Improved Bandits in Many-to-One Matching Markets with Incentive Compatibility

Fang Kong, Shuai Li[*]

John Hopcroft Center for Computer Science, Shanghai Jiao Tong University
{fangkong, shuaili8}@sjtu.edu.cn

Abstract

Two-sided matching markets have been widely studied in the literature due to their rich applications. Since participants are usually uncertain about their preferences, online algorithms have recently been adopted to learn them through iterative interactions. An existing work initiates the study of this problem in a many-to-one setting with responsiveness. However, their results are far from optimal and lack guarantees of incentive compatibility. We first extend an existing algorithm for the one-to-one setting to this more general setting and show it achieves a near-optimal bound for player-optimal regret. Nevertheless, due to the substantial requirement for collaboration, a single player's deviation could lead to a huge increase in its own cumulative rewards and a linear regret for others. In this paper, we aim to enhance the regret bound in many-to-one markets while ensuring incentive compatibility. We first propose the adaptively explore-then-deferred-acceptance (AETDA) algorithm for responsiveness setting and derive an upper bound for player-optimal stable regret while demonstrating its guarantee of incentive compatibility. This result is a significant improvement over existing works. And to the best of our knowledge, it constitutes the first player-optimal guarantee in matching markets that offers such robust assurances. We also consider broader substitutable preferences, one of the most general conditions to ensure the existence of a stable matching and cover responsiveness. We devise an online DA (ODA) algorithm and establish an upper bound for the player-pessimal stable regret for this setting.

Introduction

The problem of two-sided matching markets has been studied for a long history due to its wide range of applications in real life including the labor market and college admission (Gale and Shapley 1962; Roth 1984a; Roth and Sotomayor 1992; Abdulkadiroğlu and Sönmez 1999; Epple, Romano, and Sieg 2006; Fu 2014). There are two sides of market participants, e.g., employers and workers in the labor market, and each side has a preference ranking over the other side. The matching reflects the bilateral nature of exchange in the market. For example, a worker works for an employer and the employer employs this worker. Stability is a key concept describing the equilibrium of a matching, which ensures the current bilateral

---
[*]Corresponding author.

exchange cannot be easily broken. A rich line of works study how to find a stable matching in the market (Gale and Shapley 1962; Kelso Jr and Crawford 1982; Roth 1984a; Roth and Sotomayor 1992; Erdil and Kumano 2019). However, all of them assume the preferences of market participants are known *a priori*, which may not be satisfied in practice. For example in labor markets, workers usually have unknown preferences over employers since they do not know whether they like the task type or the employer. With the emergence of online marketplaces such as online labor market Upwork and crowdsourcing platform Amazon Mechanical Turk where employers have numerous similar tasks to delegate, workers are able to learn the uncertain preferences during the iterative matching process with employers through these tasks.

Multi-armed bandit (MAB) is a core problem that characterizes the learning process during iterative interactions when faced with uncertainty (Auer, Cesa-Bianchi, and Fischer 2002; Lattimore and Szepesvári 2020). There are also two sides of agents: a player on one side and K arms on the other side. The player has unknown preferences over arms. At each time, it selects an arm and receives a reward. The player's objective is to maximize the cumulative reward over a specified horizon. To better measure the performance of the player's strategy, an equivalent objective of minimizing the cumulative regret is widely studied, which is defined as the cumulative difference between the reward of the optimal arm and that of the selected arms.

Recently, a rich line of works study the bandit learning problem in matching markets where more than one player and arms exist. These works study the case where players have unknown preferences over arms and arms can determine their preferences over players based on some known utilities such as the profile of workers in online labor markets. To characterize the stability of the learned matching, the objective of stable regret is adopted and studied (Das and Kamenica 2005; Liu, Mania, and Jordan 2020; Liu et al. 2021; Sankararaman, Basu, and Sankararaman 2021; Basu, Sankararaman, and Sankararaman 2021; Kong, Yin, and Li 2022; Zhang, Wang, and Fang 2022; Kong and Li 2023; Wang et al. 2022). Previous works mainly focus on two types of objectives: the player-optimal stable regret and the player-pessimal stable regret. The former is defined as the cumulative difference between the reward of the arm in the players' most preferred stable matching and the accumulated reward

by the player. The latter is defined compared with the reward of the arm in the players' least preferred stable matching. Liu, Mania, and Jordan (2020) first study the centralized version where a central platform assigns an allocation of arms to players in each round and provide theoretical guarantees. Since such a platform may not always exist in real applications, the following works mainly focus on the decentralized setting where each player makes her own decision (Liu et al. 2021; Sankararaman, Basu, and Sankararaman 2021; Basu, Sankararaman, and Sankararaman 2021; Kong, Yin, and Li 2022; Maheshwari, Mazumdar, and Sastry 2022). These works only achieve guarantees on the player-pessimal stable regret (Liu et al. 2021; Kong and Li 2023) or study special markets where unique stable matching exists. Until recently, Zhang, Wang, and Fang (2022) and Kong and Li (2023) independently propose algorithms that can reach the player-optimal stable matching.

All of the above works study the one-to-one matching markets where each player proposes to one arm at a time and each arm could accept at most one player. The many-to-one setting is more general and common in real life such as in labor markets where an employer usually has a certain quota and can recruit a group of workers (Roth 1984b; Roth and Sotomayor 1992; Abdulkadiroğlu 2005; Che, Kim, and Kojima 2019). Wang et al. (2022) initiate the study in many-to-one markets by considering that arms have responsive preferences. However, their algorithm is only able to achieve player-pessimal stable matching and lacks guarantees on incentive compatibility. Incentive compatibility is a crucial property in multi-player systems as it ensures players are incentivized to act in ways that align with desired system outcomes, thereby promoting cooperation and efficiency rather than encouraging competitive or destructive behaviors. Deriving algorithms that can achieve better regret and enjoy guarantees on this property is important in matching markets.

In this paper, we aim to provide algorithms with improved regret guarantee and incentive compatibility for many-to-one markets. For the sake of the generality, we also study the decentralized setting. We propose an adaptive explore-then-DA (AETDA) algorithm for markets with responsive preferences and derive $O(N \min \{N, K\} C \log T / \Delta^2)$ upper bound for the player-optimal stable regret as well as a guarantee of incentive compatibility, where N is the number of players, K is the number of arms, C is arms' total capacities, T is the horizon, and Δ is some preference gap among players and arms. To the best of our knowledge, it is the first guarantee for the player-optimal regret in decentralized many-to-one markets and is also the first that simultaneously enjoys such robust assurance in one-to-one markets. Since arms preferences may possess a combinatorial structure which may not be well characterized by responsiveness, we also consider a more general setting with *substitutability* (Roth and Sotomayor 1992), one of the most generally known conditions to ensure the existence of a stable matching and naturally holds under responsiveness (Roth and Sotomayor 1992; Abdulkadiroğlu 2005). We design an online deferred acceptance (ODA) algorithm for this more general setting and prove that the regret against the player-pessimal stable matching is bounded by $O(NK \log T / \Delta^2)$. Table 1 provides a comprehensive comparison between our work and related results.

Related Work

The matching market model characterizes many applications such as labor market (Roth 1984a), house allocation (Abdulkadiroğlu and Sönmez 1999), college admission and marriage problems (Gale and Shapley 1962), among which the many-to-one setting is very common and widely studied (Roth and Sotomayor 1992). Responsiveness and substitutability are the most generally known conditions to guarantee the existence of a stable matching (Kelso Jr and Crawford 1982; Roth 1984b; Abdulkadiroğlu 2005) and the deferred acceptance (DA) algorithm is a classical offline algorithm to find a stable matching under this property (Kelso Jr and Crawford 1982; Roth 1984b).

For simplicity, we refer to the setting where one-side participants (players) have unknown preferences as the online setting. This line of works relies on the technique of MAB, a classical online learning framework with a single player and K arms (Lattimore and Szepesvári 2020). The explore-then-commit (ETC) (Garivier, Lattimore, and Kaufmann 2016), upper confidence bound (UCB) (Auer, Cesa-Bianchi, and Fischer 2002), Thompson sampling (TS) (Agrawal and Goyal 2012) and elimination (Auer and Ortner 2010) algorithms are common strategies to obtain $O(K \log T / \Delta)$ regret where Δ is the minimum suboptimality gap among arms.

Multiple-player MAB (MP-MAB) generalizes the standard MAB problem by considering more than one player in the environment. In this setting, each player selects an arm at each time and a player will receive nothing if it collides with others by selecting the same arm. The MP-MAB problem has been studied in both homogeneous and heterogeneous settings. The former assumes different players share the same preference over arms (Rosenski, Shamir, and Szlak 2016; Bubeck, Budzinski, and Sellke 2021) and the latter allows players to have different preferences (Bistritz and Leshem 2018; Shi et al. 2021). Both settings aim to minimize the collective cumulative regret of all players.

The matching market problem is different from the above MP-MAB framework by considering that each arm also has a preference ranking over players. When multiple players select one arm, the player preferred most by the arm would not be collided and would gain a reward. The objective in this setting is to learn a stable matching and minimize the stable regret for players. Das and Kamenica (2005) first introduce the bandit learning problem in one-to-one matching markets and explore the empirical performances of the proposed algorithms. Liu, Mania, and Jordan (2020) initiate the theoretical study on this problem. They first consider the centralized setting where a central platform assigns allocations to players in each round. Later, Sankararaman, Basu, and Sankararaman (2021), Basu, Sankararaman, and Sankararaman (2021) and Maheshwari, Mazumdar, and Sastry (2022) successively study this setting in a decentralized manner where players make their own decisions without a central platform. These works additionally assume the preferences of participants satisfy some constraints to ensure the uniqueness of the stable matching. For a general decentralized one-to-one market, Liu et al. (2021) and Kong, Yin, and Li (2022) propose UCB

	Regret bound	Setting
Liu, Mania, and Jordan (2020)	$O\left(K\log T/\Delta^2\right)*\#$ $O\left(NK\log T/\Delta^2\right)\#$	one-one, known Δ, incentive, gap_1 one-one, incentive, gap_5
Liu et al. (2021)	$O\left(\dfrac{N^5K^2\log^2 T}{\varepsilon^{N^4}\Delta^2}\right)$	one-one, gap_5
Sankararaman, Basu, and Sankararaman (2021)	$O\left(NK\log T/\Delta^2\right)$ $\Omega\left(N\log T/\Delta^2\right)$	one-one (serial dictatorship), - incentive, gap_1
Basu, Sankararaman, and Sankararaman (2021)	$O\left(K\log^{1+\varepsilon} T + 2^{\left(\frac{1}{\Delta^2}\right)^{\frac{1}{\varepsilon}}}\right)*$ $O\left(NK\log T/\Delta^2\right)$	one-one, gap_5 one-one (uniqueness), gap_1
Maheshwari, Mazumdar, and Sastry (2022)	$O\left(C'NK\log T/\Delta^2\right)$	one-one (α-reducible), gap_1
Kong, Yin, and Li (2022)	$O\left(\dfrac{N^5K^2\log^2 T}{\varepsilon^{N^4}\Delta^2}\right)$	one-one, gap_5
Zhang, Wang, and Fang (2022)	$O\left(K\log T/\Delta^2\right)*$	one-one, gap_5
Kong and Li (2023)	$O\left(K\log T/\Delta^2\right)*$	one-one, gap_4 responsiveness **(ours)**, gap_4
Wang et al. (2022)	$O\left(K\log T/\Delta^2\right)*\#$ $O\left(NK^3\log T/\Delta^2\right)\#$ $O\left(\dfrac{N^5K^2\log^2 T}{\varepsilon^{N^4}\Delta^2}\right)$	responsiveness, known Δ, gap_1 responsiveness, gap_5 responsiveness, gap_5
Ours	$O\left(\dfrac{N\min\{N,K\}C\log T}{\Delta^2}\right)*$ $O\left(NK\log T/\Delta^2\right)$	responsiveness, incentive, gap_3 substitutability, incentive, gap_2

Table 1: Comparisons of settings and regret bounds with most related works. $*$ represents the player-optimal stable regret and bounds without labeling $*$ are for player-pessimal stable regret, $\#$ represents the centralized setting. $N, K, \Delta, C, \varepsilon, C'$ are the number of players and arms, some preference gap among players and arms, the total capacities of all arms under the responsiveness condition, the hyper-parameter of algorithms which can be very small, and the parameter related to the unique stable matching condition which can grow exponentially in N, respectively. 'Incentive' means that there is a guarantee for incentive compatibility. The definition of Δ requires particular care in different results. It may be defined as the minimum preference gap between the player-optimal stable arm and the next arm among all players (labeled as gap_1); the minimum preference gap between the player-pessimal stable arm and the next arm among all players (labeled as gap_2); the minimum preference gap between any arms that have higher ranking than the arm after the player-optimal stable arm (labeled as gap_3); the minimum preference gap between any arms that have higher ranking than $\min\{N+1, K\}$ (labeled as gap_4); and the minimum preference gap between any different arms among all players (labeled as gap_5). Based on the fact that the player-optimal stable arm must be the first $\min\{N, K\}$-ranked (proved in Appendix), it holds that $\mathrm{gap}_1 > \mathrm{gap}_3 > \mathrm{gap}_4 > \mathrm{gap}_5$, and $\mathrm{gap}_2 > \mathrm{gap}_5$.

and TS-type algorithms, respectively. However, they only derive guarantees on the player-pessimal stable regret. Until recently, the theoretical analysis for the player-optimal stable regret has been derived by Zhang, Wang, and Fang (2022) and Kong and Li (2023) independently.

Due to the generality when modeling real applications, Wang et al. (2022) start to study the bandit problem in many-to-one settings. They assume arms have responsive preferences and derive algorithms in both centralized and decentralized settings. For the decentralized setting, only an upper bound for the player-pessimal stable regret is provided. Table 1 compares our results with the most related works for matching markets. As shown in the table, our results not only work under a more general setting but also achieve a great advantage over Wang et al. (2022).

Setting

The two-sided market consists of N players and K arms. Denote the player and the arm set as $\mathcal{N} = \{p_1, p_2, \ldots, p_N\}$ and $\mathcal{K} = \{a_1, a_2, \ldots, a_K\}$, respectively. Just as in common applications such as the online labor market, players have preferences over individual arms. The relative preference of player p_i for arm a_j can be quantified by a real value $\mu_{i,j} \in (0, 1]$, which is unknown and needs to be learned during interactions with arms. For each player p_i, we assume $\mu_{i,j} \neq \mu_{i,j'}$ for distinct arms $a_j, a_{j'}$ as in previous works (Kelso Jr and Crawford 1982; Roth 1984b; Liu, Mania, and Jordan 2020; Liu et al. 2021; Kong and Li 2023; Wang et al. 2022).

And $\mu_{i,j} > \mu_{i,j'}$ implies that player p_i prefers a_j to $a_{j'}$. For the other side of participants, arms are usually certain of their preferences for players based on some known utilities, e.g., the profiles of workers in the online labor markets scenario. In many-to-one markets, when faced with a set $P \subseteq \mathcal{N}$ of players, the arm can determine which subset of P it prefers most. Denote $\text{Ch}_j(P)$ as this choice of arm j when faced with P. Then for any $P' \subseteq P$, arm a_j prefers $\text{Ch}_j(P)$ to P'.

At each round $t = 1, 2, \ldots$, each player $p_i \in \mathcal{N}$ proposes to an arm $A_i(t) \in \mathcal{K}$. Let $A_j^{-1}(t) = \{p_i : A_i(t) = a_j\}$ be the set of players who propose to a_j. When faced with the player set $A_j^{-1}(t)$, arm a_j only accepts its most preferred subset $\text{Ch}_j(A_j^{-1}(t))$ and would reject others. Once p_i is successfully accepted by arm $A_i(t)$, it receives a utility gain $X_{i,A_i(t)}(t)$, which is a 1-subgaussian random variable with expectation $\mu_{i,A_i(t)}$. Otherwise, it receives $X_{i,A_i(t)}(t) = 0$. We further denote $\bar{A}_i(t)$ as p_i's matched arm at round t. Specifically, $\bar{A}_i(t) = A_i(t)$ if p_i is successfully matched and $\bar{A}_i(t) = \emptyset$ otherwise. Inspired by real applications such as labor market where workers usually update their working experience on their profiles, we also assume each player can observe the successfully matched players $\text{Ch}_j(\bar{A}_j^{-1}(t)) = \bar{A}_j^{-1}(t) = \{p_i : \bar{A}_i(t) = a_j\}$ with each arm $a_j \in \mathcal{K}$ as previous works (Liu et al. 2021; Kong, Yin, and Li 2022; Ghosh et al. 2022; Kong and Li 2023; Wang et al. 2022).

The matching $\bar{A}(t)$ at round t is the set of all pairs $(p_i, \bar{A}_i(t))$. Stability of matchings is a key concept that describes the state in which any player or arm has no incentive to abandon the current partner (Gale and Shapley 1962; Roth and Sotomayor 1992). Formally, a matching is stable if it cannot be improved by any arm or player-arm pair. Specifically, an arm a_j improves $\bar{A}(t)$ if $\text{Ch}_j(\bar{A}_j^{-1}(t)) \neq \bar{A}_j^{-1}(t)$. That's to say, arm a_j would not accept all members in $\bar{A}_j^{-1}(t)$ when faced with this set. A pair (p_i, a_j) improves the matching $\bar{A}(t)$ if p_i prefers a_j to $\bar{A}_i(t)$ and a_j would accept p_i when faced with $\bar{A}_j^{-1}(t) \cup \{p_i\}$, i.e., $p_i \in \text{Ch}_j(\bar{A}_j^{-1}(t) \cup \{p_i\})$. That's to say, p_i prefers arm a_j than its current partner and a_j would also accept p_i if p_i apply for a_j together with a_j's current partners (Kelso Jr and Crawford 1982; Abdulkadiroğlu 2005; Roth and Sotomayor 1992).

Responsive preferences are widely studied in many-to-one markets which guarantee the existence of a stable matching (Roth and Sotomayor 1992; Wang et al. 2022). Under this setting, each arm a_j has a preference ranking over individual players and a capacity $C_j > 0$. When a set of players propose to a_j, it accepts C_j of them with the highest preference ranking. This case recovers the one-to-one matching when $C_j = 1$. For convenience, define $C = \sum_{j \in [K]} C_j$ as the total capacities of all arms. Apart from responsiveness, we also consider a more general substitutability setting in Section .

In this paper, we study the bandit problem in many-to-one matching markets with responsive and substitutable preferences. Under both properties, the set M^* of stable matchings between $\mathcal{N}$ and $\mathcal{K}$ is non-empty (Roth and Sotomayor 1992; Kelso Jr and Crawford 1982). For each player p_i, let $\overline{m}_i \in [K]$ and $\underline{m}_i \in [K]$ be the index of p_i's most and least favorite arm among all arms that can be matched with p_i in a

stable matching, respectively. The objective of each player p_i is to minimize the cumulative stable regret defined as the cumulative difference between the reward of the stable arm and that the player receives during the horizon. The player-optimal and pessimal stable regret are defined as

$$\overline{R}_i(T) = \mathbb{E}\left[\sum_{t=1}^{T} \left(\mu_{i,\overline{m}_i} - X_{i,A_i(t)}(t)\right)\right], \quad (1)$$

$$\underline{R}_i(T) = \mathbb{E}\left[\sum_{t=1}^{T} \left(\mu_{i,\underline{m}_i} - X_{i,A_i(t)}(t)\right)\right], \quad (2)$$

respectively (Liu, Mania, and Jordan 2020; Zhang, Wang, and Fang 2022; Kong and Li 2023; Wang et al. 2022). The expectation is taken over by the randomness in reward gains and the players' policies. For convenience, we define the preference gaps to measure the hardness of the problem.

Definition 1. *For each player p_i and arm $a_j \neq a_{j'}$, define $\Delta_{i,j,j'} = |\mu_{i,j} - \mu_{i,j'}|$ as the preference gap of p_i between a_j and $a_{j'}$. Let ρ_i be the preference ranking of player p_i, where $\rho_{i,k}$ represents the arm ranked k-th in p_i's preference. With a little abuse of notation, denote $\rho_i(a_j)$ as the rank of a_j in p_i's preference. Define $\Delta_{\min} = \min_{i,k \in [\min\{N,K-1\}]} \Delta_{i,\rho_{i,k},\rho_{i,k+1}}$ as the minimum preference gap between the arm ranked the first $\min\{N+1, K\}$-th among all players, $\Delta_{\overline{m}} = \min_{i,k \in [\rho_i(\overline{m}_i)]} \Delta_{i,\sigma_{i,k},\sigma_{i,k+1}}$ as the minimum preference gap between the arm ranked the first $(\rho_i(\overline{m}_i) + 1)$-th among all players and $\Delta_{\underline{m}} = \min_{i,k > \rho_i(\underline{m}_i)} \Delta_{i,\underline{m}_i,\rho_{i,k}}$ as the minimum preference gap between $\underline{m}_i$ and any arm that has lower ranking than $\underline{m}_i$ among all players.*

An Extension of Kong and Li (2023)

Recall that Kong and Li (2023) provide a near-optimal bound $O(K \log T / \Delta_{\min}^2)$ for player-optimal stable regret in one-to-one markets. We first provide an extension of their algorithm, explore-then-deferred-acceptance (ETDA), for many-to-one markets with responsiveness and $N \leq K \cdot \min_{j \in [K]} C_j$.

The deferred acceptance (DA) algorithm is designed to find a stable matching when both sides of participants have known preferences. The algorithm proceeds in multiple steps. At the first step, all players propose to their most preferred arm and each arm rejects all but their favorite subset of players among those who propose to it. Such a process continues until no rejection happens. It has been shown that the final matching is the player-optimal stable matching under responsiveness (Gale and Shapley 1962; Kelso Jr and Crawford 1982; Roth and Sotomayor 1992).

Since players are uncertain about their preferences, the ETDA algorithm lets players first explore to learn this knowledge and then follow DA to find a stable matching. Specifically, each player first estimates an index in the first N rounds (phase 1); and then explores its unknown preferences in a round-robin way based on its index (phase 2). After estimating a good preference ranking, it will follow DA to find the player-optimal stable matching (phase 3). Compared with Kong and Li (2023), the difference mainly lies in the first phase of estimating indices for players where multiple players can share the same index in many-to-one markets. For

completeness, we provide the detailed algorithm in Appendix and the theoretical guarantees below.

Theorem 1. *Under the responsiveness condition, when $N \leq K \cdot \min_{j \in [K]} C_j$, the player-optimal stable regret of each player p_i by following ETDA satisfies*

$$\overline{R}_i(T) \leq O\left(K \log T / \Delta_{\min}^2\right) . \tag{3}$$

Due to the space limit, the proof of Theorem 1 is deferred to Appendix . Under the same decentralized setting, this player-optimal stable regret bound is even $O(N^5 K \log T / \varepsilon^{N^4})$ better than the weaker player-pessimal stable regret bound in Wang et al. (2022). Such a result also achieves the same order as the state-of-the-art analysis in the reduced one-to-one setting (Kong and Li 2023).

Though achieving better regret bound, the ETDA algorithm is not incentive compatible. We can consider the market where the player-optimal stable arm of a player p_i is its least preferred arm. If p_i always reports that it does not estimate the preference ranking well, then the stopping condition of phase 2 is never satisfied. In this case, all of the other players fail to find a stable matching and suffer $O(T)$ regret, while this player is always matched with more preferred arms than that in the stable matching during phase 2, resulting in $O(T)$ improvement in the cumulative rewards. Thus player p_i lacks the incentive to always act as the algorithm requires. To improve the algorithm in terms of incentive compatibility, we further propose a novel algorithm in the next section.

Adaptively ETDA (AETDA) Algorithm

In this section, we propose a new algorithm adaptively ETDA (AETDA) for many-to-one markets with responsive preferences which is incentive compatible. To ensure each player has a chance to be matched, we simply assume $N \leq C$ as existing works in many-to-one and one-to-one markets (Liu, Mania, and Jordan 2020; Liu et al. 2021; Zhang, Wang, and Fang 2022; Kong and Li 2023; Wang et al. 2022), which relaxes the requirement of ETDA in the previous section.

For simplicity, we present the main algorithm in a centralized manner in Algorithm 1, i.e., a central platform coordinates players' selections in each round. The discussion on how to extend it to a decentralized setting is provided later.

Intuitively, AETDA integrates the learning process into each step of DA instead of estimating the full preference ranking well before running DA. More specifically, each player explores arms in a round-robin manner in each step to learn its most preferred arm and then focuses on this arm before being rejected in the corresponding step of DA. For each player p_i, the algorithm maintains S_i to represent the available arm set that has not rejected p_i in previous steps and E_i to represent the exploration status. Specifically, $E_i = $ True means that p_i still needs to explore arms in a round-robin manner to find its most preferred arm in S_i, and $E_i = $ False means that p_i now focuses on its most preferred available arm. At the beginning of the algorithm, S_i is initialized as the full arm set $\mathcal{K}$ and E_i is initialized as True (Line 1).

For players with $E_i = $ True, the central platform would allocate the arm $A_i(t) \in S_i$ in a round-robin manner. And for those players with $E_i = $ False, they

---

Algorithm 1: centralized adaptively explore-then-deferred-acceptance (AETDA, from the view of the central platform)

1: Initialize: $S_i = \mathcal{K}$, $E_i = $ True for each player $p_i \in \mathcal{N}$
2: **for** round $t = 1, 2, ..., $ **do**
3:     Allocate $A_i(t) \in S_i$ to each player p_i with $E_i = $ True in a round-robin manner; Allocate $A_i(t) = \text{opt}_i$ to each player p_i with $E_i = $ False
4:     Receive the estimation status opt_i from each p_i
5:     **for** each player $p_i \in \mathcal{N}$ with $\text{opt}_i \neq -1$ **do**
6:         $E_i = $ False
7:     **end for**
8:     **for** each player $p_i \in \mathcal{N}$ and $a_j \in S_i$ with $p_i \notin \text{Ch}_j(\{p_{i'} : \text{opt}_{i'} = a_j\} \cup \{p_i\})$ **do**
9:         $S_i = S_i \setminus \{a_j\}$
10:         Set $E_i = $ True if $E_i = $ False and $a_j = \text{opt}_i$
11:     **end for**
12: **end for**

---

can just focus on the determined optimal arm opt_i (Line 3). After being matched in each round, each player p_i would update its empirical mean $\hat{\mu}_{i,A_i(t)}$ and the number of observed times $T_{i,A_i(t)}$ on arm $A_i(t)$ as $\hat{\mu}_{i,A_i(t)} = (\hat{\mu}_{i,A_i(t)} \cdot T_{i,A_i(t)} + X_{i,A_i(t)}(t))/(T_{i,A_i(t)} + 1)$, $T_{i,A_i(t)} = T_{i,A_i(t)} + 1$. For the preference value $\mu_{i,j}$ towards each arm a_j, p_i also maintains a confidence interval at t with the upper bound $\text{UCB}_{i,j} := \hat{\mu}_{i,j} + \sqrt{6 \log T / T_{i,j}}$ and lower bound $\text{LCB}_{i,j} := \hat{\mu}_{i,j} - \sqrt{6 \log T / T_{i,j}}$. If $T_{i,j} = 0$, $\text{UCB}_{i,j}$ and $\text{LCB}_{i,j}$ are set as ∞ and $-\infty$, respectively. When the UCB of a_j is even lower than the LCB of other available arms, a_j is considered to be less preferred. Based on the estimations, p_i needs to determine whether an arm can be considered as optimal in S_i and submit this status to the platform (Line 4). Specifically, if there exists an arm $a_j \in S_i$ such that $\text{LCB}_{i,j} > \max_{a_{j'} \in S_i \setminus \{a_j\}} \text{UCB}_{i,j'}$, then a_j is regarded as optimal and player p_i would submit $\text{opt}_i = a_j$ to the platform. Otherwise, no arm can be regarded as optimal, and p_i would submit $\text{opt}_i = -1$. For players who have learned their most preferred arm, the platform would mark their exploration status as False (Line 6).

To avoid conflict when players with $E_i = $ True explore arms in a round-robin manner, we introduce a detection procedure to detect whether an arm in S_i is occupied by its more preferred players (Line 8-11). Specifically, if an arm a_j does not accept player p_i when faced with the player set who regards a_j as the optimal one (Line 8), then p_i can be regarded to be rejected by a_j when exploring this arm. In this case, no matter whether this arm is the most preferred one, p_i has no chance of being matched with it. So p_i directly deletes a_j from its available arm set S_i (Line 9). And if this arm is just the estimated optimal arm of p_i, then this case is equivalent in offline DA to that p_i is rejected when proposing to its most preferred arm. In this case, p_i needs to explore to learn its next preferred arm and update E_i as True (Line 10).

For the arrangement of round-robin exploration, without loss of generality, we can convert the original set of K arms with total capacity C into a set of C new arms, each with a

capacity 1. When N players explore these C new arms: the platform let p_1 follow the ordering $1, 2, ..., C-1, C, 1, ...; p_2$ follow $2, 3, ..., C, 1, 2, ...;$ and so on. If an arm a_j is unavailable for a player p_i, p_i simply forgo the opportunity to select in the corresponding rounds. This pre-arranged ordering ensures that, in the worst case, each player can match with each available new arm, and so as to the available original arm, at least once in every C rounds.

Extension to the decentralized setting. In the decentralized setting without a central platform, each player maintains and updates their own S_i and E_i. We can define a phase version of Algorithm 1. Specifically, each phase contains a number of rounds and the size of phases grows exponentially, i.e., $2, 2^2, 2^3, \cdots$. Within each phase, each player p_i would explore arms in S_i in a round-robin manner if $E_i = \text{True}$ as discussed above and focus on arm opt_i otherwise. Players only update the status of opt_i (Line 4), E_i (Line 6), and S_i (Line 8-11) at the end of the phase based on the communication with other players and arms. If L observations on arms are enough to learn the optimal one in the centralized version, then the stopping condition (Line 4) would be satisfied at the end of the phase guaranteeing the number of observations in this decentralized version and the total number of selecting times would be at most $2L$ due to the exponentially increasing phase length. So the regret in this decentralized version is at most two times as that suffered in the centralized version. And the number of communications is at most $O(\log T)$ which is of the same order as the ETDA algorithm and also Kong and Li (2023) for the one-to-one setting.

Theoretical Analysis

Algorithm 1 presents a new perspective that integrates the learning process into each step of the DA algorithm to find a player-optimal stable matching. In the following, we will show that such a design simultaneously enjoys guarantees of player-optimal stable regret and incentive compatibility.

Theorem 2. *Under the responsiveness condition, when $N \leq C$, the player-optimal stable regret of each player p_i by following Algorithm 1 satisfies*

$$\overline{R}_i(T) \leq O\left(N \cdot \min\{N, K\} C \log T / \Delta_{\overline{m}}^2\right).$$

Theorem 3. *(Incentive Compatibility) When all of the other players follow Algorithm 1, no single player p_i can improve its final matched arm by misreporting opt_i in some rounds.*

Compared with Wang et al. (2022), our result not only achieves an $O(N^4 K \log T / (C \varepsilon^{N^4}))$ improvement over their weaker player-pessimal stable regret objective but also enjoys guarantees of incentive compatibility. Compared with the state-of-the-art result in one-to-one settings, our algorithm is more robust to players' deviation only with the cost of $O(NC)$ worse regret bound (Zhang, Wang, and Fang 2022; Kong and Li 2023). To the best of our knowledge, it is the first algorithm that simultaneously achieves guarantees of polynomial player-optimal stable regret and incentive compatibility in both many-to-one markets and previously widely studied one-to-one markets without knowing the value of Δ.

Due to the space limit, the proofs of two theorems are deferred to Appendix .

Online DA Algorithm for Substitutability

In many-to-one markets, arms may have combinatorial preferences over groups of players, which may not be well characterized by responsiveness. In this setting, we consider the markets with substitutability, which is one of the most common and general conditions that ensure the existence of a stable matching and is defined below.

Definition 2. *(Substitutability) The preference of arm a_j satisfy substitutability if for any player set $P \subseteq \mathcal{N}$ that contains p_i and $p_{i'}$, $p_i \in \text{Ch}_j(P \setminus \{p_{i'}\})$ when $p_i \in \text{Ch}_j(P)$.*

The above property states that arm a_j keeps accepting player p_i when other players become unavailable. This is the sense that a_j regards players in a team as substitutes rather than complementary individuals (in which case the arm may give up accepting the player when others become unavailable). Such a phenomenon appears in many real applications and covers responsiveness as proved below.

Remark 1. *Select a player set $P \subseteq \mathcal{N}$ which contains p_i and $p_{i'}$. Suppose $p_i \in \text{Ch}_j(P)$, i.e., p_i is one of the C_j highest-ranked players in P. Then when the available set becomes $P \setminus \{p_{i'}\}$, p_i is still one of the C_j highest-ranked players, i.e., $p_i \in \text{Ch}_j(P \setminus \{p_{i'}\})$.*

The substitutability property is more general than responsiveness as arms' preferences can have combinatorial structures. The following is an example that satisfies substitutability but not responsiveness (Roth and Sotomayor 1992).

Example 1. $\mathcal{N} = \{p_1, p_2, p_3\}$ *and* $\mathcal{K} = \{a_1, a_2\}$. *The arms' preference rankings over subsets of players are*

- $a_1 : \{p_1, p_2\}, \{p_1, p_3\}, \{p_2, p_3\}, \{p_3\}, \{p_2\}, \{p_1\}$.
- $a_2 : \{p_3\}, \emptyset$.

That is to say, $\text{Ch}_j(P)$ is the subset that ranks highest among all subsets listed above that only contain players in P. Taking the preferences of a_2 as an example, when $p_3 \in P$, then $\text{Ch}_j(P) = \{p_3\}$; otherwise, $\text{Ch}_j(P) = \emptyset$.

For many-to-one markets with substitutable preferences, we propose an online deferred acceptance (ODA) algorithm (presented in Algorithm 2). ODA is inspired by the idea of the DA algorithm with the arm side proposing, which finds a player-pessimal stable matching when players know their preferences. Specifically, the DA algorithm with the arm proposing proceeds in several steps. In the first step, each arm proposes to its most preferred subset among all players. Each player would reject all but the most preferred arm among those who propose it. In the following each step, each arm still proposes to its most preferred subset of players among those who have not rejected it and each player rejects all but the most preferred one among those who propose to it. This process stops when no rejection happens and the final matching is the player-pessimal stable matching (Kelso Jr and Crawford 1982; Roth and Sotomayor 1992).

The ODA algorithm is designed with the guidance of this procedure but players decide which arm to select in each round. Specifically, each player p_i needs to record the available player set $P_{i,j}$ for each arm a_j, which consists of players who have not rejected arm a_j and is initialized as the full player set $\mathcal{N}$. Then if a player p_i is in the choice set of a_j

Algorithm 2: online deferred acceptance (from view of p_i)

1: Input: player set $\mathcal{N}$, arm set $\mathcal{K}$
2: Initialize: $P_{i,j} = \mathcal{N}, \hat{\mu}_{i,j} = 0, T_{i,j} = 0$ for each $j \in [K]$; $S_i(1) = \{a_j \in \mathcal{K} : p_i \in \text{Ch}_j(P_{i,j})\}$
3: **for** each round $t = 1, 2, \cdots$ **do**
4:     Select $A_i(t) \in S_i(t)$ in a round-robin way
5:     Update $\hat{\mu}_{i,\bar{A}_i(t)}$ and $T_{i,\bar{A}_i(t)}$ if $\bar{A}_i(t) = A_i(t) \neq \emptyset$
6:     $S_i(t+1) = S_i(t)$
7:     **for** $a_j \in S_i(t)$ and $\text{UCB}_{i,j}(t) < \max_{a_{j'} \in S_i(t)} \text{LCB}_{i,j'}(t)$ **do**
8:         $S_i(t+1) = S_i(t+1) \setminus \{a_j\}$
9:     **end for**
10:     **if** $t \geq 2$ and $\forall p_{i'} \in \mathcal{N} : \bar{A}_{i'}(t) = \bar{A}_{i'}(t-1)$ **then**
11:         $\forall j \in [K]$, $P_{i,j} = P_{i,j} \setminus \{p_{i'} : \bar{A}_{i'}(t) \neq j, \exists t' < t-1 \text{ s.t. } \bar{A}_{i'}(t') = j\}$
12:         $S_i(t+1) = \{a_j : p_i \in \text{Ch}_j(P_{i,j})\}$
13:     **end if**
14: **end for**

when the set $P_{i,j}$ of players is available, i.e., $p_i \in \text{Ch}_j(P_{i,j})$, p_i would be accepted if it proposes to a_j together with other players in $P_{i,j}$. The main purpose of the algorithm is to let players wait for this opportunity to choose arms that will successfully accept them.

Each player p_i can further construct the plausible set S_i to contain those arms that may successfully accept it, i.e., $S_i = \{a_j : p_i \in \text{Ch}_j(P_{i,j})\}$. Here for simplicity, we additionally assume each player p_i knows whether $p_i \in \text{Ch}_j(P)$ for each possible $P \subseteq \mathcal{N}$. This assumption is only used for clean analysis and the algorithm can also be generalized to the case where this information is unavailable by letting players in $P_{i,j}$ pull a_j and observe whether it is accepted. Since arms know their own preferences and conflicts are deterministically resolved, at most 2^N rounds are needed to obtain this information. Apart from $P_{i,j}$ and S_i, each player p_i also maintains $\hat{\mu}_{i,j}$ and $T_{i,j}$ to record the estimated value for $\mu_{i,j}$ and the number of its observations. At the beginning, both values are initialized to 0.

In each round t, each player p_i proposes to the arm a_j in the plausible set $S_i(t)$ in a round-robin way (Line 4). If they are successfully matched with each other (Line 5), p_i would update the corresponding $\hat{\mu}_{i,j}, T_{i,j}$ as Section . When the UCB of a_j is even lower than the LCB of other plausible arms, a_j is considered to be less preferred. In this case, the final stable arm of player p_i must be more preferred than a_j and thus there is no need to select a_j anymore (Line 8).

Recall that the plausible sets of players are constructed based on the available sets for arms. To ensure each player successfully be accepted by arms in their own plausible set, all players need to keep the available sets for arms updated in sync. With the awareness that players always select plausible arms in a round-robin way, once p_i observes that all players focus on the same arm in the recent two rounds, it believes all players have determined the most preferred one. In this way, p_i updates the available set $P_{i,j}$ for each arm a_j by deleting players who do not consider a_j as stable arms (Line 11). Since all players have the same observations, the update

times of $P_{i,j}$ would be the same. Such a stage in which all players determine the most preferred arm in the plausible set can just be regarded as a step of the offline DA algorithm (with the arm side proposing) where each player rejects all but the most preferred one among those who propose to it. Thus the update times of $P_{i,j}$ just divide the total horizon into several stages with each corresponding to a step of DA.

Theoretical Analysis

We first provide the regret guarantee for Algorithm 2.

Theorem 4. *Under the substitutability condition, when players know arms' exact preferences, the player-pessimal stable regret of each player p_i by following Algorithm 2 satisfies*

$$\underline{R}_i(T) \leq O(NK \log T / \Delta_{\underline{m}}^2). \tag{4}$$

Apart from the regret guarantee, we also discuss the incentive compatibility of the algorithm.

Theorem 5. *(Incentive Compatibility) Suppose that all of the other players follow the ODA algorithm, then a single player p_i has no incentive to select arms beyond S_i. And if p_i misreports its estimated optimal arm in S_i towards the optimal manipulation for itself, i.e., a manipulation under which the DA algorithm would match p_i with an arm has a higher ranking than that under other manipulations, all of the other players would also benefit from this behavior.*

How to define arms' preferences over subsets of players is an interesting question. Our method provides the first attempt. When players do not know arms' preferences, a dependence on 2^N would be involved as the cost of learning arms' combinatorial preferences. Removing such dependence would be more preferred. But as a preliminary step for combinatorial preferences, understanding algorithmic performance under more comprehensive information conditions is also important as it lays the groundwork for further exploration in more generalized settings. Due to the space limit, the proofs of two theorems are provided in Appendix .

Conclusion

In this paper, we study the bandit learning problem in many-to-one markets. We first extend the result of Kong and Li (2023) to the many-to-one markets with responsive preferences and provide a player-optimal regret bound. Since such an algorithm lacks incentive compatibility, we further propose the AETDA algorithm which enjoys a guarantee of player-optimal regret and is simultaneously incentive compatible. We also consider a more general setting with substitutable preferences and provide an upper bound for player-pessimal stable regret. Compared with existing works for many-to-one markets (Wang et al. 2022), our algorithms achieve a significant improvement in terms of not only regret bound but also guarantees of incentive compatibility.

An interesting future direction is to optimize the player-optimal stable regret in the general many-to-one markets with substitutable preferences. All of the previous algorithms for the reduced settings go through based on the uniform exploration strategy. However, under substitutability, an arm may accept none of the candidates which makes it challenging for players to perform such a strategy.

Acknowledgements

References

Abdulkadiroğlu, A. 2005. College admissions with affirmative action. *International Journal of Game Theory*, 33(4): 535–549.

Abdulkadiroğlu, A.; and Sönmez, T. 1999. House allocation with existing tenants. *Journal of Economic Theory*, 88(2): 233–260.

Agrawal, S.; and Goyal, N. 2012. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, 39–1.

Auer, P.; Cesa-Bianchi, N.; and Fischer, P. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2): 235–256.

Auer, P.; and Ortner, R. 2010. UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2): 55–65.

Basu, S.; Sankararaman, K. A.; and Sankararaman, A. 2021. Beyond $\log^2(T)$ regret for decentralized bandits in matching markets. In *International Conference on Machine Learning*, 705–715.

Bistritz, I.; and Leshem, A. 2018. Distributed multi-player bandits-a game of thrones approach. *Advances in Neural Information Processing Systems*, 7222–7232.

Bubeck, S.; Budzinski, T.; and Sellke, M. 2021. Cooperative and stochastic multi-player multi-armed bandit: Optimal regret with neither communication nor collisions. In *Conference on Learning Theory*, 821–822. PMLR.

Che, Y.-K.; Kim, J.; and Kojima, F. 2019. Stable matching in large economies. *Econometrica*, 87(1): 65–110.

Das, S.; and Kamenica, E. 2005. Two-sided bandits and the dating market. In *International Joint Conference on Artificial Intelligence*, 947–952.

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

Epple, D.; Romano, R.; and Sieg, H. 2006. Admission, tuition, and financial aid policies in the market for higher education. *Econometrica*, 74(4): 885–928.

Erdil, A.; and Kumano, T. 2019. Efficiency and stability under substitutable priorities with ties. *Journal of Economic Theory*, 184: 104950.

Fu, C. 2014. Equilibrium tuition, applications, admissions, and enrollment in the college market. *Journal of Political Economy*, 122(2): 225–281.

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

Garivier, A.; Lattimore, T.; and Kaufmann, E. 2016. On explore-then-commit strategies. *Advances in Neural Information Processing Systems*, 29: 784–792.

Ghosh, A.; Sankararaman, A.; Ramchandran, K.; Javidi, T.; and Mazumdar, A. 2022. Decentralized Competing Bandits in Non-Stationary Matching Markets. *arXiv preprint arXiv:2206.00120*.

Kelso Jr, A. S.; and Crawford, V. P. 1982. Job matching, coalition formation, and gross substitutes. *Econometrica: Journal of the Econometric Society*, 1483–1504.

Kong, F.; and Li, S. 2023. Player-optimal Stable Regret for Bandit Learning in Matching Markets. In *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM.

Kong, F.; Yin, J.; and Li, S. 2022. Thompson Sampling for Bandit Learning in Matching Markets. In *International Joint Conference on Artificial Intelligence*.

Lattimore, T.; and Szepesvári, C. 2020. *Bandit algorithms*. Cambridge University Press.

Liu, L. T.; Mania, H.; and Jordan, M. 2020. Competing bandits in matching markets. In *International Conference on Artificial Intelligence and Statistics*, 1618–1628. PMLR.

Liu, L. T.; Ruan, F.; Mania, H.; and Jordan, M. I. 2021. Bandit learning in decentralized matching markets. *Journal of Machine Learning Research*, 22(211): 1–34.

Maheshwari, C.; Mazumdar, E.; and Sastry, S. 2022. Decentralized, Communication-and Coordination-free Learning in Structured Matching Markets. In *Advances in Neural Information Processing Systems*.

Rosenski, J.; Shamir, O.; and Szlak, L. 2016. Multi-player bandits–a musical chairs approach. In *International Conference on Machine Learning*, 155–163. PMLR.

Roth, A. E. 1982. The economics of matching: Stability and incentives. *Mathematics of operations research*, 7(4): 617–628.

Roth, A. E. 1984a. 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.

Roth, A. E. 1984b. Stability and polarization of interests in job matching. *Econometrica: Journal of the Econometric Society*, 47–57.

Roth, A. E.; and Sotomayor, M. 1992. Two-sided matching. *Handbook of game theory with economic applications*, 1: 485–541.

Sankararaman, A.; Basu, S.; and Sankararaman, K. A. 2021. Dominate or Delete: Decentralized Competing Bandits in Serial Dictatorship. In *International Conference on Artificial Intelligence and Statistics*, 1252–1260. PMLR.

Shi, C.; Xiong, W.; Shen, C.; and Yang, J. 2021. Heterogeneous Multi-player Multi-armed Bandits: Closing the Gap and Generalization. *Advances in Neural Information Processing Systems*, 34.

Vaish, R.; and Garg, D. 2017. Manipulating gale-shapley algorithm: preserving stability and remaining inconspicuous. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 437–443.

Wang, Z.; Guo, L.; Yin, J.; and Li, S. 2022. Bandit Learning in Many-to-One Matching Markets. In *Proceedings of*

the 31st ACM International Conference on Information & Knowledge Management, 2088–2097.

Zhang, Y.; Wang, S.; and Fang, Z. 2022. Matching in Multiarm Bandit with Collision. In *Advances in Neural Information Processing Systems*.