

当たり確率が増える多本腕バンディット問題における レーザカオスを用いた強化学習と意思決定

Reinforcement learning and decision making using laser chaos for multi-arm bandit problem with variable hit probabilities

埼玉大学¹, 情報通信研究機構²

○小田 章裕¹, 巳鼻 孝朋¹, 菅野 円隆¹, 成瀬 誠², 内田 淳史¹

Saitama University,¹ NICT²

○Akihiro Oda¹, Takatomo Mihana¹, Kazutaka Kanno¹, Makoto Naruse², and Atsushi Uchida¹

E-mails: a.oda.652@ms.saitama-u.ac.jp, auchida@mail.saitama-u.ac.jp

はじめに: 近年、人工知能における強化学習の研究が盛んに行われている。強化学習の問題例として、当たり確率が未知である複数のスロットマシンにおいて有限回の試行を通して報酬の最大化を図る多本腕バンディット問題が挙げられる。多本腕バンディット問題は一般性を失うことなく数学的問題として定式化されており、実際に無線チャンネル割り当て[1]などに用いられている。

多本腕バンディット問題の解法として綱引き理論が提唱されている[2]。半導体レーザカオスを用いた綱引き理論では、可動であるしきい値を設け、半導体レーザカオスの値との比較を行い、その大小に応じて試行するスロットマシンを決定する。更に試行したスロットマシンの結果に応じてしきい値を移動させることにより各スロットマシンを選択する頻度を变化させる方式である。この方式において各スロットマシンの当たり確率が一定の場合における調査が行われてきた[3]。しかしながら、スロットマシンの当たり確率が増える場合における調査は未報告である。

そこで本研究では、半導体レーザカオスを用いて、スロットマシンの当たり確率が増える多本腕バンディット問題における意思決定を数値計算により行うことを目的とする。

方法: 本研究では、半導体レーザカオスの時間波形を取得し、数値計算による実装を行う。半導体レーザカオスの時間波形としきい値との比較を行い、しきい値が時間波形よりも小さければスロットマシン1を選択し、大きければスロットマシン2を選択する。また、選択したスロットマシンの報酬に応じて、しきい値を移動させる。しきい値の移動量はスロットマシンの試行結果により算出される推定当たり確率を用いて決定される。また、推定当たり確率 P_i は、スロットマシン i にて当たった回数 H_i から選択した回数 C_i を割ることで算出され($P_i(t) = H_i(t)/C_i(t)$)、 H_i と C_i は以下に示されるように記憶係数 β で過去の値が重み付けされる。

$$H_i(t) = \begin{cases} 1 + \beta H_i(t-1) & (\text{hit}) \\ \beta H_i(t-1) & (\text{miss}) \end{cases} \quad (1)$$

$$C_i(t) = 1 + \beta C_i(t-1) \quad (2)$$

この記憶係数 β を用いることにより当たり確率が増える場合における推定が可能となる。

実装方法は、スロットマシン1と2の当たり確率をそれぞれ0.4と0.6に設定し、1000回ごとに当たり確率を入れ替える。この状況において4000回試行を1000ループ行う。また、本実装の評価方法として、当たり確率の高いスロットマシンを選んだ割合を示す平均正答率(Correct Decision Rate: CDR)を用いる。

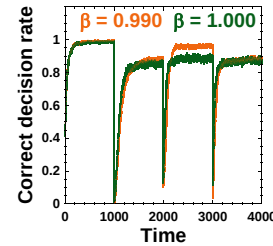


Fig. 1 Evolution of correct decision rate (CDR) for $\beta = 0.990$ (orange) and $\beta = 1.000$ (green).

結果: 本研究では記憶係数 β を变化させた時の意思決定性能の比較を行った。記憶係数 $\beta = 0.990$ と $\beta = 1.000$ の場合の平均正答率の時間変化をFig. 1に示す。 $\beta = 0.990$ の方が平均正答率は高く、当たり確率の入れ替えが生じた後も当たり確率の高いスロットマシンを多く選択できることが分かる。

さらに、記憶係数 $\beta = 0.990$ と $\beta = 1.000$ の時の推定当たり確率をFig. 2に示す。 $\beta = 1.000$ の時は過去のスロットマシンの試行結果を常に保持し続けるため、当たり確率の変化に追従できず当たり確率の推定に失敗している。一方で $\beta = 0.990$ の場合、当たり確率の変化に対して推定当たり確率も追従できている。これがしきい値の移動量に影響を及ぼし、当たり確率の高いスロットマシンを多く選択できることが分かった。

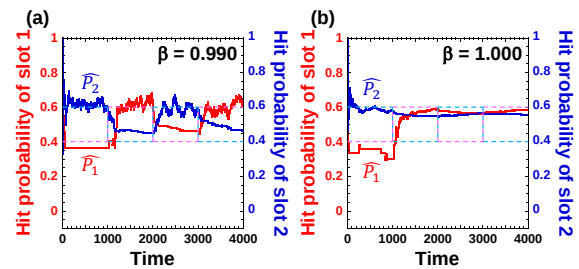


Fig. 2 Estimated hit probabilities for slot machine 1 (red) and 2 (blue) in case of (a) $\beta = 0.990$ and (b) $\beta = 1.000$.

まとめ: 本研究では、半導体レーザカオスの時間波形を用いて、スロットマシンの当たり確率が増える多本腕バンディット問題の意思決定を数値計算により調査した。記憶係数 β を用いることにより当たり確率の変化に対して正しい推定を行うことができ、その結果、平均正答率が上昇することが明らかになった。

参考文献

- [1] K. Kuroda, et al., Nonlinear Theory and Its Applications, IEICE **9**, 74 (2018).
- [2] S.-J. Kim, et al., New J. Phys. **17**, 083023 (2015).
- [3] M. Naruse, et. al., Scientific Reports **7**, 8772 (2017).