EE6106 Project Proposal

Multiple-Play Stochastic Bandits with Shareable Finite-Capacity Arms

Malyala Preethi Sravani (200070041)
Sakshi Heda (200070071)

Introduction

The Multi-Arm Bandit (MAB) problem is a fundamental problem in the field of reinforcement learning. It involves a learner making a sequence of choices from a set of K different actions, or "arms," in order to maximize the expected reward. In standard MAB problems, the learner selects one arm in each play and observes the resulting reward. Based on this reward, the learner decides which arm to select in the next play.

In our specific problem, however, the learner is allowed to select a subset of the K arms, with N being the maximum number of arms that can be selected in each time slot. Furthermore, the arms can be shared among plays, meaning that more than one play can select the same arm and obtain the reward.

To calculate the total reward from each arm, we consider each arm's per load reward as a random variable with a Gaussian distribution. The total reward from each arm is then the product of the per load reward and the number of plays that select that arm.

Problem setting

Each arm k has a per load reward X_k, which is a Gaussian random variable, and a reward capacity m_k , which is a positive integer representing the maximum number of plays an arm can accommodate. The total reward of an arm is load-dependent and is calculated as the minimum of the reward capacity and the number of times the arm was played, multiplied by its per load reward. Specifically, the reward of arm k at time t when it has been played $a_{k,t}$ times is

$$R_k(a_{k,t}) = min\{m_k, a_{k,t}\} \cdot X_k.$$

After each round, the total reward from each arm is observed, but both the values of X_k and m_k remain unknown.

Algorithm

To evaluate the performance of our algorithm, we use a metric known as regret, which measures the total accumulated loss compared to an oracle that performs optimal allocation. The optimal allocation in our case involves assigning plays to arms with per load rewards in descending order which implies $\mu_1 > \mu_2 \ldots.. > \mu_K$ where the arm with the highest per load reward is assigned the most plays and the arm with the lowest per load reward is assigned the fewest. The optimal allocation is achieved by allotting m_1 plays to arm 1, m_2 plays to arm 2, and so on, until all the plays are exhausted. Our goal is to minimize the difference between the reward from our algorithm and the reward from the optimal allocation.

To achieve this goal, we plan to use the Orchestrative Exploration algorithm, as described in the paper "Multiple-Play Stochastic Bandits with Shareable Finite-Capacity Arms." As a modification to the work presented in this paper, we will analyze the problem with subgaussian rewards for each arm and attempt to derive regret and sample complexity bounds.

Overall, our research aims to address the challenges posed by the MAB problem in our specific setting and contribute to the development of more effective algorithms for solving this problem. By incorporating subgaussian rewards into our analysis, we hope to gain further insight into the behavior of these algorithms and improve their performance in practical applications.