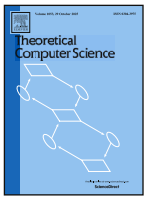
Theoretical Computer Science 1071 (2026) 115846

Contents lists available at ScienceDirect

Theoretical Computer Science

journal homepage: www.elsevier.com/locate/tcs



Bandit learning in matching markets with relative feedback

Fang Kong [a,b], Xiaoxi Zhang [c], Xiao Huang [a], Lijun Zhang [d], Shuai Li [a,*]

[a] Shanghai Jiao Tong University, Shanghai, China
[b] Southern University of Science and Technology, Shenzhen, Guangdong Province, China
[c] Columbia University, New York, NY, United States
[d] Nanjing University, Nanjing, Jiangsu, China

ARTICLE INFO

Section Editor: Dr. Vladimiro Sassone
Handling Editor: Katie Harris

Keywords:
Multi-armed bandits
Stable matching
Matching markets
Relative feedback

ABSTRACT

The two-sided matching market problem has been extensively studied in the literature. How to find a stable matching is a key focus in the field. A significant body of recent work considers scenarios where one side of the market (players) has uncertain preferences and learns them through the absolute rewards obtained during repeated interactions with the other side (arms). A common assumption in these works is that arms deterministically resolve conflicts when faced with multiple players. However, in practical applications, the arms' selection process may also be stochastic due to fluctuations in players' performances. Under such circumstances, it becomes challenging for players to observe absolute rewards that quantify the arms' satisfaction. Instead, the relative feedback about which applicant wins in a competition is often more realistic. In this paper, we investigate the pure exploration problem for bandit learning in matching markets where players need to additionally learn the uncertain preferences of arms based on more practical relative feedback. We show that given confidence level $\delta \in (0,1)$, the market can reach the player-optimal stable matching in at most $O(\max\{N,K\}\log(1/\delta)/\Delta^2 + NK\log(1/\delta)/\epsilon^2)$ rounds with probability at least $1-\delta$, where N, K correspond to market size, ϵ represents arms' relative preference gap, and Δ corresponds to the players' preference gap. We also conduct experiments to verify the performances of the proposed algorithms.

1. Introduction

The problem of the matching market has been studied on a large scale of real applications including marriage problems, college admission [1], house allocation [2], and labor market [3]. There are typically two sides of participants, such as students and colleges in the college admission scenario or workers and employers in the labor market. Each participant usually has a preference ranking over participants on the other side. If a worker applies to an employer and meanwhile the employer accepts to hire this worker, such a relationship between the worker and the employer can be viewed as a matching pair. A matching between two sides of the market is a set of matching pairs. Stability is a key concept to describe the equilibrium state of the market. It ensures that no participant has the motivation to break the existing matching relationships [1,4]. How to find a stable matching has been widely studied in the literature [1,2,4–6].

The above works mainly focus on the most idealized case where each participant has a known and deterministic preference ranking over the other side. However, in real applications, the preferences of the market participants are usually uncertain. For example, in labor markets, the workers may not know the styles of employers before being matched. So they cannot submit a complete preference

* Corresponding author.
E-mail address: shuaili8@sjtu.edu.cn (S. Li).

https://doi.org/10.1016/j.tcs.2026.115846
Received 10 June 2025; Received in revised form 6 November 2025; Accepted 20 February 2026
Available online 24 February 2026

ranking over all candidates at the start of the game process. With the emergence of online marketplaces such as the online labor market Upwork and crowdsourcing platform Amazon Mechanical Turk where employers have lots of similar tasks to delegate, they are expected to learn the uncertain preferences during iterative matching processes through these numerous tasks.

The multi-armed bandits (MAB) problem is a classic framework to characterize the learning process during iterative interactions [7,8]. The problem consists of a single player and K arms. Every time the player selects an arm, it receives an absolute reward value sampled from the arm's reward distribution. Both regret minimization and pure exploration are common objectives of a bandit algorithm. The former aims to minimize the cumulative regret in a specified horizon while the latter hopes to identify the optimal arm in as few rounds as possible [8].

A rich line of recent works study the bandit learning problem in matching markets [9–16]. They consider that one-side participants (players) have uncertain preferences and would learn them based on repeated interactions with arms. Specifically, after each match with an arm, the player would measure his satisfaction by an absolute reward. The preference value of a player over an arm is the expected satisfaction, and the stable matching in the market is defined in terms of the preference rankings induced by these values. The objective is to minimize the stable regret for each player, which is defined as the difference between the reward of the arm in the stable matching and the received reward during interactions.

To better design the players' learning strategy, this line of work assumes that arms deterministically decide which player to accept based on a pre-defined preference ranking. However, in real applications, the arms' decision process may not be fixed. Consider an employer interviewing multiple candidates, the performances of the candidates usually fluctuate over time. So the best choice of arms would also change over time. In this case, assuming arms' deterministic selection may not always be satisfied. And to identify the arms' underlying preferences, it is not realistic for players to observe the absolute rewards scored by arms. Instead, the relative feedback on which candidate is selected by the employer would be more natural.

In this paper, we study the bandit learning problem in matching markets where arms' acceptance criteria are also stochastic, and the players would learn them based on more practical relative feedback. Since the regret minimization objective may not be achieved in the stable matching [17], we study the pure exploration objective which aims to identify the stable arm for each player in as few rounds as possible. We propose a new algorithm called online Gale-Shapley (OGS) for players to learn the preferences of arms and find the stable matching. Given confidence level δ, we show that our OGS can find the stable arm for each player in at most $O(\max\{N, K\}\log(1/\delta)/\Delta^2 + NK\log(1/\delta)/\epsilon^2)$ rounds with probability at least $1-\delta$, where N is the number of players, K is the number of arms, Δ refers to players' preference gap, and ϵ corresponds to the difference of arms' relative preferences. Such a framework also extends previous works on dueling bandits [18–21] from learning with pairwise comparisons to multiple comparisons, and our sample complexity guarantee matches the lower bound when only a duel of players and 1 arm exists [18].

2. Related work

The matching market model has been studied for many years [2–4]. We refer to the case where both sides of participants have known deterministic preferences as the *offline* setting. How to find a stable matching in the offline setting has been widely studied in the literature [4–6] and the Gale-Shapley algorithm is one of the most widely known [1].

Since market participants are usually uncertain about their preferences, Das and Kamenica [22] start to consider the *online* setting where the preferences need to be learned. They propose algorithms for a special market in which all participants on the same side share the same preferences and verify their performances empirically. Liu et al. [9] later consider a variant that only participants on one side have uncertain preferences and initialize the theoretical study by providing upper bounds for the stable regret. Liu et al. [9] mainly design strategies for a centralized setting where a central platform coordinates participants' behavior to avoid conflicts. Since such a platform is usually unavailable, the following works mainly focus on a more general decentralized setting [10–16,23,24].

Apart from the players' unknown preferences, Pagare and Ghosh [25] and Zhang and Fang [26] study that arms may also have uncertain preferences. They assume the preference function and observation of the arm side is the same as the player side. Arms would learn their uncertain preferences based on the received absolute reward and would assist players in finding the players' most preferred stable matching. These works may lack a motivation for arms to assist players in finding the players' most preferred, which is the arms' least preferred, stable matching. Different from these works, we still regard arms acceptance criteria as a component of the environment and players would learn them based on more practical relative feedback.

Apart from this line, there are also other works related to learning uncertain preferences in matching markets. Ghosh et al. [27] consider the non-stationary environments where the preferences of market participants may change over time. Min et al. [28] focus on the Markov matching market with state transitions when matchings occur. Jagadeesan et al. [29] incorporate monetary transfer into the matching process.

In addition to minimizing the regret, pure exploration is another widely studied problem in bandit literature [8]. There are mainly two objectives: minimizing the time budget required for finding the optimal arm within the given confidence level; and maximizing the confidence of identifying the optimal arm in a given time budget [30–34]. Our work is more related to the former. Hosseini et al. [17] identify that the stable matching may not be the one that minimizes the stable regret and study the pure exploration problem in matching markets. This work considers the complexity of player-arm matches, i.e., the number of successful matchings between each player-arm pair, to learn the stable matching. Our work is different from theirs since we also incorporate the conflict among market participants in the sample complexity and aim to bound the number of total rounds required to find the stable matching.

All the above works adopt absolute rewards as feedback to learn unknown preferences. However, in many real applications, the preference is the result of multiple factors overlapping, making it challenging to be represented by a simple utility value. This motivates the study of learning with relative feedback. The problem of dueling bandits [18–20,35] is proposed for learning unknown

knowledge which is hard to measure by absolute values. To find the best arm, in each duel, 2 arms are compared and only one of them wins. The difference between 2 arms is the probability of one winning the duel with another. Motivated by real applications where there may be multiple players who select arms, Chen et al. [21,36,37] study the model of combinatorial dueling bandits to find the best candidate-position matching. This problem needs a central platform to sample a duel of two candidates and observe the outcome of which one wins in the duel. In contrast, we consider a more general setting where each player makes his own choices and more than two players are allowed to compete together. It is still worth noting that their best matching is defined as Borda winner which maximizes the total preference probabilities, and Condorcet winner whose probability of winning is more than half, while our goal is to find a stable matching with consideration of participants' preferences on both sides.

3. Setting

There are two sides of participants in the market: N players on one side and K arms on the other side. The player set is denoted by $\mathcal{N} = \{p_1, p_2, \dots, p_N\}$ and the arm set is denoted by $\mathcal{K} = \{a_1, a_2, \dots, a_K\}$.

For simplicity, we first assume each player $p_i \in \mathcal{N}$ has a known preference ranking $\pi_i = (\pi_{i,j})_{j\in[K]}$ over arms. The arm with a higher rank is more preferred by the player, suggesting that each player p_i prefers arm $a_{\pi_{i,j}}$ to $a_{\pi_{i,j'}}$ for any $1 \le j < j' \le K$. How to learn the players' unknown preference ranking is deferred to Section 5.

Similar to that in labor markets, arms (employers) usually hope to accept the best-performed player (worker) when faced with multiple candidates. Due to the players' fluctuating performances, the arms' selection may change over time. However, since the skill levels of players are typically within a fixed range, each performance observation can be viewed as a random realization of their underlying skill level. So the selection process of arms can be characterized by a underlying preference probability. Specifically, the preferences of each arm a_j towards the set $\mathcal{N}$ of players can be modeled by a probability set $\mathrm{P}_j := \mathrm{P}_j^{\mathcal{N}} = (\mathrm{P}_{j,i}^{\mathcal{N}})_{p_i\in\mathcal{N}}$ where $\sum_{p_i\in\mathcal{N}} \mathrm{P}_{j,i}^{\mathcal{N}} = 1$. When the set $\mathcal{N}$ of players propose to arm a_j, a_j will select one of them. Based on the Plackett-Luce model [38,39], the probability of being selected for each player $p_i \in \mathcal{N}$ is $\mathrm{P}_{j,i}^{\mathcal{N}}$ and the preference ranking of arm a_j towards different players in $\mathcal{N}$ is the ranking of this probability. The arm $\hat{i}_j^* \in argmax_{i\in\mathcal{N}} p_{j,i}^{\mathcal{N}}$ with the largest probability is regarded as a_j's most preferred player among $\mathcal{N}$. Without loss of generality, the preference probabilities of arms are assumed to be distinct, i.e., $\mathrm{P}_{j,i}^{\mathcal{N}} \neq \mathrm{P}_{j,i'}^{\mathcal{N}}$ for any arm a_j and different players $p_i, p_{i'}$ as existing works. Intuitively, this property states that arms have different preferences over players, which is a common assumption in matching markets literature [1,4,9–12,14–17]. And if a subset $\mathcal{N}' \subseteq \mathcal{N}$ of players propose to the arm a_j, the probability of $p_i \in \mathcal{N}'$ being selected by a_j is $\mathrm{P}_{j,i}^{\mathcal{N}'} = \mathrm{P}_{j,i}^{\mathcal{N}} / \sum_{p_{i'}\in\mathcal{N}'} \mathrm{P}_{j,i'}^{\mathcal{N}}$. For simplicity, denote $\mathrm{P}_j^{\mathcal{N}'}$ as the set of selecting probabilities when a_j is faced with $\mathcal{N}'$.

We study decentralized markets where players independently make decision on which arm to propose. At each round t, denote $A_i(t)$ as the selected arm of player $p_i \in \mathcal{N}$. Corresponding to the other side, let $A_j^{-1}(t) = \{p_i : A_i(t) = a_j\}$ be the set of players who propose to a_j. Each arm would only accept one player from $A_j^{-1}(t)$, which can be regarded as being sampled from preference probability $\mathrm{P}_j^{A_j^{-1}(t)}$. We denote $\bar{A}_j^{-1}(t) \in A_j^{-1}(t)$ as the player who is selected by arm a_j. Motivated by real applications such as labor markets where workers usually update their working experience on their profiles, we also assume players can observe which player is successfully selected by arms [10,12,14,27,40]. In this case, all players know which one is selected by each arm. And if no player proposes to a_j, we simply set $\bar{A}_j^{-1}(t) = \emptyset$.

Stability is a key concept characterizing the equilibrium of the market. A stable matching is a one-to-one mapping from the player set to the arm set where no player-arm pair exists such that they both prefer to be matched with each other rather than their current partner. Given players' preferences $(\pi_i)_{i\in[N]}$ and arms' preferences $(\mathrm{P}_j)_{j\in[K]}$, there may be more than one stable matching. Let $M = \{m := (i, m_i)_{i\in[N]} : m \text{ is stable}\}$ be the set of all stable matchings and denote $m^* \in M$ as the player-optimal stable matching with $\pi_{i,m_i^*} \le \pi_{i,m_i}$ for any $m \in M$. As discussed in Hosseini et al. [17], the regret minimization objective may not ensure the stability. So we study the pure exploration problem in matching markets aiming to find the player-optimal stable matching $m^* \in M$ in as few rounds as possible. Formally speaking, given a confidence level $\delta \in (0, 1)$, the aim is to minimize the sample complexity (total number of rounds) of identifying the player-optimal stable arm m_i^* for each $p_i \in \mathcal{N}$ with probability at least $1 - \delta$.

4. Online Gale-Shapley algorithm

In this section, we present our online Gale-Shapley (OGS) algorithm (Algorithm 1) for the pure exploration problem in two-sided matching markets. Our OGS algorithm is inspired by the idea of the offline Gale-Shapley (GS) algorithm [1], which finds the player-optimal stable matching when both sides of participants have known and deterministic preferences. The GS algorithm proceeds in several steps. In the first step, each player proposes to its most preferred arm according to its preference ranking. Each arm then accepts its most preferred player among those who propose to it and rejects others. In the following step, players still propose to the most preferred arm from those who did not reject it in the previous steps and arms still only accept their most preferred one. It has been shown that when all players or all arms are matched, the algorithm successfully reaches the player-optimal stable matching [1]. And since each arm can reject one player at most once, the algorithm would stop in at most NK steps.

However, since players' performances fluctuate over time, the accepted players among different rounds may not be fixed. Thus players need more observations on arms' selections to determine their preference probabilities. For any player p_i, let s denote the highest ranking of the arm that does not reject p_i in the previous steps, which is initialized to 1 (Line 2). Similar to the GS algorithm, at each step ℓ, players still propose to its most preferred one $a_{\pi_{i,s}}$ from those who did not reject it in previous steps (Line 5). To make

Algorithm 1 Online Gale-Shapley (from view of p_i).

1: Input: player set $\mathcal{N}$, arm set $\mathcal{K}$; confidence level δ
2: Initialize: $s = 1$; $\hat{\mathrm{P}}_{j,i} = 0, \forall j \in [K], i \in [N]$
3: **for** step $\ell = 1, 2, \cdots$ **do**
4: **for** round $t = 1, 2, \cdots$ **do**
5: Select $A_i(t) = a_{\pi_{i,s}}$
6: For each $a_j \in \mathcal{K}$, observe $\bar{A}_j^{-1}(t)$
7: **if** $\cup_{j\in[K]} \bar{A}_j^{-1}(t) = \mathcal{N}$ or $\cup_{i\in[N]} \bar{A}_i(t) = \mathcal{K}$ **then**
8: **return** $a_{\pi_{i,s}}$
9: **end if**
10: **for** $j \in [K], i' \in [N]$ **do**
11: Update $\hat{\mathrm{P}}_{j,i'}(t) = \frac{1}{t} \sum_{\tau=1}^{t} \mathbb{1}\left\{ \bar{A}_j^{-1}(\tau) = p_{i'} \right\}$
12: **end for**
13: Compute

$$c_t := \sqrt{\frac{8(\log(2t^2/\delta) + N \log 2 + \log NK)}{2t}} \tag{1}$$

14: **if** $\forall j$ with $\bar{A}_j^{-1}(t) \neq \emptyset : \exists \hat{i}_j^*$ s.t. $\hat{\mathrm{P}}_{j,\hat{i}_j^*}(t) - c_t > \hat{\mathrm{P}}_{j,i'}(t) + c_t, \forall i' \neq \hat{i}_j^*$ **then**
15: **if** $j = \pi_{i,s}$ and $i \neq \hat{i}_j^*$ **then**
16: $s = s + 1$
17: **end if**
18: stop step ℓ and enter the next step $\ell + 1$
19: **end if**
20: **end for**
21: **end for**

it clear, the word "an arm rejects player p_i" in our OGS algorithm means that player p_i realizes that it is not the most preferred player of the arm.

During each step ℓ, each player p_i always proposes to its most preferred arm $a_{\pi_{i,s}}$. Based on the relative feedback on who is selected by arms (Line 6), players can update the estimation of arms' preference probabilities. Let $\hat{P}_{j,i'}(t)$ represent the estimated preference probability of arm a_j toward player $p_{i'}$ at round t of the current step, which is computed as the number of times $p_{i'}$ is selected by a_j (Line 11).

We will later show that the distance between the estimated preference probability and the real preference probability can be upper bounded by c_t defined in Eq. (1). With this observation, when there exists a player whose estimated preference probability is $2c_t$ higher than all of the others for each arm (Line 14), we believe players successfully determine all arms' most preferred one among those who propose to it. Then the algorithm can proceed to the next step (Line 18). If player p_i is such a player with the highest estimated preference probability, then p_i can be regarded as being successfully accepted in this step and it would propose to the same arm in the next step. Otherwise, p_i is regarded as being rejected by $a_{\pi_{i,s}}$ and then updates $s = s + 1$ to propose to the next preferred arm (Line 16). Once all players or all arms are successfully matched (Line 7), the algorithm is considered to find the player-optimal stable matching and will return the current arm $a_{\pi_{i,s}}$ for player p_i (Line 8).

4.1. Theoretical results

In this section, we provide the theoretical analysis for our OGS algorithm. Before giving the main result, we first introduce some useful notations. For convenience, for any $\mathcal{N}' \subseteq \mathcal{N}$, we set $\mathrm{P}_{j,i}^{\mathcal{N}'} = 0$ for all players $p_i \notin \mathcal{N}'$ who are not in $\mathcal{N}'$. At each step ℓ, denote $\mathcal{N}_{j,\ell}$ as the set of players who propose to arm a_j. We first define corresponding gaps to measure the hardness of the learning task.

Definition 1. For any arm $a_j \in \mathcal{K}$ and step $\ell \in [NK]$,

$$\epsilon_{j,\ell} = \min_{i,i' \neq i'' \in \mathcal{N}_{j,\ell}} \left\{ \left| \mathrm{P}_{j,i'}^{\mathcal{N}_{j,\ell}} - \mathrm{P}_{j,i''}^{\mathcal{N}_{j,\ell}} \right|, \left| \mathrm{P}_{j,i}^{\mathcal{N}_{j,\ell}} \right| > 0 \right\} \tag{2}$$

is defined as the minimum non-negative preference probability gap that needs to be identified among the player set $\mathcal{N}_{j,\ell}$ and further define $\epsilon := \min_{j,\ell} \epsilon_{j,\ell}$ as the minimum preference probability gap among all arms.

Theorem 1 gives the sample complexity of the algorithm to find the player-optimal stable matching.

Theorem 1. *Given any confidence level $\delta \in (0, 1)$, our OGS algorithm (Algorithm 1) returns the player-optimal stable matching with sample complexity*

$$O\left(\frac{NK}{\epsilon^2} \log\left(\frac{NK \cdot 2^N}{\delta \epsilon^4} \right) \right). \tag{3}$$

When all players have the same preference rankings over arms, the above sample complexity can be further improved.

Corollary 1. *(Global preferences) Suppose all players have the same preference rankings over arms. Formally, $\pi_i = \pi_{i'}$ for any $p_i, p_{i'} \in \mathcal{N}$. Then given any confidence level $\delta \in (0,1)$, the sample complexity of Algorithm 1 to find the player-optimal stable matching is*

$$O\left(\frac{\min\{N,K\}}{\epsilon^2}\log\left(\frac{NK\cdot 2^N}{\delta\epsilon^4}\right)\right). \tag{4}$$

4.2. Proof sketch of Theorem 1 and Corollary 1

Before providing the main proof, we first introduce some useful notations. Denote $\hat{\mathrm{P}}_{j,i}^{\mathcal{N}_{j,\ell}}(t)$ as the estimated probability of arm a_j towards player p_i at round t when the set $\mathcal{N}_{j,\ell}$ of players propose to a_j. Further, let $\hat{\mathrm{P}}_{j}^{\mathcal{N}_{j,\ell}}(t) = (\hat{\mathrm{P}}_{j,i}^{\mathcal{N}_{j,\ell}}(t))_{i\in[N]}$. Define the following event

$$\mathcal{F}_{j,\ell} := \left\{\exists t>0 : \left\|\hat{\mathrm{P}}_{j}^{\mathcal{N}_{j,\ell}}(t) - \mathrm{P}_{j}^{\mathcal{N}_{j,\ell}}\right\|_1 > c_t\right\} \tag{5}$$

to represent that there exists a round t when the distance between the estimated preference probability and the real preference probability is larger than c_t, where c_t is defined in Eq. (1). The following Lemma shows that this bad event holds with probability smaller than δ/NK.

Lemma 1. *For any $a_j \in \mathcal{K}$ and step ℓ, $\mathbb{P}(\mathcal{F}_{j,\ell}) \le \delta/NK$.*

With the result of Lemma 1, we can further identify the maximum number of rounds required for players to identify the favorite player of arm a_j.

Lemma 2. *Suppose the set $\mathcal{N}_{j,\ell}$ of players propose to arm a_j at step ℓ with $|\mathcal{N}_{j,\ell}| > 1$. Then based on $\neg\mathcal{F}_{j,\ell}$, at most*

$$O\left(\frac{1}{\epsilon^2}\log\left(\frac{NK\cdot 2^N}{\delta\cdot\epsilon^4}\right)\right) \tag{6}$$

rounds are needed when $\hat{i}_j^* \in \mathcal{N}_{j,\ell}$ is identified such that $\hat{\mathrm{P}}_{j,\hat{i}_j^*} - c_t > \hat{\mathrm{P}}_{j,i'} + c_t$, for all $i' \neq \hat{i}_j^*$.

The following Lemma further illustrates that the identified player $\hat{i}_j^*$ is indeed the most preferred player of a_j in $\mathcal{N}_{j,\ell}$.

Lemma 3. *Suppose the set $\mathcal{N}_{j,\ell}$ of players propose arm a_j at step ℓ with $|\mathcal{N}_{j,\ell}| > 1$. Then based on $\neg\mathcal{F}_{j,\ell}$, the identified player $\hat{i}_j^*$ is actually the most preferred player of a_j in $\mathcal{N}_{j,\ell}$. That is, $\hat{i}_j^* \in argmax_{i\in[N]}\mathrm{P}_{j,i}^{\mathcal{N}_{j,\ell}}$.*

Due to the space limit, the proof of Lemmas 1–3 is deferred to Appendix A. We now provide proof of Theorem 1 and Corollary 1.

Proof of Theorem 1 and Corollary 1. With the result of Lemma 3, the accepted player of each arm a_j at each step ℓ is actually its most preferred one. Thus Algorithm 1 can be exactly regarded as the online version of the GS algorithm: in each step, players propose to its most preferred arms among those who did not reject it previously for several rounds until they determine which player is accepted by each arm. In general markets, since each arm rejects each player at most once, at most NK rejections happen before finding the stable matching. Based on Lemma 2 and the definition of ϵ, each rejection can be determined in $O\left(\log\left((NK2^{N+1})/(\delta\epsilon^4)\right)/\epsilon^2\right)$ rounds with probability at most $1-\delta/NK$. Considering the total NK steps, we conclude that the algorithm can achieve the stable matching in $O\left(NK\log\left((NK2^{N+1})/(\delta\epsilon^4)\right)/\epsilon^2\right)$ rounds with probability larger than $1-\delta$. Thus Theorem 1 is proved.

In markets where players have the same preferences, all players would propose to the same arm at each step ℓ before finding the stable arm. Thus at most $\min\{N,K\}$ steps proceed before the algorithm stops. Adopting similar analysis as Theorem 1, we can conclude the algorithm can achieve the stable matching in $O\left(\min\{N,K\}\log\left((NK2^{N+1})/(\delta\epsilon^4)\right)/\epsilon^2\right)$ rounds with probability larger than $1-\delta$.

4.3. Discussions

Discussion on the term 2^N. We first provide further clarification on the origin and role of the exponential term 2^N in our theoretical analysis. Although this term might appear to suggest an exponential dependence on N, it only appears inside a logarithmic factor, i.e., in $\log(2^N)$. Therefore, the overall sample complexity does *not* scale exponentially with N. The logarithmic factor $\log(2^N)$ arises naturally from the tight concentration analysis of the categorical distribution. In proof of Lemma 1, we jointly consider all N dimensions of the categorical probability vector and take the supremum over the support $\{-1,+1\}^N$. This joint treatment ensures a uniform bound over all categories and leads to the logarithmic dependence on 2^N. Alternatively, one could analyze each coordinate independently using Hoeffding's inequality. Specifically, bounding the deviation probability $\Pr(|p-\hat{p}|>\epsilon) \le \Pr\left(\exists i\in[N] : |p_i-\hat{p}_i|>\epsilon/N\right)$ and applying the Hoeffding inequality independently to each coordinate yields $\epsilon \ge N\sqrt{\log(N/\delta)/t}$, where p denotes the expectation of an N-dimension categorical distribution, $\hat{p}$ denotes the empirical mean, ϵ denotes and tolerance and t denotes the number of observations. In contrast, our joint analysis gives $\epsilon \ge \sqrt{\log(2^N/\delta)/t} = \sqrt{(\log(1/\delta)+N\log 2)/t}$, which provides a tighter upper bound on the sample complexity. Hence, although the logarithmic term involves 2^N, it reflects a sharper concentration bound. Moreover, the resulting bound in our set-dueling setting remains consistent with that of the classical one-to-one setting in its dependence on δ and t, further confirming the soundness of theoretical guarantees.

Optimality of the results. We build upon the *Optimally Stable Bandits (OSB)* setting introduced in Sankararaman et al. [23] and discuss the lower bound. Specifically, For each arm a_j, its most preferred player is p_j, i.e., $\mathtt{P}^{\mathcal{N}}_{j,j} \in argmax_{i\in[N]}\mathtt{P}^{\mathcal{N}}_{j,i}$. And all players share an identical preference ranking over arms: $a_1 \succ a_2 \succ a_3 \succ \cdots \succ a_K$. Under these preferences, the market admits a unique stable matching $m^\star = \{(p_1, a_1), (p_2, a_2), \ldots, (p_{\min(N,K)}, a_{\min(N,K)})$.

For analytical convenience, we adopt an equivalent formulation by reversing the roles of proposals-letting the *arms* propose instead of the players. At each round, each arm a_j can propose to any subset of players. Each player accepts the most preferred arm among those that propose to it and subsequently enters a competition on this arm. The arms observe which player wins the competition and update their internal estimates of the preference (or winning) probabilities. As the underlying market possesses a unique stable matching and the arm-proposing mechanism preserves the same information as the player-proposing OSB formulation, the lower bound established in this setting remains valid in our model.

For this problem instance, fix an arm a_j with $j \le \min\{N, K\}$, we can analyze the lower bound for the cumulative unstability over T rounds as

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\{\bar{A}(t) \text{ is unstable}\}\right] &\ge \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\{a_j \text{ is not accepted by } p_j\}\right]\\ &\ge \sum_{j'<j}\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\{a_{j'} \text{ is accepted by } p_j\}\right]\\ &\ge \sum_{j'<j}\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}\{a_{j'} \text{ is accepted by } p_j \text{ and } p_{j'}\}\right].\end{aligned}$$

By Lemma 1 in Komiyama et al. [41], for arm $a_{j'}$, any algorithm needs to compare p_j and $p_{j'}$ for at least $\Omega(\log T/(\mathtt{P}_{j,j'} - \mathtt{P}_{j,j})^2)$ times to be convinced that player p_j is inferior to player $p_{j'}$ and thus arm $a_{j'}$ does not propose to p_j. Thus we can derive a lower bound of order $\Omega(\min\{N, K\}\log T/\epsilon_0^2)$ where $\epsilon_0 := \min_{j\in[K]}\max_{i\neq j}\mathtt{P}_{j,j} - \mathtt{P}_{j,i}$.

Our sample-complexity upper bound can be naturally translated into an upper bound on the cumulative instability, yielding a rate of $O(NK\log T/\epsilon^2)$ for general markets and $O(\min\{N, K\}\log T/\epsilon^2)$ for global preference by setting δ as $1/T$. In the global preference setting, our upper bound matches this lower bound up to constant factors. In the general preference setting, our upper bound is worse by at most a factor of $\max\{N, K\}$. We leave whether a tight regret can be obtained as a future direction.

Observing the relative feedback on who is selected. In real applications, the available observations for workers to learn the uncertain preferences of employers are very limited. Previous works model such preferences as the expected absolute rewards, which can be observed by employers [9–14,23,24]. But from the view of the workers, they are not able to get the full information of the employer's satisfaction with itself as well as with other workers. The relative feedback on which worker is successfully selected is a more practical feedback type to learn the employers' uncertain preferences. It is worth noting that such available feedback does not violate the essence of decentralization. In previous works on the centralized setting [9,40], players and arms would communicate their current preferences estimations in each round to get a non-collision allocation. But our algorithm does not require such real-time communication. It works only based on the relative feedback of which player is successfully matched. Since workers usually update their working experience on the profile in labor markets, this feedback is available in real applications. Such observation type is also adopted in previous decentralized works [10,12,14,27,40].

5. Unknown preferences of players

In this section, we generalize the setting and introduce how to learn the players' unknown preferences. For the preferences of the player side, we adopt the framework of previous online matching markets [9–15,17,23]. Specifically, the preference ranking $\pi_i = (\pi_{i,j})_{j\in[K]}$ of each player p_i is induced by the unknown preference values $(\mu_{i,j})_{j\in[K]}$ where $\mu_{i,j} > 0$ for any j. To ensure the preference ranking is strict as in previous works, the preference values towards different arms are assumed to be distinct [1,4,9–15,17,23]. At a round t, if p_i is successfully selected by the proposed arm $A_i(t)$, it would receive a reward $X_{i,A_i(t)}(t)$, which is a 1-subgaussian random variable with expectation $\mu_{i,A_i(t)}$. If p_i is not matched with $A_i(t)$, it only receives a zero reward $X_{i,A_i(t)}(t) = 0$. Players can learn their unknown preference rankings based on these received random rewards.

To estimate players' preference rankings, we introduce a uniform-exploration algorithm which is inspired by [13,14]. Due to the space limit, the algorithm chart is deferred to Appendix. The algorithm mainly proceeds in two phases: in the first phase, each player estimates a unique index; and in the second phase, all players explore arms in a round-robin way until all of them have a good estimated ranking. This algorithm also runs in a decentralized manner.

The first phase consists of N rounds. In the 1st round, all players propose to arm a_1. The only selected player by a_1 would get the index 1 and never propose to a_1 again during this phase. In the 2nd round, the remaining set of players still propose to arm a_1. One of them is selected and gets an index 2. Each player would get a unique index after N rounds in this process.

The second phase is an exploration phase which consists of several epochs. At the ℓth epoch, players would first explore for 2^ℓ rounds and then determine whether the preferences are estimated well at the end of the epoch. Recall that all players get a unique index during the first phase. They can conduct exploration in a round-robin way to avoid conflict. Based on this design, each arm would only receive one offer at each round such that the proposing player can be successfully selected and receive reward $X_{i,A_i(t)}(t)$. Each player p_i would then update the estimated preference value $\hat{\mu}_{i,A_i(t)} = \left(\hat{\mu}_{i,A_i(t)}\cdot T_{i,A_i(t)} + X_{i,A_i(t)}(t)\right)/\left(T_{i,A_i(t)} + 1\right)$ and the

observed time $T_{i,A_i(t)} := T_{i,A_i(t)} + 1$. After the last exploration round of the ℓth epoch, each player p_i would compute a confidence interval for each preference value $\mu_{i,j}$ with the upper bound $\mathrm{UCB}_{i,j} := \hat{\mu}_{i,j} + \sqrt{2\log(2NKt^2/\delta)/T_{i,j}(t)}$ and the lower bound $\mathrm{LCB}_{i,j} := \hat{\mu}_{i,j} - \sqrt{2\log(2NKt^2/\delta)/T_{i,j}(t)}$ (when $T_{i,j} = 0$, UCB and LCB are set to be ∞ and $-\infty$, respectively). If player p_i finds that the confidence intervals of all arms are disjoint, it can determine its preference ranking has been estimated well.

To ensure that all players have already estimated their preferences well such that they can enter the OGS algorithm simultaneously, the last $\max\{N, K\}$ rounds of the epoch is used to communicate the estimation status. Specifically, each player would propose to arms based on its unique index in a round-robin manner if it estimates its preferences well; otherwise, it would propose to nothing. Recall that players can observe the successfully selected player of each arm. Once a player observes all players are successfully matched with each arm once, it believes that all players estimate their preference rankings well. Above all, all players simultaneously get an estimated ranking by following the algorithm and then run OGS with these rankings.

5.1. Sample complexity analysis

To measure the hardness of learning players' unknown preferences, we first define the preference gaps of players.

Definition 2. For any $p_i \in \mathcal{N}$ and $a_j, a_{j'} \in \mathcal{K}$, let $\Delta_{i,j,j'} = \left|\mu_{i,j} - \mu_{i,j'}\right|$ be the preference gap of player p_i between arm a_j and arm $a_{j'}$. Further define

$$\Delta := \min_{i,k\in[\min\{N,K-1\}]} \Delta_{i,\pi_{i,k},\pi_{i,k+1}}$$

as the minimum preference gap between arms ranked the first $\min\{N+1, K\}$th among all players.

The following theorem provides the sample complexity of the uniform-exploration algorithm to learn the players' preference rankings.

Theorem 2. *Given any confidence level $\delta \in (0,1)$, Algorithm 2 returns the correct preference ranking towards the first $\min\{N, K\}$ arms for each player simultaneously with sample complexity*

$$O\left(\frac{\max\{N,K\}}{\Delta^2}\log\left(\frac{NK}{\delta\Delta^4}\right)\right).$$

Above all, when players also have uncertain preferences, they can follow the uniform-exploration algorithm to learn them and run the OGS algorithm using the estimated preference rankings. Combining Theorems 2 and 1, all players would find the player-optimal stable matching in at most $O\left(\frac{\max\{N,K\}}{\Delta^2}\log\left(\frac{1}{\delta}\right) + \frac{NK}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right)$ rounds with probability larger than $1-2\delta$, where Δ and ϵ correspond to the preference gap of the player side and arm side, respectively. Due to the space limit, the proof of Theorem 2 and corresponding discussions are deferred to Appendix C.

6. Experiments

In this section, we test the performances of our proposed algorithms in markets with different sizes and preference probability gaps. For each experiment, we report the maximum sample complexity when finding the optimal stable arm among all players. All results are averaged over 50 trials. The standard errors that are computed as the standard deviations divided by $\sqrt{50}$ are shown by the error bar.

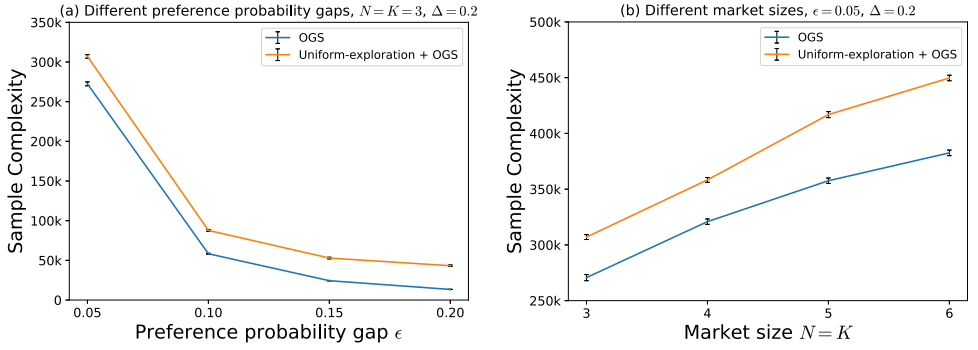


Fig. 1. Experimental results of our OGS and Uniform-exploration together with OGS in terms of sample complexity. Markets with different preference probability gaps $\epsilon \in \{0.05, 0.1, 0.15, 0.2\}$ (left) and with different sizes $N = K \in \{3, 4, 5, 6\}$ (right) are tested.

Varying Preference Probability Gaps. We first test the effect of the minimum preference probability gap ϵ on the performances of the algorithms. The market size is fixed with $N = 3$ players and $K = 3$ arms. The preference ranking of each market participant is generated as random permutations over the other side of participants. We choose the minimum preference probability gap ϵ between any adjacently ranked players among all arms from $\{0.05, 0.1, 0.15, 0.2\}$. To verify the performance of Algorithm 2, we also include this algorithm together with OGS. The minimum preference gap between any adjacently ranked arms among all players is set as $\Delta = 0.2$. The sample complexity of the algorithms is reported in Fig. 1(a). It is intuitive that when ϵ becomes smaller, the sample complexity increases as players need more time to identify the arms' most preferred one in each step.

Varying market sizes. We then test the effect of the market size on the performances of the algorithms. The preference values of participants are generated as the last experiment. The minimum preference probability gap ϵ between adjacent players is set as 0.05. The minimum preference gap Δ between adjacent arms is set as 0.2. The market size $N = K$ is chosen from $\{3, 4, 5, 6\}$. We report the sample complexity of the algorithms in Fig. 1(b). The performance dependence on the market size also coincides with the theoretical guarantees. In larger markets, players need to explore more rounds to converge to stable matching.

7. Conclusion

In this paper, we study the bandit learning problem in matching markets with relative feedback. Different from previous works which model arms' acceptance behavior as deterministic, we consider that arms' acceptance process may change due to players' fluctuating performances. Instead of using absolute rewards as feedback for learning this unknown knowledge, we consider a more natural relative feedback inspired by the Plackett-Luce model. For this framework, we propose the OGS algorithm that integrates the learning process based on relative feedback into each GS step and show its sample complexity to reach stability is $O\big(NK\log(1/\delta)/\epsilon^2\big)$. The algorithm can also be generalized to deal with the case where players also have unknown preferences. The newly introduced learning phase for players' preferences only brings additional $O(\max\{N, K\}\log(1/\delta)/\Delta^2)$ sample complexity. Our results achieve the same order as the state-of-the-art theoretical guarantee in the recovered dueling bandits [18] and matching markets setting [13,14]. A series of experiments are conducted to verify the performances of the algorithms.

Recall that our currently considered setting is based on one-to-one matching. One interesting future direction is to consider the many-to-one setting where an employer can select more than one worker. How to model the combinatorial preferences of employers and investigate a practical feedback type to learn such preferences is an interesting question. Deriving algorithms with guarantees of incentive compatibility is also an interesting future direction.

CRediT authorship contribution statement

Fang Kong: Writing – review & editing, Writing – original draft, Validation, Methodology, Formal analysis; **Xiaoxi Zhang:** Writing – original draft, Methodology, Formal analysis; **Xiao Huang:** Visualization, Validation, Data curation; **Lijun Zhang:** Writing – review & editing, Validation; **Shuai Li:** Writing – review & editing, Supervision, Project administration, Methodology, Conceptualization.

Data availability

No data was used for the research described in the article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The corresponding author Shuai Li is supported by National Key Research and Development Program of China (2022ZD0114804) and National Natural Science Foundation of China (62376154). Fang Kong is supported by National Natural Science Foundation of China (62506150) and Guangdong Basic and Applied Basic Research Foundation (2025A1515011412).

Appendix A. Proof of Lemmas

Proof of Lemma 1. Let $X_j(1), X_j(2), \dots, X_j(t)$ denote the choice of arm a_j at different round t, which are independent and identically distributed vectors taking players in $\mathcal{N}_{j,\ell}$. In particular, if arm a_j chooses player p_i at time τ, $X_j(\tau)$ is a one-hot vector of length N with only 1 at p_i and 0 at other players. Besides, the distance of the estimated preference from the real preference probabilities at round t can be written as

$$\left\| \mathrm{P}_j^{\mathcal{N}_{j,\ell}} - \hat{\mathrm{P}}_j^{\mathcal{N}_{j,\ell}}(t) \right\|_1 = \max_{\lambda \in \{-1,1\}^N} \left\langle \lambda, \mathrm{P}_j^{\mathcal{N}_{j,\ell}} - \hat{\mathrm{P}}_j^{\mathcal{N}_{j,\ell}}(t) \right\rangle .$$

Fix a $\lambda \in \{-1,1\}^N$, we have

$$\langle \lambda, \mathrm{P}_j^{\mathcal{N}_{j,\ell}} - \hat{\mathrm{P}}_j^{\mathcal{N}_{j,\ell}}(t)\rangle = \frac{1}{t}\sum_{\tau=1}^{t}\langle \lambda, \mathrm{P}_j^{\mathcal{N}_{j,\ell}} - X_j(\tau)\rangle .$$

Further, since

$$\begin{aligned}|\langle \lambda, \mathrm{P}_j^{\mathcal{N}_{j,\ell}} - X_j(\tau)\rangle| \le& \|\lambda\|_\infty \left\| \mathrm{P}_j^{\mathcal{N}_{j,\ell}} - X_j(\tau)\right\|_1 \\ \le& 1 + \max_{i\in[N]} \mathrm{P}_{j,i}^{\mathcal{N}_{j,\ell}} \le 2,\end{aligned}$$

we can conclude that

$$\mathbb{E}\left[\langle \lambda, \mathrm{P}_j^{\mathcal{N}_{j,\ell}} - X_j(\tau)\rangle\right] = 0, \text{ and}$$

$$\forall \tau, \langle \lambda, \mathrm{P}_j^{\mathcal{N}_{j,\ell}} - X_j(\tau)\rangle \in [-2,2] .$$

Then using Hoeffding's inequality presented in Lemma 7,

$$\mathbb{P}\left(\langle \lambda, \mathrm{P}_j^{\mathcal{N}_{j,\ell}} - \hat{\mathrm{P}}_j^{\mathcal{N}_{j,\ell}}(t)\rangle > \sqrt{\frac{8\log(2NKt^2/\delta)}{2t}}\right) \le \frac{\delta}{2NKt^2} .$$

Since there are totally 2^N possible values of λ, it holds that

$$\mathbb{P}\left(\left\| \mathrm{P}_j^{\mathcal{N}_{j,\ell}} - \hat{\mathrm{P}}_j^{\mathcal{N}_{j,\ell}}(t)\right\|_1 > \sqrt{\frac{8\log(2^{N+1}NKt^2/\delta)}{2t}}\right) \le \frac{\delta}{2NKt^2} .$$

Finally, by the union bound,

$$\mathbb{P}\left(\mathcal{F}_{j,\ell}\right) \le \sum_{t=1}^{\infty}\mathbb{P}\left(\left\| \mathrm{P}_j^{\mathcal{N}_{j,\ell}} - \hat{\mathrm{P}}_j^{\mathcal{N}_{j,\ell}}(t)\right\|_1 > c_t\right) \le \frac{\delta}{NK} .$$

□

Proof of Lemma 2. Recall that the best player $\hat{i}_j^*$ of arm a_j is identified when $\hat{\mathrm{P}}_{j,\hat{i}_j^*}^{\mathcal{N}_{j,\ell}}(t) - c_t > \hat{\mathrm{P}}_{j,i'}^{\mathcal{N}_{j,\ell}}(t) + c_t$ for all $i' \neq \hat{i}_j^*$. Additionally, the event $\neg\mathcal{F}_{j,\ell}$ guarantees

$$\left\| \mathrm{P}_j^{\mathcal{N}_{j,\ell}} - \hat{\mathrm{P}}_j^{\mathcal{N}_{j,\ell}}(t)\right\|_1 \le c_t,$$

which implies that

$$\begin{aligned}&\hat{\mathrm{P}}_{j,\hat{i}_j^*}(t) - c_t \le \mathrm{P}_{j,\hat{i}_j^*} \le \hat{\mathrm{P}}_{j,\hat{i}_j^*}(t) + c_t,\\ &\hat{\mathrm{P}}_{j,i'}(t) - c_t \le \mathrm{P}_{j,i'} \le \hat{\mathrm{P}}_{j,i'}(t) + c_t, \forall i' \neq \hat{i}_j^*\end{aligned}$$

based on the definition of ℓ_1 norm. In the following, we remove the notation $\mathcal{N}_{j,\ell}$ in the definition of preference probabilities when the context is clear.

Considering the above concentration inequalities, the identification condition $\hat{\mathrm{P}}_{j,\hat{i}_j^*} - c_t > \hat{\mathrm{P}}_{j,i'} + c_t$ for all $i' \neq \hat{i}_j^*$ is guaranteed to occur as long as

$$\mathrm{P}_{j,\hat{i}_j^*} - 2c_t \ge \mathrm{P}_{j,i'} + 2c_t .$$

Based on the definition of ϵ, the above inequality would occur if

$$\epsilon \ge 4c_t = 4\sqrt{\frac{8\log(2^{N+1}NKt^2/\delta)}{2t}} .$$

To guarantee this inequation, the minimum number of rounds is of order

$$O\left(\frac{64}{\epsilon^2}\log\left(NK\cdot\frac{2^N}{\delta\epsilon^4}\right)\right).$$

The rigorous proof is as follows. Let T^* be the unique solution for t in the equation:

$$t = \frac{64}{\epsilon^2}\log\left(\frac{t^2}{\eta}\right),$$

where $\eta = \delta/\left(2^{N+1}NK\right)$, then we will prove that T^* is bounded by

$$L_t = \frac{64}{\epsilon^2}\log\left(\frac{16}{\eta\epsilon^4}\right) \le T^* \le \frac{608}{\epsilon^2}\log\left(\frac{1}{\eta\epsilon^4}\right) = U_t .$$

First, put the upper bound U_t in t the right of equation

$$\begin{aligned}
t &= \frac{64}{\epsilon^2}\log\left(\frac{t^2}{\eta}\right)\\
&\le \frac{64}{\epsilon^2}\log\left(\frac{\frac{C_0^2}{\epsilon^4}\left(\log\left(\frac{1}{\eta\epsilon^4}\right)\right)^2}{\eta}\right)\\
&= \frac{64}{\epsilon^2}\left(2\log C_0 + \log\left(\frac{1}{\eta\epsilon^4}\right) + 2\log\log\left(\frac{1}{\eta\epsilon^4}\right)\right)\\
&\le \frac{64}{\epsilon^2}\left(2\log C_0 + \left(1+\frac{2}{e}\right)\log\left(\frac{1}{\eta\epsilon^4}\right)\right)\\
&\le \frac{64}{\epsilon^2}\left(2\log C_0 + 1 + \frac{2}{e}\right)\log\left(\frac{1}{\eta\epsilon^4}\right),
\end{aligned}$$

The penultimate inequality uses the fact that

$$x + a\log x \le \left(1+\frac{1}{e}\right)x\,.$$

If the number of players or arms is more than 1, the last inequality holds due to

$$\frac{1}{\eta\epsilon^4} = \frac{NK2^{N+1}}{\delta\epsilon^4} > e,$$

and choose the constant $C_0 = 608$ satisfies:

$$4\left(2\log C_0 + 1 + \frac{2}{e}\right) < C_0,$$

then there comes $T^* \le \frac{64}{\epsilon^2}\left(2\log C_0 + 1 + \frac{2}{e}\right)\log\left(\frac{1}{\eta\epsilon^4}\right) < \frac{C_0}{\epsilon^2}\log\left(\frac{1}{\eta\epsilon^4}\right)$, we prove that T^* is upper bounded by U_t. Second, define a sequence $T_{n+1} = \frac{64}{\epsilon^2}\log\left(\frac{T_n^2}{\eta}\right)$, then

$$\begin{aligned}
T_0 &= 1,\\
T_1 &= \frac{64}{\epsilon^2}\log\left(\frac{1}{\eta}\right),\\
T_2 &= \frac{64}{\epsilon^2}\log\left(\frac{\frac{16}{\epsilon^4}\log^2\left(\frac{1}{\eta}\right)}{\eta}\right)\\
&= \frac{64}{\epsilon^2}\log\left(\frac{16}{\eta\epsilon^4}\right) + \frac{64}{\epsilon^2}2\log\log\left(\frac{1}{\eta}\right).
\end{aligned}$$

As n grows to infinity, the sequence T_n converges to solution T^* in the equation. Because the map $n \longmapsto T_n$ is increasing, there comes $T^* \ge T_2 \ge \frac{64}{\epsilon^2}\log\left(\frac{16}{\eta\epsilon^4}\right)$. To end with, we prove that T^* is lower bounded by L_t. □

Proof of Lemma 3. Based on the event $\neg\mathcal{F}_{j,\ell}$, for any player $p_i \in \mathcal{N}$, it holds that

$$\left|\mathrm{P}_{j,i}^{\mathcal{N}_{j,\ell}} - \hat{\mathrm{P}}_{j,i}^{\mathcal{N}_{j,\ell}}(t)\right| \le \left\|\mathrm{P}_{j}^{\mathcal{N}_{j,\ell}} - \hat{\mathrm{P}}_{j}^{\mathcal{N}_{j,\ell}}(t)\right\|_1 \le c_t\,.$$

And the identified player $\hat{i}_j^*$ satisfies

$$\forall i \ne \hat{i}_j^*, \hat{\mathrm{P}}_{j,\hat{i}_j^*}^{\mathcal{N}_{j,\ell}}(t) - c_t > \hat{\mathrm{P}}_{j,i}^{\mathcal{N}_{j,\ell}}(t) + c_t\,.$$

Combining $\neg\mathcal{F}_{j,\ell}$, we have

$$\mathrm{P}_{j,\hat{i}_j^*}^{\mathcal{N}_{j,\ell}} \ge \hat{\mathrm{P}}_{j,\hat{i}^*}^{\mathcal{N}_{j,\ell}}(t) - c_t > \hat{\mathrm{P}}_{j,i}^{\mathcal{N}_{j,\ell}}(t) + c_t \ge \mathrm{P}_{j,i}^{\mathcal{N}_{j,\ell}}\,.$$

Thus it concludes that

$$\mathrm{P}_{j,\hat{i}_j^*}^{\mathcal{N}_{j,\ell}} > \mathrm{P}_{j,i}^{\mathcal{N}_{j,\ell}}, \forall p_i \in \mathcal{N} \text{ and } \hat{i}_j^* \in argmax_{i\in[N]}\mathrm{P}_{j,i}^{\mathcal{N}_{j,\ell}}\,.$$

□

Appendix B. Algorithm Chart in Section 5

Algorithm 2 Uniform-exploration (from view of p_i).

Input: player set $\mathcal{N}$, arm set $\mathcal{K}$
Output: the estimated preference ranking $\hat{\pi}_i$
Initialize: $\hat{\mu}_{i,j} = 0, T_{i,j} = 0, \forall j \in [K]$
For $t = 1, 2, \ldots, N$: get an index Index
for $\ell = 1, 2, \ldots,$ **do**
  for $t = N + 2^\ell - 1, \ldots, N + 2^\ell - 1 + 2^\ell$ **do**
    $A_i(t) = a_{(\text{Index}+t-1)\% \max\{N,K\}+1}$
    if $\bar{A}_i(t) = A_i(t)$ **then**
      Update $T_{i,A_i(t)}$ and $\hat{\mu}_{i,A_i(t)}$ with $X_{i,A_i(t)}(t)$
    end if
  end for
  Compute $\text{UCB}_{i,j}$ and $\text{LCB}_{i,j}$ for each $j \in [K]$
  if $\exists \sigma$ such that $\text{LCB}_{i,\sigma_k} > \text{UCB}_{i,\sigma_{k+1}}$ for any $k \in [\min\{N-1, K-1\}]$ and (if $N < K$) $\text{LCB}_{i,\sigma_N} > \text{UCB}_{i,\sigma_k}$ for any $k = N+1, N+2, \ldots, K$ **then**
    Flag = True
  else
    Flag = False
  end if
  for $t = N + 2^\ell + 2^\ell, \ldots, N + 2^\ell + 2^\ell + \max\{N, K\}$ **do**
    if Flag = True **then**
      $A_i(t) = a_{(\text{Index}+t-1)\% \max\{N,K\}+1}$
    else
      $A_i(t) = \emptyset$
    end if
  end for
  Return $\hat{\pi}_i = \sigma$ if all players have been matched with each arm for once during the previous $\max\{N, K\}$ rounds
end for

Appendix C. Proof of Theorem 2

Before providing the main proof, we first define the following bad event when learning the unknown preferences of the player side,

$$\mathcal{F}^P := \left\{ \exists t > 0, i \in [N], j \in [K] : \left|\hat{\mu}_{i,j}(t) - \mu_{i,j}\right| > \sqrt{\frac{2\log(2NKt^2/\delta)}{T_{i,j}(t)}} \right\} \tag{C.1}$$

to represent that some preference of player p_i is not very accurately estimated at some round t. The following lemma shows that this bad event holds with a probability smaller than δ.

Lemma 4. *It holds that*

$$\mathbb{P}\left(\mathcal{F}^P\right) \le \delta\,.$$

Proof. By using Lemma 8, there is

$$\begin{aligned}
\mathbb{P}\left(\mathcal{F}^P\right) =& \mathbb{P}\left(\exists t > 0, i \in [N], j \in [K] : \left|\hat{\mu}_{i,j}(t) - \mu_{i,j}\right| > \sqrt{\frac{2\log(2NKt^2/\delta)}{T_{i,j}(t)}} \right) \\
\le & \sum_{i\in[N],j\in[K]} \sum_{t} \mathbb{P}\left(\left|\hat{\mu}_{i,j}(t) - \mu_{i,j}\right| > \sqrt{\frac{2\log(2NKt^2/\delta)}{T_{i,j}(t)}} \right) \\
= & \sum_{i\in[N],j\in[K]} \sum_{t} \sum_{w=1}^{t} \mathbb{P}\left(T_{i,j}(t) = w, \left|\hat{\mu}_{i,j}(t) - \mu_{i,j}\right| > \sqrt{\frac{2\log(2NKt^2/\delta)}{T_{i,j}(t)}} \right) \\
= & \sum_{i\in[N],j\in[K]} \sum_{t} \sum_{w=1}^{t} \mathbb{P}\left(T_{i,j}(t) = w\right) \cdot \\
& \qquad \mathbb{P}\left(\left|\hat{\mu}_{i,j}(t) - \mu_{i,j}\right| > \sqrt{\frac{2\log(2NKt^2/\delta)}{T_{i,j}(t)}} \mid T_{i,j}(t) = w \right) \\
\le & \sum_{i\in[N],j\in[K]} \sum_{t} \sum_{w=1}^{t} \mathbb{P}\left(T_{i,j}(t) = w\right) \frac{\delta}{NKt^2}
\end{aligned}$$

$$\leq \delta\,.$$

□

The following lemma shows that the estimated preference ranking is correct conditional on $\neg\mathcal{F}^P$.

Lemma 5. *Conditional on $\neg\mathcal{F}^P$, $\mathrm{UCB}_{i,j}(t) < \mathrm{LCB}_{i,j'}(t)$ at a time t implies $\mu_{i,j} < \mu_{i,j'}$.*

Proof. At any time t, Conditional on $\neg\mathcal{F}^P$, for each $i \in [N], j \in [K]$, we have

$$\begin{aligned}\mathrm{LCB}_{i,j}(t) = \hat{\mu}_{i,j}(t) - \sqrt{\frac{2\log(2NKt^2/\delta)}{T_{i,j}(t)}} \leq \mu_{i,j} \leq & \hat{\mu}_{i,j}(t) + \sqrt{\frac{2\log(2NKt^2/\delta)}{T_{i,j}(t)}} \\ = & \mathrm{UCB}_{i,j}(t)\,.\end{aligned}$$

Thus if $\mathrm{UCB}_{i,j}(t) < \mathrm{LCB}_{i,j'}(t)$, there would be

$$\mu_{i,j} \leq \mathrm{UCB}_{i,j}(t) < \mathrm{LCB}_{i,j'}(t) \leq \mu_{i,j'}\,.$$

□

The following lemma further illustrates the sample complexity of each arm to ensure that the estimation condition is true.

Lemma 6. *For any time t, let $T_i(t) = \min\left\{T_{i,j}(t) : j \in [K]\right\}$, $\bar{T}_i = \frac{32}{\Delta^2}\log\left(\frac{NK}{\delta\Delta^4}\right)$. Conditional on $\neg\mathcal{F}^P$, if $T_i(t) > \bar{T}_i$, the condition in Line 20 of Algorithm 2 holds.*

Proof. By contradiction, suppose there exists pair j, j' where j' is the first $\min\{N,K\}$-ranked arm and $\mu_{i,j} < \mu_{i,j'}$ such that $\mathrm{UCB}_{i,j}(t) \geq \mathrm{LCB}_{i,j'}(t)$. Conditional on $\neg\mathcal{F}^P$, we have

$$\mu_{i,j'} - 2\sqrt{\frac{2\log(2NKt^2/\delta)}{T_{i,j}(t)}} \leq \mathrm{LCB}_{i,j'}(t) \leq \mathrm{UCB}_{i,j}(t) \leq \mu_{i,j} + 2\sqrt{\frac{2\log(2NKt^2/\delta)}{T_{i,j}(t)}}\,.$$

We can then conclude $\Delta_{i,j,j'} \leq 4\sqrt{\frac{2\log(2NKt^2/\delta)}{T_{i,j}(t)}}$. Adopting similar proof of Lemma 2, we can conclude $T_i(t) \leq \frac{32}{\Delta^2}\log\left(\frac{NK}{\delta\Delta^4}\right)$, which contradicts $T_i(t) > \bar{T}_i$. □

Based on the above useful lemmas, we then provide the main proof.

Proof of Theorem 2. Recall that Algorithm 2 lets players explore arms in a round-robin way based on their unique indices. Thus every $\max\{N,K\}$ rounds, each player can obtain an observation on each arm once. Lemma 6 shows that the number of observation for each arm is $O\left(\frac{1}{\Delta^2}\log\left(NK/(\delta\Delta^4)\right)\right)$ to ensure that the stopping condition is true. And the number of exploration rounds of each epoch ℓ is 2^ℓ. So the number of epochs before stopping is $O\left(\log\left(\frac{\max\{N,K\}}{\Delta^2}\log\left(NK/(\delta\Delta^4)\right)\right)\right)$ and the total number of explorations would be $2 * O\left(\frac{\max\{N,K\}}{\Delta^2}\log\left(NK/(\delta\Delta^4)\right)\right) = O\left(\frac{\max\{N,K\}}{\Delta^2}\log\left(NK/(\delta\Delta^4)\right)\right)$. Above all, the total sample complexity before stopping is

$$\begin{aligned}& O\left(N + \frac{\max\{N,K\}}{\Delta^2}\log\left(\frac{NK}{\delta\Delta^4}\right) + \max\{N,K\}\log\left(\frac{\max\{N,K\}}{\Delta^2}\log\left(\frac{NK}{\delta\Delta^4}\right)\right)\right) \\ =& O\left(\frac{\max\{N,K\}}{\Delta^2}\log\left(\frac{NK}{\delta\Delta^4}\right)\right).\end{aligned}$$

where the first term comes from the first phase of N rounds to estimate players' indices, the second term corresponds to the exploration rounds and the last term corresponds to the communication rounds.

And Lemma 5 shows that once the stopping condition is satisfied, the returned estimated ranking over the most preferred $\min\{N,K\}$ arms is correct. Above all, we show that with probability $1-\delta$, Algorithm 2 will return the correct preference towards the first $\min\{N,K\}$ arms for all players simultaneously (as all players have the same observations) with sample complexity $O\left(\frac{\max\{N,K\}}{\Delta^2}\log\left(\frac{NK}{\delta\Delta^4}\right)\right)$.

And according to the offline GS algorithm procedure, once an arm has been proposed by players, this arm has a temporary partner. Above all, once N arms have been proposed, they will occupy N players and the algorithm stops. So before the algorithm stops, at most $N-1$ arms have been previously proposed. This indicates that the player-optimal stable arm must be the first $\min\{N,K\}$-ranked of each player. So the returned ranking of Algorithm 2 which is correct for the first $\min\{N,K\}$th together with Algorithm 1 would finally return the player-optimal stable matching.

Appendix D. Technical Lemmas

Lemma 7. *(Hoeffding's inequality in [42]) If $X_1, X_2, \ldots, X_n$ are independent and zero-mean random variables such that $X_t \in [a_t, b_t]$ almost surely with $a_t \leq b_t$ for all t, then for any $\epsilon \geq 0$,*

$$\mathbb{P}\left(\frac{1}{n}\sum_{t=1}^{n}\left(X_t - \mathbb{E}\left[X_t\right]\right) \geq \epsilon\right) \leq \exp\left(\frac{-2n^2\epsilon^2}{\sum_{t=1}^{n}(b_t - a_t)^2}\right).$$

Lemma 8. *(Corollary 5.5 in [8]) Assume that $X_1, X_2, \dots, X_n$ are independent, σ-subgaussian random variables centered around μ. Then for any $\varepsilon > 0$,*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq \mu + \varepsilon\right) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right), \quad \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \leq \mu - \varepsilon\right) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right).$$

References

[1] D. Gale, L.S. Shapley, College admissions and the stability of marriage, Am. Math. Mon. 69 (1) (1962) 9–15.
[2] A. Abdulkadiroğlu, T. Sönmez, House allocation with existing tenants, J. Econ. Theory 88 (2) (1999) 233–260.
[3] A.E. Roth, The evolution of the labor market for medical interns and residents: a case study in game theory, J. Polit. Econ. 92 (6) (1984) 991–1016.
[4] A.E. Roth, M. Sotomayor, Two-sided matching, Handb. Game Theory Econ. Appl. 1 (1992) 485–541.
[5] D. Gusfield, R.W. Irving, The stable marriage problem - structure and algorithms, in: Foundations of Computing Series, 1989.
[6] A. Roth, Deferred acceptance algorithms: history, theory, practice, and open questions, Int. J. Game Theory 36 (2008) 537–569. https://doi.org/10.1007/s00182-008-0117-6
[7] P. Auer, N. Cesa-Bianchi, P. Fischer, Finite-time analysis of the multiarmed bandit problem, Mach. Learn. 47 (2) (2002) 235–256.
[8] T. Lattimore, C. Szepesvári, Bandit Algorithms, Cambridge University Press, 2020.
[9] L.T. Liu, H. Mania, M. Jordan, Competing bandits in matching markets, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2020, pp. 1618–1628.
[10] L.T. Liu, F. Ruan, H. Mania, M.I. Jordan, Bandit learning in decentralized matching markets, J. Mach. Learn. Res. 22 (211) (2021) 1–34.
[11] S. Basu, K.A. Sankararaman, A. Sankararaman, Beyond $\log^2(T)$ regret for decentralized bandits in matching markets, in: International Conference on Machine Learning, 2021, pp. 705–715.
[12] F. Kong, J. Yin, S. Li, Thompson sampling for bandit learning in matching markets, in: International Joint Conference on Artificial Intelligence, 2022.
[13] Y. Zhang, S. Wang, Z. Fang, Matching in multi-arm bandit with collision, in: Advances in Neural Information Processing Systems, 2022.
[14] F. Kong, S. Li, Player-optimal stable regret for bandit learning in matching markets, in: Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), SIAM, 2023.
[15] F. Kong, Z. Wang, S. Li, Improved analysis for bandit learning in matching markets, in: The Thirty-eighth Annual Conference on Neural Information Processing Systems, 2024.
[16] S. Li, Z. Wang, F. Kong, A survey on bandit learning in matching markets, in: Proceedings of the 34th International Joint Conference on Artificial Intelligence, 2025, pp. 10546–10554.
[17] H. Hosseini, S. Roy, D. Zhang, Putting gale & shapley to work: guaranteeing stability through learning, in: The Thirty-eighth Annual Conference on Neural Information Processing Systems, 2024.
[18] Y. Yue, J. Broder, R. Kleinberg, T. Joachims, The K-armed dueling bandits problem, J. Comput. Syst. Sci. 78 (5) (2012) 1538–1556.
[19] Y. Sui, M. Zoghi, K. Hofmann, Y. Yue, Advancements in dueling bandits, in: Proceedings of the 27th International Joint Conference on Artificial Intelligence, 2018, pp. 5502–5510.
[20] A. Saha, A. Gopalan, Combinatorial bandits with relative feedback, Adv. Neural Inf. Process. Syst. 32 (2019,) 985–995.
[21] W. Chen, Y. Du, L. Huang, H. Zhao, Combinatorial pure exploration for dueling bandit, in: International Conference on Machine Learning, PMLR, 2020, pp. 1531–1541.
[22] S. Das, E. Kamenica, Two-sided bandits and the dating market, in: International Joint Conference on Artificial Intelligence, 2005, pp. 947–952.
[23] A. Sankararaman, S. Basu, K.A. Sankararaman, Dominate or delete: decentralized competing bandits in serial dictatorship, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2021, pp. 1252–1260.
[24] C. Maheshwari, E. Mazumdar, S. Sastry, Decentralized, communication-and coordination-free learning in structured matching markets, in: Advances in Neural Information Processing Systems, 2022.
[25] T. Pagare, A. Ghosh, Two-sided bandit learning in fully-decentralized matching markets, in: ICML 2023 Workshop the Many Facets of Preference-Based Learning, 2023.
[26] Y. Zhang, Z. Fang, Decentralized two-sided bandit learning in matching market, in: The 40th Conference on Uncertainty in Artificial Intelligence, 2024.
[27] A. Ghosh, A. Sankararaman, K. Ramchandran, T. Javidi, A. Mazumdar, Decentralized competing bandits in non-stationary matching markets,(2022). arXiv preprint arXiv:2206.00120
[28] Y. Min, T. Wang, R. Xu, Z. Wang, M.I. Jordan, Z. Yang, Learn to match with no regret: reinforcement learning in Markov matching markets, in: Advances in Neural Information Processing Systems, 2022.
[29] M. Jagadeesan, A. Wei, Y. Wang, M. Jordan, J. Steinhardt, Learning equilibria in matching markets from bandit feedback, Adv. Neural Inf. Process. Syst. 34 (2021),3323–3335.
[30] E. Even-Dar, S. Mannor, Y. Mansour, Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems, J. Mach. Learn. Res. 7 (2006) 1079–1105.
[31] K. Jamieson, M. Malloy, R. Nowak, S. Bubeck, lil' UCB : an optimal exploration algorithm for multi-armed bandits, J. Mach. Learn. Res. 35 (2013), 423–439.
[32] E. Kaufmann, O. Cappé, A. Garivier, On the complexity of best-arm identification in multi-armed bandit models, J. Mach. Learn. Res. 17 (1) (2016) 1–42.
[33] L. Chen, J. Li, On the optimal sample complexity for best arm identification, (2015). arXiv preprint arXiv:1511.03774
[34] L. Chen, J. Li, Open problem: best arm identification: almost instance-wise optimality and the gap entropy conjecture, in: Conference on Learning Theory, PMLR, 2016, pp. 1643–1646.
[35] B. Chen, P.I. Frazier, Dueling bandits with weak regret, in: International Conference on Machine Learning, PMLR, 2017, pp. 731–739.
[36] S. Chen, T. Lin, I. King, M.R. Lyu, W. Chen, Combinatorial pure exploration of multi-armed bandits, Adv. Neural Inf. Process. Syst. 27 (2014),379–387.
[37] L. Chen, A. Gupta, J. Li, M. Qiao, R. Wang, Nearly optimal sampling algorithms for combinatorial pure exploration, in: Conference on Learning Theory, PMLR, 2017, pp. 482–534.
[38] R.L. Plackett, The analysis of permutations, J. R. Stat. Soc. Ser. C 24 (2) (1975) 193–202.
[39] R.D. Luce, Individual Choice Behavior: A Theoretical Analysis, Courier Corporation, 2012.
[40] Z. Wang, L. Guo, J. Yin, S. Li, Bandit learning in many-to-one matching markets, in: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, 2022, pp. 2088–2097.
[41] J. Komiyama, J. Honda, H. Kashima, H. Nakagawa, Regret lower bound and optimal algorithm in dueling bandit problem, in: Conference on Learning Theory, PMLR, 2015, pp. 1141–1154.
[42] W. Hoeffding, Probability inequalities for sums of bounded random variables, J. Am. Stat. Assoc. 58 (301) (1963) 13–30.