非定常多腕バンディットゲームと Rogers' paradox

SP13143　　量子物理学研究室　　中村隆造

1.はじめに

我々人間は他の動物とは異なり、知識という形で先駆者の経験や知恵をもらい受け、自分の学習に活かすことが出来る。他人の経験から学習できる社会的学習(コピー)は自分だけで学ばなければならない独自学習(ランダムサーチ)に比べて有利である。それゆえ、社会的学習は独自学習より有利ではないという Rogers の発見は直感に反している。これを Rogers' paradox という。本研究では、非定常多腕バンディットゲーム(restless multiarmed bandit Game, 以下 rMAB と略す。)における社会学習者に対するモデルについて報告する。

2.モデル

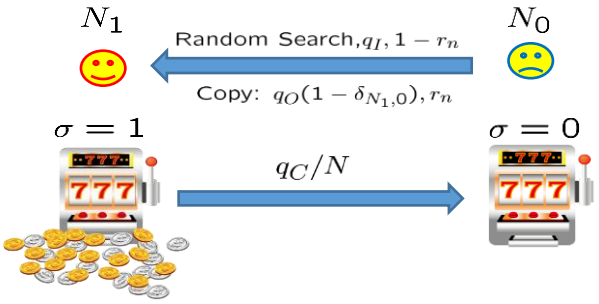

Fig1:ゲームの状況説明

rMAB は 1 つの当たりとハズレのバンディット(スロットマシン)を持つ。N 人のエージェントを n=1,2,…,N とし、当たりを知っているエージェント数を N_1、知らないエージェント数を N_0 とする。それぞれのターンに、あるエージェント n が選ばれる。当たりのバンディットを知っていればそれを選び、報酬 1 を得る。当たりのバンディットを知らなければ、確率 $1-r_n$ で独自学習するか、確率 r_n で社会的学習（他人の当たりのバンディットの情報をコピー)する。ランダムサーチが成功する確率は q_I である。当たりのバンディットを知っているエージェントが 1 人でもいればコピーは確率 q_O で成功し、当たりを知っているエージェントがいなければ失敗する。その後、当たりのバンディットは確率 q_C/N でハズレに変わり、新しい当たりのバンディットが現れる。当たりがハズレに変わったら、当たりを知っているエージェントはそれを忘れなければならない。そして、それがハズレに変わったことを知る。

エージェント n が当たりのバンディットを知っているか知らないかを $\sigma_n \in (1,0)$ という 2 つの値で表すことにする。$\sigma_n = 1$ は当たりを知っている状態、$\sigma_n = 0$ は当たりを知らない状態である。定常状態での $(\sigma_1, \sigma_2, \cdots, \sigma_N)$ の確率分布を $P(\sigma_1, \sigma_2, \cdots, \sigma_N)$ とする。σ_n の期待値をエージェントの適応度 w_n と定義すると、w_n は r_n と $\bar{r}_n = \frac{1}{N-1}\sum_{k\neq n} r_k$ の関数である。

3.ゲーム構造

エージェントは社会的学習と独自学習の混合戦略を採用し、社会的学習の確率 r_n は 0~1 の任意の数だと仮定する。それぞれのエージェントは他人の r_n を知っており、固定された $r_k (k \neq n)$ に対し、w_n を最大化することを考える。この時の r_n の値を $f(\bar{r_n})$ と書く。そして、$(r_1, r_2, \ldots, r_N)$ から $(r_1, r_2, \ldots, r_N)$ への写像 $F(r_1, \ldots, r_N) = (f(\bar{r}_1), \ldots, f(\bar{r}_N))$ を定義する。これは全てのエージェントの r_n を $f(\bar{r}_n)$ に変える写像である。F の固定点はただひとつ $(r_{Nash}, r_{Nash}, \ldots, r_{Nash})$ に存在し、次の不等式を満たすので、それが Nash 均衡点であることが分かる。

$$w(r_{Nash}, r_{Nash}) > w(r, r_{Nash}), \qquad \text{for all } r \neq r_{Nash}$$

ここで、r_{Nash} は N, q_I, q_O, q_C の関数である。さらに、次の不等式を満たすので、これは進化論的に安定な戦略(ESS)である。

$$w(r_{Nash}, r) > w(r, r), \qquad \text{for all } r \neq r_{Nash}$$

どのエージェントも戦略 r_{Nash} を採用すれば、自分の適応度を最大化することができる。言い換えれば、自分の戦略を変えることによって、今まで以上の適応度を得ることはできず、戦略を変える意味がないのである。また、他の戦略 r のエージェントが戦略 r_{Nash} の集団に侵入しても、この集団の適応度は進化論的に安定である。また、上記の不等式において $r = 0$ とすれば、戦略 r_{Nash} の適応度が独自学習より高いことが分かり、Rogers' paradox は解消される。

4.参考文献

[1]S.Mori K.Nakayama M.Hisakado Phys.Rev. E vol. 94, 052301 (2016).

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