生物種間競争原理を利用した MAB 型強化学習問題の最適解探索法

Identification of optimal solution in multi-armed banddit problems by interspecific competition dynamics

金沢大理工 [1]，埼玉大工 [2]，東京大情報理工 [3]，JST さきがけ [4]，
○ 新山 友暁 [1]，古畑 玄貴 [1]，内田 淳史 [2]，成瀬 誠 [3]，砂田 哲 [1,4]
Kanazawa Univ.[1], Saitama Univ.[2], Univ. of Tokyo[3], JST PREST[4],
○Tomoaki Niiyama[1], Genki Furuhata[1], Atsushi Uchida[2], Makoto Naruse[3], Satoshi Sunada[1,4]
E-mail: niyama@se.kanazawa-u.ac.jp

Multi-armed bandit 問題（MAB 問題）は，異なる報酬期待値 μ_i $(i = 1, ..., N)$ をもつスロットマシン（選択肢）が N 個存在するときに，少ない時間投資損失の中でいかにして最大の期待値をもつマシン（最適解）を探し出すかという問題で，強化学習の初歩的な典型的問題設定のひとつである [1]。これに対して UCB 法や TOW 法といった様々な手法が提案されているが，本講演では体積保存則にヒントを得た我々の新規手法が，生物種間競争という物理現象のメカニズムに裏付けられていることを述べる [2]。

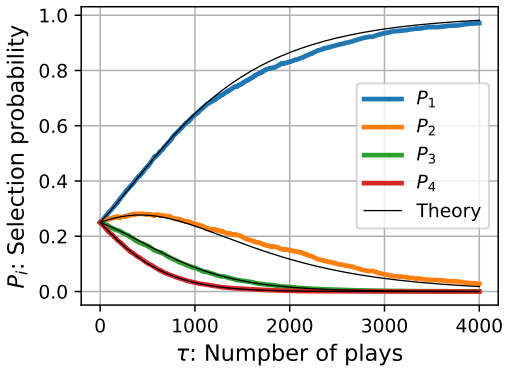

Fig. 1　報酬期待値 $\mu_i = 0.6 - (i-1)/10$ $(i = 1, ..., 4)$ の MBA 問題に対して，本手法を適用した結果（標準偏差 0.5）。

本提案手法では，τ 回目において N 個のうちどのスロットマシンをプレイするかを選択確率 $P_i(\tau)$ $(i = 1, 2, ..., N)$ に従って決定し，その結果選ばれたマシン j をプレイして得られた報酬 x_j をもとに，全マシンの選択確率を次式のように更新していく。

$$P_j(\tau+1) = P_j(\tau) + bx_j(1 - P_j(\tau)), \quad P_k(\tau+1) = P_k(\tau) - bx_jP_k(\tau) \quad (1 \le k \ne j \le N) \tag{1}$$

ここで，b は $bx_i < 1$ となるような十分小さなパラメーターである。

Fig. 1 に示したように，この手法を用いることで報酬期待値が最大であるスロットマシン 1 の選択確率が試行の繰り返しによって増大し，最適解を見出していることがわかる。興味深い点として，$b \ll 1$ の極限において $t = b\tau$ として時間連続系に接続することで上記更新則が以下の微分方程式

$$\frac{\mathrm{d}P_i}{\mathrm{d}t} = P_i\left(\mu_i - \sum_j \mu_j P_j\right) \tag{2}$$

に帰着し，Lotka-Volterra 型生物種間競争方程式と一致することを発見した（Fig. 1 の黒線はこの微分方程式の数値解である）。このことは，環境に適合した生物種がその個体数を増やすという自然の摂理 (winner-take-all) が MAB 問題において有効に働きうることを示唆している。

[1] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction* (MIT press, 2018).
[2] T. Niiyama, G. Furuhata, A. Uchida, M. Naruse, and S. Sunada, J.Phys. Soc. J. **89**, 014801 (2020).