Competing Bandits in Matching Markets

Lydia T. Liu [0] lydiatliu@berkeley.edu
Horia Mania [0] hmania@berkeley.edu
Michael I. Jordan jordan@cs.berkeley.edu
Department of Electrical Engineering and Computer Science
University of California, Berkeley

Abstract

Stable matching, a classical model for two-sided markets, has long been studied assuming known preferences. In reality agents often have to learn about their preferences through exploration. With the advent of massive online markets powered by data-driven matching platforms, it has become necessary to better understand the interplay between learning and market objectives. We propose a statistical learning model in which one side of the market does not have a priori knowledge about its preferences for the other side and is required to learn these from stochastic rewards. Our model extends the standard multi-armed bandits framework to multiple players, with the added feature that arms have preferences over players. We study both centralized and decentralized approaches to this problem and show surprising exploration-exploitation trade-offs compared to the single player multi-armed bandits setting.

1 Introduction

Decision making under uncertainty is a fundamental problem in machine learning. Literature on bandits and reinforcement learning, though active, mainly study problem settings involving single decisions or decision-makers. In real-world settings, individual decisions must be made in the context of other related decisions. Moreover, such decisions often involve scarcity, with competition among multiple decision-makers. To study such settings we need to blend economics with learning.

In this paper, we present a formal study of such a blend. We focus on the multi-arm bandit (MAB) problem, a core machine-learning problem in which there are K actions giving stochastic rewards, and the learner must discover which action gives maximal expected reward (Bubeck and Cesa-Bianchi, 2012; Lai and Robbins, 1985; Lattimore and Szepesvari, 2019; Thompson, 1933). The bandit problem highlights the fundamental tradeoff between exploration and exploitation, quantified via *regret* analysis. We study an economic version of the problem in which there are *multiple* agents solving a bandit problem, and there is competition—if two or more agents pick the same arm, only one of the agents is given a reward.[1] We assume that the arms have a preference ordering over the agents—a key point of departure from the line of work on multi-player bandits with collisions (Bubeck et al., 2019; Cesa-Bianchi et al., 2016; Liu and Zhao, 2010; Shahrampour et al., 2017)—and this ordering is unknown a priori to the agents.

We are motivated by problems involving two-sided markets that link producers and consumers or workers and employers, where each side sees the other side via a recommendation system, and where there is scarcity on the supply side (for example, a restaurant has a limited number of seats, a street has a limited capacity, or a worker can attend to one task at a time). The overall goal is an economic one—we wish to find a stable matching between producers and consumers. To study the core mathematical problems that arise in such a setting, we have abstracted away the recommendation systems on the two sides, modeling them via the preference orderings and the differing reward functions. Several massive online labor and service markets can be captured by this abstraction; see the end of this section for an illustration of an application. In the context of two-sided markets the arms' preferences can be explict, e.g. when the arms represent entities in the

[0] Equal contribution

Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

[1] Note that Mansour et al. (2018) and Aridor et al. (2019) have used the term "competing bandits" for a different problem formulation where a user can choose between two different bandit algorithms; this differs from our setting where multiple learners compete over scarce resources.

market with their own utilities for the other side of the market, or implicit, e.g. when the arms represent resources their "preferences" encode the skill levels of the agents in securing those resources.

To determine the appropriate notions of equilibria in our multi-agent MAB model, we turn to the literature on stable matchings in two-sided markets Gale and Shapley (1962); Gusfield and Irving (1989); Roth and Sotomayor (1990); Knuth (1997); Roth (2008). Since its introduction by Gale and Shapley (1962), the stable matching problem has had high practical impact, leading to improved matching systems for high-school admissions and labor markets Roth (1984), house allocations with existing tenants Abdulkadiro**u**glu and Sönmez (1999), content delivery networks Maggs and Sitaraman (2015), and kidney exchanges Roth et al. (2005).

In spite of these advances, standard matching models tend to assume that entities in the market know their preferences over the other side of the market. Models that allow unknown preferences usually assume that preferences can be discovered through one or few interactions Ashlagi et al. (2017), e.g., one interview per candidate in the case of medical residents market Roth (1984); Roth and Sotomayor (1990). These assumptions do not capture the statistical uncertainty inherent in problems where data informs preferences. We discuss related work in further detail in Section 4.

In contrast, our work is motivated by modern matching markets which operate at scale and require repeated interactions between the two sides of the market, leading to exploration-exploitation tradeoffs. We consider two-sided markets in which entities on one side of the market do not know their preferences over the other side, and develop matching and learning algorithms that can provably attain a stable market outcome in this setting. Our contributions are as follows:

- We introduce a new model for understanding two-sided markets in which one side of the market does not know its preferences over the other side, but is allowed multiple rounds of interaction. Our model combines work on multi-armed bandits with work on stable matchings. In particular, we define two natural notions of regret, based on stable matchings of the market, which quantify the exploration-exploitation trade-off for each individual agent.

- We extend the Explore-then-Commit (ETC) algorithm for single agent MAB to our multi-agent setting. We consider two versions of ETC: centralized and decentralized. For both versions we prove $\mathcal{O}(\log(n))$ problem-dependent upper bounds on the regret of each agent.

- In addition to the known limitations of ETC for single agent MAB, in Section 3.2 we discuss other issues with ETC in the multi-agent setting. To address these issues we introduce a centralized version of the well-known upper confidence bound (UCB) algorithm. We prove that centralized UCB achieves $\mathcal{O}(\log(n))$ problem-dependent upper bounds on the regret of each agent. Moreover, we show that centralized UCB is incentive compatible.

Most of the above results can be extended to the case where arms also have uncertain preferences over agents in a straightforward manner. For the sake of simplicity, we focus on the setting where one side of market initiates the exploration and leave extensions of our results to future work.

Online labor markets Our model is applicable to matching problems that arise in online labor markets (e.g., Upwork and Taskrabbit for freelancing, Handy for housecleaning) and online crowdsourcing platforms (e.g., Amazon Mechanical Turks). In this case, the employers, each with a stream of similar tasks to be delegated, can be modeled as the players, and the workers can be modeled as the arms. For an employer, the mean reward received from each worker when a task is completed corresponds to how well the task was completed (e.g., did the Turker label the picture correctly?). This differs for each worker due to differing skill levels, which the employer does not know a priori and must learn by exploring different workers. A worker has preferences over different types of tasks (e.g., based on payment or prior familiarity the task) and can only work on one task at a time; hence they will pick their most preferred task to complete out of all the tasks that are offered to them.

2 Problem setting

We denote the set of N agents by $\mathcal{N} = \{p_1, p_2, \ldots, p_N\}$ and the set of K arms by $\mathcal{K} = \{a_1, a_2, \ldots, a_K\}$. We assume $N \leq K$. At time step t, each agent p_i selects an arm $m_t(i)$, where $m_t \in \mathcal{K}^N$ is the vector of all agents' selections.

When multiple agents select the same arm only one agent is allowed to pull the arm, according to the arm's preferences via a mechanism we detail shortly. Then, if player p_i successfully pulls arm $m_t(i)$ at time t, they are said to be *matched* to $m_t(i)$ at time t and they receive a stochastic reward $X_{i,m_t}(t)$ sampled from a 1-sub-Gaussian distribution with mean $\mu_i(m_t(i))$.

Each arm a_j has a fixed known ranking π_j of the agents, where $\pi_j(i)$ is the rank of player p_i. In other words, π_j is a permutation of $[N]$ and $\pi_j(i) < \pi_j(i')$ implies

that arm a_j prefers player p_i to player $p_{i'}$. If two or more agents attempt to pull the same arm a_j, there is a *conflict* and only the top-ranked agent successfully pulls the arm to receive a reward; the other agent(s) $p_{i'}$ is said to be *unmatched* and does not receive any reward, that is, $X_{i',m_t}(t) = 0$. As a shorthand, the notation $p_i \succ_j p_{i'}$ means that arm a_j prefers player p_i over $p_{i'}$. When arm a_j is clear from context, we simply write $p_i \succ p_{i'}$. Similarly, the notation $a_j \succ_i a_{j'}$ means that p_i prefers arm a_j over $a_{j'}$, i.e. $\mu_i(j) > \mu_i(j')$.

We now proceed to develop suitable notions of *regret* for the agents. Recall that when preferences are known on both sides, the goal is to attain a *stable matching*, where no pair of agent and arm prefer each other over their respective matches and hence no pair has the incentive to deviate. Given the full preference rankings of the arms and players, arm a_j is called a *valid match* of player p_i if there exists a stable matching according to those rankings such that a_j and p_i are matched.

We say a_j is the *optimal match* of agent p_i if it is the most preferred valid match. Similarly, we say a_j is the *pessimal match* of agent p_i if it is the least preferred valid match. Given complete preferences, the Gale-Shapley (GS) algorithm (Gale and Shapley, 1962) finds a stable matching after repeated proposals from one side of the market to the other. The matching returned by the GS algorithm is always optimal for the proposing side and pessimal for the non-proposing side (Knuth, 1997). We denote by $\overline{m}$ and $\underline{m}$ the functions from $\mathcal{N}$ to $\mathcal{K}$ that define the optimal and pessimal matchings of the players according to the true preferences of the players and arms.

In the online matching problem where agent preferences are a priori unknown, agents aim to perform well relative to their best action in hindsight[2], as is typical in online learning. Thus, it is natural to define the *agent-optimal stable regret* of agent p_i as

$$\overline{R}_i(n) := n\mu_i(\overline{m}(i)) - \sum_{t=1}^{n} \mathbb{E}X_{i,m_t}(t), \qquad (1)$$

because when the arms' mean rewards are known the GS algorithm outputs the optimal matching $\overline{m}$. However, as we show in Section 3.2, there are natural algorithms which cannot achieve sublinear agent-optimal stable regret. Therefore, we also consider the *agent-pessimal stable regret*, defined as

$$\underline{R}_i(n) := n\mu_i(\underline{m}(i)) - \sum_{t=1}^{n} \mathbb{E}X_{i,m_t}(t). \qquad (2)$$

[2]When the stable matching is unique, it is not hard to see that the best action for any agent—if all agents knew their own preferences and are maximizing reward—is to choose their unique valid match.

When the stable matching is unique, the agent-optimal and agent-pessimal stable regrets coincide. Also, we note that when $N > K$ stable matches will not match all players with arms. Then, we denote $m(i) = \emptyset$ if player p_i does not have a match in m and we let $\mu_i(\emptyset)$ be the reward player p_i receives when not matched. Then, the results presented below extend to this case. For simplicity, we assume $N \leq K$ throughout.

The central question of our investigation is as follows:

How to achieve a *sequence of matchings* where all agents have *low stable regret*?

Several interaction settings are of interest:

Centralized: At each time step the agents are required to send a ranking of the arms to a matching platform. Then, the platform decides the action vector m_t. In this work we consider two platforms. The first platform (shown in on the left of Table 1) outputs a random assignment for a number of time steps and then computes the agent-optimal stable matching according to the agents' preferences. The second platform (shown on the right of Table 1) takes in the agent's preferences at each time step and outputs a stable matching between the agents and arms. Both platforms ensure that there will be no conflicts between the agents. The first platform corresponds to an explore-then-commit strategy. When the second platform is used the agents must rank arms in a way which enables exploration and exploitation. We show that ranking according to upper confidence bounds yields $\mathcal{O}(\log(n))$ agent-pessimal stable regret.

Decentralized: Agents observe each other's actions and the outcomes of the ensuing conflicts, but do not have a platform for coordination and communication. We can also ask what happens if, after selecting an arm, agents observe whether they lost a conflict and if they successfully pull an arm they observe their own reward, but they do not observe any other information. Both decentralized cases are interesting and we leave their study for future work.

3 Multi-agent bandits with a platform

3.1 Centralized Explore-then-Commit

In this section we give a guarantee for the explore-then-commit planner defined in Algorithm 1(left). At each iteration, each agent p_i updates their mean reward for arm j to be

$$\widehat{\mu}_{i,j}(t) = \frac{1}{T_{i,j}(t)} \sum_{s=1}^{t} \mathbb{1}\{m_s(i) = j\} X_{i,m_s}(s), \qquad (3)$$

where $T_{i,j}(t) = \sum_{s=1}^{t} \mathbb{1}\{m_t(i) = j\}$ is the number of times agent p_i successfully pulled arm a_j. At each time step, player p_i ranks the arms in decreasing order according to $\widehat{\mu}_{i,j}(t)$ and sends the resulting ranking $\widehat{r}_{i,t}$ to the platform. As seen in Table 1, for the first hK time steps, the platform assigns players to arms cyclically, ensuring that each agent samples every arm h times. We now provide a regret analysis of centralized ETC. The proof is deferred to Appendix A.

Theorem 1. *Suppose all players rank arms according to the empirical mean rewards* (3) *and submit their rankings to the explore-then-commit platform. Let* $\overline{\Delta}_{i,j} = \mu_i(\overline{m}(i)) - \mu_i(j)$, $\overline{\Delta}_{i,\max} = \max_j \overline{\Delta}_{i,j}$, *and* $\Delta = \min_{i\in[N]} \min_{j:\,\overline{\Delta}_{i,j}>0} \overline{\Delta}_{i,j} > 0$. *Then, the expected agent-optimal regret of player* p_i *is upper bounded by*

$$\overline{R}_i(n) \le h\sum_{j=1}^{K} \overline{\Delta}_{i,j} + (n - hK)\overline{\Delta}_{i,\max} NK \exp\left(-\frac{h\Delta^2}{4}\right).$$

In particular, if $h = \max\left\{1, \frac{4}{\Delta^2}\log\left(1 + \frac{n\Delta^2 N}{4}\right)\right\}$, *we have*

$$\overline{R}_i(n) \le \max\left\{1, \frac{4}{\Delta^2}\log\left(1 + \frac{n\Delta^2 N}{4}\right)\right\}\sum_{j=1}^{K}\overline{\Delta}_{i,j} + \frac{4K\overline{\Delta}_{i,\max}}{\Delta^2}\log\left(1 + \frac{n\Delta^2 N}{4}\right).$$

This result shows that centralized ETC achieves $\mathcal{O}(\log(n))$ agent-optimal stable regret when the number of exploration rounds is chosen appropriately. As is the case for single agent ETC, centralized ETC requires knowledge of both the horizon n and the minimum gap Δ (see, e.g., Lattimore and Szepesvari, 2019, Chapter 6). However, a glaring difference between the settings is that in the latter the regret of each agent scales with $1/\Delta^2$, where Δ is the minimum reward gap between the optimal match and a suboptimal arm across all agents. In other words, the regret of an agent might depend on the suboptimality gap of other agents. Example 2 shows that this dependence is real in general and not an artifact of our analysis. Moreover, while single agent ETC achieves $\mathcal{O}(\sqrt{n})$ problem-independent regret, Example 2 shows that centralized ETC does not have this desirable property. Finally, $\sum_{j=1}^{K}\overline{\Delta}_{i,j}$ could be negative for some agents. Therefore, some agents can have negative agent-optimal regret, an effect that never occurs in the single agent MAB problem.

Example 2 (The dependence on $1/\Delta^2$ cannot be improved in general)**.** *Let* $\mathcal{N} = \{p_1, p_2\}$ *and* $\mathcal{K} = \{a_1, a_2\}$ *with true preferences:*

$$p_1 \colon a_1 \succ a_2 \qquad a_1 \colon p_1 \succ p_2$$
$$p_2 \colon a_2 \succ a_1 \qquad a_2 \colon p_1 \succ p_2.$$

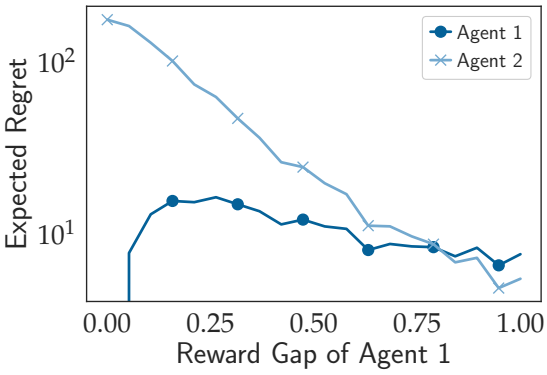

Figure 1: The empirical performance of centralized UCB in the setting described in Example 2

The agent-optimal stable matching is given by $\overline{m}(1) = 1$ *and* $\overline{m}(2) = 2$. *Both* a_1 *and* a_2 *prefer* p_1 *over* p_2. *Therefore, at the end of the exploration stage* p_1 *is matched to their top choice arm while* p_2 *is matched to the remaining arm. In order for* p_2 *to be matched to their optimal arm,* p_1 *must correctly determine that they prefer* a_1 *over* a_2. *The number of exploration rounds would then have to be* $\Omega(1/\overline{\Delta}_{1,2}^2)$ *where* $\overline{\Delta}_{1,2} = \mu_1(1) - \mu_1(2)$. *Hence, when* $\overline{\Delta}_{1,2} \le 1/\sqrt{n}$, *the regret of* p_2 *is* $\Omega(n\overline{\Delta}_{2,1})$. *Figure 1 depicts this effect empirically; we observe that a smaller gap* $\overline{\Delta}_{1,2}$ *causes* p_2 *to have larger regret.*

In Figure 1, there are two agents and two arms. Player p_2 *receives Gaussian rewards from the arms* a_1, a_2 *with means* 0 *and* 1 *respectively and variance* 1. *Player* p_1 *receives Gaussians rewards* Δ *and* 0 *from the arms* a_1 *and* a_2. *Both arms prefer* p_1 *over* p_2. *Figure 1 shows the regret of each agent as a function of* Δ *when we run centralized UCB with horizon* 400 *and average over* 100 *trials.*

3.2 Centralized UCB

We saw that centralized ETC achieves $\mathcal{O}(\log(n))$ agent-optimal regret for all agents. However, centralized ETC must know the horizon n and the minimum gap Δ between an optimal arm and a suboptimal arm. While knowing the horizon n is feasible in certain scenarios, knowing Δ is not plausible. It is known that single agent ETC achieves $\mathcal{O}(n^{2/3})$ when the number of exploration rounds is chosen deterministically without knowing Δ, and there are also known methods for adaptively choosing the number of exploration rounds so that single agent ETC achieves $\mathcal{O}(\log(n))$ Lattimore and Szepesvari (2019). However, in our setting, the $\mathcal{O}(n^{2/3})$ guarantee does not hold because the suboptimality gaps of one agent affect the regret of other agents, and the known adaptive stopping times cannot

input: h, and the preference ranking π_j of all arms $a_j \in \mathcal{K}$, the horizon length n

for $t = 1, \dots, T$ **do**
  if $t \leq hK$ **then**
    $m_t(i) \leftarrow a_{t+i-1 \pmod K)+1}$, $\forall i$.
  else if $t = hK + 1$ **then**
    Receive rankings $\widehat{r}_{i,t}$ from all p_i.
    Compute agent-optimal stable matching $m_t(i)$ according to $\widehat{r}_{i,t}$ and π_j.
  else
    $m_t(i) \leftarrow m_{hK+1}(i)$, $\forall i$.

input: the preference ranking π_j of all arms $a_j \in \mathcal{K}$

for $t = 1, \dots, T$ **do**
  Receive rankings $\widehat{r}_{i,t}$ from all p_i.
  Compute agent-optimal stable matching m_t according to all $\widehat{r}_{i,t}$ and π_j.

Table 1: (*left*) Explore-then-Commit Platform. (*right)* Gale-Shapley Platform.

be implemented because the platform does not observe the agents' rewards. Therefore, it is necessary to find methods which do not need to know Δ.

Another drawback of centralized ETC is that it requires agents to learn concurrently, i.e. players must explore randomly at the same time. Hence, even if a player knew their preferences a priori, they would still be required to explore randomly in order to guarantee low regret for all players. The Gale-Shapley Platform shown in Table 1(right) resolves this problem, always outputting a matching that is stable—in fact, agent-optimal—according to the rankings received from the agents. We derive an upper bound on the agent-pessimal stable regret in this setting when all agents use upper confidence bounds to rank arms. In Section 3.3 we show this method is incentive compatible.

Before proceeding with the analysis we define more precisely the UCB method employed by each agent and also introduce several technical concepts. At each time step the platform matches agent p_i with arm $m_t(i)$. Each player p_i successfully pulls arm $m_t(i)$, receives reward $X_{i,m_t}(t)$, and updates their empircal mean for $m_t(i)$ as in (3). They then compute the upper confidence bound

$$u_{i,j}(t) = \begin{cases} \infty & \text{if } T_{i,j}(t) = 0, \\ \widehat{\mu}_{i,j}(t) + \sqrt{\frac{3\log t}{2T_i(t-1)}} & \text{otherwise.} \end{cases} \quad (4)$$

Finally, each player p_i orders the arms according to $u_{i,j}(t)$ and computes the ranking $\widehat{r}_{i,t+1}$ so that a higher upper confidence bound means a better rank, e.g. $\arg\max_j u_{i,j}(t)$ is ranked first in $\widehat{r}_{i,t+1}$.

Let m be an injective function from the set of players $\mathcal{N}$ to the set of arms $\mathcal{K}$; hence m is the matching where $m(i)$ is the match of agent i. Then, let $T_m(t)$ be the number of times matching m is played by time t. We say a matching is *truly stable* if it is stable according to the true preferences induced by the mean rewards of the arms, and *non-truly stable*, otherwise.

For agent p_i and arm a_ℓ we consider the set $M_{i,\ell}$ of non-truly stable matchings m such that $m(i) = \ell$. Let $\underline{\Delta}_{i,\ell} = \mu_i(\underline{m}(i)) - \mu_i(\ell)$.

Then, since any truly-stable matching yields agent-pessimal regret smaller or equal than zero for all agents, we can upper-bound the agent-pessimal regret of agent i as follows:

$$\underline{R}_i(n) \leq \sum_{\ell:\, \underline{\Delta}_{i,\ell} > 0} \underline{\Delta}_{i,\ell} \left(\sum_{m \in M_{i,\ell}} \mathbb{E} T_m(n) \right). \quad (5)$$

For any matching m that is non-truly stable there must exist an agent p_j and an arm a_k, different from arm $m(j)$, such that the pair (p_j, a_k) is a *blocking pair* according to the true preferences μ, i.e. $\mu_j(k) > \mu_j(m(j))$ and arm a_k is either unmatched or $\pi_k(j) > \pi_k(m^{-1}(k))$. We say a triplet $(p_j, a_k, a_{k'})$ blocks a matching when p_j is matched with $a_{k'}$ and (p_j, a_k) is a blocking pair. Let $B_{j,k,k'}$ be the set of all matchings blocked by the triplet $(p_j, a_k, a_{k'})$. Given a set S of matchings, we say a set Q of triplets $(p_j, a_k, a_{k'})$ is a *cover* of S if

$$\bigcup_{(p_j, a_k, a_{k'}) \in Q} B_{j,k,k'} \supseteq S.$$

Let $\mathcal{C}(S)$ denote the set of covers of S. Also, let $\Delta_{j,k,k'} = \mu_j(k) - \mu_j(k')$. Now we state our result.

Theorem 3. *When all agents rank arms according to the upper confidence bounds* (4) *and submit their preferences to the Gale-Shapley Platform, the agent-pessimal regret of agent p_i up to time n, $\underline{R}_i(n)$, is upper-bounded by*

$$\sum_{\ell:\, \underline{\Delta}_{i,\ell} > 0} \underline{\Delta}_{i,\ell} \left[\min_{Q \in \mathcal{C}(M_{i,\ell})} \sum_{\substack{(p_j, a_k, a_{k'}) \\ \in Q}} \left(5 + \frac{6\log(n)}{\Delta^2_{j,k,k'}} \right) \right].$$

Theorem 3 offers a problem-dependent $\mathcal{O}(\log(n))$ upper bound guarantee on the agent-pessimal stable regret of

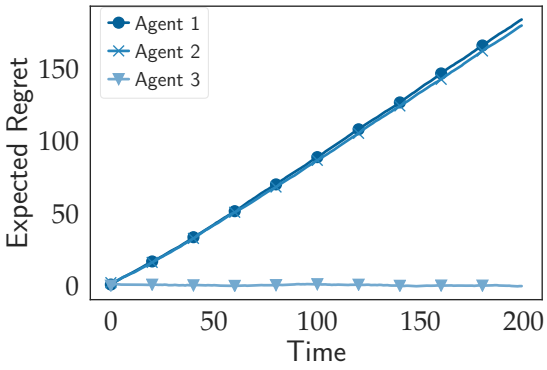

Figure 2: The empirical performance of centralized UCB in the setting described in Example 4

each agent p_i. Similarly to the case of centralized ETC, the regret of one agent depends on the suboptimality gaps of other agents. However, we saw in Section 3.1 that centralized ETC achieves $\mathcal{O}(\log(n))$ agent-optimal stable regret, a stronger notion of regret. Example 4 shows that centralized UCB cannot yield sublinear agent-optimal stable regret in general. While centralized ETC has stronger regret guarantees, it requires knowledge of the reward gaps and of the horizon of the problem. Also, centralized ETC requires all players to have synchronized exploration rounds. UCB with the Gale-Shapley platform does not have these drawbacks.

Example 4 (Centralized UCB does not achieve sublinear agent-optimal stable regret)**.** *Let $\mathcal{N} = \{p_1, p_2, p_3\}$ and $\mathcal{K} = \{a_1, a_2, a_3\}$, with true preferences given by:*

$$p_1 : a_1 \succ a_2 \succ a_3 \qquad a_1 : p_2 \succ p_3 \succ p_1$$
$$p_2 : a_2 \succ a_1 \succ a_3 \qquad a_2 : p_1 \succ p_2 \succ p_3$$
$$p_3 : a_3 \succ a_1 \succ a_2 \qquad a_3 : p_3 \succ p_1 \succ p_2.$$

The agent-optimal stable matching is (p_1, a_1), (p_2, a_2), (p_3, a_3). When p_3 incorrectly ranks $a_1 \succ a_3$ and the other two players submit their correct rankings, the Gale-Shapley Platform outputs the matching (p_1, a_2), (p_2, a_1), (p_3, a_3). In this case p_3 will never correct their mistake because they never get matched with a_1 again, and hence their upper confidence bound for a_1 will never shrink. Figure 2 illustrates this example; the optimal regret for p_1 and p_2 is seen to be linear in n.

In Figure 2, the rewards of the arms for each agent are Gaussian with variance 1. The mean rewards of the arms are set so that the preference structure shown in Example 4 is satisfied. For agents p_1 and p_2, the gap in mean rewards between consecutive arms is 1. For agent p_3 the gap between arms a_1 and a_3 is 0.05. Figure 2 shows the performance of centralized UCB, averaged over 100 trials, as a function of the horizon.

Proof of Theorem 3. Let $L_{j,k,k'}(n)$ be the number of times agent p_j pulls arm $a_{k'}$ when the triplet $(p_j, a_k, a_{k'})$ is blocking the matching selected by the platform. Then, by definition

$$\sum_{m \in B_{j,k,k'}} T_m(n) = L_{j,k,k'}(n). \tag{6}$$

By the definition of a blocking triplet we know that if p_j pulls $a_{k'}$ when $(p_j, a_k, a_{k'})$ is blocking, they must have a higher upper confidence bound for $a_{k'}$ than for a_k. In other words, we are trying to upper bound the expected number of times the upper confidence bound on $a_{k'}$ is higher than that of the better arm a_k when we have the guarantee that each time this event occurs $a_{k'}$ is successfully pulled. Therefore, standard analysis for the single agent UCB (e.g., Bubeck and Cesa-Bianchi, 2012, Chap. 2) shows that

$$\mathbb{E}L_{j,k,k'}(n) \leq 5 + \frac{6\log(n)}{\Delta^2_{j,k,k'}}. \tag{7}$$

The conclusion follows from equations (5) and (6). □

To better understand the guarantee of Theorem 3 we consider two examples in which the markets have a special structure which enables us to simplify the upper bound on the regret. Moreover, in Corollary 7 we consider the worst case upper bound over possible coverings of matchings.

Example 5 (Global preferences)**.** *Let $\mathcal{N} = \{p_1, \cdots, p_N\}$ and $\mathcal{K} = \{a_1, \cdots, a_K\}$. We assume the following preferences: $p_i : a_1 \succ \cdots \succ a_K$ and $a_j : p_1 \succ \cdots \succ p_N$. In other words all agents have the same ranking over arms, and all arms have the same ranking over agents. Hence, the unique stable matching is (p_1, a_1), (p_2, a_2), $\ldots$, (p_N, a_N). Moreover, for any p_i and a_ℓ we can cover the set of matchings $M_{i,\ell}$ with the triplets (p_i, a_k, a_ℓ) for all k with $1 \leq k \leq i$. Then, Theorem 3 implies (8) once we observe that $\Delta_{i,k,\ell} \geq \underline{\Delta}_{i,\ell}$ for all $k \leq i$.*

$$\underline{R}_i(n) \leq 5i \sum_{\ell=i+1}^{K} \underline{\Delta}_{i,\ell} + \sum_{\ell=i+1}^{K} \frac{6i\log(n)}{\underline{\Delta}_{i,\ell}}. \tag{8}$$

Figure 3 illustrates this example empirically, displaying the pessimal stable regret of 5 out of 20 agents. In this experiment, there are 20 agents and 20 arms. The rewards of the arms are Guassian with variance 1. The mean reward gap between consecutive arms is 0.1. Figure 3 shows the performance of centralized UCB, averaged over 50 trials, as a function of the horizon.

As one can see, the 1st-ranked agent has sublinear regret, consistent with (8), while the 20th-ranked agent has negative regret and our upper bound is indeed 0.

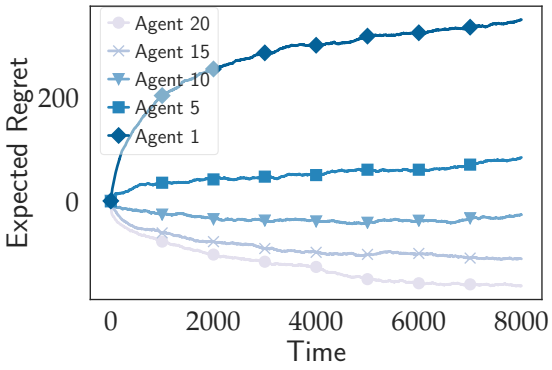

Figure 3: The empirical performance of centralized UCB in the setting described in Example 5

Example 6 (Unique pairs). *Let $\mathcal{N} = \{p_1, \cdots, p_N\}$ and $\mathcal{K} = \{a_1, \cdots, a_N\}$ and assume that agent p_i prefers arm a_i the most and that arm a_i prefers agent p_i the most. Therefore, the unique stable matching is (p_1, a_1), (p_2, a_2), $\ldots$, (p_N, a_N). Then, we can cover each set $M_{i,\ell}$ with the triplet (p_i, a_i, a_ℓ). Therefore, Theorem 3 implies (9); note that the right-hand side is identical to the guarantee for single agent UCB:*

$$\underline{R}_i(n) \leq 5 \sum_{\ell \neq i}^{K} \underline{\Delta}_{i,\ell} + \sum_{\ell \neq i}^{K} \frac{6 \log(n)}{\underline{\Delta}_{i,\ell}}. \tag{9}$$

Corollary 7. *Let $\Delta = \min_i \min_{j,j'} |\mu_i(j) - \mu_i(j')|$. When all players follow the centralized UCB method, the regret of p_i can be upper bounded as follows*

$$\underline{R}_i(n) \leq \max_{\ell} \underline{\Delta}_{i,\ell} \left(6NK^2 + 12 \frac{NK \log(n)}{\Delta^2} \right).$$

Proof. We consider the covering (j, k, k') composed of all possible triples with $\mu_j(k) > \mu_j(k')$. Then, Theorem 3 implies the result because $\sum_{k' \colon \mu_j(k') < \mu_j(k)} \frac{1}{\Delta^2_{j,k,k'}} \leq \sum_{\ell=1}^{K} \frac{1}{\ell^2 \Delta^2} \leq \frac{2}{\Delta^2}$. □

3.3 Honesty and Strategic Behavior

Classical results show that in the agent-proposing GS algorithm, no single agent can improve their match by misrepresenting their preferences, assuming that the other agents and arms submit their true preferences (Roth, 1982; Dubins and Freedman, 1981). The result generalizes to coalition of agents. Moreover, when there is a unique stable matching, the Dubins-Freedman Theorem says that no arms or agents can benefit from misrepresenting their preferences (Dubins and Freedman, 1981).

The ETC Platform does not allows agents to choose which arms to explore. In this case, the classical results on honesty in agent-proposing GS apply; the agents are incentivized to submit the rankings according to their current mean estimates. When agents have some degree of freedom to explore over multiple rounds, it is no longer clear if any agents, or arms, can benefit from misrepresenting their preferences in some of the rounds. In general, one agent's preferences can influence not only the matches of other agents, but also their reward estimates. One might be able to improve their regret by capitalizing on the ranking mistakes of other agents. The possibilities for long-term strategic behavior are more diverse than in the single-round setting.

In general, the optimal regret for a player can be negative if the player is on average getting rewards higher than its optimal stable arm, as seen in Figure 3.

We now show that when all agents except one submit their UCB-based preferences to the GS Platform, the remaining agent has an incentive to also submit preferences based on their UCBs, so long as they do not have multiple stable arms. This result is a lower bound on the optimal regret of the remaining agent, hence establishing that they have limited gains from deviating from their UCB-based preferences.

First, we establish the following lemma, which is an upper bound on the expected number of times the remaining agent can pull an arm that is better than their optimal match, regardless of what preferences they might have submitted to the platform.

Lemma 8. *Let $T_l^i(n)$ be the number of times an agent i pulls an arm l such that the mean reward of l for i is greater than i's optimal match. Then*

$$\mathbb{E}[T_l^i(n)] \leq \min_{Q \in \mathcal{C}(M_{i,\ell})} \sum_{(j,k,k') \in Q} \left(5 + \frac{6 \log(n)}{\Delta^2_{j,k,k'}} \right). \tag{10}$$

Proof. If agent i is matched with arm l in any round, the matching m must be unstable according true preferences. We claim that there must exist a blocking triplet (j, k, k') where $j \neq i$.

Arguing by contradiction, we suppose otherwise, that all blocking triplets in m only involve agent i. By Theorem 4.2 in Abeledo and Rothblum (1995), we can go from the matching m to a μ-stable matching, by iteratively *satisfying* block pairs in a 'gender consistent' order O. To satisfy a blocking pair (k, j), we break their current matches, if any, and match (k, j) to get a new matching. Doing so, agent i can never get a worse match than l or become unmatched as the algorithm proceeds, so the matching remains unstable—a contradiction. Hence there must exist a $j \neq i$ such that j is part of a blocking triplet in m. In particular, agent j must be submitting its UCB preferences.

The result follows from Equation (7) and the identity

$$\mathbb{E}[T_l^i(n)] = \sum_{m \in M_{i,\ell}} \mathbb{E}T_m(n).$$

□

Lemma 8 directly implies the following lower bound on the remaining agent's optimal regret.

Proposition 9. *Suppose all agents other than p_i submit preferences according to the UCBs* (4) *to the GS Platform. Then the following lower bound on agent i's optimal regret holds:*

$$\overline{R}_i(n) \geq \sum_{\ell \colon \overline{\Delta}_{i,l} < 0} \overline{\Delta}_{i,l} \left[\min_{Q \in \mathcal{C}(M_{i,\ell})} \sum_{(j,k,k') \in Q} \left(5 + \frac{6 \log(n)}{\Delta^2_{j,k,k'}} \right) \right].$$

Therefore, there is no sequence of preferences that an agent can submit to the GS Platform that would give them negative optimal regret greater than $\mathcal{O}(\log n)$ in magnitude. When there is a unique stable matching, Proposition 9 shows that no agent can gain significantly above and beyond the mean reward of their optimal stable arm by submitting preferences other than their UCB rankings. When there exist multiple stable matchings, however, Proposition 9 leaves open the question of whether any agent can submit a sequence of preferences that achieves super-logarithmic negative *pessimal* regret for themselves, when all other agents are playing their UCB preferences. In other words, can an agent do significantly better than its pessimal stable arm, by possibly deviating from their UCB rankings?

4 Related work

Since its introduction by Thompson (1933), the stochastic multi-armed bandit problem has inspired a rich body of work spanning different settings, algorithms, and guarantees (Lai and Robbins, 1985; Bubeck and Cesa-Bianchi, 2012; Lattimore and Szepesvari, 2019).

There has been recent interest in the MAB literature in problems with multiple, interacting players (Cesa-Bianchi et al., 2016; Shahrampour et al., 2017). In one popular formulation known as *bandits with collision*, multiple players choose from the same set of arms, and if two or more players choose the same arm, no reward is received by any player (e.g. Liu and Zhao, 2010; Anandkumar et al., 2011; Avner and Mannor, 2014; Bubeck et al., 2019; Lugosi and Mehrabian, 2018; Rosenski et al., 2016). This differs from our formulation, in which arms have preferences and the most preferred player receives a reward, while the other players selecting the arm do not.

A variant of this problem is where agents have different preferences over arms. Then, Bistritz and Leshem (2018)'s algorithm approximately finds the maximum matching of players to arms with $\mathcal{O}(\log(n)^{2+\kappa})$ regret. However, stable matching does not reduce to maximum matching in general, so such guarantees do not apply to matching with two-sided preferences.

The two-sided matching problem has also been studied in sequential settings. Das and Kamenica (2005) performed an empirical study of a two-sided matching problem with uncertain preferences. Johari et al. (2017) studied a sequential matching problem, where participants are one of several types and the goal is to learn the type of agents on one side of the market.

Ashlagi et al. (2017) considered the communication and preference learning cost of stable matching. Their model formulates preference learning as querying a noiseless choice function, rather than obtaining noisy observations of one's underlying utility. Different players can query their choice functions independently; hence congestion in the preference learning stage is not captured by this model. In many markets, obtaining information about the other side of the market itself can lead to congestion and thus the need for strategic decision. For example, Roth and Sotomayor (1990, chap. 10) noted that graduating medical students go to interviews to ascertain their own preferences for hospitals, but the interviews that a student can schedule are limited. Our model begins to capture such tradeoffs by introducing statistical uncertainty in the preferences of one side of the market and providing a natural mode of interaction between the learning agents.

5 Discussion

We have proposed a new model for dynamic matching in markets under uncertain preferences. The model blends learning and competition, and captures two desirable notions: stability and sample efficiency. We presented two natural algorithms which combine classical ideas from multi-armed bandits and stable matching. Our focus in the current paper was the centralized UCB method, which we proved enjoys small regret for each player and ensures that the market converges quickly to a stable configuration.

There are many additional questions that can be studied in this model, including problems with incomplete information, decentralized matching protocols and shared reward structures. We have already seen that the uncertainty of one agent can depress the long-term utility of other agents, and we expect to uncover other interactions between learning and strategic decision making in this model.

References

Atila Abdulkadiro**u**glu and Tayfun Sönmez. House allocation with existing tenants. *Journal of Economic Theory*, 88(2):233–260, 1999.

Hernan Abeledo and Uriel G. Rothblum. Paths to marriage stability. *Discrete Applied Mathematics*, 63: 1–12, 10 1995.

A. Anandkumar, N. Michael, A. K. Tang, and A. Swami. Distributed algorithms for learning and cognitive medium access with logarithmic regret. *IEEE J.Sel. A. Commun.*, 29(4):731–745, April 2011.

Guy Aridor, Kevin Liu, Aleksandrs Slivkins, and Zhiwei Steven Wu. Competing bandits: The perils of exploration under competition. *The 20th ACM conference on Economics and Computation (EC)*, 2019.

Itai Ashlagi, Mark Braverman, Yash Kanoria, and Peng Shi. Communication requirements and informative signaling in matching markets. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, EC '17, pages 263–263, New York, NY, USA, 2017. ACM.

Orly Avner and Shie Mannor. Concurrent bandits and cognitive radio networks. In Toon Calders, Floriana Esposito, Eyke Hüllermeier, and Rosa Meo, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 66–81, Berlin, Heidelberg, 2014. Springer Berlin Heidelberg.

Ilai Bistritz and Amir Leshem. Distributed multi-player bandits - a game of thrones approach. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 7222–7232. Curran Associates, Inc., 2018.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

Sébastien Bubeck, Yuanzhi Li, Yuval Peres, and Mark Sellke. Non-Stochastic Multi-Player Multi-Armed Bandits: Optimal Rate With Collision Information, Sublinear Without. *arXiv e-prints*, art. arXiv:1904.12233, Apr 2019.

Nicolò Cesa-Bianchi, Claudio Gentile, Yishay Mansour, and Alberto Minora. Delay and cooperation in nonstochastic bandits. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 605–622, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR.

Sanmay Das and Emir Kamenica. Two-sided bandits and the dating market. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, IJCAI'05, pages 947–952, San Francisco, CA, USA, 2005. Morgan Kaufmann Publishers Inc.

L. E. Dubins and D. A. Freedman. Machiavelli and the Gale-Shapley Algorithm. *The American Mathematical Monthly*, 88(7):485–494, 1981.

D. Gale and L. S. Shapley. College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15, 1962.

Dan Gusfield and Robert W. Irving. *The Stable Marriage Problem: Structure and Algorithms*. MIT Press, Cambridge, MA, USA, 1989.

Ramesh Johari, Vijay Kamble, and Yash Kanoria. Matching while learning. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, EC '17, pages 119–119, New York, NY, USA, 2017. ACM.

Donald E. Knuth. *Stable Marriage and Its Relation to Other Combinatorial Problems*, volume 10 of *CRM Proceedings and Lecture Notes*. American Mathematical Society, 1997.

T.L Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.*, 6(1): 4–22, March 1985.

Tor Lattimore and Csaba Szepesvari. *Bandit Algorithms*. Cambridge University Press (To Appear), 2019.

Keqin Liu and Qing Zhao. Distributed learning in multi-armed bandit with multiple players. *Trans. Sig. Proc.*, 58(11):5667–5681, November 2010.

Gábor Lugosi and Abbas Mehrabian. Multiplayer bandits without observing collision information. *CoRR*, abs/1808.08416, 2018.

Bruce M Maggs and Ramesh K Sitaraman. Algorithmic nuggets in content delivery. *ACM SIGCOMM Computer Communication Review*, 45(3):52–66, 2015.

Yishay Mansour, Aleksandrs Slivkins, and Zhiwei Steven Wu. Competing bandits: Learning under competition. In *9th Innovations in Theoretical Computer Science Conference, ITCS 2018, January 11-14, 2018, Cambridge, MA, USA*, pages 48:1–48:27, 2018.

Jonathan Rosenski, Ohad Shamir, and Liran Szlak. Multi-player bandits – a musical chairs approach. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 155–163, New York, New York, USA, 20–22 Jun 2016. PMLR.

Alvin E. Roth. The economics of matching: Stability and incentives. *Mathematics of Operations Research*, 7(4):617–628, 1982.

Alvin E Roth. The evolution of the labor market for medical interns and residents: a case study in game theory. *Journal of political Economy*, 92(6):991–1016, 1984.

Alvin E. Roth. Deferred acceptance algorithms: History, theory, practice, and open questions. Working Paper 13225, National Bureau of Economic Research, July 2007.

Alvin E. Roth. Deferred acceptance algorithms: history, theory, practice, and open questions. *International Journal of Game Theory*, 36(3):537–569, Mar 2008.

Alvin E. Roth and Marilda A. Oliveira Sotomayor. *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*. Econometric Society Monographs. Cambridge University Press, 1990. doi: 10.1017/CCOL052139015X.

Alvin E Roth, Tayfun Sönmez, and M Utku Ünver. Pairwise kidney exchange. *Journal of Economic theory*, 125(2):151–188, 2005.

S. Shahrampour, A. Rakhlin, and A. Jadbabaie. Multi-armed bandits in multi-agent networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2786–2790, March 2017.

William R. Thompson. On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika*, 25(3-4): 285–294, 12 1933.