Bandit Learning in Decentralized Matching Markets

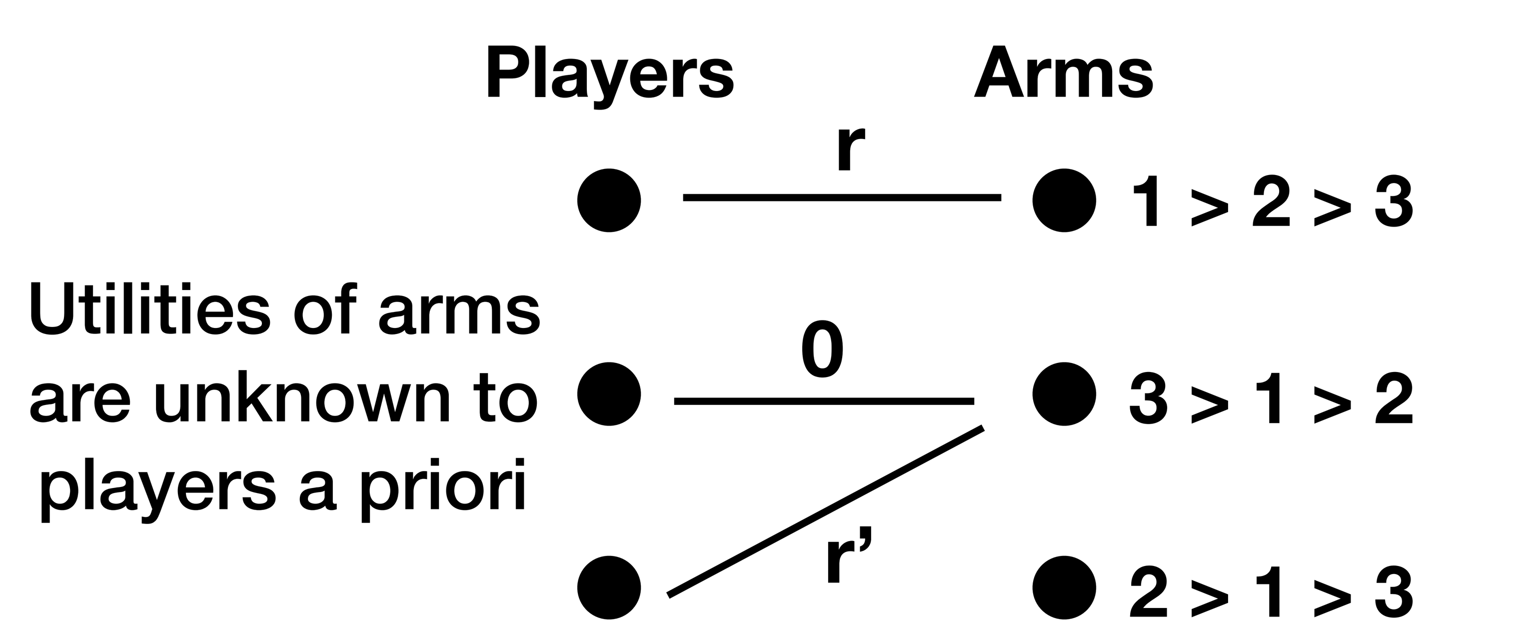
Berkeley UNIVERSITY OF CALIFORNIA

Lydia T. Liu, Feng Ruan, Horia Mania, Michael Jordan

Overview. We study exploration-exploitation tradeoffs in a two-sided matching market where preferences are learned from noisy observations in an *online* manner (Liu et al 2020a).
We focus on the setting where players are *decentralized,* that is, their actions cannot be coordinated by a matching platform, but they can observe past matchings.

Our contributions:

- Introduce a new low-regret algorithm based on randomized conflict avoiding
  - $O(\log(T))$ regret when preferences of the arms over players are shared
  - $O(\log(T))^2$ regret when there are no assumptions on the preferences.
- Where a single player may deviate, the algorithm is incentive-compatible whenever the arms' preferences are shared, but not necessarily so when preferences are general.



Competition: When multiple players pull the same arm only the most preferred player is successful and gets a reward.

Goal: converge to <u>stable</u> matchings despite the players' uncertainty about preferences.

Agent-optimal stable regret of player i at time n:

$$\overline{R}_i(n) := n\underbrace{\mu_i(\overline{m}(i))}_{\text{Mean reward of optimal stable match}} - \sum_{t=1}^{n} \underbrace{\mathbb{E}X_{i,m_t}(t)}_{\text{Reward at time t}}$$

Agent-pessimal stable regret of player i at time n:

$$\underline{R}_i(n) := n\underbrace{\mu_i(\underline{m}(i))}_{\text{Mean reward of pessimal stable match}} - \sum_{t=1}^{n} \underbrace{\mathbb{E}X_{i,m_t}(t)}_{\text{Reward at time t}}$$

Algorithm: Conflict Avoiding UCB with random delays (CA-UCB)

Additional randomness key to reaching a stable matching

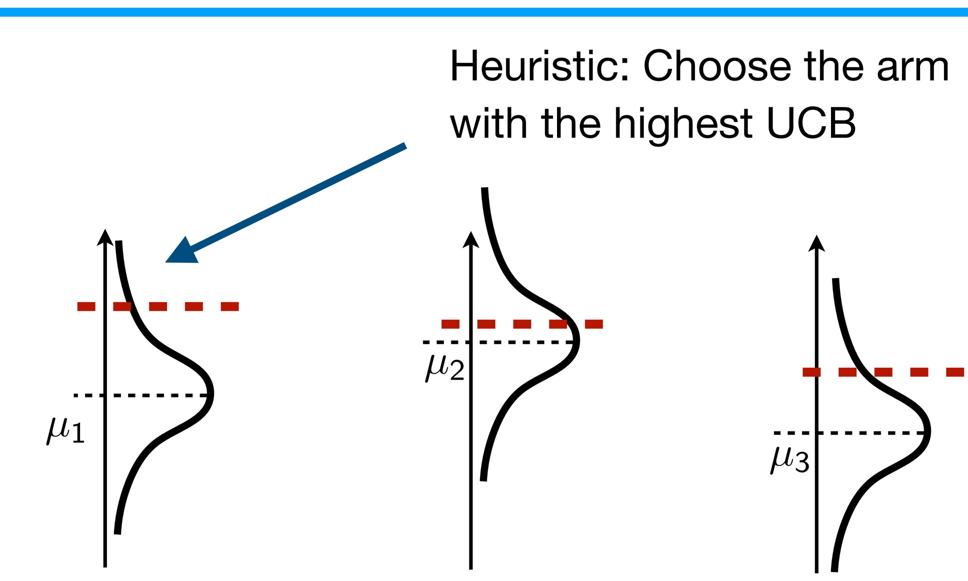
Arms that the player can pull successfully if all other players repeated their actions at *t-1*

At time *t*:

1. <u>Players</u> construct a **plausible set** of arms by looking at the successful matches at time *t-1*
2. Each player, <u>independently,</u>
   - <u>with probability *p*</u> attempts the same arm as time *t-1,*
   - <u>with probability *1-p*</u> attempts the arm in the **plausible set** with the highest UCB.
3. <u>Players</u> receive rewards from matched Arms and update their UCB for the Arm.

The Upper Confidence Bound (UCB)

(Lai and Robbins [1985], Agarwal [1995])



Regret of CA-UCB

Theorem (informal): If there are *N* players and *N* arms and CA-UCB is run for *T* rounds with *0<p<1*, the *pessimal* stable regret of player *i* satisfies, **for arbitrary two-sided preferences,**

$$\underline{R}_i(T) = \mathcal{O}\left(\frac{\log(T)^2 \cdot \boxed{\exp(N^4)}}{\boxed{\Delta^2}}\right)$$

Depends on hyper-parameter *p*

Minimum gap of arms' rewards for all players.

This rate can be improved under assumptions on the preferences. E.g. When **all arms have the same preferences** over players, CA-UCB with *p=0* attains

$$\underline{R}_i(T) = \mathcal{O}\left(\frac{\log(T)^2 \cdot N^3}{\Delta^2}\right)$$

and the algorithm is incentive-compatible for any player.

Convergence of CA-UCB on Random Markets

Is the exponential dependence on N tight? Not for randomly sampled markets.