Cooperative Stochastic Bandits with Asynchronous Agents and Constrained Feedback

Lin Yang[1], Yu-Zhen Janice Chen[1], Stephen Pasteris[2], Mohammad Hajiesmaili[1],
John CS Lui[3], Don Towsley[1]

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

NEURAL INFORMATION PROCESSING SYSTEMS

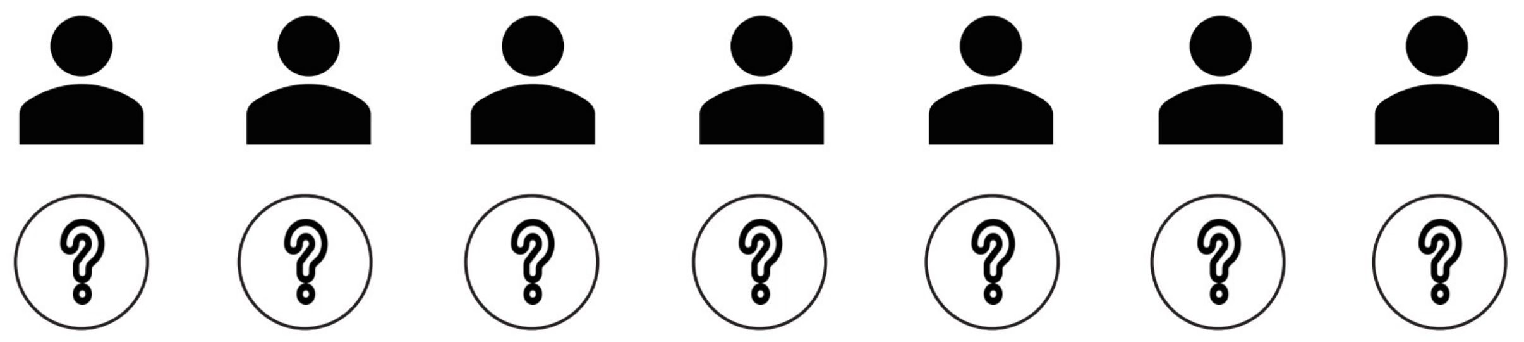
Summary of Our Work and Contributions

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):

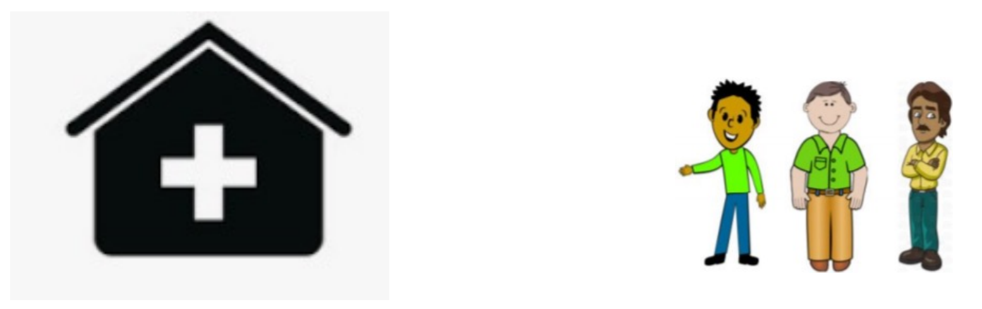
- Each agent solves an instance of MAB problem and share observations with others
- **(Main Difference) Agents are assigned different action rates and constraints in feedback collection**

Problem Background

Motivation and Application Examples

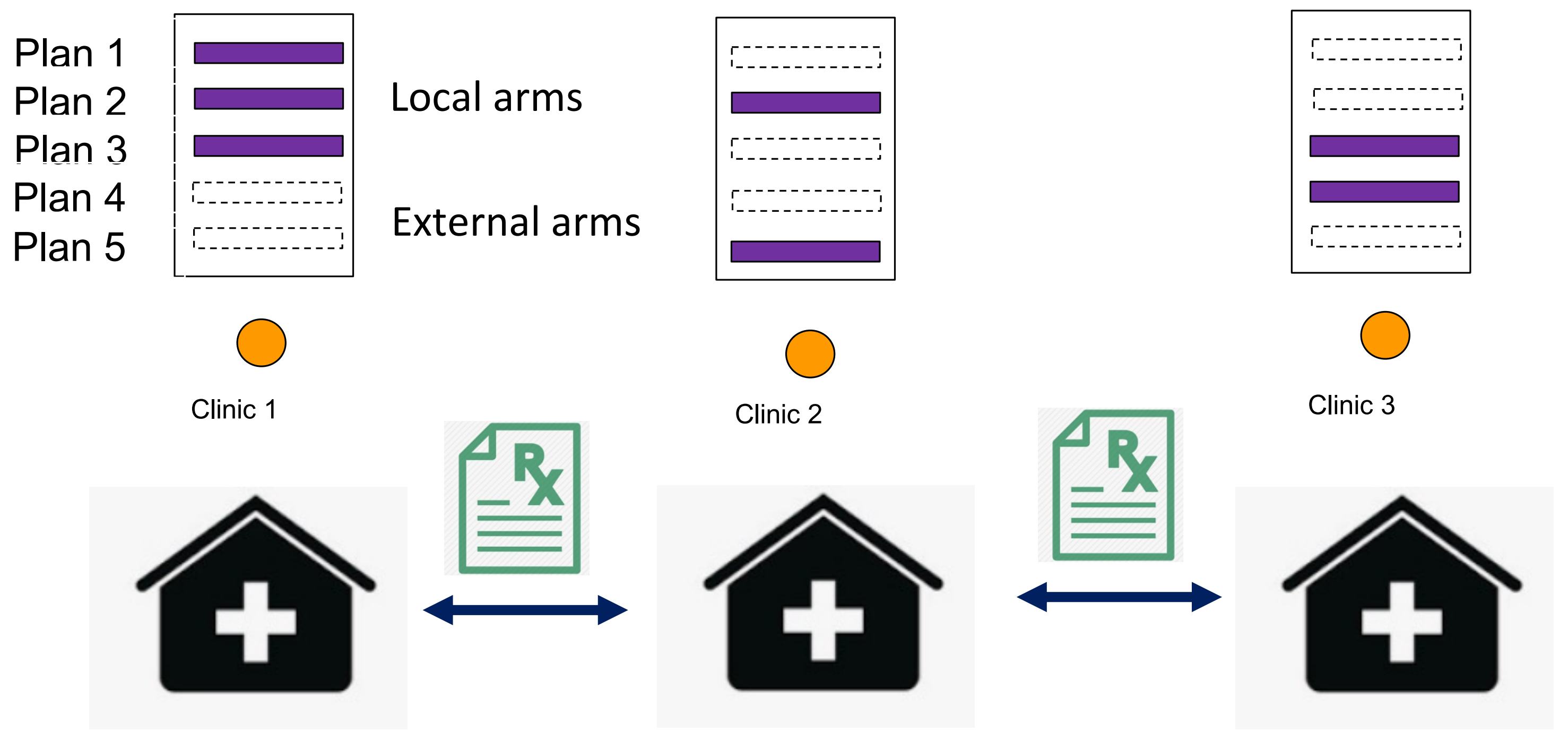
Our model is motivated by distributed setting in bandit-related applications, such as cooperative clinic trails and all kinds of cooperative learning applications in a multi-agent system.

Motivating Example: Mobile Ad Fraud



1. A clinic earns reputation when recommending correct meal plan to the patient.

2. Clinics have different access to the feedback of suggested treatment plans.



- The treatment plans with an outcome (e.g., customers come back for follow-up treatment) is modeled as local arms; others are taken as external arms.
- Associated with each clinic, an action rate is assigned, which corresponds to the popularity of the clinic.

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:

$$O\left(\sum_{i\in\mathcal{K}} \frac{K\Theta_i \log T}{\Theta_{i^*}\Delta_i}\right)$$ **(By AAE-LCB)**

$$\Omega\left(\frac{\Theta}{\Theta_{\min}} \log T\right)$$ **(By AAE algorithm in two stages)**

Key Notations:
i^*: index of the optimal arm
Δ_i: optimality gap
K: the number or arms
Θ_i: Aggregate action rate of agents containing arm i
Θ: 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-CMA2B`**)** *For any consistent algorithm π and any $0 < \sigma < 1$, its expected regret satisfies*

$$\liminf_{T\to+\infty,\Theta/\Theta_{i^*}\to+\infty} \frac{\mathbb{E}\left[R_T(\pi)\right]}{(\Theta/\Theta_{i^*})^\sigma \log(T\Theta)} = \Omega\Big(\sum_{i:\Delta_i>0} \frac{\Delta_i}{\mathrm{KL}(\mu_i, \mu_i + \Delta_i)}\Big).$$

References

Y. Bar-On and Y. Mansour (2019). “Individual regret in cooperative nonstochastic multi-armed bandits” In: Advances in Neural Information Processing Systems, pages 3116–3126.

D. Basu, C. Dimitrakakis, and A. Y. Tossou (2019). “Privacy in multi-armed bandits: Fundamentaldefinitions and lower bounds.” In: arXiv:1905.12298v2, 2019.

I. Bistritz and N. Bambos (2020). “Cooperative multi-player bandit optimization.” In: Advances in Neural Information Processing Systems, pages 3116–3126.

I. Bistritz and A. Leshem (2018). “Distributed multi-player bandits - a game of thrones approach.” In: Advances in Neural Information Processing Systems, pages 7222–7232.