Cooperative Stochastic Bandits with Asynchronous Agents and Constrained Feedback

Lin Yang*, Janice Yu-zhen Chen*, Stephen Pasteris^, Mohammad Hajiesmaili*, John CS Lui #, Don Towsley*

* University of Massachusetts, Amherst

^ University College London

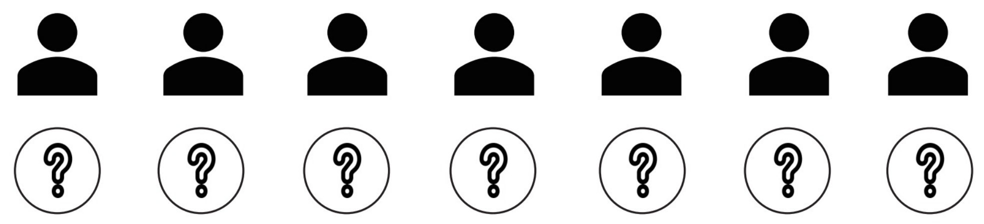
The Chinese University of Hong Kong

2021/10/15

The Multi-Armed Bandit Problem

Core Properties of MAB:

1. Sequentially taking actions of unknown quality
2. Feedback only involves information on selected action
3. Regret: gap of cumulative rewards between the optimal arm and the algorithm

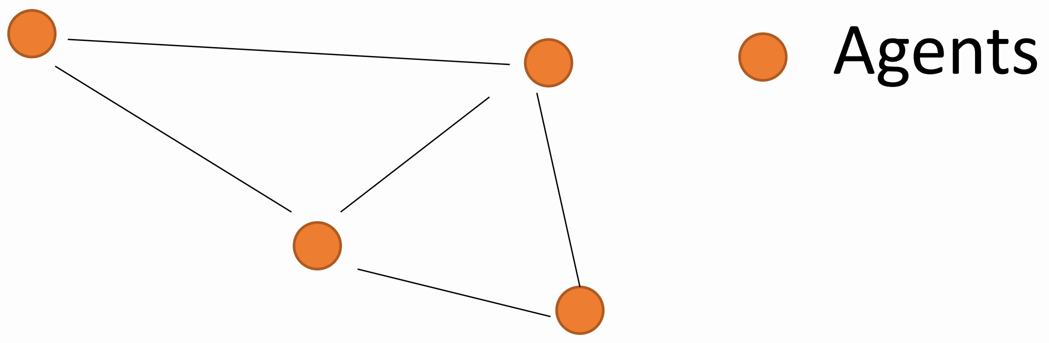


Adversarial Bandits: No assumptions on the rewards

Stochastic Bandits: Rewards subject to identical and independent distribution

MAB in Multi-Agent Systems

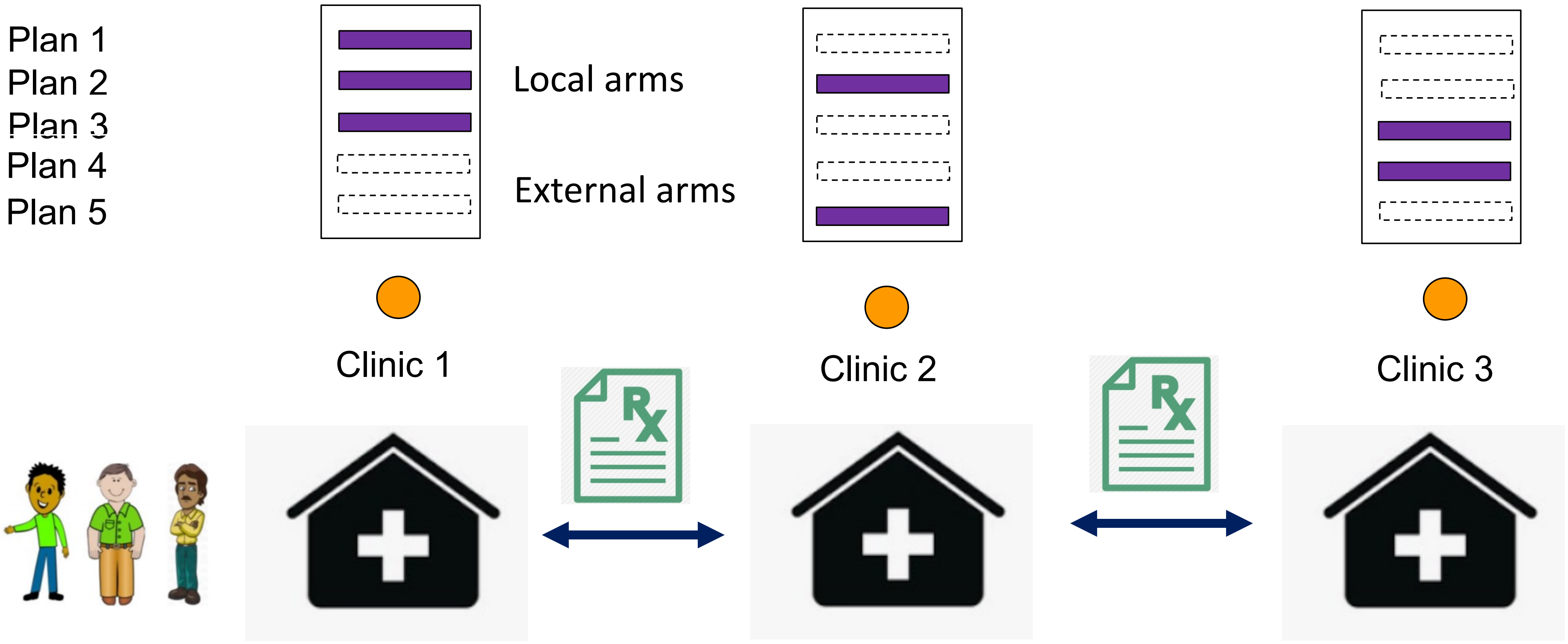
Each agent solves an instance of MAB problem and share observations with others



Homogeneous Agents – synchronized actions and non feedback constraints

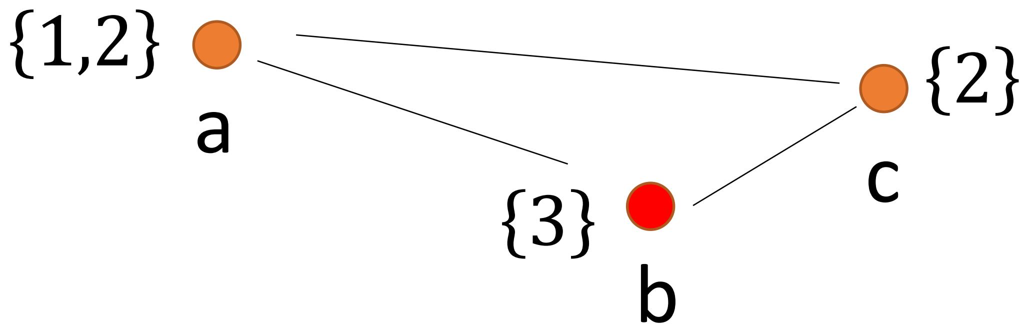
Heterogeneous Agents (new in our work) - agents are assigned different action rates and constraints in feedback collection

Multi-Agent Model for Cooperative Clinical Trials



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

An Example to Show Drawbacks of Traditional Algorithm



arms	Reward mean
1	0.8
2	0.4
3	0.6

agents	Action rate
a	1
b	0
c	0.5

- Arms are associated with Bernoulli rewards
- Agent b only takes action at the first slot
- With probability 0.6, the observed reward for arm 3 is 1
- There are only one observation, so other agents will select arm 3 constantly

Performance Degradation with Slow Agents

Strategies	UCB	Elimination-based	ϵ-greedy
Influenced by slow agents	Yes	Yes	Yes

Reasons that traditional algorithms suffer poor performance:

1. Fail to guarantee enough observations
2. Selection rules ignore the impact of action rate

A Two-Stage Cooperative Algorithm: AAE-LCB

Core Ideas:

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

   - Select the external arm with the largest lower confidence bound (LCB is large only if the arm is well-observed)

Theoretical Results

Regret by AAE-LCB:

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

Regret by Cooperative UCB:

$$\Omega\left(\frac{\Theta}{\Theta_{\min}} \log T\right)$$

K - number of arms

i^* - the optimal arm

Θ_i - aggregate action rate of agents containing arm i

Θ - aggregate action rate of all agents

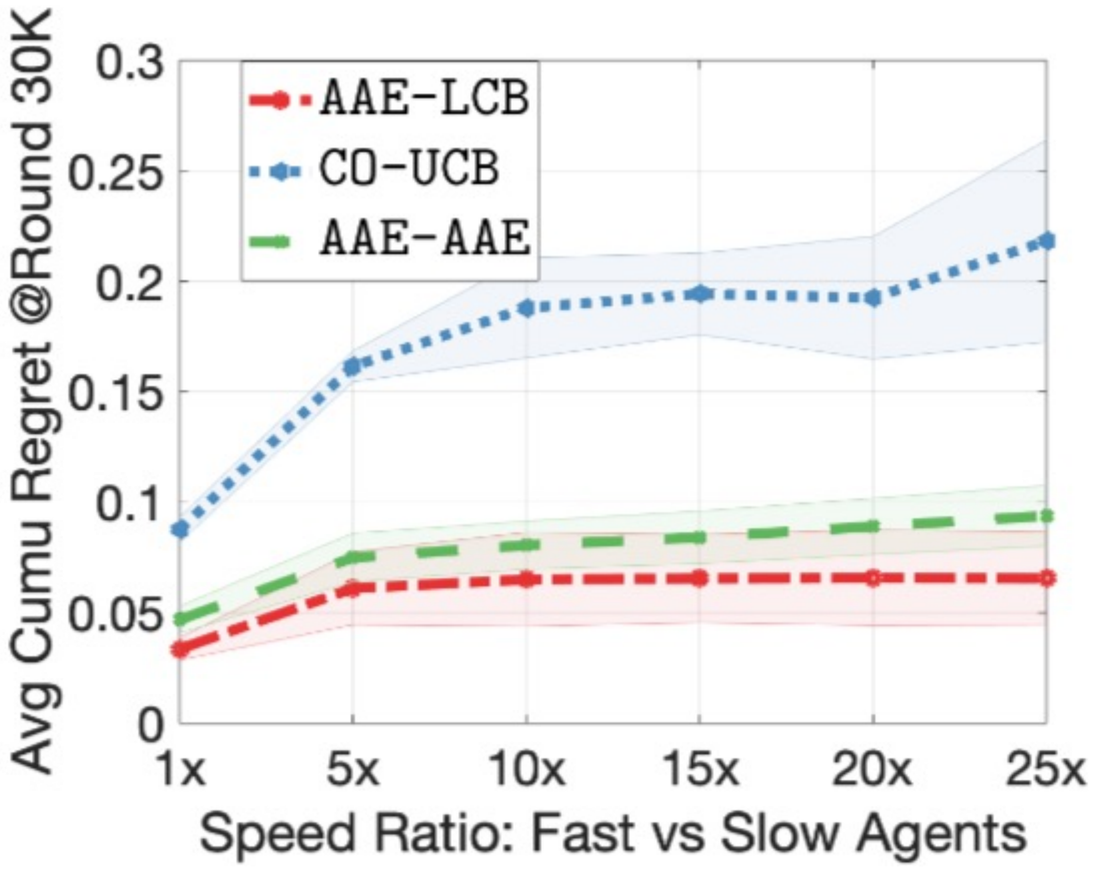
Δ_i - gap of reward means between the optimal arm and arm i

Numerical Results

- 20 agents (10 fast and 10 slow)
- 100 arms, randomly allocated to agents, each having 12
- 30K rounds and 10 simulations for each data point

AAE-AAE: Use AAE to eliminate both local and external suboptimal arms

CO-UCB: Select the arm with largest UCB



AAE-LCB outperforms others with different ratios of action rate between fast and slow agents

Thanks!