Two-Sided Bandit Learning in Fully-Decentralized Matching Markets

Tejas Pagare [1] **Avishek Ghosh** [2]

Abstract

Online learning in a decentralized two-sided matching markets, where the demand-side (players) compete to match with the supply-side (arms), has received substantial interest because it abstracts out the complex interactions in matching platforms (e.g. UpWork, TaskRabbit). However, past works (Liu et al., 2020; 2021; Sankararaman et al., 2021; Basu et al., 2021; Kong & Li) assume that the supply-side arms know their preference ranking of demand-side players (one-sided learning), and the players aim to learn the preference over arms through successive interactions. Moreover, several structural (and often impractical) assumptions on the problem are usually made for theoretical tractability. For example (Liu et al., 2020; 2021; Kong & Li) assume that when a player and an arm is matched, the information of the matched pair becomes a common knowledge to all the players whereas (Sankararaman et al., 2021; Basu et al., 2021; Ghosh et al., 2022) assume a serial dictatorship (or its variant) model where the preference rankings of the players are uniform across all arms. In this paper, we study the *first* fully decentralized two sided learning, where we do not assume that the preference ranking over players are known to the arms apriori. Furthermore, we do not have any structural assumptions on the problem. We propose a multi-phase explore-then-commit type algorithm namely Epoch-based CA-ETC (collision avoidance explore then commit) (`CA-ETC` in short) for this problem that does not require any communication across agents (players and arms) and hence fully decentralized. We show that the for the initial epoch length of T_0 and subsequent epoch-lengths of $2^{l/\gamma}T_0$ (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 $\mathcal{O}[T_0 \left(\frac{K \log T}{T_0 (\Delta^{(i)})^2}\right)^{1/\gamma} + T_0 (T/T_0)^{\gamma}]$ for the i-th player, where T is the learning horizon, K is the number of arms and $\Delta^{(i)}$ is an appropriately defined problem gap. Furthermore, we propose several other baselines for two-sided learning for matching markets.

1. Introduction

Online matching markets (e.g. Mechanical Turk, Upwork, Uber, Labour markets, Restaurant) are economic platforms that connect demand side, (e.g. businesses in Mechanical Turk or Upwork, customers wanting Uber ride or restaurant reservations), to the supply side (e.g. freelancers in Upwork, or crowdworkers in Mechanical Turk, drivers in Uber, restaurant availability). In these platforms, the demand side (also known as the agent or player side) makes repeated decisions to obtain (match) the resources in the supply side (also known as arms side) according to their preference. The supply side is usually resource constrained, and hence it is possible that more than one players compete for a particular resource. Given multiple offers, the supply side arm chooses the agent of its choice. This agent is given the non-zero random reward, while all other players participated in the collision gets a deterministic zero reward. Hence, in this framework, the players need to simultaneously compete as well as estimate the uncertainty in the quality of resource. Economic markets have been studied empirically in (Johari et al., 2021; Das & Kamenica, 2005). Usually the interaction between the supply and demand side is modelled as a bipartite graph, with players and resources in both sides have a preference (or ranking) over the other side, which is unknown apriori. Each players' task is to learn this preference through successive but minimal interaction between the sides and thereafter obtain an optimal stable matching between the demand and the supply side.

Multi-Armed Bandits (MAB) is a popular framework that balances exploration and exploitation and while navigating amongst uncertainty in the system (Lattimore & Szepesvári, 2020). Classical algorithms like Upper Confidence Bound (Auer et al., 2002), Explore-Then-Commit (ETC) (Latti-

[1] Department of Electrical Engineering, IIT Bombay, India [2] Systems and Control Engineering and Centre for Machine Intelligence and Data Sciences (CMInDS), IIT Bombay, India. Correspondence to: Tejas Pagare <tejaspagare2002@gmail.com>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

more & Szepesvári, 2020) are well-known MAB algorithms that yield sub-linear regret. Learning in matching markets has received considerable interest in the recent past, especially from the lens of a multi-agent MAB framework (Liu et al., 2021; Sankararaman et al., 2021; Kong & Li; Basu et al., 2021). In this formulation, the demand side corresponds to multiple players and the supply side resources correspond to multiple arms. The additional complexity here is the presence of competition among players. The objective of this problem is to learn the preference ranking *for both the players and arms simultaneously* through successive interactions. Once the preferences are learnt, matching algorithm like Gale Shapley (Gale & Shapley, 1962; Roth & Sotomayor, 1990) may be used to obtain the optimal stable matching.

Although the MAB framework in markets and bandits have been popular in the recent past, starting with a centralized matching markets (i.e., there exists a centralized authority who helps in the matching process) (Liu et al., 2020; Jagadeesan et al., 2021) to a more recent decentralized framework (Lattimore & Szepesvári, 2020; Sankararaman et al., 2021; Basu et al., 2021; Kong & Li; Maheshwari et al., 2022), *all* the previous works have several strong (and often impractical) assumptions on the market for theoretical tractability. We now discuss them in detail.

One-sided learning: One of the main assumptions made in all the previous works (Liu et al., 2020; 2021; Sankararaman et al., 2021; Basu et al., 2021; Maheshwari et al., 2022; Ghosh et al., 2022), including the state-of-the-art (Kong & Li) is that of *one-sided learning*. It is assumed that the arms know their preferences over players apriori before the start of the learning, and hence the crux of the problem is to learn the preferences for the players only. So, the problem of two sided learning reduces to one sided learning. Here, algorithms based on Upper-Confidence-Bound (Sankararaman et al., 2021; Liu et al., 2021) as well as Explore-Then-Commit (Kong & Li; Basu et al., 2021) are both analyzed and sublinear regret guarantees are obtained. With the knowledge of the arms' ranking apriori, the system can correctly resolve the conflicts every time a collision occurs, which is used crucially in the analysis of such algorithms.

Additional structural market assumptions: Apart from the one-sided learning assumption, several additional structural assumptions on markets are usually being made to obtain sub-linear regret. In (Liu et al., 2021; Kong & Li; Ghosh et al., 2022), it is assumed that when an agent successfully receives a reward (or in other words, an agent and an arm are successfully matched), the matched pair becomes a common knowledge to all the players. This assumption is used as a form of communication across players to calculate the regret of the players in the system. Moreover, (Sankararaman et al., 2021; Ghosh et al., 2022) assumed a *serial dictatorship* model where the preference ranking the players are assumed to be same for all the arms. In (Liu et al., 2021), the same assumption is termed as *global ranking*. In (Maheshwari et al., 2022), this assumption is weakened and termed as α-*reducible* condition. Moreover, in (Basu et al., 2021), the authors propose a *uniqueness consistency* assumption on the market, implying that the leaving participants do not alter the original stable matching of the problem.

1.1. Summary of Contributions

We study a completely decentralized learning algorithm, namely Epoch based CA-ETC (collision avoidance explore then commit, `CA-ETC` in short) which works for a *2-sided* market and with *no restrictive assumptions*. We now explain our contributions in detail.

Fully decentralized 2 sided learning algorithm: Our main contribution is to propose (the first) provable 2 sided learning algorithm for matching markets. To be concrete, we do not assume that the preference of the players are known to the arms apriori. Our proposed scheme, `CA-ETC` is based on Explore Then Commit (ETC) algorithm, similar to (Kong & Li). However, `CA-ETC` is able to learn the preferences for both the agent and the arm side simultaneously through obtained samples in the exploration phase. To this end, we propose a two-sided reward model, i.e., when a player is matched with an arm, the player as well as the arm receives a (random) reward, which is used to estimate the preferences. Note that in real world applications, like 2 sided labor markets (Upwork, Uber, Restaurant), crowd-sourcing platforms (Mechanical Turk), scheduling jobs to servers (AWS, Azure) (Dickerson et al., 2019; Even et al., 2009), the preference of players side are apriori unknown and a 2 sided learning algorithm is necessary. Moreover, when an agent and an arm interacts, two sided reward is usually generated (e.g., in Mechanical Turk, Uber, Yelp the customers as well as the resource (arm) are both rated simultaneously). Note that, without generating rewards (or related information) from both sides (agent as well as arms side), it is not possible for any 2-sided learning algorithm to simultaneously learn both the preferences.

Our algorithm, `CA-ETC` is a completely communication free algorithm, i.e., the players do not communicate during the entire learning duration and fully decentralized. Moreover, `CA-ETC` is *completely collision free*.

`CA-ETC` is a multi-epoch algorithm. At the beginning of each epoch, it spends a fixed amount of time in exploration, and uses the collected samples upto that instant (including exploration rounds of all previous epochs as well) to build the Lower Confidence Bound (LCB) and Upper Confidence Bound (UCB) index for both players and arms simultaneously. Using these, CA-ETC estimates the preference ordering (rank) for both the players as well as for the arms side. The rest of the epoch is spent on exploitation, where Gale Shapley (Gale & Shapley, 1962) algorithm is played with the learned preference so far. The idea behind the design of CA-ETC is that in the first few epochs, the total amount of exploration might not be sufficient to properly learn the preference ordering and hence Gale Shapley might output an incorrect stable matching leading to a higher regret (to be defined formally shortly). However, after a finitely many epochs, the total exploration period becomes sufficient in order to estimate the preference ranking and so the exploit stage (Gale Shapley) incurs zero regret with high probability.

We would like to point out that very recently, (Jagadeesan et al., 2021) proposes an algorithm for 2-sided learning. However, the algorithms in (Jagadeesan et al., 2021) are centralized. Moreover, the problem framework allows utility transfer (monetary transfers), with convergence to a weaker notion of stability.

No structural assumptions on markets: We emphasize that CA-ETC does not require any additional assumption on the economic market, like serial dictatorship, global knowledge of matching player arm pair, unique consistency etc. This makes CA-ETC more practical. In applications like labor markets (Upwork, Taskrabbit) (Massoulié & Xu, 2016), crowd-sourcing platforms (Mechanical Turk), scheduling jobs to servers in an online marketplace (AWS, Azure) (Dickerson et al., 2019; Even et al., 2009), question answering platform (Quora, Stack Overflow) (Shah et al., 2020) the structural assumptions mentioned above are naturally not satisfied. So, there is a gap between theory and practice, and our proposed algorithm, CA-ETC aims to close this gap.

Summary: To summarize, previous papers on matching markets, through theoretically tractable and obtains near-optimal regret, the strong market assumptions and the 1 sided learning makes it practically un-usable. On the other hand, CA-ETC is a practical algorithm for 2 sided learning, with no assumptions. As a result, the regret bounds of CA-ETC is weaker compared to the existing works. CA-ETC needs to estimate the ranking for both arms and players side, and thus the amount of exploration needed turns out to be large, leading to high regret. The additional regret can be thought of the price of 2 sided learning without assumptions. Of course, we problem of obtaining the regret lower bound in 2 sided learning is still open. Table 1 gives a comprehensive summary and comparison of our results with the existing works.

2. Problem Setting

We now explain the problem formulation. Consider a market with N players and K arms with $N \leq K$. Denote $\mathcal{N} = \{p_1, p_2, \dots, p_N\}$ as the set of players (or agents) and $\mathcal{K} = \{a_1, a_2, \dots, a_K\}$ as the set of arms. In general, not all players and arms will participate in the market, but we consider that the participating agents include all the players and arms. At time step t, each player p_i attempts to pull an arm $A_i(t) \in \mathcal{K}$. When multiple players pull the same arm, only one player will successfully pull the arm, based on arm's preferences which are also learned over time.

Two sided reward: Since our goal is to learn the preferences of the players and arms simultaneously, we propose a two-sided reward model in the following way. If player p_i successfully pulls an arm $A_i(t) = a_j$ then p_i receives a stochastic reward $X_j^{(i)}(t) \sim \text{SubGaussian}(\mu_j^{(i)})$, and arm a_j receives a stochastic reward $Y_i^{(j)}(t) \sim \text{SubGaussian}(\eta_i^{(j)})$. We remark that for two sided learning, reward information for both the player and the arm side are necessary, and hence we propose this two sided reward model. If $\mu_j^{(i)} > \mu_{j'}^{(i)}$, we say that player p_i truly prefers arm a_j over $a_{j'}$. Similarly, if $\eta_i^{(j)} > \eta_{i'}^{(j)}$, we say that arm a_j truly prefers player p_i over $p_{i'}$. We denote $A_j^{-1}(t) := \{p_i : A_i(t) = a_j\}$ as the set of players proposing arm a_j, $\bar{A}_i(t)$ as the successfully matched arm of player p_i, $\bar{A}_j^{-1}(t)$ as the successfully matched player of arm a_j i.e. $\bar{A}_j^{-1}(t) \in \text{argmax}_{p_i \in A_j^{-1}(t)} \eta_i^{(j)}$. Then $\bar{A}_i(t) = A_i(t)$ when p_i is successfully accepted by arm $A_i(t)$ else if rejected $\bar{A}_i(t) = \emptyset$. When two or more players propose an arm a_j then only the most preferred player $\bar{A}_j^{-1}(t)$ among $A_j^{-1}(t)$ gets an reward $X_j^{(\bar{A}_j^{-1}(t))}(t)$ and other get a zero reward. Also denote $M(t) := \{(i, \bar{A}_i(t) : i \in [N]\}$ as the final matching at round t. We also assume that all the participating agents have strict preference ranking i.e. $\mu_j^{(i)} \neq \mu_{j'}^{(i)}$ and $\eta_i^{(j)} \neq \eta_{i'}^{(j)}$ for all arms $a_j \neq a_{j'}$ and players $p_i \neq p_{i'}$.

2.1. Stable matching

We now define the stable matching (Gale & Shapley, 1962), which is closely related to the Nash equilibrium of the corresponding game. We say a market matching is stable if no pair of players and arms would prefer to be matched with each other over their respective matches. Formally, $M(t)$ is stable if there exists no player-arm pair (p_i, a_j) such that $\mu_j^{(i)} > \mu_{\bar{A}_i(t)}^{(i)}$ and $\eta_i^{(j)} > \eta_{\bar{A}_j^{-1}(t)}^{(j)}$, where we simply define $\mu_\emptyset^{(i)} = -\infty$ and $\eta_\emptyset^{(j)} = -\infty$ for each $i \in [N], j \in [K]$. Let $M := \{m : m \text{ is a stable matching}\}$ be the set of all stable matching and define player-optimal stable matching $\bar{m}_p = \{(i, \bar{m}_i) : i \in [N]\} \in M$ as the players' most preferred match i.e. $\mu_{\bar{m}_i}^{(i)} \geq \mu_{m_i}^{(i)}$ for any $m \in M$ and for all $i \in [N]$. One can similarly define arm-optimal stable matching $\bar{m}_a = \{(\bar{m}_j, j) : j \in [K]\} \in M$ as

	2-sided	Assumption	Regret Type	Regret
(Liu et al., 2020)	No	Centralized	Player-pessimal	$\mathcal{O}(NK^3\log T/\Delta^2)$
(Liu et al., 2021)	No	(player, arm) broadcast	Player-pessimal	$\mathcal{O}\left(\frac{N^5K^2\log^2 T}{\epsilon^{N^4}\Delta^2}\right)$
(Sankararaman et al., 2021)	No	Serial Dictatorship	Player-optimal	$\mathcal{O}\left(NK\log T/\Delta^2\right)$
(Basu et al., 2021)	No	Uniqueness consistency	Player-optimal	$\mathcal{O}\left(NK\log T/\Delta^2\right)$
(Maheshwari et al., 2022)	No	α-reducible	Player-optimal	$\mathcal{O}\left(\mathcal{C}NK\log T/\Delta^2\right)$
(Kong & Li)	No	(player, arm) broadcast	Player-optimal	$\mathcal{O}\left(K\log T/\Delta^2\right)$
This paper	Yes	No assumptions	Player-optimal	$\mathcal{O}\left[T_0\left(\frac{K\log T}{T_0\Delta^2}\right)^{1/\gamma}+T_0(T/T_0)^\gamma\right]$

Table 1. Table comparing the regret bound of CA-ETC with existing works. Here, N is the number of players, K is the number of arms, T is the learning horizon, Δ is the minimum gap (to be defined later). Also, ϵ (in (Liu et al., 2021)) and $\mathcal{C}$ (in (Maheshwari et al., 2022) are problem dependent hyper-parameters. For our algorithm, CA-ETC, T_0 is the initial epoch length and $\gamma \in (0,1)$ is an input parameter related to the subsequent epoch lengths.

Algorithm 1 Index Estimation (view of player p_i)

```
Input :arbitrary preference ranking of players for arm $a_1$
for round $t$=1,2,..., $N$ do
    $A_i(t) = a_1$

    if $\bar{A}_i(t) = A_i(t) = a_1$ then
        Index = $t$;
    end
end
```

the arms' most preferred one match i.e. $\eta^{(j)}_{\bar{m}_j} \geq \eta^{(j)}_{m_j}$ for any $m \in M$ and for all $j \in [K]$. Previous works (Kong & Li) has also defined the notion of player-pessimal stable matching defined as the players' least preferred match $\underline{m}_p = \{(i, \underline{m}_i) : i \in [N]\} \in M$ i.e. $\mu^{(i)}_{\underline{m}_i} \leq \mu^{(i)}_{m_i}$ for any $m \in M$ and for all $i \in [N]$. Similarly we define arm-pessimal stable matching $\underline{m}_a = \{(\underline{m}_j, j) : j \in [K]\} \in M$ as the arms' least preferred match i.e. $\eta^{(j)}_{\underline{m}_j} \leq \eta^{(j)}_{m_j}$ for any $m \in M$ and for all $j \in [K]$.

2.2. Regret Definition

Based on the above definitions one can define player-optimal (arm-optimal) regret for each player $i \in [N]$ (arm $j \in [K]$) over T rounds as

$$\overline{RP}_i(t) = \sum_{t=1}^{T} \mu^{(i)}_{\bar{m}_i} - \mathbb{E}\left[\sum_{t=1}^{T} X^{(i)}_j(t)\right],$$

$$\overline{RA}_j(t) = \sum_{t=1}^{T} \eta^{(j)}_{\bar{m}_j} - \mathbb{E}\left[\sum_{t=1}^{T} Y^{(j)}_i(t)\right].$$

Similarly, we define the player-pessimal and arm-pessimal regret as follows

$$\underline{RP}_i(t) = \sum_{t=1}^{T} \mu^{(i)}_{\underline{m}_i} - \mathbb{E}\left[\sum_{t=1}^{T} X^{(i)}_j(t)\right],$$

$$\underline{RA}_j(t) = \sum_{t=1}^{T} \eta^{(j)}_{\underline{m}_j} - \mathbb{E}\left[\sum_{t=1}^{T} Y^{(j)}_i(t)\right].$$

Note that, player-pessimal (arm-pessimal) regret is upper bounded by player-optimal (arm-optimal) regret, and hence any upper bound on player-optimal regret automatically serves as an upper bound on player pessimal regret. When there are more than one stable matching i.e. $|M| > 1$, the difference between player-pessimal and player-optimal regret can be $\mathcal{O}(T)$ due to a constant difference between $\mu^{(i)}_{\underline{m}_i}$ and $\mu^{(i)}_{\bar{m}_i}$, similarly for arms. Throughout we give guarantees for player-optimal regret, but the same hold for arm-optimal regret.

3. Algorithms for 2 sided matching markets

We present the learning algorithms in this section. We start with a simple setup, where the user has knowledge of the problem gap. We then make an attempt to remove this by introducing a blackboard. Finally we present out main algorithm, CA-ETC. We define the gap of player p_i as $\Delta^{(i)} = \min_{j \neq j'} |\mu^{(i)}_{j'} - \mu^{(i)}_{j}|$ and gap of arm a_j as $\Delta'^{(j)} = \min_{i \neq i'} |\eta^{(j)}_{i'} - \eta^{(j)}_{i}|$. The universal gap is defined as the minimum of all the player and arm gap i.e. $\Delta = \min_{i \in [N], j \in [K]}\{\Delta^{(i)}, \Delta'^{(j)}\}$.

3.1. Warm up: Learning with Known Gap, Δ

We present following algorithm 2 for learning the preferences of players in full information setting where all the participating agents have the knowledge of universal gap Δ, horizon T and arms K.

Index estimation: Here, the first step is to obtain a distinct index for each player, which from Algorithm (1) will be the player's preference rank for arm a_1. In the first round, all

Algorithm 2 Expore-then-Gale-Shapley (view of player p_i) with Known Gap Δ

```
Input  :Exploration rounds $t_{\text{exp}} = \left\lceil \frac{32K \log T}{\Delta^2} \right\rceil$
Perform Index estimation (Algorithm 1)
    // Learning Preferences
for $t = 1, \dots, t_{\text{exp}}$ do
    $A_i(t) = a_{(\text{Index}+t-1)\%K+1}$ // Round-robin exploration
    Observe $X^{(i)}_{A_i(t)}(t)$ and update $\hat{\mu}^{(i)}_{A_i(t)}, T^{(i)}_{A_i(t)}$ if $\bar{A}_i(t) = A_i(t)$
end
Compute $\text{UCB}^{(i)}_k(t_{\text{exp}})$ and $\text{LCB}^{(i)}_k(t_{\text{exp}})$ for each $k \in [K]$
if $\exists \sigma$ such that $\text{LCB}^{(i)}_{\sigma_k} > \text{UCB}^{(i)}_{\sigma_{k+1}}$ for any $k \in [K-1]$ then
    Preferences = $\sigma$ // this happens with high probability
else
    Preferences = arbitrary but fixed
end
// Perform Gale-Shapley
Propose using $\sigma$ till acceptance
Initialize $s = 1$
for $t = t_{\text{exp}} + 1, \dots, T$ do
    $A_i(t) = a_{\sigma^{(i)}_s}$
    $s = s + 1$ if $\bar{A}_i(t) == \emptyset$
end
```

players propose arm a_1 and in the next round all players except the accepted player propose arm a_1 and so on. In the beginning arm a_1 will have arbitrary but distinct preference ranking over players, hence leading to a distinct index for each player.

After index estimation, each player performs a round-robin exploration for t_{exp} number of rounds by pulling an distinct arm based on the distinct indices obtained from (1) at each round. This ensures that at each round every player is matched with the proposing arm. From the view of player p_i, after observing the reward from the matched arm $A_i(t)$, it updates an estimate of mean reward $\hat{\mu}^{(i)}_{A_i(t)}$ and the observed time $T^{(i)}_{A_i(t)}$ which is defined as the number of times player p_i is matched with arm $A_i(t)$ initialized as $T^{(i)}_{A_i(0)} = 0$. At time t this is updated using

$$\hat{\mu}^{(i)}_{A_i(t)} = \frac{\hat{\mu}^{(i)}_{A_i(t)} T^{(i)}_{A_i(t)} + X^{(i)}_{A_i(t)}}{T^{(i)}_{A_i(t)} + 1}, \quad T^{(i)}_{A_i(t)} = T^{(i)}_{A_i(t)} + 1$$

After t_{exp} exploration rounds player p_i updates the UCB and LCB estimates of $\mu^{(i)}_k$ for each arm a_k as

$$\text{UCB}^{(i)}_k(t) = \mu^{(i)}_k + \sqrt{\frac{2 \log t}{T^{(i)}_k}} \text{ and } \text{LCB}^{(i)}_k(t) = \mu^{(i)}_k - \sqrt{\frac{2 \log t}{T^{(i)}_k}}$$

where $\text{UCB}^{(i)}_k(t)$ and $\text{LCB}^{(i)}_k(t)$ are initialized as ∞ and $-\infty$ resp. for all arms a_k. After this, player p_i checks for all possible permutations of arm preferences, and selects the preference which has disjoint confidence intervals i.e. a preference σ such that $\text{LCB}^{(i)}_{\sigma_k} > \text{UCB}^{(i)}_{\sigma_{k+1}}$ for any $k \in [K-1]$. We define $\Delta^{(i)}_{\max}$ as the maximum stable regret suffered by p_i in all rounds which is $\Delta^{(i)}_{\max} = \mu_{i,\bar{m}_i}$.

Theorem 3.1 (Regret of Algorithm 2)**.** *Suppose algorithm 2 is played for T iterations. Then, player p_i incurs the regret of*

$$\overline{RP}_i(T) \leq \left(N + \frac{64K \log T}{\Delta^2} + K^2 + \frac{NK\pi^2}{3} \right) \Delta^{(i)}_{\max}.$$

Remark 3.2. The regret order-wise matches with the regret in the 1 sided learning case (Kong & Li; Basu et al., 2021).

Remark 3.3. This is not realistic, knowledge of Δ is not usually, available apriori.

3.2. Main Algorithm: Epoch-based CA-ETC

We now present the main algorithm, namely CA-ETC which doesn't know Δ apriori and is fully decentralized and communication-free. For this, CA-ETC is an epoch-based explore then commit type algorithm. Here (Algorithm 3) we describe the player's learning procedure, the arm's learning procedure is similar which we describe in the Appendix.

The algorithm first uses the Index estimation subroutine (1) to get a distinct index for each arm. We then proceed with an epoch-based learning where in each epoch l, every player performs round-robin exploration for $2^{l/\gamma} T_0$ rounds. After the exploration, every agent checks if there exists a preference ranking by permuting over all possible rankings such that the confidence intervals are disjoint. If true, this assures that the ranking found is a correct preference ranking else every agent uses an arbitrary but fixed ranking over all epochs. We then perform Gale-Shapley algorithm using the preference ranking found which lasts atmost K^2 rounds. Since we do not know Δ, we aim to collect samples in the exploration rounds of each epoch. Note that, when the total sample size exceeds $\lceil \frac{32K \log T}{\Delta^2} \rceil$ (Kong & Li), with high probability, player i can estimate the correct preference. In Algorithm 3, this happens after a finitely many epochs.

4. Theoretical Guarantees

In this section, we present the regret bounds for CA-ETC for player p_i. Similar regret can be obtained for arm a_j (we defer this to the appendix). Note that these bounds are

Algorithm 3 Main Algorithm: Epoch-based CA-ETC (view of player p_i)

```
Input : Epoch length $T_0$, Parameter $\gamma \in (0, 1)$ with $b = 2^{1/\gamma}$
Run Algorithm 1 for Index Estimation
for $l = 1, 2, \ldots$ do
    Base Algorithm (Exploration rounds = $2^l T_0$,
     Horizon length = $b^l T_0$)
end
```

Algorithm 4 Base Algorithm: Expore-then-Gale-Shapley (view of player p_i)

```
Input   :Exploration rounds $2^l T_0$, Horizon $b^l T_0$
// Learning Preferences
for $t = \sum_{l'=1}^{l-1} 2^{l'} T_0 + 1, \ldots, \sum_{l'=1}^{l-1} 2^{l'}_T 0 + 2^l T_0$ do
    // $t$ is the global time
    $A_i(t) = a_{(\text{Index}+t-1)\%K+1}$ // Round-robin
       exploration
    Observe $X^{(i)}_{A_i(t)}(t)$ and update $\hat{\mu}^{(i)}_{A_i(t)}, T^{(i)}_{A_i(t)}$ if $\bar{A}_i(t) =$
     $A_i(t)$
end
Compute UCB$^{(i)}_k$ and LCB$^{(i)}_k$ for each $k \in [K]$
if $\exists\sigma$ such that LCB$^{(i)}_{\sigma_k}$ > UCB$^{(i)}_{\sigma_{k+1}}$ for any $k \in [K-1]$
 then
    Preferences = $\sigma$
else
    Preferences = arbitrary but fixed
end
// Perform Gale-Shapley
Propose using $\sigma$ till acceptance
Initialize $s = 1$
for $t = 1, 2, \ldots, b^l T_0 - 2^l T_0$ do
    $A_i(t) = a_{\sigma^{(i)}_s}$
     $s = s + 1$ if $\bar{A}_i(t) == \emptyset$
end
```

player-optimal. Recall the definition of the gap of the i-th player $\Delta^{(i)}$ and the universal gap Δ. We have the following regret upper bound.

Theorem 4.1. *Suppose* CA-ETC *is run with initial epoch length T_0 and input parameter $\gamma \in (0, 1)$. Then, provided the initial epoch satisfies*

$$T_0 \gtrsim \left[\frac{32K \log T}{\Delta^2 (T-N)^\gamma}\right]^{\frac{1}{1-\gamma}},$$

the (player-optimal) regret for player p_i is given by

$$\begin{aligned}\overline{RP}_i(T) \leq N\Delta^{(i)}_{\max} + T_0 \left(\frac{64K \log T}{T_0 (\Delta^{(i)})^2} + 4\right)^{1/\gamma} \Delta^{(i)}_{\max} \\ + 2T_0 \left(\frac{T-N}{T_0} + 1\right)^\gamma \Delta^{(i)}_{\max} \\ + K^2 \gamma \log\left(\frac{T-N}{T_0} + 1\right) \Delta^{(i)}_{\max} + \frac{NK\pi^2}{3} \Delta^{(i)}_{\max}.\end{aligned}$$

Remark 4.2. CA-ETC takes γ and T_0 as input parameters to be chosen by the learner. Typical values of γ would be $\{1/3, 1/4, 1/5\}$, which would imply a polynomial dependence on $\log T$ and a weakly increasing function of T.

Remark 4.3 (Different terms). First term in the regret comes from the index estimation subroutine. The second term results from the round-robin exploration before player p_i estimate the correct preference ranking. The third and fourth term comes from the round-robin exploration and Gale-Shapley after the rank estimation. The last term results from the SubGaussian concentration bound.

Remark 4.4 (Choice of T_0). We have a condition that restricts T_0 to be too small. However, note that, unless the gap Δ is too small, the condition is rather-mild. Of course, the optimal choice of T_0 depends on the gap, Δ and hence not known to the learner apriori.

Remark 4.5 (Regret Comparison). In comparison to 1 sided learning papers (Sankararaman et al., 2021; Liu et al., 2021; Kong & Li), the regret incurred by CA-ETC is high. This can be attributed to the cost of 2 sided assumption free learning. The dependence on γ comes from the multi-phase nature of CA-ETC.

5. Conclusion and Future Direction

In this paper, we propose practical algorithms for 2-sided learning in matching markets without restrictive assumptions. We analyze a multi-epoch ETC type algorithm and obtain sub-linear regret. Note that we only leverage the Explore-Then-Commit (ETC) algorithm for 2-sided learning. We refer to some preliminary experimental results in the Appendix for proof of concept. An immediate future work will be to study the UCB based algorithms for two sided markets. Moreover, we would like to study the market setup, with transferable utilities (i.e., monetary transfer) in a 2-sided setting. Furthermore, markets are seldom static, and the preference ranking changes over time. Capturing the dynamic behavior of markets in an assumption free 2-sided setting is certainly challenging. We keep these as our future endeavors.

References

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

Basu, S., Sankararaman, K. A., and Sankararaman, A. Beyond $\hat{log}2(t)$ regret for decentralized bandits in matching markets. In *International Conference on Machine Learning*, pp. 705–715. PMLR, 2021.

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

Dickerson, J., Sankararaman, K., Sarpatwar, K., Srinivasan, A., Wu, K.-L., and Xu, P. Online resource allocation with matching constraints. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2019.

Even, G., Halldórsson, M. M., Kaplan, L., and Ron, D. Scheduling with conflicts: online and offline algorithms. *Journal of scheduling*, 12(2):199–224, 2009.

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

Ghosh, A., Sankararaman, A., Ramchandran, K., Javidi, T., and Mazumdar, A. Decentralized competing bandits in non-stationary matching markets. *arXiv preprint arXiv:2206.00120*, 2022.

Jagadeesan, M., Wei, A., Wang, Y., Jordan, M., and Steinhardt, J. Learning equilibria in matching markets from bandit feedback. *Advances in Neural Information Processing Systems*, 34:3323–3335, 2021.

Johari, R., Kamble, V., and Kanoria, Y. Matching while learning. *Operations Research*, 69(2):655–681, 2021.

Kong, F. and Li, S. *Player-optimal Stable Regret for Bandit Learning in Matching Markets*, pp. 1512–1522. doi: 10.1137/1.9781611977554.ch55. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611977554.ch55.

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

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

Liu, L. T., Ruan, F., Mania, H., and Jordan, M. I. Bandit learning in decentralized matching markets. *The Journal of Machine Learning Research*, 22(1):9612–9645, 2021.

Maheshwari, C., Sastry, S., and Mazumdar, E. Decentralized, communication-and coordination-free learning in structured matching markets. *Advances in Neural Information Processing Systems*, 35:15081–15092, 2022.

Massoulié, L. and Xu, K. On the capacity of information processing systems. In *Conference on Learning Theory*, pp. 1292–1297. PMLR, 2016.

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

Sankararaman, A., Basu, S., and Sankararaman, K. A. Dominate or delete: Decentralized competing bandits in serial dictatorship. In *International Conference on Artificial Intelligence and Statistics*, pp. 1252–1260. PMLR, 2021.

Shah, V., Gulikers, L., Massoulié, L., and Vojnović, M. Adaptive matching for expert systems with uncertain task types. *Operations Research*, 68(5):1403–1424, 2020.

A. Appendix

A.1. Experiments

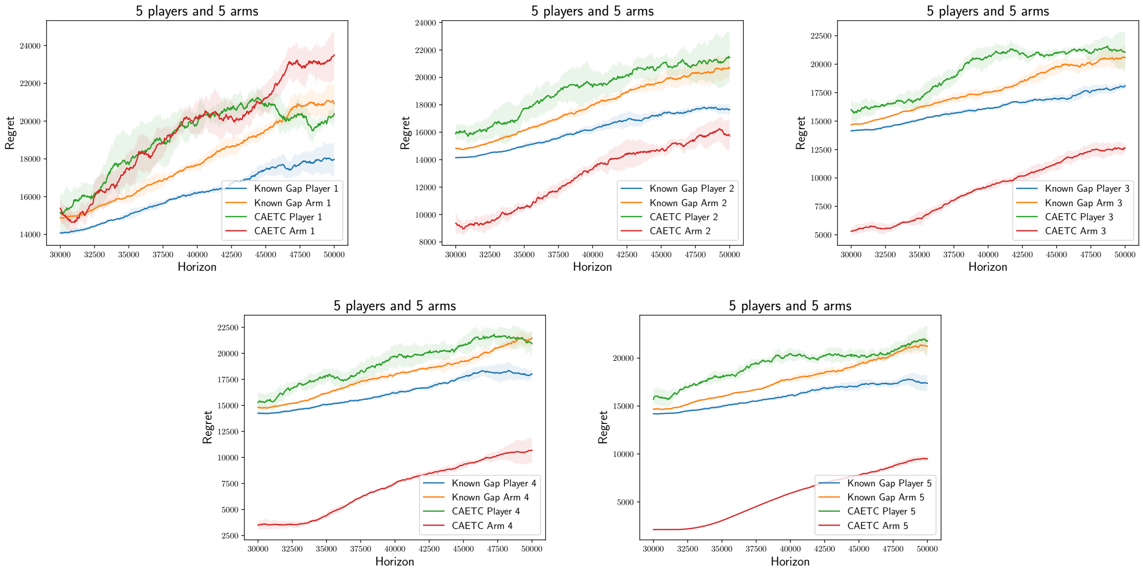

Figure 1. Experiment on Synthetic Matching Market (CA−ETC with $l_1 = 1$)

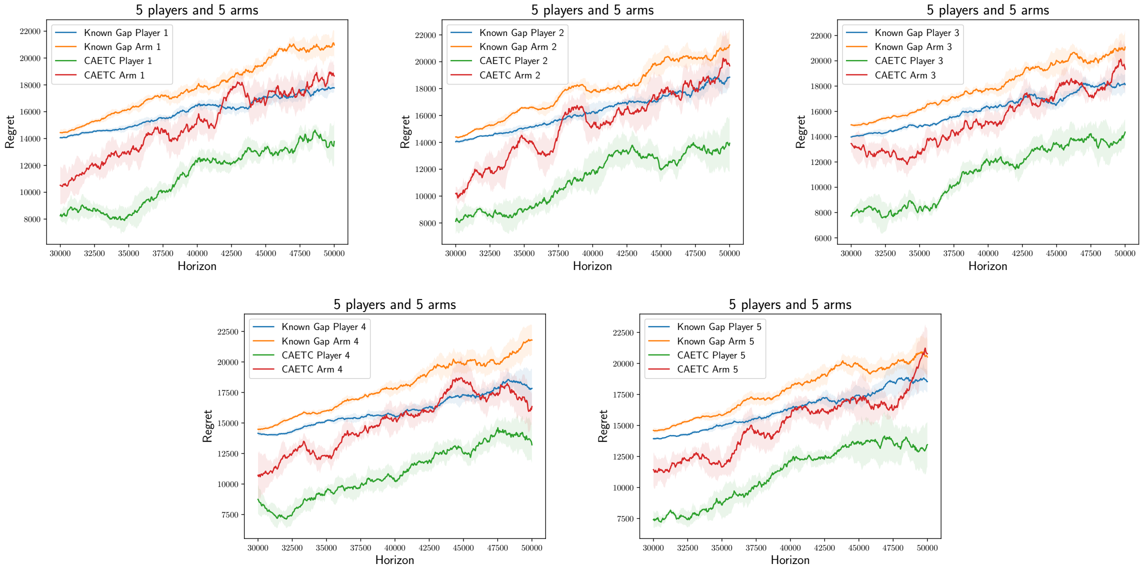

Figure 2. Experiment on Synthetic Matching Market (CA−ETC with $l_1 = 0$)

We experiment on a synthetic market with 5 players and 5 arms. We consider that the reward distributions are Gaussian with mean randomly selected without repetition from $[0, 0.25, 0.5, 0.75, 1]$ with standard deviation σ as 0.1. We plot the player-optimal regret as well as the arm-optimal regret for all the agents (players and arms) in the system.

We experiment on two cases where CA−ETC starts with epoch l_1 as 0 and 1. Surprisingly, one can note that for $l_1 = 0$, CA−ETC suffers less regret as compared to the algorithm with known gap. This is because t_{exp} is an overestimate of the number of exploration rounds.

Parameters:

- $\gamma = 0.25$
- T_0 is chosen based on $\Delta = 0.25$ for this market, however in general one can use an highly optimistic value.

In general one can use $T_0 \in \{1000, 2000, \ldots\}$. One can also use the following scheme in which one starts with lower value of T_0 say 1000 and then increment by factors of 10 in each epoch. However, in general `CA-ETC` performance does not change much with different values of T_0 as long as it is optimistically high.

Experiments are done on 5 different seeds and the plots show the line plot of regret vs horizon with shaded region being the 95% confidence interval.

A.2. Proof of Theorem 3.1

First we present lemmas, which will support our theorems. We note that the this analysis goes more or less along the same lines of (Kong & Li).

Let us define the event $\mathcal{F}_p(t) = \left\{ \exists i \in [N],\ j \in [K] : |\hat{\mu}_j^{(i)} - \mu_j^{(i)}| > \sqrt{\frac{2\log t}{T_j^{(i)}}} \right\}$, a player's bad event that some preference of the arms are not estimated by the players correctly at time t. We present the following Lemma 5.1 of (Kong & Li)

Lemma A.1.

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{F}_p(t)\}\right] \leq \frac{NK\pi^2}{3} \tag{1}$$

Lemma A.2. *In round t, let $T^{(i)}(t) = \min_{j\in[K]} T_j^{(i)}(t)$ and $\bar{T}^{(i)} = 32\log T/(\Delta^{(i)})^2$ where $\Delta^{(i)} = \min_{j\neq j'} \Delta_{j,j'}^{(i)} = \min_{j\neq j'} |\mu_j^{(i)} - \mu_{j'}^{(i)}|$. Conditional on $\mathcal{F}_p(t)^\complement$, if $T^{(i)} > \bar{T}^{(i)}$ we have $\mathrm{UCB}_j^{(i)}(t) < \mathrm{LCB}_{j'}^{(i)}(t)$ for any $j, j' \in [K]$ with $\mu_j^{(i)} < \mu_{j'}^{(i)}$.*

Proof. We will prove this by contradiction i.e. suppose there exists $j, j' \in [K]$ with $\mu_j^{(i)} < \mu_{j'}^{(i)}$ such that $\mathrm{UCB}_j^{(i)}(t) \geq \mathrm{LCB}_{j'}^{(i)}$. Conditioned on $\mathcal{F}_p(t)^\complement$ we have that

$$|\hat{\mu}_j^{(i)} - \mu_j^{(i)}| \leq \sqrt{\frac{2\log t}{T_j^{(i)}}},\ \ |\hat{\mu}_{j'}^{(i)} - \mu_{j'}^{(i)}| \leq \sqrt{\frac{2\log t}{T_{j'}^{(i)}}}$$

and using the definition of $\mathrm{UCB}_j^{(i)}(t)$ and $\mathrm{LCB}_{j'}^{(i)}(t)$ we have

$$\mu_{j'}^{(i)} - 2\sqrt{\frac{2\log t}{T_j^{(i)}}} \leq \mathrm{LCB}_{j'}^{(i)}(t) \leq \mathrm{UCB}_j^{(i)}(t) \leq \mu_j^{(i)} + 2\sqrt{\frac{2\log t}{T_j^{(i)}}}$$

This implies

$$\Delta_{j,j'}^{(i)} = \mu_{j'}^{(i)} - \mu_j^{(i)} \leq 4\sqrt{\frac{2\log t}{T_j^{(i)}}}$$
$$\implies T^{(i)}(t) \leq \frac{32\log T}{(\Delta_{j,j'}^{(i)})^2} \leq \frac{32\log T}{(\Delta^{(i)})^2}.$$

This contradicts the fact that $T^{(i)}(t) > \bar{T}^{(i)}(t)$ □

Lemma A.3. *Conditional on $\mathcal{F}_p(t)^\complement$, $\mathrm{UCB}_j^{(i)}(t) < \mathrm{LCB}_{j'}^{(i)}(t)$ implies $\mu_j^{(i)} < \mu_{j'}^{(i)}$.*

Proof. Using the LCB and UCB definiton we have

$$\mathrm{LCB}_j^{(i)}(t) = \hat{\mu}_j^{(i)} - \sqrt{\frac{2\log t}{T_j^{(i)}}} \leq \mu_j^{(i)} \leq \hat{\mu}_j^{(i)} + \sqrt{\frac{2\log t}{T_j^{(i)}}} = \mathrm{UCB}_j^{(i}(t)$$

where the inequalities are consequences of the conditional $\mathcal{F}_p(t)^{\complement}$. Thus if $\mathrm{UCB}_j^{(i)}(t) < \mathrm{LCB}_{j'}^{(i)}(t)$, we would have

$$\mu_j^{(i)} \leq \mathrm{UCB}_j^{(i)}(t)) < \mathrm{LCB}_{j'}^{(i)}(t) \leq \hat{\mu}_{j'}^{(i)}$$

which proves the lemma. □

Lemma A.4. *Conditional on $\cap_{t=1}^{T}\mathcal{F}_p(t)^{\complement}$, after t_{exp} number of exploration rounds a non-zero regret may be incurred for at most K^2 rounds where*

$$t_{\text{exp}} = \left\lceil \frac{32K\log T}{\Delta^2} \right\rceil \tag{2}$$

Proof. Since in each exploration period all players propose to distinct arms using a round-robin fashion, no collision occurs and all players are accepted at each round of the exploration period in each epoch. Thus after t_{exp} number of exploration rounds it holds that $T_j^{(i)} \geq 32\log T/\Delta^2 \geq 32\log T/(\Delta^{(i)})^2$. for any $i \in [N]$ and $j \in [K]$.

Now according to Lemma A.2, when $T_j^{(i)} \geq 32\log T/(\Delta^{(i)})^2$ for any arm a_j, player p_i finds a permutation $\sigma^{(i)}$ over arms such that $\mathrm{LCB}_{\sigma_k^{(i)}}^{(i)} < \mathrm{UCB}_{\sigma_{k+1}^{(i)}}^{(i)}$ for any $k \in [K-1]$.

This implies that after t_{exp}, player p_i finds a permutation of arms with disjoint confidence intervals. Since, Gale-Shapley algorithm will last at most K^2 steps, the regret afterwards will be zero conditioned on the event $\cap_{t=1}^{T}\mathcal{F}_p(t)^{\complement}$. □

Proof of Theorem 3.1.

$$\begin{aligned}
\overline{RP}_i(T) &= \sum_{t=1}^{T} \mu_{\bar{m}_i}^{(i)} - \mathbb{E}\left[\sum_{t=1}^{T} X_j^{(i)}(t)\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\bar{A}(t) \neq \bar{m}\}\Delta_{\max}^{(i)}\right] \\
&\leq N\Delta_{\max}^{(i)} + \mathbb{E}\left[\sum_{t=N+1}^{T} \mathbb{1}\{\bar{A}(t) \neq \bar{m}, \mathcal{F}_p(t)^{\complement}\}\Delta_{\max}^{(i)}\right] + \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\mathcal{F}_p(t)\}\Delta_{\max}^{(i)}\right] \\
&\leq \left(N + \frac{64K\log T}{\Delta^2} + K^2 + \frac{NK\pi^2}{3}\right)\Delta_{\max}^{(i)}.
\end{aligned}$$

□

A.3. Proof of Theorem 4.1

Lemma A.5. *Conditional on $\cap_{t=1}^{T}\mathcal{F}_p(t)^{\complement}$, after epoch $l_{\max}$, the regret for the period $b^l T_0 - 2^l T_0 - K^2$ in epoch $l > l_{\max}$ is zero where*

$$l_{\max} = \min\left\{ l : \sum_{l'=1}^{l} 2^{l'} T_0 \geq 32K\log T/(\Delta^{(i)})^2 \right\}. \tag{3}$$

Proof. Since in each exploration period all players propose to distinct arms using a round-robin fashion, no collision occurs and all players are accepted at each round of the exploration period in each epoch. Thus at the end of epoch $l_{\max}$ it holds that $T_j^{(i)} \geq 32\log T/(\Delta^{(i)})^2$ for any $i \in [N]$ and $j \in [K]$.

Now according to Lemma A.2, when $T_j^{(i)} \geq 32\log T/(\Delta^{(i)})^2$ for any arm a_j, player p_i finds a permutation $\sigma^{(i)}$ over arms such that $\mathrm{LCB}_{\sigma_k^{(i)}}^{(i)} < \mathrm{UCB}_{\sigma_{k+1}^{(i)}}^{(i)}$ for any $k \in [K-1]$.

This implies that after $l > l_{\max}$, player p_i finds a permutation of arms with disjoint confidence intervals. Since, Gale-Shapley algorithm will last at most K^2 steps, we will have zero regret in the period $b^l T_0 - 2^l T_0 - K^2$ conditioned on the event $\cap_{t=1}^{T}\mathcal{F}_p(t)^{\complement}$. □

We will find an equation of $l_{\max}$, from the definition we have and the fact that $b > 2$

$$l_{\max} = \left\lceil \log\left(\frac{32K\log T}{T_0(\Delta^{(i)})^2} + 2\right) - 1 \right\rceil = \left\lceil \log\left(\frac{16K\log T}{T_0(\Delta^{(i)})^2} + 1\right) \right\rceil$$

$$\implies \log\left(\frac{16K\log T}{T_0(\Delta^{(i)})^2} + 1\right) \le l_{\max} < \log\left(\frac{16K\log T}{T_0(\Delta^{(i)})^2} + 1\right) + 1 = \log\left(\frac{32K\log T}{T_0(\Delta^{(i)})^2} + 2\right)$$

Proof of Theorem 4.1. The optimal stable regret for player p_j is the sum of regret before and after $l_{\max}$.

$$\begin{aligned}
\overline{RP}_i(T) &= \mathbb{E}\left[\sum_{t=1}^{T} \mu^{(i)}_{\bar{m}_i} - X^{(i)}(t)\right] \\
&\le \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\bar{A}(t) \neq \bar{m}_i\}\Delta^{(i)}_{\max}\right] \\
&\le N\Delta^{(i)}_{\max} + \mathbb{E}\left[\sum_{N+1}^{T} \mathbb{1}\{\bar{A}(t) \neq \bar{m}_i, \mathcal{F}_p(t)^{\complement}\}\right]\Delta^{(i)}_{\max} + \mathbb{E}\left[\sum_{N+1}^{T} \mathbb{1}\{\mathcal{F}_p(t)\}\right]\Delta^{(i)}_{\max} \\
&\le N\Delta^{(i)}_{\max} + \mathbb{E}\left[\sum_{N+1}^{T} \mathbb{1}\{\bar{A}(t) \neq \bar{m}_i, \mathcal{F}_p(t)^{\complement}\}\right]\Delta^{(i)}_{\max} + \frac{NK\pi^2}{3}\Delta^{(i)}_{\max}
\end{aligned}$$

Let $T_1 = \sum_{l=1}^{l_{\max}} b^l T_0$ and $T_2 = \sum_{l=l_{\max}+1}^{\tilde{l}} b^l T_0$ be the time-period before and after epoch $\tilde{l}$ with $T - N = T_1 + T_2$. Thus we have

$$\mathbb{E}\left[\sum_{N+1}^{T} \mathbb{1}\{\bar{A}(t) \neq \bar{m}_i, \mathcal{F}_p(t)^{\complement}\}\right]\Delta^{(i)}_{\max} = \mathbb{E}\left[\sum_{l=1}^{l_{\max}} b^l T_0\right]\Delta^{(i)}_{\max} + \mathbb{E}\left[\sum_{l=l_{\max}+1}^{\tilde{l}} 2^l T_0 + K^2\right]\Delta^{(i)}_{\max}$$

The regret before epoch $l_{\max}$ i.e. for $l \le l_{\max}$ is $b^l T_0$. Thus the total regret till the end of the epoch $l_{\max}$ is

$$\begin{aligned}
R_1(l_{\max}) &= \mathbb{E}\left[\sum_{l=1}^{l_{\max}} b^l T_0\right]\Delta^{(i)}_{\max} \\
&= T_0 \frac{b^{l_{\max}+1} - b}{b-1}\Delta^{(i)}_{\max} \\
&\le T_0 \frac{b^{\log\left(\frac{64K\log T}{T_0(\Delta^{(i)})^2}+4\right)} - b}{b-1}\Delta^{(i)}_{\max} \\
&< T_0\left(\frac{64K\log T}{T_0(\Delta^{(i)})^2} + 4\right)^{\log b}\Delta^{(i)}_{\max}
\end{aligned}$$

We have

$$\begin{aligned}
\sum_{l=1}^{\tilde{l}} b^l T_0 &= T - N \\
\frac{b^{\tilde{l}+1} - b}{b-1} T_0 &= T - N \\
b^{\tilde{l}+1} &= \frac{(T-N)}{T_0}(b-1) + b \\
\tilde{l} &= \log_b\left(\frac{b-1}{b}\frac{(T-N)}{T_0} + 1\right) < \log\left(\frac{T-N}{T_0} + 1\right)^{\log_b 2}
\end{aligned}$$

Now we will check the regret for some epoch $l > l_{\max}$ is given by $2^l T_0 + K^2$. Hence total regret after epoch $l_{\max}$ is

$$\begin{aligned}\mathbb{E}\left[\sum_{l=l_{\max}+1}^{\tilde{l}} 2^l T_0 + K^2\right]\Delta_{\max}^{(i)} &\leq 2^{\tilde{l}+1}T_0\Delta_{\max}^{(i)} - 2^{l_{\max}+1} + K^2(\tilde{l} - l_{\max}) \\ &\leq 2T_0\left(\frac{T-N}{T_0}+1\right)^{\log_b 2}\Delta_{\max}^{(i)} - C_1\Delta_{\max}^{(i)} + K^2\left(\log_b\left(\frac{T-N}{T_0}+1\right) - C_2\right)\Delta_{\max}^{(i)} \\ &< 2T_0\left(\frac{T-N}{T_0}+1\right)^{\gamma}\Delta_{\max}^{(i)} + K^2\gamma\log\left(\frac{T-N}{T_0}+1\right)\Delta_{\max}^{(i)}\end{aligned}$$

C_1 and C_2 are positive constants independent of b. Thus the total regret is

$$\begin{aligned}\overline{RP}_i(T) \leq N\Delta_{\max}^{(i)} &+ T_0\left(\frac{64K\log T}{T_0(\Delta^{(i)})^2}+4\right)^{1/\gamma}\Delta_{\max}^{(i)} \\ &+ 2T_0\left(\frac{T-N}{T_0}+1\right)^{\gamma}\Delta_{\max}^{(i)} + K^2\gamma\log\left(\frac{T-N}{T_0}+1\right)\Delta_{\max}^{(i)} + \frac{NK\pi^2}{3}\Delta_{\max}^{(i)}.\end{aligned}$$

where $\Delta_{\max}^{(i)}$ is the maximum stable regret suffered by p_i in all rounds which is $\Delta_{\max}^{(i)} = \mu_{i,\bar{m}_i}$ where $\bar{m}_i$ is the optimal stable match of player p_i. □

For T_2 to be defined correctly we need that

$$\begin{aligned}\log\left(\frac{T-N}{T_0}+1\right)^{\gamma} > \tilde{l} \geq l_{\max}+1 &\geq \log\left(\frac{32K\log T}{T_0(\Delta^{(i)})^2}+2\right) \\ \implies \left(\frac{T-N}{T_0}\right)^{\gamma} &\gtrsim \frac{32K\log T}{T_0(\Delta^{(i)})^2} \\ T_0 &\gtrsim \left(\frac{32K\log T}{(\Delta^{(i)})^2(T-N)^{\gamma}}\right)^{\frac{1}{1-\gamma}} \quad \forall i \\ \implies T_0 &\gtrsim \left(\frac{32K\log T}{\Delta^2(T-N)^{\gamma}}\right)^{\frac{1}{1-\gamma}}\end{aligned}$$

A.4. Generalized CA-ETC

In this section, we consider a generalization of CA-ETC where in the l-th epoch, we have $c^l T_0$ exploration and $b^l T_0$ total rounds with $b > c > 1$ and $\beta = c/b \in (1/b, 1)$. With this, the new condition for T_0 is as follows

$$T_0 \gtrsim \left(\frac{32K\log T}{\Delta^2}(b\beta-1)\right)^{-\log_\beta b}\left(\frac{1}{T-N}\right)^{-(\log_\beta b+1)} \tag{4}$$

Theorem A.6. *For the generalized* CA-ETC *algorithm with $b > 1$ and $\beta \in (1/b, 1)$ as input with T_0 satisfying* (4) *the regret for player p_i is*

$$\begin{aligned}\overline{RP}_i(T) \leq N\Delta_{\max}^{(i)} &+ \frac{T_0}{b-1}\left(\frac{32K\log T}{T_0(\Delta^{(i)})^2}(b\beta)^2 + (b\beta)^2\right)^{\frac{1}{1+\log_b\beta}}\Delta_{\max}^{(i)} \\ &+ \frac{T_0}{b\beta-1}\left(b\frac{(T-N)}{T_0}+b\right)^{1+\log_b\beta}\Delta_{\max}^{(i)} + K^2\log_b\left(\frac{T-N}{T_0}+1\right) + \frac{NK\pi^2}{3}\Delta_{\max}^{(i)}\end{aligned}$$

Remark A.7. Generalized CA-ETC gives us one extra degree of freedom, i.e., the parameter c. For the CA-ETC algorithm of Algorithm 3, we have $c = 2$ and $b = 2^{1/\gamma}$.

Remark A.8. In general one can also choose the explorations rounds as 2^l and total rounds as $2^l + 2^{-cl} + K^2$ were c must be appropriately chosen. The intuition behind this is that the rounds after Gale-Shapley of atmost K^2 rounds should decrease with epoch l.

Lemma A.9. *Conditional on $\cap_{t=1}^{T}\mathcal{F}_p(t)^{\complement}$, after epoch $l_{\max}$, the regret for the period $b^l T_0 - 2^l T_0 - K^2$ in epoch $l > l_{\max}$ is zero where*

$$l_{\max} = \min\left\{ l : \sum_{l'=1}^{l} c^{l'} T_0 \geq 32K \log T/(\Delta^{(i)})^2 \right\}. \tag{5}$$

Proof. Since in each exploration period all players propose to distinct arms using a round-robin fashion, no collision occurs and all players are accepted at each round of the exploration period in each epoch. Thus at the end of epoch $l_{\max}$ it holds that $T_j^{(i)} \geq 32 \log T/(\Delta^{(i)})^2$ for any $i \in [N]$ and $j \in [K]$.

Now according to Lemma A.2, when $T_j^{(i)} \geq 32 \log T/(\Delta^{(i)})^2$ for any arm a_j, player p_i finds a permutation $\sigma^{(i)}$ over arms such that $\text{LCB}^{(i)}_{\sigma^{(i)}_k} < \text{UCB}^{(i)}_{\sigma^{(i)}_{k+1}}$ for any $k \in [K-1]$.

This implies that after $l > l_{\max}$, player p_i finds a permutation of arms with disjoint confidence intervals. Since, Gale-Shapley algorithm will last at most K^2 steps, we will have zero regret in the period $b^l T_0 - 2^l T_0 - K^2$. □

We will find an equation of $l_{\max}$, from the definition we have and the fact that $b > 2$

$$l_{\max} \geq \log_c \left(\frac{32K \log T}{T_0 (\Delta^{(i)})^2} \frac{(c-1)}{c} + 1 \right)$$

$$l_{\max} = \left\lceil \log_c \left(\frac{32K \log T}{T_0 (\Delta^{(i)})^2} \frac{(c-1)}{c} + 1 \right) \right\rceil$$

$$\implies \log_c \left(\frac{32K \log T}{T_0 (\Delta^{(i)})^2} \frac{(c-1)}{c} + 1 \right) \leq l_{\max} < \log_c \left(\frac{32K \log T}{T_0 (\Delta^{(i)})^2} \frac{(c-1)}{c} + 1 \right) + 1 = \log_c \left(\frac{32K \log T}{T_0 (\Delta^{(i)})^2} (c-1) + c \right)$$

Proof of Theorem A.6. The optimal stable regret for player p_i is the sum of regret before and after $l_{\max}$.

$$\begin{aligned}
\overline{RP}_i(T) &= \mathbb{E}\left[\sum_{t=1}^{T} \mu^{(i)}_{\bar{m}_i} - X^{(i)}(t) \right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{\bar{A}(t) \neq \bar{m}_i\} \Delta^{(i)}_{\max} \right] \\
&\leq N\Delta^{(i)}_{\max} + \mathbb{E}\left[\sum_{N+1}^{T} \mathbb{1}\{\bar{A}(t) \neq \bar{m}_i, \mathcal{F}_p(t)^{\complement}\} \right] \Delta^{(i)}_{\max} + \mathbb{E}\left[\sum_{N+1}^{T} \mathbb{1}\{\mathcal{F}_p(t)\} \right] \Delta^{(i)}_{\max} \\
&\leq N\Delta^{(i)}_{\max} + \mathbb{E}\left[\sum_{N+1}^{T} \mathbb{1}\{\bar{A}(t) \neq \bar{m}_i, \mathcal{F}_p(t)^{\complement}\} \right] \Delta^{(i)}_{\max} + \frac{NK\pi^2}{3} \Delta^{(i)}_{\max}
\end{aligned}$$

Let $T_1 = \sum_{l=1}^{l_{\max}} b^l T_0$ and $T_2 = \sum_{l=l_{\max}+1}^{\tilde{l}} b^l T_0$ be the time-period before and after epoch $\tilde{l}$ with $T - N = T_1 + T_2$. Thus we have

$$\mathbb{E}\left[\sum_{N+1}^{T} \mathbb{1}\{\bar{A}(t) \neq \bar{m}_i, \mathcal{F}_p(t)^{\complement}\} \right] \Delta^{(i)}_{\max} = \mathbb{E}\left[\sum_{l=1}^{l_{\max}} b^l T_0 \right] \Delta^{(i)}_{\max} + \mathbb{E}\left[\sum_{l=l_{\max}+1}^{\tilde{l}} 2^l T_0 + K^2 \right] \Delta^{(i)}_{\max}$$

The regret before epoch $l_{\max}$ i.e. for $l \leq l_{\max}$ is $b^l T_0$. Thus the total regret till the end of the epoch $l_{\max}$ is

$$\begin{aligned}
R_1(l_{\max}) &= \mathbb{E}\left[\sum_{l=1}^{l_{\max}} b^l T_0\right] \Delta_{\max}^{(i)} \\
&= T_0 \frac{b^{l_{\max}+1} - b}{b-1} \Delta_{\max}^{(i)} \\
&\leq T_0 \frac{b^{\log_c\left(\frac{32K\log T}{T_0(\Delta^{(i)})^2} c(c-1)+c^2\right)} - b}{b-1} \Delta_{\max}^{(i)} \\
&< \frac{T_0}{b-1}\left(\frac{32K\log T}{T_0(\Delta^{(i)})^2} b\gamma(b\gamma-1) + (b\gamma)^2\right)^{\frac{1}{1+\log_b \gamma}} \Delta_{\max}^{(i)}
\end{aligned}$$

We have

$$\begin{aligned}
\sum_{l=1}^{\tilde{l}} b^l T_0 &= T - N \\
\frac{b^{\tilde{l}+1} - b}{b-1} T_0 &= T - N \\
b^{\tilde{l}+1} &= \frac{(T-N)}{T_0}(b-1) + b \\
\tilde{l} &= \log_b\left(\frac{b-1}{b}\frac{(T-N)}{T_0} + 1\right) < \log_b\left(\frac{T-N}{T_0} + 1\right)
\end{aligned}$$

Now we will check the regret for some epoch $l > l_{\max}$ is given by $2^l T_0 + K^2$. Hence total regret after epoch $l_{\max}$ is

$$\begin{aligned}
\mathbb{E}\left[\sum_{l=l_{\max}+1}^{\tilde{l}} c^l T_0 + K^2\right] \Delta_{\max}^{(i)} &\leq \frac{c^{\tilde{l}+1}}{c-1} T_0 \Delta_{\max}^{(i)} + K^2 \tilde{l} \\
&\leq \frac{c^{\log_b\left((b-1)\frac{(T-N)}{T_0}+b\right)}}{c-1} T_0 \Delta_{\max}^{(i)} + K^2 \log_b\left(\frac{T-N}{T_0} + 1\right) \\
&\leq \frac{T_0}{c-1}\left((b-1)\frac{(T-N)}{T_0} + b\right)^{\log_b c} \Delta_{\max}^{(i)} + K^2 \log_b\left(\frac{T-N}{T_0} + 1\right) \\
&= \frac{T_0}{b\gamma-1}\left((b-1)\frac{(T-N)}{T_0} + b\right)^{1+\log_b \gamma} \Delta_{\max}^{(i)} + K^2 \log_b\left(\frac{T-N}{T_0} + 1\right)
\end{aligned}$$

Thus the total regret is

$$\begin{aligned}
\overline{RP}_i(T) \leq N\Delta_{\max}^{(i)} &+ \frac{T_0}{b-1}\left(\frac{32K\log T}{T_0(\Delta^{(i)})^2}(b\beta)^2 + (b\beta)^2\right)^{\frac{1}{1+\log_b \beta}} \Delta_{\max}^{(i)} \\
&\frac{T_0}{b\beta-1}\left(b\frac{(T-N)}{T_0} + b\right)^{1+\log_b \beta} \Delta_{\max}^{(i)} + K^2 \log_b\left(\frac{T-N}{T_0} + 1\right) + \frac{NK\pi^2}{3}\Delta_{\max}^{(i)}
\end{aligned}$$

where $\Delta_{\max}^{(i)}$ is the maximum stable regret suffered by p_i in all rounds which is $\Delta_{\max}^{(i)} = \mu_{j,\bar{m}_i}$ where $\bar{m}_i$ is the optimal stable match of player p_i. Note that C_1 and C_2 depends on T_0 and hence can be further optimized for different values of T_0. This is the total regret for player p_i. □

For T_2 to be defined correctly we need that

$$\begin{aligned}
&\log_b\left(\frac{T-N}{T_0}+1\right) > \tilde{l} \geq l_{\max}+1 \geq \log_c\left(\frac{32K\log T}{T_0(\Delta^{(i)})^2}(c-1)+c\right) > \log_c\left(\frac{32K\log T}{T_0(\Delta^{(i)})^2}(c-1)\right)\\
\implies &\log\left(\frac{T-N}{T_0}+1\right)^{\log_b 2} > \log\left(\frac{32K\log T}{T_0(\Delta^{(i)})^2}(c-1)\right)^{\log_c 2}\\
&\implies \left(\frac{T-N}{T_0}\right) \gtrsim \left(\frac{32K\log T}{T_0(\Delta^{(i)})^2}(c-1)\right)^{\frac{\log b}{\log c}}\\
&\qquad T_0 \gtrsim \left(\frac{32K\log T}{(\Delta^{(i)})^2}(c-1)\right)^{\frac{\log_c b}{\log_c b-1}} \left(\frac{1}{T-N}\right)^{\frac{1}{\log_c b-1}}\\
&\qquad T_0 \gtrsim \left(\frac{32K\log T}{(\Delta^{(i)})^2}(b\beta-1)\right)^{-\log_\beta b} \left(\frac{1}{T-N}\right)^{-(\log_\beta b+1)}
\end{aligned}$$

$$1 < c < b \text{ with } \beta = \frac{c}{b} \implies \frac{1}{b} < \beta < 1$$

A.5. Arm's Learning

A.5.1. Known Gap

Algorithm 5 Expore-then-Gale-Shapley (view of arm a_j) with Known Gap Δ

```
Input :Exploration rounds $t_{\text{exp}} = \left\lceil \frac{32K \log T}{\Delta^2} \right\rceil$
// Learning Preferences
for $t = 1, \ldots, t_{\text{exp}}$ do
    Accept arm with estimated preferences
     Observe $Y^{(j)}_{A_j^{-1}(t)}(t)$ and update $\hat{\mu}^{(j)}_{A_j^{-1}(t)}, T^{(j)}_{A_j^{-1}(t)}$ if $\bar{A}_j^{-1}(t) = A_j^{-1}(t)$
end
Compute UCB$_n^{(j)}$ and LCB$_n^{(j)}$ for each $n \in [N]$
 if $\exists\beta$ such that LCB$^{(j)}_{\beta_n} >$ UCB$^{(j)}_{\beta_{n+1}}$ for any $n \in [N-1]$ then
    Preferences = $\beta$
else
    Preferences = arbitrary but fixed
end
// Perform Gale-Shapley using the Preferences $\beta^{(j)} = \beta = (\beta^{(j)}_1, \beta^{(j)}_2, \ldots, \beta^{(j)}_N)$ found
Propose using $\beta$ till acceptance
 Initialize $s = 1$
 for $t = t_{\text{exp}} + 1, \ldots, T$ do
    Accept player using the estimated preference $\beta^{(j)}$
     $s = s + 1$ if $\bar{A}_j^{-1}(t) == \emptyset$
end
```

A.5.2. CA-ETC

Algorithm 6 Base Algorithm: Expore-then-Gale-Shapley (view of arm a_j)

```
Input :Exploration rounds $2^l T_0$, Horizon $b^l T_0$
// Learning Preferences
for $t = \sum_{l'=1}^{l-1} 2^{l'} T_0 + 1, \ldots, \sum_{l'=1}^{l-1} 2^{l'} T_0 + 2^l T_0$ do
    Accept arm with estimated preferences
     Observe $Y^{(j)}_{A_j^{-1}(t)}(t)$ and update $\hat{\mu}^{(j)}_{A_j^{-1}(t)}, T^{(j)}_{A_j^{-1}(t)}$ if $\bar{A}_j^{-1}(t) = A_j^{-1}(t)$
end
Compute UCB$_n^{(j)}$ and LCB$_n^{(j)}$ for each $n \in [N]$
 if $\exists\beta$ such that LCB$^{(j)}_{\beta_n} >$ UCB$^{(j)}_{\beta_{n+1}}$ for any $n \in [N-1]$ then
    Preferences = $\beta$
else
    Preferences = arbitrary but fixed
end
// Perform Gale-Shapley using the Preferences $\beta^{(j)} = \beta = (\beta^{(j)}_1, \beta^{(j)}_2, \ldots, \beta^{(j)}_N)$ found
Propose using $\beta$ till acceptance
 Initialize $s = 1$
 for $t = 1, 2, \ldots, b^l T_0 - 2^l T_0$ do
    Accept player using the estimated preference $\beta^{(j)}$
     $s = s + 1$ if $\bar{A}_j^{-1}(t) == \emptyset$
end
```

Algorithm 7 Main Algorithm: Epoch-based `CA-ETC` (view of player a_j)

```
Input : Epoch length $T_0$, Parameter $\gamma \in (0,1)$ with $b = 2^{1/\gamma}$
for $l = 1, 2, \dots$ do
    Base Algorithm (Exploration rounds = $2^l T_0$, Horizon length = $b^l T_0$)
end
```

A.5.3. Analysis

Theorem A.10 (Regret of Algorithm 5)**.** *Suppose algorithm 5 is played for T iterations. Then, arm a_j incurs the regret of*

$$\overline{RA}_j(t) \leq \left(\frac{64K \log T}{\Delta^2} + K^2 + \frac{NK\pi^2}{3} \right) \bar{\Delta}^{(j)}_{\max}.$$

Theorem A.11 (Regret of Algorithm 7)**.** *Suppose `CA-ETC` is run with initial epoch length T_0 and input parameter $\gamma \in (0,1)$. Then, provided the initial epoch satisfies*

$$T_0 \gtrsim \left[\frac{32K \log T}{\Delta^2 (T-N)^\gamma} \right]^{\frac{1}{1-\gamma}},$$

the (arm-optimal) regret for arm a_j is given by

$$\begin{aligned} \overline{RA}_j(t) \leq\ & T_0 \left(\frac{64K \log T}{T_0 (\Delta^{(j)})^2} + 4 \right)^{1/\gamma} \Delta^{(j)}_{\max} \\ & + 2T_0 \left(\frac{T-N}{T_0} + 1 \right)^\gamma \Delta^{(j)}_{\max} + K^2 \gamma \log \left(\frac{T-N}{T_0} + 1 \right) \Delta^{(j)}_{\max} + \frac{NK\pi^2}{3} \Delta^{(j)}_{\max}. \end{aligned}$$

The proof of these follow the same steps as the proof for the players, and hence skipped.