

非定常な報酬確率のバンディット問題における 半導体レーザカオスを用いた意思決定

Decision making using chaotic semiconductor laser in bandit problem with variable hit probabilities

埼玉大学¹, 東京大学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, The University of Tokyo²

°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。強化学習とは、与えられた環境の中で学習を行い、環境に適応した動作を行うことである。その問題を数学的に定式化したのがバンディット問題である[1]。バンディット問題の例として、報酬確率が未知である複数台のスロットマシンを仮定した時に、その中から1台を選択し、報酬を得ることを有限回行う中で報酬を最大化することを目的とする。これを解く方法として、報酬確率が未知であるスロットマシンを引いて情報を得る「探索」と、得られた情報から報酬を最大化するスロットマシンを選択する「知識利用」が挙げられる。この2つはどちらにも偏ってはならず、バランスが重要である。

このバランスを1つの方策に示した方式として、綱引き理論が提唱されている[2]。綱引き理論とはあるスロットマシンを選択した時に、その報酬に応じてそれぞれの選択頻度を逐次変化させる方法である。この綱引き理論による半導体レーザカオスを用いた意思決定が報告されている[3]。これは半導体レーザカオスの時間波形の高速性を活かし、数 GHz の超高速な意思決定を実現している。さらに、本手法を階層構造により拡張した64台のスロットマシンにおける意思決定も報告されている[4]。しかしながら、先行研究では報酬確率が固定されている状態で進行しており、報酬確率が変化する環境の下での意思決定の報告例は少ないのが現状である。

そこで本研究では、報酬確率が変化する非定常なバンディット問題を、半導体レーザカオス波形を用いて解くことを目的とする。

方法: ここでは、半導体レーザカオスを用いた意思決定方法について説明する。2台のスロットマシン S_1, S_2 を仮定した時、半導体レーザカオスの時間波形と可動であるしきい値を比較することにより意思決定を行う[3]。もしも時間波形がしきい値よりも高い場合は S_1 を選択し、低い場合は S_2 を選択する。さらに、 S_1 を選択して当たった場合は S_1 が選択されやすくなるようにしきい値を下げ、はずれた場合は S_2 が選択されにくくなるようにしきい値を上げる。その時の移動量は、しきい値分解能 b と報酬に応じて決まる移動量 Δ, Ω により定められる。 b が小さいほど1回におけるしきい値の移動量が大きく、 b が大きいほど移動量は小さくなる。また、移動量 Δ, Ω は推定報酬確率 $\hat{P}_i$ により決定される[3,4]。 $\hat{P}_i$ は当たり回数 H_i から選択回数 C_i を割ることで算出される ($\hat{P}_i = H_i/C_i$)。

本研究では、報酬確率の変化に対応するために、推定報酬確率に記憶係数 β を導入する。

$$H_i = \begin{cases} 1 + \beta H_i(t-1) & \text{for win} \\ \beta H_i(t-1) & \text{for lose} \end{cases} \quad (1)$$

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

また、報酬確率の正しい推定を行うために、報酬確率変化後にしきい値を初期化し、しきい値分解能 b を大きくする。これにより、確率変化後にどちらのスロットマシンも選択される可能性が高くなり、十分な探索が行われると期待できる。

結果: 本研究での問題設定は、報酬確率を $\{P_1, P_2\} = \{0.1, 0.3\}$ または $\{0.9, 0.7\}$ とし、報酬確率が1000回ごとに入れ替わるとする。これは報酬確率の和が大きく変化するため問題が難しくなる。

この時の1サイクルにおける推定報酬確率を Fig. 1(a) に示す。推定報酬確率は実際の報酬確率の変化に追従して推定できていることが分かる。また、平均正答率を Fig. 1(b) に示す。従来の方法と比べ平均正答率は報酬確率変化後も上昇し、1付近まで到達していることが分かる。これは、報酬確率変化後も報酬を最大にするスロットマシンを選択できるためである。

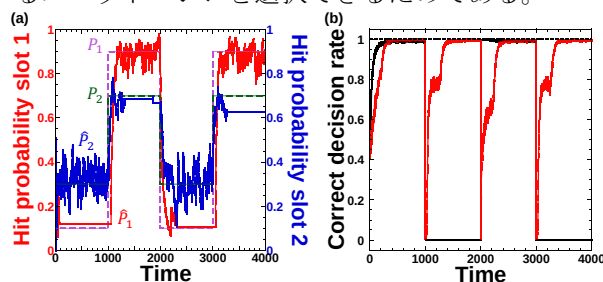


Fig.1 (a) Evolution of estimated hit probabilities for the slot machine 1 (red) and 2 (blue). (b) Evolution of correct decision rate (CDR) for previous method (black)[3], and proposed method (red).

まとめ: 本研究では、報酬確率が変化する非定常な多腕バンディット問題を、半導体レーザカオスを用いた意思決定により解いた。これにより、報酬確率が変化する環境においても、適応的な意思決定が可能であることを示した。

参考文献

- [1] H. Robbins, et al., Bulletin of the AMS **58**, 527 (1952).
- [2] S.-J. Kim., et al., New J. Phys. **17**, 083023 (2015).
- [3] M. Naruse., et al., Sci. Rep. **7**, 8772 (2017).
- [4] M. Naruse., et al., Sci. Rep. **8**, 10890 (2018).