Thompson Sampling for Bandit Learning in Matching Markets

Fang Kong[1], **Junming Yin**[2] and **Shuai Li**[1*]
[1]John Hopcroft Center for Computer Science, Shanghai Jiao Tong University
[2]Tepper School of Business, Carnegie Mellon University
{fangkong, shuaili8}@sjtu.edu.cn, junmingy@cmu.edu

Abstract

The problem of two-sided matching markets has a wide range of real-world applications and has been extensively studied in the literature. A line of recent works have focused on the problem setting where the preferences of one-side market participants are unknown *a priori* and are learned by iteratively interacting with the other side of participants. All these works are based on explore-then-commit (ETC) and upper confidence bound (UCB) algorithms, two common strategies in multi-armed bandits (MAB). Thompson sampling (TS) is another popular approach, which attracts lots of attention due to its easier implementation and better empirical performances. In many problems, even when UCB and ETC-type algorithms have already been analyzed, researchers are still trying to study TS for its benefits. However, the convergence analysis of TS is much more challenging and remains open in many problem settings. In this paper, we provide the first regret analysis for TS in the new setting of iterative matching markets. Extensive experiments demonstrate the practical advantages of the TS-type algorithm over the ETC and UCB-type baselines.

1 Introduction

The model of matching markets has been studied for several decades [Gale and Shapley, 1962; Roth and Xing, 1997; Haeringer and Wooders, 2011]. It has a wide range of applications including labor employment [Roth, 1984], house allocation [Abdulkadiroğlu and Sönmez, 1999], and college admission [Epple *et al.*, 2006; Fu, 2014]. Typically there are two sides of players, such as employers and workers in the labor market, and each player of one side has a preference ranking over players on the other side. Stability is a key concept in matching markets, which ensures market participants have no incentive to abandon the current partner [Gale and Shapley, 1962; Roth and Sotomayor, 1992]. Many researchers study how to find a stable matching in the markets [Gale and Shapley, 1962; Roth and Sotomayor, 1992]. However, most of them assume the full preferences of players are known *a priori*, which is not realistic in many real-world applications. For example, the demand-side players in online matching markets (such as employers in UpWork or Mechanical Turk) are likely to be uncertain about the qualities of supply-side players (such as workers). Many such online platforms usually have matching processes happen iteratively, which allows the participating players to adaptively learn their unknown preferences [Liu *et al.*, 2020; Liu *et al.*, 2021].

Multi-armed bandits (MAB) is a common approach to modeling this type of learning process [Auer *et al.*, 2002; Lattimore and Szepesvári, 2020]. The most basic framework considers a single player and K arms, where the player does not have prior knowledge over arms and will learn it through iteratively collected rewards. The objective of the player is to maximize the cumulative reward over a specified horizon, or equivalently minimize the cumulative regret, which is defined as the difference between the cumulative reward of the optimal arm and that of the chosen arms. To achieve this goal, the player needs to make a trade-off between exploration and exploitation: the former tries arms that have not been observed enough times to get potential higher rewards, and the latter focuses on arms with the highest observed rewards so far to maintain high profits.

There are many types of strategies to balance such trade-offs with theoretical guarantees, among which explore-then-commit (ETC), upper confidence bound (UCB), and Thompson sampling (TS) are widely adopted in the literature [Lattimore and Szepesvári, 2020]. The TS algorithm was introduced in the 1930s [Thompson, 1933] but has not been theoretically proven until the 2010s [Kaufmann *et al.*, 2012; Agrawal and Goyal, 2012; Agrawal and Goyal, 2013]. Algorithms of this type form a competitive family in the bandit area because of many advantages such as easier implementations and better practical performances. Though ETC and UCB-type algorithms have solved many bandit problems, there are still accompanying works trying to analyze TS [Cheung *et al.*, 2019; Kong *et al.*, 2021].

Das and Kamenica [2005] first introduce the bandit learning problem in two-sided matching markets and show the empirical performances of algorithms via simulation tests. Liu *et al.* [2020] later study a refinement of the problem and give the first algorithm with theoretical guarantees. They propose

[*]Corresponding author. The full version is available at http://arxiv.org/abs/2204.12048.

both ETC and UCB-type algorithms with upper bounds on the *stable regret*, which is defined as the difference between the cumulative reward of a stable matching[1] and the cumulative reward collected. Both algorithms adopt a central platform to collect players' preferences and assign allocations to participants. Since such platforms do not always exist in real-world applications, the following works extend to the decentralized setting [Liu *et al.*, 2021; Sankararaman *et al.*, 2021; Basu *et al.*, 2021]. All these works still focus on ETC and UCB-type algorithms and do not consider TS.

In a standard TS analysis, one needs to bound the inaccurate estimations for the optimal arms due to the randomness caused by the posterior samples, which is easily controlled because eventually the optimal arms can be observed enough times. However, this property may not hold anymore in matching markets since the observations are not simply decided by the number of selections and can be blocked if the selected arm rejects this player. Such a difficulty does not exist in UCB-type algorithms since it mainly needs to bound inaccuracies for sub-optimal arms, and ETC-type algorithms can control the blockings easily in the design.

In this work, we present the first TS-type algorithm for decentralized matching markets. For the stable regret analysis, though the observations might be blocked, we prove that its success probabilities could be bounded by separately analyzing the influence of other players in a fine division of events with a different number of selections on the optimal stable arms. The result guarantees the regret of TS with an order of $O(\log^2 T/\Delta^2)$, where T is the horizon and Δ is the minimum preference gap. A series of experiments are conducted to show that the practical performances of the TS-type algorithm are better than other baselines.

Related work. The MAB problem has been studied for many decades, which captures the learning process of a single player in an unknown environment. The study on stochastic MAB is an important line of works where the reward of arms are drawn from fixed distributions. ϵ-greedy, ETC, UCB and TS are all classical algorithms for this setting and are widely followed in the literature [Lattimore and Szepesvári, 2020].

Motivated by real applications including cognitive radio where multiple players compete and cooperate in the unknown environment, the problem extends to multi-player MAB. These works [Liu and Zhao, 2010; Rosenski *et al.*, 2016; Bistritz and Leshem, 2018] mainly consider the case where a player receives no reward if it conflicts with others by selecting the same arm and measure algorithms performances by the collective cumulative regret of all players.

The model of combinatorial dueling bandits [Chen *et al.*, 2020] can also be regarded as a multi-player problem, where the preferences of players are defined on pairs of arms. They study a pure exploration problem for this model with the objective to find the best matching after a specified period.

The two-sided matching market problem is different from previous multi-player MAB by considering arms' preferences [Gale and Shapley, 1962; Roth and Xing, 1997]. In this problem, not only do players have arbitrary preferences over arms, arms also have arbitrary preferences over players. When multiple players collide at the same arm, the player preferred most by this arm will receive the corresponding reward while others receive no feedback.

The bandit learning problem in two-sided matching markets was first introduced by Das and Kamenica [2005]. They assume the preferences of both sides are unknown and study the empirical performances of ϵ-greedy in a special market where all participants on one side share the same preferences. Later, Liu *et al.* [2020] study a variant of the problem by considering one-side unknown preferences. They propose both ETC and UCB-type algorithms and show the convergence analysis of the stable regret. Both algorithms adopt Gale-Shapley (GS) [Gale and Shapley, 1962] to assign allocations for players. However, in real applications, players usually prefer to independently make decisions. Liu *et al.* [2020] then propose Decentralized ETC in the decentralized setting, which lets players to explore arms for a fixed number of rounds. However, the exploration budget H needs to depend on the preference gap, which is usually not known beforehand. Sankararaman *et al.* [2021] and Basu *et al.* [2021] successively study the decentralized setting to remove this assumption but under special assumptions to guarantee a unique stable matching. For general markets, Basu *et al.* [2021] propose the phasedETC algorithm and Liu *et al.* [2021] propose a UCB-type algorithm to avoid conflicts among players and minimize the stable regret. Dai and Jordan [2021] also study to learn players' unknown preferences in a decentralized matching market. However, their objective is to estimate the unknown preferences using a statistical model but involves no cumulative regret. To the best of our knowledge, we are the first to study the TS-type algorithm for two-sided matching markets.

2 Setting

There are N players and K arms in the market. The player set is denoted by $\mathcal{N} = \{p_1, \dots, p_N\}$ and the arm set is denoted by $\mathcal{K} = \{a_1, \dots, a_K\}$. To ensure each player can be matched with an arm, we assume $N \le K$ as [Liu *et al.*, 2020; Liu *et al.*, 2021; Basu *et al.*, 2021; Sankararaman *et al.*, 2021].

For each player p_i, its preference for arm a_j is quantified by an unknown value $\mu_{i,j} \in [0, 1]$. For two different arms a_j and $a_{j'}$, $\mu_{i,j} > \mu_{i,j'}$ implies that player p_i *truly* prefers arm a_j to $a_{j'}$. Similarly, each arm a_j has preferences over players. Let $\pi_{j,i}$ to represent the preference value of arm a_j for player p_i. For two different players p_i and $p_{i'}$, $\pi_{j,i} > \pi_{j,i'}$ implies that arm a_j prefers player p_i to $p_{i'}$. The ranking for each arm a_j's preferences $(\pi_{j,i})_{i\in[N]}$ is assumed to be known since it usually can be estimated by some known utilities such as the payments given by employers in the labor market.

At each round $t = 1, 2, \dots$, each player $p_i \in \mathcal{N}$ attempts to pull an arm $A_i(t) \in \mathcal{K}$. Let $A(t) = (A_i(t))_{i\in[N]}$. When multiple players attempt to pull the same arm, there will be a conflict and only the player preferred most by this arm is accepted. If a player p_i wins the conflict, it will receive a random reward $X_{i,A_i(t)}(t) \in [0, 1]$ with expectation $\mu_{i,A_i(t)}$. Other players $p_{i'}$ who fail the conflict are unmatched in this round and receive $X_{i',A_{i'}(t)}(t) = 0$. Following the observa-

[1]Different stable matchings give different kinds of stable regrets.

tions in the scenario of the labor market and college admission where the arm side (e.g., workers or colleges) usually announces the successfully matched players (e.g., employers or students), we assume the successfully matched player for each arm is public at the end of the round as in previous work [Liu *et al.*, 2021].

To measure the status of the market, stable matching [Gale and Shapley, 1962] is introduced to define an equilibria. A stable matching is a one-to-one mapping from players to arms without a pair of player and arm such that they both prefer being matched with each other over the current partner. With the true preference rankings of players and arms, there may be more than one stable matching. Denote m_i as player p_i's least favorite arm among matched arms from all stable matchings. The objective is to find a stable matching between players and arms and minimize the cumulative stable regret for each player p_i [Liu *et al.*, 2020; Liu *et al.*, 2021; Basu *et al.*, 2021; Sankararaman *et al.*, 2021], which is defined as

$$R_i(T) = T\mu_{i,m_i} - \mathbb{E}\left[\sum_{t=1}^{T} X_{i,A_i(t)}(t)\right], \tag{1}$$

where the expectation is taken over the randomness in the reward payoffs and the algorithm.

3 Algorithm

In this section, we introduce our TS-type algorithm, referred to as conflict-avoiding TS (CA-TS, Algorithm 1), for two-sided matching markets. Some design ideas are motivated by Liu *et al.* [2021] and the main difference is that CA-TS uses parameters sampled from posterior distributions as estimations for preferences to select arms. However, this raises new analysis difficulties since CA-TS additionally requires accurate estimations for 'optimal arms' to ensure a stable matching, which will be made clearer later in Section 4.

The CA-TS algorithm takes the player set $\mathcal{N}$ and arm set $\mathcal{K}$ as input (line 1). For each player p_i and arm a_j, the algorithm maintains a Beta distribution $\mathrm{Beta}(a_{i,j}, b_{i,j})$ for the preference value. In the beginning, this distribution is initialized to $\mathrm{Beta}(1,1)$, the uniform distribution on $[0,1]$ (line 2). It will be later updated based on observed feedback and tend to concentrate on the real value $\mu_{i,j}$. In round t, the algorithm samples an index $\theta_{i,j}(t)$ from $\mathrm{Beta}(a_{i,j}, b_{i,j})$ to represent the current estimation (line 5).

One may consider letting each player p_i follow independent TS strategies, namely choose $A_i(t) \in \mathrm{argmax}_{j\in[K]}\, \theta_{i,j}(t)$. A simple example shows that frequent conflicts would happen under this strategy. Suppose there are 2 players and 2 arms, and $\mu_{i,1} = \max_{j\in[K]} \mu_{i,j}$ for each p_i. Then at each round t, $\theta_{i,1}(t) = \max_{j\in[K]} \theta_{i,j}(t)$ holds for each p_i with at least constant probability. The reason is that if enough observations are collected, then the sampled indices approach the real preference value and $\theta_{i,1}(t)$ would be the largest; otherwise, the Beta distributions tend to be uniform and $\theta_{i,1}(t)$ would still be the largest with constant probability. Thus both players attempt to pull a_1 with constant probability at each round. However, a_1 only accepts one player and the other one will be rejected and receive no feedback. The stable regret of the latter player is thus of order $O(T)$.

Algorithm 1 Conflict-Avoiding TS (CA-TS)

1: Input: Player set $\mathcal{N}$, arm set $\mathcal{K}$, parameter $\lambda \in (0,1)$;
2: Initialize: $\forall i \in [N], j \in [K], a_{i,j} = b_{i,j} = 1$
3: **for** $t = 1, 2, \cdots$ **do**
4:   **for** $p_i \in \mathcal{N}$ **do**
5:     $\forall a_j$, sample $\theta_{i,j}(t) \sim \mathrm{Beta}(a_{i,j}, b_{i,j})$
6:     Independently draw $D_i(t) \sim \mathrm{Bernoulli}(\lambda)$
7:     **if** $D_i(t) == 0$ **then**
8:       Construct plausible set

$$S_i(t) := \left\{ j : \pi_{j,i} \geq \pi_{j,i'} \text{ where } \bar{A}_{i'}(t-1) = j \right\}$$

9:       Pull $A_i(t) \in \mathrm{argmax}_{j\in S_i(t)}\, \theta_{i,j}(t)$
10:     **else**
11:       Pull $A_i(t) = A_i(t-1)$
12:     **end if**
13:     **if** p_i wins conflict **then**
14:       $\bar{A}_i(t) = A_i(t)$
15:       $Y_i(t) \sim \mathrm{Bernoulli}\left(X_{i,A_i(t)}(t)\right)$
16:       Update $a_{i,A_i(t)} = a_{i,A_i(t)} + Y_i(t)$
        $b_{i,A_i(t)} = b_{i,A_i(t)} + (1 - Y_i(t))$
17:     **else**
18:       $\bar{A}_i(t) = -1$
19:     **end if**
20:   **end for**
21: **end for**

To avoid frequent conflicts in the above case, we construct a plausible set for each player to exclude arms that may reject it (line 8). Recall that the successfully matched players and the preference rankings of arms are known. If arm a_j accepts a player $p_{i'}$ but prefers p_i more, p_i will not be rejected by a_j when the same set of players attempt to pull a_j. Following this observation, the plausible set for p_i at time t is constructed to contain arms who accept a player less preferred than p_i at $t-1$ (line 8). Player p_i then selects the arm with the highest index in the plausible set (line 9). Here we use $\bar{A}_i(t)$ to represent the arm successfully pulled by p_i at t. If p_i fails the conflict, then $\bar{A}_i(t) = -1$.

However, players can still simultaneously pull same arms. Consider an example with 2 players and 2 arms. Both players have $\mu_{i,1} = \mathrm{argmax}_{j\in[K]}\, \mu_{i,j}$, arm a_1 prefers p_1 to p_2 and arm a_2 prefers p_2 to p_1. It is possible that both players attempt to pull a_1 at $t = 1$. Then p_2 is rejected and its plausible set only contains a_2 at $t = 2$. For player p_1, its plausible set is still $\{a_1, a_2\}$, and its Beta distribution for arm a_2 is still uniform. Thus with constant probability, $\theta_{1,2}(t) > \theta_{1,1}(t)$ and both players attempt to pull arm a_2 at time $t = 2$. In this case, player p_1 will then be rejected. The same analysis shows that both players can always pull the same arm and be rejected in turn. Thus the stable regrets of these two are both of the order $O(T)$.

A random delay mechanism is further introduced to keep the effectiveness of the plausible set and avoid such rejections. The intuition is that if all players except p_i follow the last-round choice, then p_i will not be rejected by selecting arms in the plausible set. To be specific, each player first draws a Bernoulli random variable $D_i(t)$ with expectation λ (line 6), which is a hyper-parameter. If $D_i(t) = 0$, p_i still attempts the arm with the largest index in the plausible set (line 7-9); otherwise, it follows the last-round choice (line 11).

When all players decide which arm to pull in this round, arms will determine which player to accept according to their rankings. If player p_i wins the conflict and successfully pulls arm $A_i(t)$, the algorithm marks $\bar{A}_i(t)$ as $A_i(t)$ and updates the corresponding Beta distribution (line 13-19).

4 Theoretical Results

In this section, we provide the theoretical result of CA-TS. The corresponding gaps are defined to measure the performance of the algorithm.

Definition 1. *(Gaps) For each player $p_i \in \mathcal{N}$, denote $\Delta_{i,\max} = \mu_{i,m_i}$ as the maximum stable regret that player p_i needs to pay in unstable matchings. For any pair of arms a_j and $a_{j'}$, define*

$$\Delta_{i,j,j'} = |\mu_{i,j} - \mu_{i,j'}| \tag{2}$$

as the reward difference between arm a_j and $a_{j'}$ for player p_i. Let $\Delta = \min_{i,j,j':\mu_{i,j}\neq\mu_{i,j'}} |\mu_{i,j} - \mu_{i,j'}|$.

Theorem 1. *Let $\rho = \lambda^{N-1}(1-\lambda)$. Following Algorithm 1, the stable regret of each player p_i satisfies*

$$R_i(T) \leq \left\{ \frac{2N^5K^2\log T}{\rho^{N^4}} \left(4 + \frac{6\log T}{\rho(\Delta-2\varepsilon)^2} + \frac{C}{\rho\varepsilon^6}\right) + 6 + \frac{6N^4}{\rho^{N^4}} \right\} \cdot \Delta_{i,\max} \tag{3}$$

$$= O\left(\frac{N^5K^2\log^2 T}{\rho^{N^4}\Delta^2} \cdot \Delta_{i,\max}\right), \tag{4}$$

for any ε such that $\Delta - 2\varepsilon > 0$, where C is a universal constant.

Due to the space limit, we provide the proof sketch as well as the full proof in the Appendix. In addition to Beta priors shown in Algorithm 1, we also analyze the CA-TS algorithm with Gaussian priors for 1-subgaussian rewards, which achieves the same order of regret upper bound as Theorem 1. For completeness, we provide the full algorithm and analysis in the Appendix.

4.1 Discussions

Hardness. The TS-type algorithm faces new challenge for analysis in the setting of matching markets. Note in this setting, once a player wrongly estimate arms, both this player and others in the market could suffer non-negative stable regret. In the following, we take a player p_i and two arms $a_j, a_{j'}$ with $\mu_{i,j'} > \mu_{i,j}$ for further analysis. Specifically, we need to bound the number of times when p_i incorrectly attempts a_j instead of $a_{j'}$. If the attempt is due to inaccurate estimations of $\mu_{i,j}$, then it can be easily guaranteed since with more selections, $\mu_{i,j}$ would be finally observed enough times and estimated well. However, if $\mu_{i,j}$ has already been estimated accurately, the analysis gets more complicated.

For this important part, we first investigate what properties of $\theta_{i,j'}$ are implied by the event. We show that the sample $\theta_{i,j'}$ must be inaccurate and once an accurate sample of it is drawn, p_i will select $a_{j'}$. In standard analysis of MAB, such cases can be guaranteed since when $\theta_{i,j'}$ is not accurate, the variance of the posterior distribution would push it to generate a larger sample in some round and the observations can thus be obtained. But in matching markets, even when an accurate sample for $\mu_{i,j'}$ is drawn, the observation may still be unavailable. To deal with this case, at each time with a good sample of $\mu_{i,j'}$, we analyze the influence of all other players and compute the exact probability of obtaining an observation. Based on this probability, we construct a new counter to estimate the number of observations on $\mu_{i,j'}$. The total horizon can then have a fine division of slices based on this counter and within each slice, the posterior distribution of $\mu_{i,j'}$ has some common properties and we are able to bound the number of rounds to wait for a good event of being matched with arm $a_{j'}$.

Note this difficulty does not need to be dealt with by UCB and ETC. In UCB, due to the monotonicity between UCB index and real parameter, inaccurate estimations of $\mu_{i,j'}$ would not contribute to the regret, while ETC forces players to collect enough observations on every arm without considering other participants' influence.

Regret bound. Our CA-TS is the first TS-type algorithm for general decentralized matching markets. Two comparable algorithms in the same setting with regret guarantees are CA-UCB [Liu *et al.*, 2021] and PhasedETC [Basu *et al.*, 2021]. The former has the same main order of regret upper bound as ours. The latter, though its upper bound has better dependence on T (of order $O(\log^{1+\varepsilon} T + 2^{(1/\Delta^2)^{1/\varepsilon}})$, $\varepsilon > 0$ is a hyper-parameter), suffers from the problem of cold-start and only works for a huge horizon $T = \Omega\left(\exp\left(N/\Delta^2\right)\right)$. Compared with CA-UCB, our regret upper bound has an additional constant term $1/\varepsilon^6$. This term comes from the special nature of TS as discussed above and also appears in many other TS-based works [Agrawal and Goyal, 2013; Wang and Chen, 2018; Perrault *et al.*, 2020; Kong *et al.*, 2021]. As stated in the theorem, if we let ε take value of $\Delta/3$, then this term would become $729/\Delta^6$, which is a constant relative to $\log T$. It is not known whether such a term is unavoidable and we would leave this as future work. It is worth noting that the additional constant term does not imply bad practical performances. As shown in following Section 5, our TS-type algorithm performs better than these baselines.

Player-pessimal stable regret. In this paper, we study the player-pessimal stable regret which is defined with respect to players' least favorite stable matching. Except for basic ETC, the analyses of all existing works also focus on this type of stable regret [Liu *et al.*, 2020; Liu *et al.*, 2021; Sankararaman *et al.*, 2021; Basu *et al.*, 2021]. In the market shown in [Liu *et al.*, 2020, Example 6], both UCB and TS cannot achieve sublinear player-optimal stable regret, compared with players' favorite stable matching. This is because once a player mistakenly ranks two arms, other players' behavior can force it to have no chance for more observations and learn a correct preference ranking for optimal stable

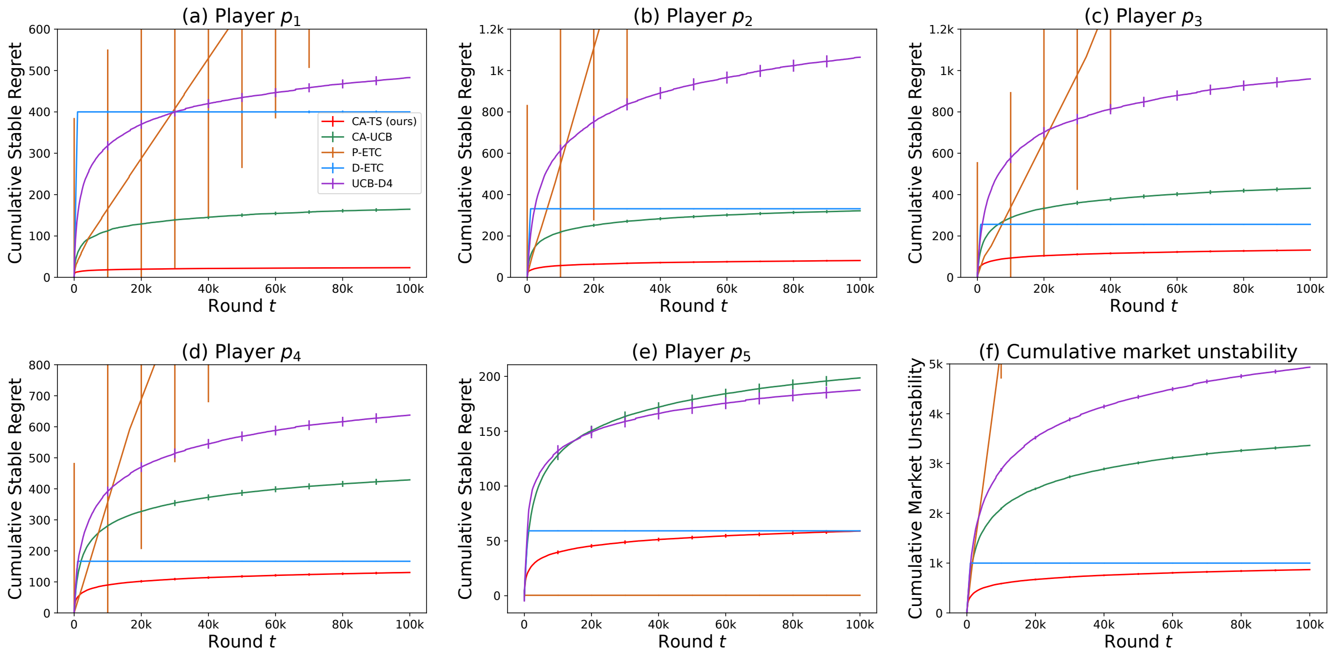

Figure 1: Experimental comparisons of our CA-TS with CA-UCB, P-ETC, D-ETC, and UCB-D4 in the market of $N = 5$ players and $K = 5$ arms with global preferences.

matching. How to minimize the player-optimal stable regret in general matching markets is still an open problem.

5 Experiments

In this section, we compare the performances of our CA-TS with other related baselines in different environments[2] where all players and arms share the same preferences (Section 5.1), the minimum reward gap Δ is varied (Section 5.2), and the market size is varied (Please see the full version for experiments in this setting). The baselines include CA-UCB [Liu *et al.*, 2021], PhasedETC (P-ETC) [Basu *et al.*, 2021], Decentralized ETC (D-ETC) [Liu *et al.*, 2020] and UCB-D4 [Basu *et al.*, 2021]. Since UCB-D4 requires the market to satisfy uniqueness consistency, we only test it in global-preference case where only unique stable matching exists (Section 5.1). The hyper-parameters of all baselines are set as their original paper with details in the Appendix.

We first report the cumulative stable regrets for all experiments. Also since the goal is to learn stable matchings and is not well reflected from cumulative regrets, we also report the cumulative market unstability, which is defined as the number of unstable matchings over T rounds, as in Liu *et al.* [2021]. In each market, we run all algorithms for $T = 100k$ rounds and all results are averaged over 50 independent runs. The error bars correspond to standard errors, which are computed as standard deviations divided by $\sqrt{50}$.

[2]The code is available at https://github.com/fangkongx/TSforMatchingMarkets.

5.1 Global Preferences

In this section, we construct a market of $N = 5$ players and $K = 5$ arms with global preferences, where all players share the same preference over arms and all arms share the same preference over players. Specifically, we set $\mu_{i,1} > \mu_{i,2} > \mu_{i,3} > \mu_{i,4} > \mu_{i,5}$ for each player p_i and $\pi_{j,1} > \pi_{j,2} > \pi_{j,3} > \pi_{j,4} > \pi_{j,5}$ for each arm a_j. The preference towards the least favorite arm is set as $\mu_{i,5} = 0.1$ for any p_i and the reward gap between any two consecutively ranked arms is set as $\Delta = 0.2$. In this market, the unique stable matching is $\{(p_1, a_1), (p_2, a_2), (p_3, a_3), (p_4, a_4), (p_5, a_5)\}$, thus the assumption of uniqueness consistency required by UCB-D4 is satisfied. The feedback $X_{i,j}(t)$ for player p_i on arm a_j at time t is drawn independently from Bernoulli$(\mu_{i,j})$.

We report the cumulative stable regret of each player in Figure 1 (a-e) and the cumulative market unstability for each algorithm in Figure 1 (f).

Our CA-TS pays the least stable regret on player p_1, p_2, p_3, and p_4 among all algorithms, while only pays higher regret than P-ETC on player p_5. However, under P-ETC, both the market unstability and the stable regrets of other players still do not converge, which phenomenon also coincides with the theoretical result that P-ETC suffers from cold-start and only works when T is huge [Basu *et al.*, 2021]. The reason for lower stable regret of p_5 is that P-ETC avoids rejections by following allocations of the GS algorithm and p_5 would suffer zero regret in this process since all other arms are better than its partner a_5 in the stable matching. For other algorithms, the stable regrets are not consistent among different players since they explore differently to find a stable matching. So we look into market unstability for further analysis. As shown

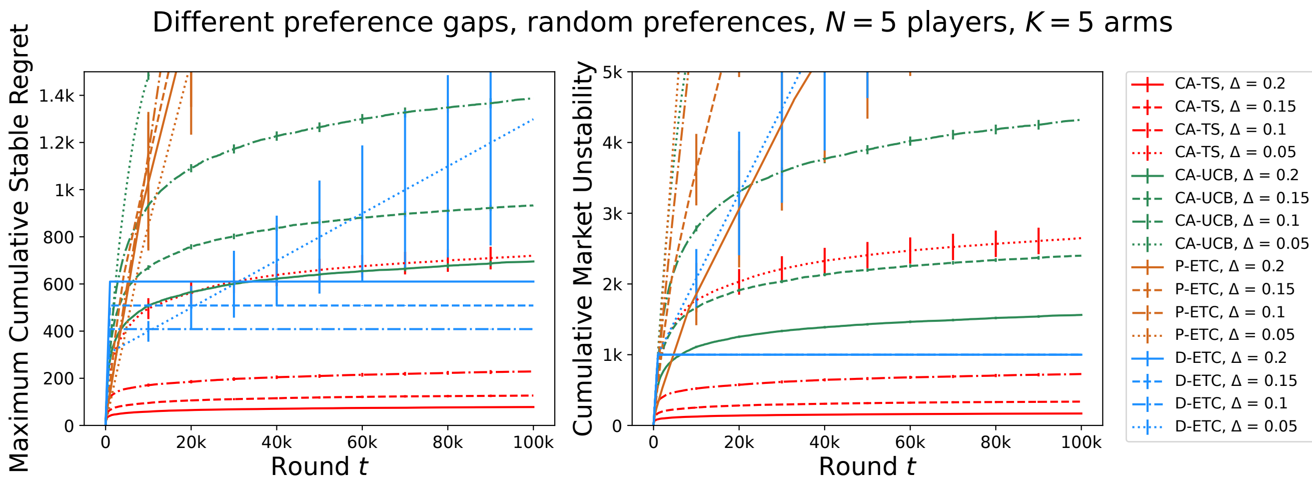

Figure 2: Experimental comparisons of our CA-TS with CA-UCB, P-ETC, and D-ETC in terms of maximum cumulative stable regret among players (left) and cumulative market unstability (right). Markets of $N = 5$ players and $K = 5$ arms with different Δ are tested.

in Figure 1 (f), our CA-TS shows the least market unstability. By carefully choosing an appropriate hyper-parameter H representing the exploration budget, D-ETC performs slightly worse than ours. CA-UCB and UCB-D4 converge much slower and explore more to find a stable matching.

5.2 Varying Gaps for Random Preferences

In this experiment, we test the effect of the minimum reward gap Δ on the stable regret and market unstability. The market size is fixed with $N = 5$ players and $K = 5$ arms. The preference rankings of all players (arms) are generated as random permutations of arms (players, respectively). We set the preference value towards the least favorite arm as 0.1 for all players and test four different choices of reward gap between any two consecutively ranked arms $\Delta \in \{0.2, 0.15, 0.1, 0.05\}$.

Figure 2 shows the maximum cumulative stable regret among all players and the cumulative market unstability for each algorithm under different values of Δ.

Our CA-TS pays the least stable regret among algorithms under each value of Δ and performs the most robust when Δ is changed. The D-ETC performs second to ours when $\Delta \in \{0.2, 0.15, 0.1\}$. However, its regret does not converge when $\Delta = 0.05$, which illustrates its high dependence on the selection of H and confirms the theoretical result that H needs to depend on Δ to guarantee good performances [Liu *et al.*, 2020]. The CA-UCB algorithm pays higher stable regret than our CA-TS and its performances also change more drastically with different Δ. The P-ETC algorithm still suffers from cold-start and does not converge in given periods.

Similar observations can also be found from the perspective of the market unstability. As shown in Figure 2 (right), the cumulative market unstability of algorithms gets higher when Δ is small. This is as expected since algorithms need to explore more to get accurate estimations on preference rankings and thus find the true stable matching.

One may also concern that the performances of algorithms could be worse when the preferences of all players tend to be the same since players may always attempt same arms thus leading to more conflicts. In the Appendix, we test the performances of algorithms in markets with different heterogeneity of players' preferences and find our CA-TS still shows the best and the most robust performances.

6 Conclusion

In this paper, we show the first TS-type algorithm, CA-TS, for the two-sided decentralized matching markets. Compared with previous UCB and ETC-type algorithms, the special nature of TS and the setting of matching markets bring additional analysis difficulties, since TS requires enough observations on arms in stable matchings but market participants may force a player to observe no feedback. We overcome the difficulty by bounding the success probability of observing these arms with a fine division of the horizon based on the number of previous attempts and the analysis of other players' influence. Our stable regret upper bound achieves the same order as previous UCB-type algorithms. Extensive experiments on markets with different properties are conducted to verify the practical performances of the TS-type algorithm. Compared with all baselines, our CA-TS learns stable matchings much faster and shows more robust performances.

We study a decentralized setting where players independently choose arms. An interesting future direction is to design TS-type algorithms for scenarios with more communication such as the centralized setting in Liu *et al.*[2020] where players can communicate with a central platform. New design ideas are needed as the requirement of TS is more restrictive than UCB. An example in the Appendix shows that if we only replace the UCB index in Liu *et al.* [2020] with the sampled parameters from posterior distributions as TS, players can suffer a stable regret of order $O(T)$. This example further illustrates the difference between TS and UCB and the challenge of TS analysis in matching markets.

Acknowledgements

The corresponding author Shuai Li is supported by National Natural Science Foundation of China (62006151, 62076161). This work is sponsored by Shanghai Sailing Program.

References

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

[Agrawal and Goyal, 2012] Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pages 39–1, 2012.

[Agrawal and Goyal, 2013] Shipra Agrawal and Navin Goyal. Further optimal regret bounds for thompson sampling. In *AISTATS*, pages 99–107, 2013.

[Auer *et al.*, 2002] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, 2002.

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

[Bistritz and Leshem, 2018] Ilai Bistritz and Amir Leshem. Distributed multi-player bandits-a game of thrones approach. In *Advances in Neural Information Processing Systems*, pages 7222–7232, 2018.

[Chen *et al.*, 2020] Wei Chen, Yihan Du, Longbo Huang, and Haoyu Zhao. Combinatorial pure exploration for dueling bandit. In *ICML*, pages 1531–1541, 2020.

[Cheung *et al.*, 2019] Wang Chi Cheung, Vincent Tan, and Zixin Zhong. A thompson sampling algorithm for cascading bandits. In *AISTATS*, pages 438–447, 2019.

[Dai and Jordan, 2021] Xiaowu Dai and Michael I Jordan. Learning strategies in decentralized matching markets under uncertain preferences. *Journal of Machine Learning Research*, 22(260):1–50, 2021.

[Das and Kamenica, 2005] Sanmay Das and Emir Kamenica. Two-sided bandits and the dating market. In *International Joint Conference on Artificial Intelligence*, pages 947–952, 2005.

[Epple *et al.*, 2006] Dennis Epple, Richard Romano, and Holger Sieg. Admission, tuition, and financial aid policies in the market for higher education. *Econometrica*, 74(4):885–928, 2006.

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

[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.

[Haeringer and Wooders, 2011] Guillaume Haeringer and Myrna Wooders. Decentralized job matching. *International Journal of Game Theory*, 40(1):1–28, 2011.

[Kaufmann *et al.*, 2012] Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *International Conference on Algorithmic Learning Theory*, pages 199–213. Springer, 2012.

[Kong *et al.*, 2021] Fang Kong, Yueran Yang, Wei Chen, and Shuai Li. The hardness analysis of thompson sampling for combinatorial semi-bandits with greedy oracle. In *NeurIPS*, volume 34, 2021.

[Lattimore and Szepesvári, 2020] Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

[Liu and Zhao, 2010] Keqin Liu and Qing Zhao. Distributed learning in multi-armed bandit with multiple players. *IEEE Transactions on Signal Processing*, 58(11):5667–5681, 2010.

[Liu *et al.*, 2020] Lydia T Liu, Horia Mania, and Michael Jordan. Competing bandits in matching markets. In *AISTATS*, pages 1618–1628, 2020.

[Liu *et al.*, 2021] Lydia T Liu, Feng Ruan, Horia Mania, and Michael I Jordan. Bandit learning in decentralized matching markets. *Journal of Machine Learning Research*, 22(211):1–34, 2021.

[Perrault *et al.*, 2020] Pierre Perrault, Etienne Boursier, Vianney Perchet, and Michal Valko. Statistical efficiency of thompson sampling for combinatorial semi-bandits. In *NeurIPS*, 2020.

[Rosenski *et al.*, 2016] Jonathan Rosenski, Ohad Shamir, and Liran Szlak. Multi-player bandits–a musical chairs approach. In *ICML*, pages 155–163. PMLR, 2016.

[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 and Xing, 1997] Alvin E Roth and Xiaolin Xing. Turnaround time and bottlenecks in market clearing: Decentralized matching in the market for clinical psychologists. *Journal of Political Economy*, 105(2):284–329, 1997.

[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.

[Sankararaman *et al.*, 2021] Abishek Sankararaman, Soumya Basu, and Karthik Abinav Sankararaman. Dominate or delete: Decentralized competing bandits in serial dictatorship. In *AISTATS*, pages 1252–1260, 2021.

[Thompson, 1933] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

[Wang and Chen, 2018] Siwei Wang and Wei Chen. Thompson sampling for combinatorial semi-bandits. In *ICML*, pages 5114–5122, 2018.