CS 747: Project Proposal

Algorithm for Adversarial Bandits with Dynamic Exploration and Exploitation Rate

Group Members: 154190002, 163190014, 163190026

Introduction

In the course, we discussed algorithms like UCB [1], ϵ_t-greedy [1], Thompson Sampling [3] for stochastic multi-armed bandits. In this project, we will work on important variant of the multi-armed bandit problem where no stochastic assumption is made on the generation of rewards/losses. In the worst case of this setting, learning agent (algorithm) is playing against an adversarial environment (adversary). The adversary knows how the algorithm playing decisions and tries to change the actual outcome in order to penalize the algorithm. A good example for this is the game of chess, in which the players are adversarial environment to each other. The main goal of learning agent is to achieve sublinear regret bounds on the regret uniformly over all possible adversarial assignments of rewards. Due to adversarial nature, this problem setting is also known as *Adversarial Multi-Armed Bandits* problem. In this project, we will implement different algorithms (Exp3, Exp3.P & Exp3-IX) which are designed to work in adversarial settings. These algorithms work with static exploration and exploitation rate so we are proposing a variant which uses dynamic exploration and exploitation rate like ϵ_t-greedy [1].

Related Work: Algorithms for Adversarial Multi-Armed Bandits

In Exp3 (Exponential weights for Exploration and Exploitation) [2] algorithm, the learning agent chooses a decision-arm according to a probability distribution over arms. The learning agent then earns a reward through the drawn of arm, which then updates the probability distribution used to choose the arms as such the arm giving higher rewards will be played more in future. The Exp3 algorithm is not adequate for high probability bound on the regret but modified version of Exp3 that is Exp3.P [2] (P stands for Probability) is used for this. Exp3.P algorithm tries to tone down the variance of regret achieved by increasing the exploration factor. One more variant of Exp3, Exp3-IX [4] algorithm focuses on controlling the exploitation characteristic of the learning agent through something known as '*Implicit Exploration*' (IX). In the Exp3-IX paper [4], author proves that Exp3-IX has lower regret bound than Exp3 and Exp3.P algorithm.

Scope of the Project

This project is divided into two parts. First part, we will implement Exp3, Exp3.P and Exp3-IX algorithms and verify the claims of paper [4]. All these algorithms have static exploration and exploitation parameter. Like ϵ_t-greedy [1], in the second part, we will implement the variant of Exp3-IX with *Dynamic Exploration and Exploitation Rate* and compare the result against those with *static exploration and exploitation Rate*. If we achieve good experimental results, then we will try to give the theoretical regret bound for our algorithm.

References

[1] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer, *Finite-time analysis of the multiarmed bandit problem*, Machine learning **47** (2002), no. 2-3, 235–256.

[2] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire, *The nonstochastic multiarmed bandit problem*, SIAM journal on computing **32** (2002), no. 1, 48–77.

[3] Emilie Kaufmann, Nathaniel Korda, and Rémi Munos, *Thompson sampling: An asymptotically optimal finite-time analysis.*, ALT, vol. 12, Springer, 2012, pp. 199–213.

[4] Gergely Neu, *Explore no more: Improved high-probability regret bounds for non-stochastic bandits*, Advances in Neural Information Processing Systems, 2015, pp. 3168–3176.