n=1$ のときは、そのまま、$\sigma_n=0$ のときは、確率 $r_n$ で社会的学習、確率 $1-r_n$ で独自学習を行うとする。エージェントのアクションののち、リターン 1 のバンディットの状態が確率 $q_C/N$ で変化する。バンディットが変化したとき、$\sigma_n=1$ のエージェント $n$ の記憶は消去され、$\sigma_n=0$ と変化する。$N$ ターンで 1 モンテカルロステップ (MCS) とし、1[MCS]=N ターンでバンディットが変化する確率は $N$ が十分大きいとき $1-(1-q_C/N)^N=1-e^{-q_C}$ と評価できる。$N\to\infty$ の極限操作のためにバンディットの変化の確率を $q_c/N$ と $N$ でスケールしている。

## 3. 均衡状態から振動状態への変化

全エージェントが同じ混合戦略 $r_n=r, n=1,\cdots,N$ を採用するとする。$N$ が十分大きいとき、$\sigma_n=1$ となるエージェント数 $N_1$ と $\sigma_n=0$ のエージェント数 $N_0$ が均衡する。つまり、$\sigma=1$ の状態と $\sigma=0$ の状態の間のエージェントの学習とバンディットの変化によるエージェントの出入りが釣り合う。この釣り合い条件を解くことで、$\sigma=1$ のエージェントの比率 $x=N_1/N$ を求めると、

$$x=\frac{A+B}{q_C+A+B}, A=r\cdot q_O, B=(1-r)\cdot q_I$$

となる。$q_O>q_I$ のとき $x$ は $r$ の単調増加関数となる。

一方、$N$ が有限のとき、バンディットが変化しない時間が長くなるとエージェントはリターン 1 のバンディットに集中する。そのバンディットが確率 $q_C/N$ でリターン 0 のバンディットに変化すると全エージェントはリターン 0 のバンディットを記憶することになる。学習によりエージェントはリターン 1 のバンディットを探すことになるが、$r$ が大きいときは社会的学習を主に使い、$N_1=0$ では社会的学習の成功確率 $q_O(1-\delta_{N_1,0})$ はゼロなので、リターン 1 のバンディットを見つけるまでのターンが非常に長くなる。その結果、確率変数 $\hat{x}=N_1/N$ の定常状態での期待値は $N$ が非常に大きく釣り合った状態での値 $x$ に比べて大幅に低下する。$\hat{x}N$ の期待値 $\mathrm{E}(x)$ は次の self-consistent 方程式を満たす。

$$\mathrm{E}(x)=(1-\frac{q_C}{N})\frac{A+B-AP(0)}{q_C+(1-\frac{q_C}{N})(A+B)}$$

この式で $P(0)$ は $N_1=0$ となる確率を表し、具体的には次の式で計算される。

$$P(0)=\frac{q_C}{q_C+(N-q_C)B}.$$

$q_I=0$ または $r=1$ のとき $B=0$ となり、$P(0)=1$ となり系は時間変化しなくなるので、$q_I>0, r<1$ とする。$N\to\infty$ の極限では $P(0)\to 0, q_C/N\to 0$ となり、$\mathrm{E}(x)=x$ となることが分かる。一方、$N, B$ が小さくなると $P(0)$ が大きくなり、その結果、$\mathrm{E}(x)$ と $x$ の差が大きくなる。

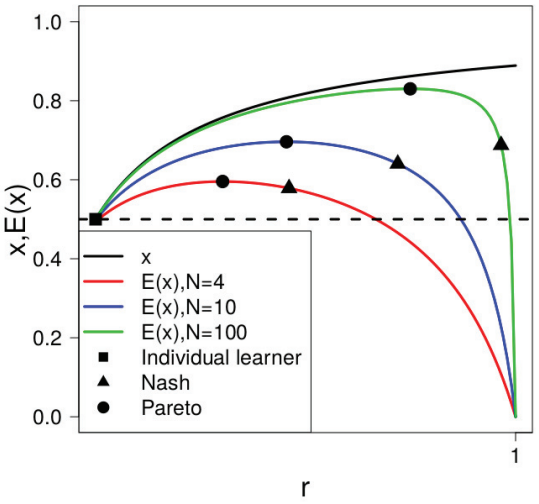


図 1: $r$ に対して $x$ および $\mathrm{E}(x)$ をプロットした。$x$ は $N\to\infty$ での $N_1/N$ の値、$\mathrm{E}(x)$ は $N_1/N$ の定常状態での期待値である。$N=4,10,10^2$ とし、また $q_I=0.1, q_O=0.8, q_C=0.1$ とした。■ は $r=0$ の独自学習エージェントの $\mathrm{E}(x)$ を示している。● はそれぞれの $N$ に対するパレート解 $r_{Pareto}$ 、▲ はナッシュ解 $r_{Nash}$ を示す。$r_{Pareto}, r_{Nash}$ については後に説明する。

$x$ および $\mathrm{E}(x)$ を $r$ の関数としてプロットしたものが図 1 である。パラメータは $q_I=0.1, q_O=0.8, q_C=0.1$ とし、エージェント数は $N=4,10,10^2$ とした。$r$ が小さい時は差は小さいが $r$ が増加し 1 に近づくにつれて差が大きくなることが分かる。特に、ある値以上では $r=0$ の独自学習エージェントの $\mathrm{E}(x)$ を下回るようになる。このときエージェント間にリターン 0 のバンディット情報が駆け巡るエコー・チェンバー状態となっている。

リターン 1 のバンディットの個数 $M$ が 1 個のとき ($M=1$)、釣り合った均衡状態から振動状態への変化は相転移ではない。しかし、$r$ が小さい時の均衡状態と $r$ が大きいときの $N_1=0$ と $N_1\simeq xN$ の間の振動状態の間の変化はマクロな変化なので、相転移的な振る舞いと呼ぶことにした。[Mori 2016] で示したように、$M/N=\delta$ として $N, M\to\infty$ の極限では系は相転移をする。この相転移は $r_c=\frac{q_I}{q_I+q_O}$ を閾値として起こり、$r<r_c$ では $N_1$ は通常の分布をし、$N_1$ の分散も $N$ に比例する。一方、$r\geq r_c$ ではエージェントがリターン 1 のある 1 つのバンディットに集中し、エージェントの分布関数のべき指数が 2 以下となって $N_1$ の分散が $N$ よりも早く発散するようになる。また、エージェントがあるリターン 1 のバンディットに集中した結果、そのバンディットの状態変化によって $N_1=0$ の状態になり、$M=1$ の場合と同様に $N_1=0$ の状態から抜け出せなくなる。$M=1$ のときはエージェントは唯一のリターン 1 のバンディットに集中するしかないので、$N_1$ の分散に相転移的な振る舞いは見られないが、$\mathrm{E}(x)$ の均衡状態での値 $x$ からの低下の物理は $M=1$ と $M>1$ で同一である。

## 4. ナッシュ均衡 $r_{Nash}$ と Rogers パラドックス

前節では、$N$ が小さく $r$ が大きいとエージェントはリターン 1 のあるひとつのバンディットに集中し、そのバンディットの変化のあと、$N_1 = 0$ の状態からなかなか抜け出せず、その結果、$\mathrm{E}(x)$ が低下すること、また、この状態ではエージェント間にリターン 0 のバンディット情報が駆け巡るのでエコー・チェンバー状態であることを説明した。では、エージェントはどのような $r$ の値を採用するのであろうか？ゲーム論を用いて考えてみる [Nakayama 2017]。ゲームは混合戦略の $r_n$ を選択して期待リターンを利得表とするものである。エージェントの期待リターンとして $\sigma_n$ の定常状態での期待値 $\mathrm{E}(\sigma_n)$ を用いる。これは、エージェントがリターン 1 のバンディットを記憶している確率に他ならない。エージェント $n$ は全エージェントの混合戦略（社会的学習の確率）$r_n$ を知っているとし、その上で $r_n$ を自分の期待リターンが最大になるように選択する。

$\sigma_n$ の定常状態の期待値 $\mathrm{E}(\sigma_n)$ は $r_n$ と $r_m, m \neq n$ の平均値 $\bar{r}_n \equiv \sum_{m \neq n} r_m/(N-1)$ の関数 $\omega(r_n, \bar{r}_n)$ となる。

$$\mathrm{E}(\sigma_n) = \omega(r_n, \bar{r}_n)$$

$\omega(r_n, \bar{r}_n)$ を区間 $[0,1]$ で $r_n$ について最大化したときの $r_n$ は $\bar{r}_n$ の関数 $f(\bar{r}_n)$ となる。

$$f(\bar{r}_n) \equiv \mathrm{argmax}_{r_n} \omega(r_n, \bar{r}_n)$$

全エージェントの $r_n, n = 1, \cdots, N$ のそれぞれの $r_n$ を $f(\bar{r}_n)$ で写像する $N$ 次元空間 $J = \vec{r}$ から $N$ 次元空間 $J = \vec{r}$ の写像を $\vec{F}(\vec{r}) = (f(\bar{r}_1), \cdots, f(\bar{r}_N))$ と書くと、ナッシュ解は写像 $F$ の固定点となる。固定点は唯一であり、$N$ 次元空間 $J$ の対角線上に存在する。それを $(r_{Nash}, r_{Nash}, \cdots, r_{Nash})$ と書くことにする。$r_{nash}$ は次の 2 つの不等式を満たすことが証明できる。

- $\omega(r_{Nash}, r_{Nash}) > \omega(r, r_{nash})$ for $\forall r \neq r_{Nash} \in [0,1]$
- $\omega(r_{Nash}, r) > \omega(r, r)$ for $\forall r \neq r_{Nash} \in [0,1]$

最初の不等式は $r_{Nash}$ がナッシュ解であり、他のエージェントがすべてナッシュ解 $r_{Nash}$ を採用しているとき、他の $r$ を選ぶと期待リターンが $r_{Nash}$ と選んだときより減少するので $r_{Nash}$ から戦略を変更しないことを意味する。つまり抜け駆けで損するので $r_{Nash}$ は安定である。2 つめの不等式は他のエージェントがすべて $r$ を採用しているとき、$r_{Nash}$ 戦略を選ぶとリターンが増加することを意味する。つまり、$r_{Nash}$ の戦略が $r$ 戦略の集団に侵入可能である。これらの 2 つの不等式を満たす戦略を ESS（進化論的に安定な戦略）と呼ぶ。ESS 戦略は他の戦略の侵入を抑止し、かつ ESS 戦略の侵入が可能であることを意味するので、進化論的に安定な戦略としてメイナード・スミスが導入した概念である [Smith 1982]。

ESS ナッシュ解は進化論的に安定なナッシュ解であり、自分のリターンのみを最大化する利己的なエージェント系で採用されると考えられる戦略である。図 1 で ▲ はナッシュ解を示す。同じ図からナッシュ解は対称解:$r_n = r, \forall n = 1, \cdots, N$ で $\omega(r_n, \bar{r}_n)$ が最大となる解ではないことが分かる。$N = 4, 10, 10^2$ の各々のケースで $\mathrm{E}(x) = \mathrm{E}(\sigma_n)$ を最大にする解をパレート最適解と呼び、$r_{Pareto}$ と書くものとする。$r_{Pareto}$ は対称解で $\omega(r, r)$ を最大にする解でもあるが、より一般的には各エージェントの期待リターンの和 $\sum_n \omega(r_n, \bar{r}_n)$ を最大にする解でもある。エージェントがお互いを出し抜かずに協力する場合、個体および集団のリターンが最大になる戦略となっている。蟻やハチなど、共通の遺伝子を持つ集団では抜け駆けする誘因がないため $r_{Pareto}$ 戦略を採用してもおかしくない。一方、ヒトを含む共通の遺伝子を持たない社会的動物集団では抜け駆けする誘因があり、$r_{Nash}$ を採用する可能性がある。$r_{Nash}, r_{Pareto}$ を $N, q_I, q_O, q_c$ の関数で表す計算式については文献 [Nakayama 2017] を参照のこと。

以上の結果から、最適な戦略は $r_{Nash}, r_{Pareto}$ で与えられることが分かった。重要な点は、これらの戦略での期待リターンは $r = 0$ の独自学習の期待リターンよりも $q_I < q_O$ の条件のもとではよいことである。

$$\omega(r_{Pareto}, r_{Pareto}) > \omega(r_{Nash}, r_{Nash}) > \omega(0, r)$$

つまり Rogers' パラドックスは混合戦略の枠組みでは存在しないことが分かる [Nakayama 2017]。

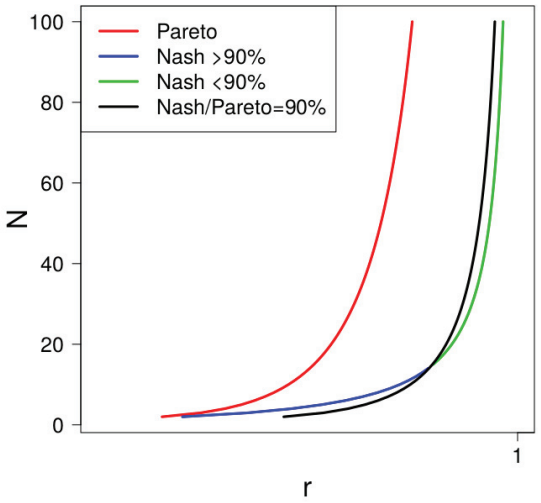


図 2: $N$ に対して $r_{Pareto}, r_{Nash}, r_{EC}$ をプロットした。$r_I, q_O, q_C$ は図 1 と同じ値を用いている。$r_{Nash}$ のうち、期待リターンが $r_{Pareto}$ での期待リターンの 90%以上なら青、90%以下なら緑の実線で、また、黒の実線で $r_{EC}$ を示した。

一方、エコー・チェンバー状態を $r > r_{Pareto}$ で期待リターン $\omega(r, r)$ が $r_{Pasreto}$ での期待リターン $\omega(r_{Pareto}, r_{Pareto})$ から 10%以上低下する $r$ の領域 $r > r_{EC}$ と定義してみる。$r_{EC}$ はエコー・チェンバー領域の境界である。図 1 より、$N = 4$ では $r_{Nash}$ と $r_{Pareto}$ での期待リターンの差は小さいので、$r_{Nash}$ はエコー・チェンバー状態ではない ($r_{Nash} < r_{EC}$)。一方、$N = 10^2$ のとき、$r_{Nash}$ の期待リターンは低く、$r_{Nash}$ はエコー・チェンバー状態にある ($r_{Nash} > r_{EC}$)。そこで、$(r, N)$ 面で $r_{Pareto}, r_{Nash}$ そして $r_{EC}$ をプロットしたものが図 2 である。$N = 14$ を境界とし、$N < 14$ なら $r_{Nash} < r_{EC}$ でナッシュ解 $r_{Nash}$ はエコー・チェンバー状態ではない。一方、$N \geq 14$ では $r_{Nash} > r_{EC}$ となりうナッシュ解 $r_{Nash}$ はエコー・チェンバー状態であることが分かる。

## 5. 戦略 $r$ のダイナミクスと $r_{Nash}$ への収束

エージェントが戦略 $r$ を $\omega(r_n, \bar{r}_n)$ を改善する方向に変化させたときの戦略のダイナミクスを調べる。時間変数を $\tau$ とし、

$r_n(\tau)$ に対する次の微分方程式である。

$$\frac{d}{d\tau}r_n(\tau) = \frac{\partial}{\partial r_n}\omega(r_n, \bar{r}_n), n = 1, \cdots, N$$

ただし、$r_n$ が領域 $[0, 1]$ を飛び出ないよう、$r_n = 0$ で右辺は負の場合と $r_n = 1$ で右辺が正のとき、右辺をゼロと修正する。

図 3 の上図は $N = 3, M = 1$ で $q_I, q_O, q_C$ は図 1 と同じとし、初期条件として $r_1 = r_2 = 0, r_3 = 1$ のときの $r_1$ と $r_3$ を $\tau$ に対してプロットしたものである。各エージェントが $\omega(r_n, \bar{r}_n)$ を増加させる方向に $r_n$ を変化させたとき、$r_{Nash}$ に収束することが分かる。

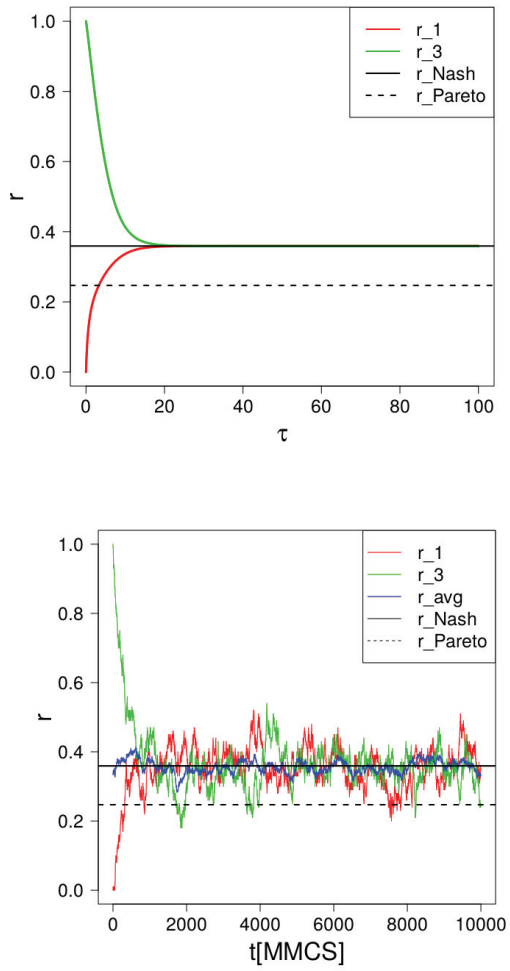


図 3: $r_n$ の微分方程式の数値解（上図）とモンテカルロ・シミュレーションの結果 (下図).$N = 3$ とし、$q_I, q_O, q_C$ は図 1 と同じ。初期条件は $r_1(0) = r_2 = 0, r_3(0) = 1$ とした。1[MMCS]=$10^6$[MCS]. $r_{avg}$ は $r_1, r_2, r_3$ の平均値である。

図 3 の下図はモンテカルロ・シミュレーションでの $r_n$ のダイナミクスを示したものである。$10^6$[MCS]=1[MMCS] 毎にランダムに選ばれたエージェント $n$ が $\omega(r_n, \bar{r}_n)$ を過去 $10^6[MCS]$ での $\sigma_n = 1$ の比率を計算し、比率が増加したなら過去の $r_n$ の変化と同じ方向に $r_n$ を 0.01 変化させ、減少した場合は反対方向に $r_n$ を 0.01 させる。初期条件は微分方程式と同じである。モンテカルロ・シミュレーションでは $r$ はナッシュ解 $r_{Nash}$ のまわりでゆらぐことが分かる。

## 6. 結論

非定常多腕バンディットゲームの学習エージェント系を研究した。バンディットはリターンありとなしの 2 値の状態をとり、エージェントは独自学習か社会的学習をそれぞれ確率 $(1-r, r)$ で行う混合戦略を採用し、リターンありのバンディット情報の獲得を目的とする。全エージェントが共通の $r$ を採用するとき $r$ が大きく $N$ が小さいと、エージェントの期待リターンは悪化する。全エージェントがリターンのないバンディットに存在する状態が長く続き、エージェント間にリターン 0 のバンディット情報が駆け巡るエコー・チェンバー状態となるためである。戦略 $r$ の変更を行い定常状態での期待リターンを利得表とするゲームをゲーム理論の枠組みで議論し、ESS ナッシュ解 $r_{Nash}$、パレート最適解 $r_{Pareto}$ を求めた。特に $N$ が大きいとき、$r_{Nash}$ での期待リターンとパレート最適解の期待リターンの差が大きくなり、$r_{Nash}$ が期待リターンが $r_{Pareto}$ での期待リターンの 90%未満として便宜的に決めたエコーチェンバ領域の境界 $r_{EC}$ を超えることが分かった。最後に、エージェントが戦略パラメータ $r$ を期待リターンを改善する方向に変化させたときの戦略 $r$ のダイナミクスを微分方程式、確率モデルでモデル化しナッシュ解 $r_{Nash}$ への収束（のまわりでゆらぐこと）を確かめた。

## 参考文献


[Pentland 2014] Pentland, A., Social Physics: How Good Ideas Spread, Penguin Press, London(2014).

[Rogers 1988] Rogers, A. R., Does biology constrain culture?, Am. Anthropol. 90, 819831 (1988).

[Rendell 2010] Rendell, L. et al. Why copy others? insights from the social learning strate- gies tournament. Science 328, 208213 (2010).

[Mori 2016] Mori, S., Nakayama, K. and Hisakado, M., Phase transition of social learning collectives and "Echo chamber", Phys.Rev.E.vol.94,No.5,052301-052309 .

[Nakayama 2017] Nakayama, K., Hisakado, M. and Mori, S., Nash Equilibrium of Social-Learning Agents in a Restless Multiarmed Bandit Game, preprint.

[Smith 1982] Maynard-Smith, J. Evolution and the Theory of Games (Cambridge University Press, Cambridge, 1982).