Multiple-Play Stochastic Bandits with Shareable Finite-Capacity Arms

Malyala Preethi Sravani(200070041)
Sakshi Heda(200070071)

March 2023

Introduction

- In the multi-arm bandit (MAB) problem that we are exploring in the course, a learner pulls one arm out of $K \in \mathbb{N}_+$ arms and receives a reward corresponding to it that is sampled from a probability distribution whose parameters are unknown.
- In a multi-play multi-arm bandit (MPMAB) setting, the learner can select $N \in 2, 3, ..., K-1$ **different** arms out of the K arms. The rewards of different arms are sampled from different distributions as above.
- [1] lifts the constraint of having to choose N different arms. The concept of *sharable arms* is introduced.

[1] Wang, Xuchuang, Hong Xie, and John CS Lui. "Multiple-Play Stochastic Bandits with Shareable Finite-Capacity Arms." International Conference on Machine Learning. PMLR, 2022

Problem Setting and Algorithm

$K \in \mathbb{N}_+$ arms are indexed by $[K] := 1, 2, ...K$
Each arm, $k \in [K]$ is modelled as (m_k, X_k), where $m_k \in 1, 2, ...N$ is the reward capacity - a positive integer representing the maximum number of plays an arm can accommodate and X_k is the per-load reward, which is a Gaussian random variable
Reward of an arm at time t, when it is played $a_{i,t}$ times is given by

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

The paper uses **Orchestrative Exploration Algorithm** to minimize the difference between the obtained reward and the reward from the optimal allocation.

Proposed Modification

As mentioned in the previous slide, the rewards of the arms are assumed to be sampled from Gaussian random variables.
In our project, we will perform the regret analysis for the same algorithm assuming subgaussian rewards and will try to get tighter bounds on the regret.