Cooperative Stochastic Bandits with Asynchronous Agents and Constrained Feedback

Lin Yang¹, Yu-Zhen Janice Chen¹, Stephen Pasteris², Mohammad Hajiesmaili¹, John CS Lui³, Don Towsley¹

1. University of Massachusetts Amherst 2. University College London 3. The Chinese University of Hong Kong

Summary of Our Work and Contributions
Motivated by the scenario of large-scale learning in distributed systems, this paper studies a scenario where multiple agents cooperate together to solve the same instance of a K-armed stochastic bandit problem. The agents have limited access to a local subset of arms and are asynchronous with different gaps between decision-making rounds. The goal is to find the global optimal arm and agents are able to pull any arm, however, they can only observe the reward when the selected arm is local. In this work, we made the following contributions:

• First, We propose AAE-LCB, a two-stage algorithm that prioritizes pulling local arms following an active arm elimination policy, and switches to other arms only if all local arms are dominated by some external arms.
• Second, we analyze the regret of AAE-LCB and show it matches the regret lower bound up to a small factor.

The Model
Basic bandit and distributed bandit models

The basic bandit model:
• At each round, the learner selects one arm to pull observing the reward on the selected arm
• The goal of the learning algorithm is to maximize the cumulative reward. The performance of a learning algorithm is measured by “regret”, which is defined as the difference of the cumulative reward between the optimal arm and the learning algorithm
• Reward on each arm could be either stochastic (stochastic bandit model) or non-stochastic (non-stochastic/adversarial bandit model)

The corrupted bandit model (studied in this paper):
• Each agent solves an instance of MAB problem and share observations with others
• (Main Difference) Agents are assigned different action rates and constraints in feedback collection

Problem Background
Motivation and Application Examples
Our model is motivated by distributed setting in bandit-related applications, such as cooperative clinical trials and all kinds of cooperative learning applications in a multi-agent system.

Motivating Example: Mobile Ad Fraud

Two-Stage Cooperative Algorithm: AAE-LCB

1. Pull local arms as much as possible (first stage)
   - Use AAE to eliminate local arms, switch to select external arms only when an external arm dominates all local arms
2. Avoid selecting external arms with low-confidence estimates (second stage)
   - Select the external arm with the largest lower confidence bound (LCB is large only if the arm is well-observed)

Our Results
1. Regret Upper Bound
   \[
   O\left(\sum_{i=1}^{K} E_i \log T \right) \quad (By \ AAE-LCB)
   \]
   \[
   O\left(\frac{\Theta}{\theta_{\min}} \log T \right) \quad (By \ AAE \ algorithm \ in \ two \ stages)
   \]

Key Notations:
1: index of the optimal arm
\Delta_i: optimality gap
K: the number or arms
\Theta_i: Aggregate action rate of agents containing arm i
\theta: Aggregate action rate of all agents
\theta_{\min}: Smallest aggregate action rate of agents

2. Regret Lower Bound

Theorem 1 (Asymptotic Regret Lower Bound for FC-CH2EB): For any consistent algorithm \( \pi \) and any \( 0 < \gamma < 1 \), the expected regret satisfies

\[
E[\pi] T \sup_{\alpha, \gamma, \tau} \alpha \gamma \tau \sum_{i=1}^{K} \Delta_i^2
\]

References