A Survey on Bandit Learning in Matching Markets

Shuai Li[1*], **Zilong Wang**[1], **Fang Kong**[2]
[1]Shanghai Jiao Tong University
[2]Southern University of Science and Technology
{shuaili8, wangzilong}@sjtu.edu.cn, kongf@sustech.edu.cn

Abstract

The two-sided matching market problem has attracted extensive research in both computer science and economics due to its wide-ranging applications in multiple fields. In various online matching platforms, market participants often have unclear preferences. As a result, a growing area of research focuses on the online scenario. Here, one-side participants (players) gradually figure out their unknown preferences through multiple rounds of interactions with the other-side participants (arms). This survey comprehensively reviews and systematically organizes the abundant literature on bandit learning in matching markets. It covers not only existing theoretical achievements but also various other related aspects. Based on the current research, several distinct directions for future study have emerged. We are convinced that delving deeper into these directions could potentially yield theoretical algorithms that are more suitable for real-world situations.

1 Introduction

The two-sided matching market, a classic model, has diverse applications in areas such as school admissions and labor markets. It has been extensively explored in literature, as seen in [Gale and Shapley, 1962; Roth, 1984; Roth and Sotomayor, 1992; Roth, 2002]. In this market, two sets of participants exist, each having preferences for those on the opposite side. Stability is a key property defining the matching's equilibrium. The problem of finding stable matchings in a given market has been studied for ages, with [Gale and Shapley, 1962] and [Roth and Sotomayor, 1992] making significant contributions. These studies usually assume that preferences are fixed and known. However, this assumption is often unrealistic in real-world applications, such as online labor markets where employers are uncertain about their preferences for workers and can learn through multiple rounds of iterative matching.

Multi-armed bandit (MAB) is an essential mechanism for decision - making when faced with uncertainty across multiple rounds. This framework typically involves a single player and K arms, where each arm has its own unique reward distribution that remains unknown to the player. The player's objective is to minimize their cumulative regret, defined as the expectation of the difference in cumulative rewards between the arm with the highest reward and the player's selected arm over T rounds. Given that the player can only observe the reward of the arm they choose to pull, they need to strike a balance between exploration, which means gathering information about the various arms, and exploitation, which involves choosing the arm that appears to be the best based on the available information.

The problem of bandit learning in matching markets is first introduced by [Das and Kamenica, 2005] and has been studied by a rich line of work [Liu *et al.*, 2020; Liu *et al.*, 2021; Basu *et al.*, 2021; Sankararaman *et al.*, 2021; Kong *et al.*, 2022; Kong and Li, 2023; Zhang *et al.*, 2022; Wang *et al.*, 2022; Kong and Li, 2024; Kong *et al.*, 2024]. Players and arms correspond to the participants on two sides of markets. Each player has an unknown preference, while arms are certain about their preferences over players. For example, employers are unsure of their preferences because they lack knowledge about workers' capabilities, which can only be understood through interactions. This problem aims to minimize the stable regret, defined as the reward difference between the stable matched arm and the player's selected arm over T rounds. When the stable matching is not unique, there are mainly two types of regret: one is the player-optimal stable regret concerning the player's most preferred stable matching; the other one is the player-pessimal stable regret with respect to the player's least preferred stable matching.

In addition to the fundamental one-to-one matching markets focused on minimizing stable regret, a series of research efforts explore other variations. One strand of work expands the market framework to the many-to-one scenario [Wang *et al.*, 2022; Kong and Li, 2024]. Here, each arm has the capacity to accept multiple players. Another line of inquiry delves into non-stationary markets [Muthirayan *et al.*, 2022; Ghosh *et al.*, 2022]. In such markets, the preferences of each player do not remain constant but change over time. This dynamic nature adds a layer of complexity to the matching process, as algorithms must adapt to these evolving preferences. Some studies concentrate on the setting of two-sided unknown preferences [Pagare and Ghosh, 2023; Zhang and Fang, 2024b]. In this case, not only are the

*Corresponding author.

players' preferences a factor for the learner, but the preferences on the arm side are also unknown and appear random to the learner. Finally, there are works that shift the focus away from merely minimizing regret [Hosseini *et al.*, 2024; Athanasopoulos *et al.*, 2025]. Instead, they aim to minimize the sample complexity required to identify a stable matching. This approach emphasizes the efficiency of the matching process in terms of the amount of data needed, rather than just the quality of the matching in terms of regret.

This survey situates and structures the extensive body of literature on bandit learning for matching markets. In Section 2 we introduce the basic framework of bandit learning in matching markets. In Section 3 the problem lower bound is introduced. Section 4 introduces the existing algorithm for centralized environment and decentralized environment respectively. In Section 5, we introduce the works beyond the basic one-to-one matching markets. Section 6 summarizes the potential future directions.

2 Preliminaries

This section introduces the model of bandit learning in the one-to-one matching markets with one-sided unknown preferences, which is a basic model studied by most works.

Suppose there are N players and K arms. Denote $\mathcal{N} = \{p_1, p_2, \cdots, p_N\}$ as the set of players and denote $\mathcal{K} = \{a_1, a_2, \cdots, a_K\}$ as the set of arms. Most works assume $N \leq K$ to ensure no player will be unmatched. Each arm has a fixed preference rank $\pi_{k,i}$ over players. $\pi_{k,i} > \pi_{k,i'}$ means arm a_k prefers player p_i to $p_{i'}$. The preference of player p_i over arm a_k is modeled by the utility $\mu_{i,k} > 0$. $\mu_{i,k} > \mu_{i,k'}$ implies that player p_i prefers arm a_k rather than $a_{k'}$. The preferences of players are random and unknown, and can be learned through interactive matching iterations.

For each player p_i and arm $a_k \neq a_{k'}$, let $\Delta_{i,k,k'} = |\mu_{i,k} - \mu_{i,k'}|$ be the reward gap between arm a_k and $a_{k'}$ for player p_i. Define $\Delta = \min_{i,k,k'} \Delta_{i,k,k'} > 0$ as the minimum reward gap across all players and arms, which measures the hardness of the learning problem.

At each round $t = 1, 2, \ldots$, each player p_i proposes to an arm $A_i(t)$. For each arm a_k, denote $A_k^{-1}(t) = \{p_i : A_i(t) = a_k\}$ as the set of players who selects arm a_k at round t. When more than one player selects a_j, it accepts its most-preferred one in $A_k^{-1}(t)$, i.e. a_k will match with $p_i \in \arg\max_{p_i \in A_k^{-1}(t)} \pi_{k,i}$. If a player p_i is successfully matched with arm $A_i(t)$, it will receive a random reward $X_i(t)$ characterizing its matching experience, which we assume is a 1-subgaussian random variable with expectation $\mu_{i,A_i(t)}$. Otherwise, p_i is rejected by its proposed arm and only gets reward $X_i(t) = 0$.

Stability is a key property of a matching in two-sided markets to prevent the system from collapse [Gale and Shapley, 1962; Roth and Sotomayor, 1992]. A matching $\bar{A}(t) = \{(i, \bar{A}_i(t)) : i \in [N]\}$ is stable if no market participant wants to break up its current matching relationship and find a new partner. Formally speaking, there is no player-arm pair (p_i, a_k) such that $\mu_{i,k} > \mu_{i,\bar{A}_i(t)}$ and $\pi_{k,i} > \pi_{k,\bar{A}_k^{-1}(t)}$. It is worth noting that there may be multiple stable matchings in the market. Denoted $M = \{m : m \text{ is stable}\}$ as the set of all stable matchings. It is shown that there exists a stable matching $m^* \in M$ such that all players are matched with their most preferred stable arm [Gale and Shapley, 1962], i.e., $\mu_{i,m_i^*} \geq \mu_{i,m_i}$ for any $m \in M, i \in [N]$, which is called the player-optimal stable matching. Meanwhile there also exists a stable matching $\underline{m}^* \in M$ such that all players are matched with their least preferred stable arm, i.e., $\mu_{i,\underline{m}_i^*} \leq \mu_{i,m_i}$ for any $m \in M, i \in [N]$, which is called the player-pessimal stable matching.

Here we introduce a classic offline algorithm called the Gale-Shapley (GS) algorithm that efficiently finds a stable matching when preferences on both sides are known. This algorithm is also implemented by many online matching works thus we introduce it in this section. It proceeds in rounds: in each round, each unmatched participant proposes to the most preferred participant on their list who has not yet rejected them. The participants receiving proposals tentatively accept the best proposal they have received so far and reject the rest. This process continues until all participants are matched. The key feature of the GS algorithm is that it always terminates and produces a player-optimal stable matching, providing an efficient and reliable way to achieve stable pairings in various matching scenarios.

Given a specified horizon T, the learning objective is to minimize the stable regret for each player p_i. Since there may exist multiple stable matchings, there are two different definitions of regret with respect to player-optimal stable matching m^* and player-pessimal stable matching $\underline{m}^*$ respectively. The player-optimal stable regret is defined as the difference between the cumulative reward received by being matched with m_i^* and the cumulative reward received by p_i over T rounds:

$$\mathrm{Reg}_i(T) = \mathbb{E}\left[\sum_{t=1}^{T} \left(\mu_{i,m_i^*} - X_i(t)\right)\right].$$

Here, the expectation is taken over by the randomness of the reward generation and the randomness inherent in the player's strategy. Similarly, the player-pessimal stable regret is defined as the difference between the cumulative reward received by being matched with $\underline{m}_i^*$ and the cumulative reward received by p_i over T rounds.

3 Lower Bound

[Sankararaman *et al.*, 2021] provide the regret lower bound of $\Omega(\max\{N \log T/\Delta^2, K \log T/\Delta\})$. They study the special market named Optimally stable bandits (OSB) where all arms share the same preferences and each player's stable matched arm is exactly its optimal arm, i.e., $m_i^* = \arg\max_k \mu_{i,k}$. In this instance, the $\Omega(N \log T/\Delta^2)$ comes from the collisions when other players select the stable matched arm of player i, and $\Omega(K \log T/\Delta)$ comes from the necessary explorations.

Theorem 1. *(Theorem 7 in [Sankararaman* et al.*, 2021]) For any agent $i \in [N]$, under any decentralized universally consistent algorithm π on a OSB instance ν satisfies*

$$\mathrm{Reg}_i(T) \geq \max\left\{\frac{(i-1)\log T}{\Delta^2}, \frac{K \log T}{\Delta}\right\}.$$

	Regret bound	Setting
[Liu *et al.*, 2020]	$O\left(K\log T/\Delta^2\right)^*$ $O\left(NK\log T/\Delta^2\right)$	known Δ, gap_1 gap_2
[Liu *et al.*, 2021]	$O\left(\frac{N^5K^2\log^2 T}{\varepsilon^{N^4}\Delta^2}\right)$	gap_2
[Sankararaman *et al.*, 2021]	$O\left(NK\log T/\Delta^2\right)$ $\Omega\left(\max\left\{N\log T/\Delta^2, K\log T/\Delta\right\}\right)$	serial dictatorship, gap_1
[Basu *et al.*, 2021]	$O\left(K\log^{1+\varepsilon}T+2^{(\frac{1}{\Delta^2})^{\frac{1}{\varepsilon}}}\right)^*$ $O\left(NK\log T/\Delta^2\right)$	gap_2 α-condition, gap_1
[Maheshwari *et al.*, 2022]	$O\left(CNK\log T/\Delta^2\right)$	α-reducible condition, communication-free, gap_1
[Kong *et al.*, 2022]	$O\left(\frac{N^5K^2\log^2 T}{\varepsilon^{N^4}\Delta^2}\right)$	gap_2
[Zhang *et al.*, 2022]	$O\left(K\log T/\Delta^2\right)^*$	gap_2
[Kong and Li, 2023]	$O\left(K\log T/\Delta^2\right)^*$	gap_3
[Wang and Li, 2024]	$O\left(N\log T/\Delta^2+K\log T/\Delta\right)$	serial dictatorship, gap_3
[Kong *et al.*, 2024]	$O\left(N^2\log T/\Delta^2+K\log T/\Delta\right)^*$ $O\left(N\log T/\Delta^2+K\log T/\Delta\right)$	gap_4 α-condition, gap_3

Table 1: Comparisons of settings and regret bounds with works of basic one-to-one matching with one-sided unknown preference setting, $*$ represents the player-optimal stable regret and bounds without labeling $*$ are for player-pessimal stable regret or the unique stable matching. N and K are the number of players and arms with $N \leq K$, T is the total horizon, Δ corresponds to some preference gap, ε depends on the hyper-parameter of algorithms, and C is related to the unique stable matching condition which can grow exponentially in N. The definition of Δ in different works has different notions: gap_1 is the minimum preference gap between the (player-optimal) stable arm and the next arm in the preference ranking among all players; gap_2 is the minimum preference gap between any two different arms among all players; gap_3 is the minimum preference gap between the first $N+1$ ranked arms among all players; gap_4 is the minimum preference gap between arms that are more preferred than the next arm after the player-optimal stable arm among all players. The following inequality holds $\mathrm{gap}_1 \geq \mathrm{gap}_4 \geq \mathrm{gap}_3 \geq \mathrm{gap}_2$.

From the lower bound analysis, the regret is lower bounded by two terms: the necessary explorations for those sub-optimal arms, and the unavoidable collisions caused by other players' explorations. This analysis also guides the algorithm design idea to attain the better regret upper bound.

4 Existing Results

In this section we review existing algorithms with their conditions and the corresponding regret upper bound.

4.1 Centralized Market

[Liu *et al.*, 2020] first theoretically analyze the online matching market with the bandit problem. They consider the centralized scenario, in which a platform receives the preferences of both sides and assign matching for each player. They apply both the ETC algorithm and the UCB algorithm to estimate the ranking.

In the centralized ETC algorithm, players first explore all arms in a round-robin way in the first h rounds. After that platforms collect players' estimated preferences for each arm based on empirical mean reward, and then the platform computes the player-optimal stable matching by running Gale-Shapley algorithm and each player keeps selecting its stable matched arm in the remaining rounds. Centralized ETC requires the knowledge of reward gap Δ and time horizon T to determine the explore horizon h, and it achieves the $O\left(K\log(T)/\Delta^2\right)$ player-optimal regret.

In the centralized UCB algorithm, each player estimates the reward of arm by UCB index. At each round the centralized platform runs Gale-Shapley algorithm based on the preferences ranked by UCB index. Each player follows platform's assigned estimated player-optimal stable matching at each round and updates their UCB indices. This algorithm does not require the knowledge of Δ and T, but it may fail to converge to player-optimal stable matching and only converge to a stable matching in some preference structures. They prove the $O\left(NK\log(T)/\Delta^2\right)$ player-pessimal regret for the centralized UCB algorithm.

Note that though the centralized UCB does not need to know the parameters T, Δ beforehand, it can only achieve the sub-linear player-pessimal regret, which is a weaker notion of

regret compared with player-optimal regret. This is because since each player ranks its preferences over arm based on the UCB index, it does not imply that the player will be able to select the arm with higher UCB index since the player might be rejected by running GS algorithm. Thus there may exist some arms with high UCB values but never be matched, which leads to the failure of the centralized UCB to attain sub-linear player-optimal stable regret.

Note that in real-world case, the centralized setting is rarely satisfied since it requires a platform that assigns actions for all players and is hard to achieve when the market size is large. Thus only few works study the centralized setting and most of the works study the more general decentralized setting.

4.2 Decentralized Market

This section introduces the literature of studying the decentralized market, where each player make decisions based on its own observation without the coordination from the central platform.

Known Matching Result at Each Round

In this subsection, we introduce decentralized algorithms that assume players can observe the entire matching result at each round, i.e., $\bar{A}_i(t)$ for every $i \in [N]$.

[Liu *et al.*, 2021] and [Kong *et al.*, 2022] propose the UCB and TS-type algorithm for the decentralized market, respectively. They both obtain $O\left(\exp(N^4)N^5K^2\log^2(T)/\Delta^2\right)$ player-pessimal stable regret. The algorithm design idea is that each player tries to select the best possible successfully matched arm based on the last round's matching information, which means they will select the best possible arm where no collisions happen with high probability. And this selection approach will converge to a stable matching when each player has the correct estimation of the best possible arm, which is guaranteed by the UCB algorithm [Liu *et al.*, 2021] and the TS algorithm [Kong *et al.*, 2022], respectively.

Subsequently, [Kong and Li, 2023] carries out the in-depth research. They put forward the explore-then-Gale-Shapley (ETGS) algorithm. At first, each player explores all arms in a round-robin way. When all players have identified their full preference rankings over all arms, they terminate the exploration and run the GS algorithm to obtain the player-optimal stable matching. To make each player aware of other players exploration status, the exploration process is divided by phases with exponentially growing length. After each phase, each player communicate with each other by selecting certain arms and observing the matching result. The work effectively improves the player-optimal stable regret bound. Specifically, it has enhanced the bound to $O\left(K\log(T)/\Delta^2\right)$.

Not Observing Matching Result

In this subsection, we introduce the decentralized algorithms that assume players can only observe their matching information.

Since not observing the matching result is more challenging, most of the works study the market satisfying certain uniqueness conditions as a beginning. The uniqueness condition means the preferences over arms and players satisfy some constraints to ensure that the stable matching is unique.

The work of [Sankararaman *et al.*, 2021] proposes the UCB-D3 algorithm based on the assumption of serial dictatorship, where each arm has the same preference over players. Intuitively, each player selects the arm with the highest UCB index while not facing colliding. Since the market satisfies the serial dictatorship where all arms have the same preferences over players, each player will converge to its stable matching sequentially from the most preferred player to the least preferred one. An $O(NK\log T/\Delta^2)$ regret is derived. They also provide the lower bound analysis.

The work of [Basu *et al.*, 2021] assumes the market satisfies α-condition, which generalizes the serial dictatorship [Sankararaman *et al.*, 2021] and is the weakest sufficient condition to guarantee the uniqueness. They propose the UCB-D4 algorithm modified based on UCB-D3 algorithm while maintaining an additional deleting arm set. Each player will also converge to stable matching by the order determined by α-condition. They also achieve the $O\left(NK\log(T)/\Delta^2\right)$ regret bound.

[Maheshwari *et al.*, 2022] study the market satisfying α-reducible and proposes a communication-free algorithm, where each player does not communicate with other players to synchronous their information. Note that other works focusing on the same decentralized setting all permit the communication process that exchange the information among players, and this work only allows each player make decisions without information from any other players. α-reducible is also a uniqueness condition weaker than serial dictatorship but stronger than α-condition. In their work, each player runs a single-player adversarial bandit algorithm, which means other players' actions are treated as an adversary and take actions based on that player's strategy. Each player tunes its stochastic policy based on collision and reward information observed at each step. The technique of instance-dependent regret for adversarial bandit is applied for obtaining the $O\left(NK\log(T)/\Delta^2\right)$ stable regret.

As for the general matching markets, the work of [Zhang *et al.*, 2022] improved the player-optimal stable regret. They have proposed the ML-ETC algorithm. By the structure of arms' preferences, players are divided into different levels. The choices of players with lower levels do not influence those with higher levels. Then the algorithm finds the player-optimal stable matching from the highest level to the lowest. In the same level, each player explores all arms in a round-robin way until they identify their full rankings of all arms, then they run the GS algorithm to find the stable matching, similar with [Kong and Li, 2023]. The difference is that in this work players communicate their information by collisions rather than the total matching result, which avoids observing the matching result at each round. Through this algorithm, they have achieved the player-optimal stable regret bound of $O\left(K\log(T)/\Delta^2\right)$.

To improve the regret upper bound to reach the lower bound [Sankararaman *et al.*, 2021], [Wang and Li, 2024] first study the market satisfying serial dictatorship. Since players have an order from the most preferred to the least preferred, each player runs the elimination algorithm to find the stable matched arm sequentially. Here elimination algorithm is a classic bandit algorithm where the learner explores

arms in a round-robin way and eliminates the sub-optimal arm when it is identified not the best. Player p_1 first runs the elimination algorithm to find its best arm, while player p_2 runs the elimination algorithm for other arms except p_1 selects. Then similarly for player p_i, she runs the elimination algorithms for arms of those not selected by player $p_1, \ldots, p_{i-1}$. Then the regret for player p_i is bounded by $O((i-1)\log T/\Delta^2 + K\log T/\Delta)$. Here $O(K\log T/\Delta)$ is derived from the number of explorations for sub-optimal arms. $O((i-1)\log T/\Delta^2)$ is obtained from the number of times player $p_1, \ldots, p_{i-1}$ explores the stable arm of p_i and thus she can not select its stable matched arm. It should be noted that Δ is the minimum gap among all K arms and is smaller than that in the lower bound analysis, which is the minimum reward gap between sub-optimal arms and the stable matched arm. Thus there is still a gap on the dependence of Δ.

Recently [Kong *et al.*, 2024] provides the improved algorithm for the decentralized general matching markets. To remove the dependence of K on the leading term, they use the elimination algorithm to explore the sub-optimal arm, similar to [Wang and Li, 2024]. Unlike identifying all rankings among arms and running the full steps of GS algorithms [Kong and Li, 2023; Zhang *et al.*, 2022]. This work proposes the adaptive online GS algorithm, which divides each step of GS algorithm into exploration. At first, all players run the elimination algorithm to identify their best arms. Note that each player eliminates arms until the number of non-eliminated arms equals to the number of players, i.e., N. This design is to ensure all players can explore arms in a round-robin way without collisions. When all players have identified their best arms, they run one step of the GS algorithm. Those rejected players turn to explore their second best arms. And those accepted players keep selecting their matched arms. The next step of GS is performed when those rejected players have identified their second best arms. The algorithm finds the player-optimal stable matching when full steps of GS have been performed. An $O(N^2\log T/\Delta^2 + K\log T/\Delta)$ player-optimal regret is obtained. This work removes the dependence on K in the leading term for general markets.

Theorem 2. *[Kong* et al.*, 2024] Following the algorithm proposed in [Kong* et al.*, 2024], the player-optimal stable regret for each player p_i satisfies*

$$\mathrm{Reg}_i(T) \leq O\left(N^2\log T/\Delta^2 + K\log T/\Delta\right),$$

where Δ is the minimum preference gap between arms that are more preferred than the next of the player-optimal stable arm among all players.

Note that there is still a gap between the problem lower bound $\Omega(N\log T/\Delta^2 + K\log T/\Delta)$ [Sankararaman *et al.*, 2021] and the state-of-the-art upper bound $O(K\log T/\Delta^2)$ [Kong and Li, 2023; Zhang *et al.*, 2022], $O(N^2\log T/\Delta^2 + K\log T/\Delta)$ [Kong *et al.*, 2024] in terms of both market size and definition of Δ. It remains an open question to close the gap for designing the optimal algorithm.

5 Other Variants

In this section, we introduce a line of works that study the setting beyond the basic one-to-one matching markets with one-sided unknown preference, which generalize the literature of bandit learning in matching markets.

5.1 Many-to-one Matching Market

In many real-world examples like school admissions, one side participants (schools) are able to match with multiple participants on the other side (students). This motivates the following extending setting of many-to-one matching markets.

[Wang *et al.*, 2022] first study the problem of bandit learning in many-to-one matching markets, where each arm can accept multiple players rather than a single player. Specifically, each arm $k \in [K]$ has a fixed capacity c_k and it accept its most preferred top c_k players at each round. This preference structure is called the responsiveness. This setting makes the system more complicated and players can be matched with an arm even though other players also propose to it. They study the decentralized matching market and propose the MOCA-UCB algorithm, which extends the CA-UCB algorithm to the many-to-one setting [Liu *et al.*, 2021]. Similarly they obtain the $O\left(\exp(N^4)N^5K^2\log^2(T)/\Delta^2\right)$ player-pessimal stable regret.

[Zhang and Fang, 2024a] extend the work of [Zhang *et al.*, 2022] to many-to-one setting with responsiveness preference. By the structure of arms' preferences, players are divided into different levels. The choices of players with lower levels do not influence those with higher levels. Then the algorithm finds the player-optimal stable matching from the highest level to the lowest. In the same level, each player explores all arms in a round-robin way until they identify their full rankings of all arms, then they run the GS algorithm to find the stable matching. Through this algorithm, they have achieved the player-optimal stable regret bound of $O\left(K\log(T)/\Delta^2\right)$.

[Kong and Li, 2024] take an algorithm initially designed for the one-to-one setting and expand it to the more general many-to-one case. Through this expansion, they manage to reach a near-optimal bound for player-optimal stable regret. Nevertheless, single-player deviation can pose problems because of the requirements for collaboration. The primary objective in this research is to enhance the regret bound in many-to-one markets while maintaining incentive compatibility. To begin with, for the responsiveness setting, the adaptively explore-then-deferred-acceptance (AETDA) algorithm is put forward. This algorithm enables the derivation of an upper bound for player-optimal stable regret and also proves its incentive compatibility. It offers a polynomial player-optimal guarantee in matching markets without the need to have prior knowledge of Δ.

[Li *et al.*, 2024] study the many-to-one matching market with complementary preferences and quota constraints. Here arms are divided into M types, and each type m includes K_m arms. Each player has fixed but unknown preferences over each arm's type, and at each time each player has a quota constraint that each type's arm has to be matched over a certain numbers, and the total matched number is also bounded.

Real world examples include: firms seeking certain number of workers with skills that complement their existing workforce, sports teams forming teams with certain number of players that have complementary roles. A centralized Multi-agent Multi-type Thompson Sampling (MMTS) algorithm is proposed and it achieves an $O(\sqrt{T})$ Bayesian regret with high probability.

5.2 Non-stationary Matching

Since participants' preference might shift, there are some studies focusing on non-stationary rewards, where the reward distribution is not fixed and may vary over time.

The work of [Muthirayan *et al.*, 2022] specifically examines learning when players' preferences are time-varying and unknown. They assume that the number of preference changes is upper bounded by a constant L and this is known beforehand. The algorithm design idea is that the algorithm restart after each $L^{-1/2}T^{1/2}$ rounds and the algorithm runs centralized UCB at each phase. It is demonstrated that with the proposed algorithm, each player gets a uniform sub-linear regret of $O(L^{1/2}T^{1/2})(1+\Delta^{-2})$. The article also discusses the extensions of the algorithm to situations where the number of changes does not need to be known beforehand.

[Ghosh *et al.*, 2022] focus on a different non-stationary setting where at each time the reward change for an arm does not exceed a small term δ. It introduces the framework of a decentralized two-sided matching market in non-stationary (dynamic) environments under the serial dictatorship setting. The authors propose and analyze a decentralized and asynchronous learning algorithm called Decentralized Non-stationary Competing Bandits (DNCB). In this algorithm, agents use successive elimination type learning algorithms to learn their preferences over the arms. The complexity of understanding the system comes from the asynchronous action selection of competing bandits and the situation where lower ranked agents can only learn from a set of arms not dominated by higher ranked agents, resulting in "forced exploration". By carefully defining complexity parameters, the paper characterizes this "forced exploration" and achieves sub-linear (logarithmic) regret for DNCB. Additionally, the theoretical findings are validated through experiments.

5.3 Two-sided Unknown Preference

When considering the market where preferences of participants on two sides are both unknown and random, it is more general than the basic one-sided unknown setting and needs more careful technique balancing the exploration on both sides.

[Pagare and Ghosh, 2023] propose a multi-phase explore-then-commit type algorithm namely Epoch-based CA-ETC (collision avoidance explore then commit) for this problem that does not require any communication across agents (players and arms) and hence fully decentralized. They show that the for the initial epoch length of T_0 and subsequent epoch-lengths of $2^{\ell/\gamma}T_0$ (for the ℓ-th epoch with $\gamma \in (0,1)$ as an input parameter to the algorithm), CA-ETC yields a player optimal expected regret of $O(T_0(\frac{K\log T}{T_0\Delta^2})^{1/\gamma} + T_0(T/T_0)^{\gamma})$ for each player. Furthermore, we propose several other baselines for two-sided learning for matching markets.

[Zhang and Fang, 2024b] model the arm side, with a reasonable "Rational Condition", where their objective is to maximize their individual rewards. Then, on the player side, they introduce the Round-Robin ETC algorithm, incorporating various techniques to tackle challenges arising from unreliable feedback from arms and the absence of information and communication. Through rigorous analysis, we demonstrate that the optimal matching for the proposing side can be achieved with high probability Their algorithm achieves an $O(\log T)$ player-optimal stable regret.

5.4 Objectives Beyond Minimizing Regret

There is a line of works studying the objectives beyond minimizing regret. Specifically, they mainly study the sample complexity of finding a stable matching.

[Hosseini *et al.*, 2024] showcase crucial techniques within the realm of learning preferences. These techniques are centered around leveraging the structure of stable solutions. Specifically, they made use of the known preferences of arms in the arm-proposing variant of the Deferred Acceptance (DA) algorithm. By doing so, and by eliminating arms at an early stage, they were able to provably decrease the sample complexity associated with finding stable matchings. Moreover, from an experimental perspective, this approach had minimal influence on optimality, which was gauged by the metric of regret.

[Athanasopoulos *et al.*, 2025] focus on the centralized case where, at each time step, an online platform matches agents and gets a noisy evaluation of their preferences. They introduce the concept of a probably correct optimal stable matching, a special type of probably approximately correct (PAC) solutions requiring the output matching to be optimal with high probability. First, they analyze an algorithm that uniformly samples all available agent pairs, similar to the ETC strategy, demonstrating its ability to produce the optimal stable matching with high probability and providing a bound on its sample complexity. Next, they explore an action elimination-based algorithm, which improves sample efficiency and reduces dependence on instance - specific parameters. Additionally, they enhance sample complexity by modifying the stopping criterion, enabling the algorithm to terminate when enough information is gathered.

6 Future Direction

6.1 Optimal Analysis for Matching Markets

There still exists a gap between the existing best upper bound $O(N^2\log T/\Delta^2 + K\log T/\Delta)$ when $N^2 \leq K$ [Kong *et al.*, 2024], $O(K\log T/\Delta^2)$ when $N^2 > K$ [Kong and Li, 2023; Zhang *et al.*, 2022], and the problem lower bound $\Omega(\max\{N\log T/\Delta^2, K\log T/\Delta\})$ [Sankararaman *et al.*, 2021]. Thus a natural remaining open question is to design an optimal algorithm achieves the best regret bound.

A possible approach for designing the optimal algorithm is to refine the procedure of the adaptive Gale-Shapley algorithm proposed by [Kong *et al.*, 2024]. The key bottleneck to attain the $O(N\log T/\Delta^2)$ regret is because at each

step of GS, the number of collisions is only bounded by $O(N \log T/\Delta^2)$, and this multiplies the maximum number of steps N contributes to the final regret bound. To improve the final regret bound, it is likely to modify the algorithm to be more adaptive for reducing the bound of collision times at each step.

6.2 Incentive Compatibility Analysis

Incentive compatibility is a key yet under-explored aspect in matching markets with bandit learning. Agents in these markets often have private information and self-interested motives. To ensure bandit-based matching mechanisms are incentive-compatible, agents must be motivated to truthfully disclose their preferences and characteristics.

Most existing studies on bandit learning in matching markets prioritize matching efficiency over agents' strategic behavior [Kong and Li, 2024]. Future research could delve into how different bandit algorithms interact with incentive compatibility. This could involve analyzing agent incentives in various matching market types using game-theoretic models. Additionally, there's a need to design mechanisms that are both incentive compatible and computationally efficient, perhaps by exploring approximate incentive compatibility concepts. Empirical studies, whether in a lab or using real-world data, are also essential to validate theoretical results and refine these mechanisms for practical use.

6.3 Robustness Analysis Against Attack

In the context of bandit learning in matching markets, robustness against attack has emerged as a critical area that demands in-depth exploration in future research. With the increasing digitization and reliance on automated matching systems, these markets become vulnerable to various malicious attacks that can disrupt the normal operation of bandit-based algorithms and undermine the fairness and efficiency of the matching process.

One of the primary future research directions is to comprehensively identify and categorize the potential attack models. For example, in a matching market for online auctions, malicious bidders could launch shill-bidding attacks. They create fake identities to inflate the prices artificially, tricking the bandit-learning algorithm into making sub-optimal matching decisions. In a ride-sharing matching market, attackers might manipulate location data to disrupt the efficient pairing of drivers and passengers. By precisely defining these attack models, researchers can then develop targeted defense mechanisms.

6.4 Other Possible Settings

There are several other possible settings in bandit learning for matching markets that merit future research attention.

Dynamically Changing Market Structures. Most current studies assume relatively static market structures in the context of bandit learning for matching. However, in real-world scenarios, markets are often dynamic. For example, new participants may continuously enter the market, and existing ones may leave. In a job-matching market, new companies may be established, creating new job openings, while some existing firms might downsize or go out of business. Future research could explore how bandit learning algorithms can adapt to such dynamic market structures. This may involve developing algorithms that can quickly re-evaluate and adjust the matching strategy as the market composition changes. One approach could be to use time-series analysis techniques to predict the inflow and outflow of market participants and then incorporate these predictions into the bandit-learning framework.

Contextual Matching Markets. Contextual bandit for online matching market holds great significance as it enables more accurate and efficient matching in various real-world scenarios, from e-commerce platforms connecting buyers and sellers to ride-sharing services pairing passengers with drivers, where the preference of a participant is related to its contextual information. Specifically, each arm a_k is associated with a fixed context $X_k \in \mathbb{R}^d$, where d is the number of dimension. The preference of player p_i over arm a_k is determined by a reward function $R_i(X_k)$, such as the linear model with $R_i(X_k) = \theta_i^\top X_k$. The learner aims to learn the reward model of each player through multiple matchings. This model can well describe how the player's preference is determined by various attributes of each arm.

7 Conclusion

The study of bandit learning in two-sided matching markets has emerged as a rapidly evolving research area. This survey comprehensively reviews and systematically organizes the abundant literature on bandit learning in matching markets. These works have shown great promise in improving the efficiency of matching, whether it is in job-seeker-employer match-ups, ride-sharing driver-passenger pairings, or online dating platform matches. This survey covers the existing theoretical achievements and various related aspects.

There are still numerous avenues for future research. The optimal algorithm design is not sufficiently explored and remains open. Incentive compatibility has also been recognized as a crucial factor, ensuring that market participants are motivated to truthfully reveal their preferences is essential for the proper functioning of bandit-based matching mechanisms. Robustness analysis against attacks, is becoming increasingly important in the face of potential malicious disruptions in digital matching markets. Additionally, the study of bandit learning in dynamically changing market structures, and contextual matching market structure, also hold great potential for advancing the field. Overall, the future of bandit learning in matching markets is bright, with the potential to bring about more efficient and robust matching systems across a wide range of industries.

Acknowledgements

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 the Guangdong Basic and Applied Basic Research Foundation 2025A1515011412.

References

[Athanasopoulos *et al.*, 2025] Andreas Athanasopoulos, Anne-Marie George, and Christos Dimitrakakis. Probably correct optimal stable matching for two-sided markets under uncertainty. *arXiv preprint arXiv:2501.03018*, 2025.

[Basu *et al.*, 2021] Soumya Basu, Karthik Abinav Sankararaman, and Abishek Sankararaman. Beyond $\log^2$ (t) regret for decentralized bandits in matching markets. In *International Conference on Machine Learning*, pages 705–715. PMLR, 2021.

[Das and Kamenica, 2005] Sanmay Das and Emir Kamenica. Two-sided bandits and the dating market. In *Proceedings of the 19th international joint conference on Artificial intelligence*, pages 947–952, 2005.

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

[Ghosh *et al.*, 2022] Avishek Ghosh, Abishek Sankararaman, Kannan Ramchandran, Tara Javidi, and Arya Mazumdar. Decentralized competing bandits in non-stationary matching markets. *arXiv preprint arXiv:2206.00120*, 2022.

[Hosseini *et al.*, 2024] Hadi Hosseini, Sanjukta Roy, and Duohan Zhang. Putting gale & shapley to work: Guaranteeing stability through learning. *arXiv preprint arXiv:2410.04376*, 2024.

[Kong and Li, 2023] Fang Kong and Shuai Li. Player-optimal stable regret for bandit learning in matching markets. In *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1512–1522. SIAM, 2023.

[Kong and Li, 2024] Fang Kong and Shuai Li. Improved bandits in many-to-one matching markets with incentive compatibility. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38:13256–13264, 03 2024.

[Kong *et al.*, 2022] Fang Kong, Junming Yin, and Shuai Li. Thompson sampling for bandit learning in matching markets. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 3164–3170. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Main Track.

[Kong *et al.*, 2024] Fang Kong, Zilong Wang, and Shuai Li. Improved analysis for bandit learning in matching markets. *Advances in Neural Information Processing Systems*, 2024.

[Li *et al.*, 2024] Yuantong Li, Guang Cheng, and Xiaowu Dai. Two-sided competing matching recommendation markets with quota and complementary preferences constraints. In *Forty-first International Conference on Machine Learning*, 2024.

[Liu *et al.*, 2020] Lydia T Liu, Horia Mania, and Michael Jordan. Competing bandits in matching markets. In *International Conference on Artificial Intelligence and Statistics*, pages 1618–1628. PMLR, 2020.

[Liu *et al.*, 2021] Lydia T Liu, Feng Ruan, Horia Mania, and Michael I Jordan. Bandit learning in decentralized matching markets. *J. Mach. Learn. Res.*, 22:211–1, 2021.

[Maheshwari *et al.*, 2022] Chinmay Maheshwari, Shankar Sastry, and Eric Mazumdar. Decentralized, communication-and coordination-free learning in structured matching markets. In *Advances in Neural Information Processing Systems*, 2022.

[Muthirayan *et al.*, 2022] Deepan Muthirayan, Chinmay Maheshwari, Pramod P Khargonekar, and Shankar Sastry. Competing bandits in time varying matching markets. *arXiv preprint arXiv:2210.11692*, 2022.

[Pagare and Ghosh, 2023] Tejas Pagare and Avishek Ghosh. Two-sided bandit learning in fully-decentralized matching markets. In *ICML 2023 Workshop The Many Facets of Preference-Based Learning*, 2023.

[Roth and Sotomayor, 1992] Alvin E Roth and Marilda Sotomayor. Two-sided matching. *Handbook of game theory with economic applications*, 1:485–541, 1992.

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

[Roth, 2002] Alvin E Roth. The economist as engineer: Game theory, experimentation, and computation as tools for design economics. *Econometrica*, 70(4):1341–1378, 2002.

[Sankararaman *et al.*, 2021] Abishek Sankararaman, Soumya Basu, and Karthik Abinav Sankararaman. Dominate or delete: Decentralized competing bandits in serial dictatorship. In *International Conference on Artificial Intelligence and Statistics*, pages 1252–1260. PMLR, 2021.

[Wang and Li, 2024] Zilong Wang and Shuai Li. Optimal analysis for bandit learning in matching markets with serial dictatorship. *Theoretical Computer Science*, 1010:114703, 2024.

[Wang *et al.*, 2022] Zilong Wang, Liya Guo, Junming Yin, and Shuai Li. Bandit learning in many-to-one matching markets. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 2088–2097, 2022.

[Zhang and Fang, 2024a] Yirui Zhang and Zhixuan Fang. Decentralized competing bandits in many-to-one matching markets. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pages 2603–2605, 2024.

[Zhang and Fang, 2024b] Yirui Zhang and Zhixuan Fang. Decentralized two-sided bandit learning in matching market. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, UAI '24. JMLR.org, 2024.

[Zhang *et al.*, 2022] Yirui Zhang, Siwei Wang, and Zhixuan Fang. Matching in multi-arm bandit with collision. In *Advances in Neural Information Processing Systems*, 2022.