CONTEXTUAL LINEAR BANDITS FOR MATCHING MARKETS

Anonymous authors
Paper under double-blind review

ABSTRACT

An extensive line of work studies bandit learning in two-sided matching markets, where the goal is to learn unknown preferences and converge to a stable matching through repeated interactions. Existing approaches predominantly focus on the tabular setting, where regret guarantees scale with the number of arms K. This dependence severely limits scalability in large markets commonly encountered in real-world applications. In this paper, we study contextual matching markets and leverage arm features to accelerate preference learning. For the strict preference setting, we propose a decentralized G-optimal design-based algorithm that achieves a player-optimal regret of $O\big(\max\{d\log T/\Delta^2, Nd\ \max_j \pi_i(j)\log T/\Delta^2\}\big)$, where N denotes the number of players, d denotes the feature dimension, T is the time horizon, $\pi_i(j)$ denotes the sampling probability of arm j and Δ is the preference gap. This result significantly improves scalability when $d \ll K$. We further extend our framework to matching markets with indifference and develop an arm-proposing exploration algorithm that achieves a regret of $O\left(Nd^2\log^2 T/\Delta^2\right)$. To the best of our knowledge, this is the first work that simultaneously addresses both contextual information and tied preferences. In addition, to characterize the hardness of our problem, we establish an instance-dependent regret lower bound of $\Omega\big(\max\{N\log T/\Delta^2,\ d\log T/\Delta\}\big)$. Our framework also recovers existing guarantees in the tabular setting as a special case. Experiments demonstrate that our methods substantially outperform existing approaches in large-scale markets.

1 INTRODUCTION

The two-sided matching market problem has been a widely investigated area in operations research due to its broad applications in college admissions, recommendation systems and labor markets (Fleiner, 2003; Gale & Shapley, 1962; Pagare & Ghosh, 2023; Wang et al., 2022; Jagadeesan et al., 2021). Specifically, the matching market comprises two sides, such as employers and employees in a labor market. Agents on either side have a personalized preference ranking over the other side. The goal of the two-sided matching market problem is to find a stable matching in which no pair of agents has an incentive to abandon their current matches and pursue a new one, hence promoting the market's viability and efficiency in the long term.

The seminal work of Gale & Shapley (1962) introduces an iterative matching strategy to find a stable matching, followed by works of Roth (1984); Roth & Sotomayor (1992); Kelso Jr & Crawford (1982). Most of these works assume that the preference ranking of both sides are known in advance, which may not hold in real-world applications. For example, in a labor market, employers cannot precisely evaluate employees' ability before they are hired. Recently, with the emergence of online market platforms such as Amazon Mechanical Turk and TaskRabbit, agents on the two sides repeatedly interact and learn their preferences from feedback.

The multi-armed bandit (MAB) framework is a classic framework featuring the iterative learning process of agents (Slivkins, 2019; Lattimore & Szepesvári, 2020). An MAB instance consists of a single player and K arms with associated rewards. The player has no prior knowledge about the preferences among arms and aims to find the optimal arm through an iterative learning process. In each time slot, the player pulls an arm and receives a noisy reward; subsequent decisions are updated based on the observed feedback.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Inspired by the success of multi-armed bandits in handling uncertainty, a large body of work has studied two-sided matching markets under bandit feedback (Liu et al., 2020; 2021; Ghosh et al., 2022; Basu et al., 2021). Existing research spans both centralized and decentralized settings, with the goal of learning preferences and converging to stable matchings. Most works focus on the strict preference setting and establish regret guarantees for player-optimal stable matchings. More recent efforts further extend the model to allow for indifference (i.e., preference ties) and design algorithms that converge to stable outcomes in this more general setting. Due to the space limitations, we compare the settings and regret guarantees of these works in Table 1 of Appendix A.

Despite these advances, most existing approaches operate in the tabular setting, where each player–arm pair is treated independently and must be explored separately. As a result, their regret bounds scale with the number of arms K (Kong et al., 2024; Lin et al., a; Kong et al., 2025). This inherent dependence on K severely limits scalability, especially in large markets where the number of arms is massive.

To mitigate the dependence on K in the upper bound and enhance exploration, we integrate linear contextual bandits into the matching market. Beyond rewards, this framework incorporates a context vector C with dimensionality d that yields additional information for each arm and a preference vector $\theta \in \mathbb{R}^d$ for each player (Lattimore & Szepesvári, 2020). The reward of arms can be represented as the inner product of the player's preference vector and the arm's context. Contextual features (e.g., an employee's age, education, or skills) usually satisfy $d \ll K$ in practice. By leveraging contextual linear bandits, each player-arm interaction fundamentally serves a dual purpose: beyond revealing the specific expected reward of the selected arm, the feedback(i.e., reward and context) concurrently contributes to estimating the preference vector θ, thereby sharing additional information of the expected rewards of other arms. Thus, information is utilized more efficiently, and the regret upper bound can scale with the context dimension d rather than the number of arms K.

Despite the potential of linear contextual bandits to improve information efficiency by generalizing across arms, extending this framework to matching markets is non-trivial. In particular, players must explore the shared linear feature space to eliminate the dependence on K, while simultaneously competing for arms in a multi-agent environment, leading to inevitable collisions. This coupling between efficient linear exploration and multi-player coordination makes it challenging to design exploration strategies that are both information-efficient and collision-robust.

In this paper, we overcome the above challenges and derive improved regret upper bounds in decentralized matching markets using the linear contextual bandit framework. Our contributions are summarized as: (1) We introduce a linear contextual bandit formulation for matching markets, in which players can learn their underlying preference vectors during interactions. (2) We design a decentralized G-optimal-design-based linear contextual explore-then-Gale-Shapley algorithm that enables players to learn their preference vectors by only exploring arms from a small subset of the original arm set, thereby facilitating exploration. We also employ a routing scheme inspired by the edge-coloring problem to avoid collisions during exploration, which completes the exploration using the minimum number of global time slots. This algorithm achieves a regret upper bound of $O\big(\max\{d\log T/\Delta^2, Nd\ \max_j \pi_i(j)\log T/\Delta^2\}\big)$, which outperforms state-of-the-art results (Zhang et al., 2022; Kong & Li, 2023; Kong et al., 2024) in large-scale markets. (3) We extend our approach to the indifference setting, as the original algorithm may incur linear regret when tied preferences occur. In this extended algorithm, we eliminate the stopping condition to enter Gale Shapley and enumerate all possible stable matchings (i.e., semi-stable matchings) at the beginning of each exploration, and only calculate the G-optimal design on the arms that can yield a semi-stable matching. In this way, once all players' underlying preferences are determined, any semi-stable matching becomes equivalent to a stable matching, ensuring stability in all subsequent rounds. This approach guarantees a stable matching with a regret bound of $O(Nd\log T/\Delta^2)$. (4) Since the extended algorithm requires enumerating all matchings and leads to exponential complexity, we further derive an adaptive LinearUCB-based algorithm utilizing the arm-proposing idea from (Kong et al., 2025). We integrate this idea with contextual linear bandits by setting the arms as the proposing side, players select the arm with the largest variance among proposing arms that have not been identified as suboptimal. Immediate communication upon preference updates reduces the additional regret incurred by phase exploration. These interactions eventually reveal players' preference rankings, facilitating optimal arm selection and a guaranteed stable matching with $O\left(Nd^2\log^2 T/\Delta^2\right)$ regret. Algorithms in (3) and (4) achieve K-independent regret, improving upon the results in (Kong et al., 2025). Notably, our

results for both strict and indifferent preferences recover existing tabular guarantees as special cases. (5) We establish a lower bound of $\Omega\big(\max\{N\log T/\Delta^2,\ d\log T/\Delta\}\big)$ to investigate the hardness of our problem. Extensive experiments show that our algorithms achieve lower regret comparing to existing works.

2 Related Work

Bandit Learning in Matching Markets. Das & Kamenica (2005) pioneered bandit learning in two-sided matching markets, providing empirical results under the strong assumption of homogeneous preferences. Liu et al. (2020) first theoretically addressed this via two paradigms: an explore-then-commit (ETC) algorithm that relies on prior knowledge of the minimum preference gap Δ to achieve $O(K\log T/\Delta^2)$ player-optimal stable regret, and a UCB algorithm yielding $O(NK\log T/\Delta^2)$ player-pessimal stable regret. Liu et al. (2021); Kong et al. (2022) extended this to decentralized settings, but yielded a player-pessimal regret of $O(\exp(N^4)N^5K^2\log^2 T/\Delta^2)$ due to communication difficulties. For markets with specific structures guaranteeing unique stable matchings, such as serial dictatorship (Sankararaman et al., 2021), the α-condition (Basu et al., 2021), and α-reducibility (Maheshwari et al., 2022), regret bounds of $O(NK\log T/\Delta^2)$ or $O(CNK\log T/\Delta^2)$ have been established. For generalized markets, state-of-the-art approaches include ETC-based algorithms (Kong & Li, 2023; Zhang et al., 2022) and the adaptive Explore-then-Gale-Shapley (ETGS) algorithm (Kong et al., 2024), achieving player-optimal stable regrets of $O(K\log T/\Delta^2)$ and $O(N^2\log T/\Delta^2 + K\log T/\Delta)$, respectively. However, as these stochastic bandit methods learn arm preferences independently, the K dependency in the main regret term remains inevitable.

Matching with Indifference.In practice, markets often exhibit tied preferences with zero gaps. Previous works (Das & Kamenica, 2005; Liu et al., 2020; 2021; Kong et al., 2022; 2024) typically assume distinct preferences, resulting in linear regret in these settings. Lin et al. (a) first investigated this via an offline α-approximate oracle and a centralized algorithm achieving an instance-independent α-approximate player-optimal stable regret. Aligned with our focus, Kong et al. (2025) proposed an adaptive exploration algorithm for indifference markets, achieving a state-of-the-art stable regret bound of $O(NK\log T/\Delta^2)$. However, due to individual exploration of player-arm pairs, the NK term limits its scalability in large markets.

Contextual Matching Markets. Parikh et al. (2024) incorporate contextual bandits into matching markets. They study the decentralized market with shifting environments and achieve a $O\left(d^2L^2\log T/(\kappa\Delta^2)\right)$ regret upper bound, where L and κ are context-dependent variables. However, they rely on a strong spectral assumption on the data matrix. While Lin et al. (b) propose a centralized batched adaptive regret-balancing algorithm, they use a probabilistic minimum preference gap different from the line of prior works (Liu et al., 2020; Sankararaman et al., 2021; Basu et al., 2021; Kong et al., 2024) that align with our objectives. Moreover, neither work (Parikh et al., 2024; Lin et al., b) directly applies to markets with indifference without an approximation factor. A detailed comparison with these works (Parikh et al., 2024; Lin et al., b) is provided in Appendix B.

3 Problem Setting

In this section, we introduce the problem setting of bandit learning in matching markets within the linear contextual bandit framework. The matching market comprises N players and K arms denoted by $\mathcal{N} = \{p_1, p_2, \ldots, p_N\}$ and $\mathcal{K} = \{a_1, a_2, \ldots, a_K\}$ respectively, where $N \leq K$. Each arm a_j is associated with a context vector $C_j \in \mathbb{R}^d$ with $\|C_j\|_2 \leq L$ (Abbasi-Yadkori et al., 2011). Each player p_i has an unknown preference parameter vector $\theta_i^* \in \mathbb{R}^d$ with $\|\theta_i^*\|_2 \leq M$. We assume that $\mathrm{span}\{C_1, C_2, ..., C_K\} = \mathbb{R}^d$ and $d \ll K$. Notice that this assumption is assumed only for illustration convenience, as it can be generalized into the non-spanning case following Chapter 21, Exercise 21.6 of Lattimore & Szepesvári (2020). The expected reward of arm a_j to player p_i is defined as the inner product of the player's preference vector and the arm's context vector $\mu_{i,j} = \langle \theta_i^*, C_j \rangle \in (0, 1]$ (Abbasi-Yadkori et al., 2011; Parikh et al., 2024). A larger expected reward indicates a stronger preference, i.e., $\mu_{i,j} < \mu_{i,j'}$ implies p_i prefers $a_{j'}$ to a_j, where $j \neq j'$. Since θ_i^* is unknown, preferences must be estimated through iterative interactions with arms (Abbasi-Yadkori et al., 2011; Lattimore & Szepesvári, 2020).

It is worth noting that most existing works (Liu et al., 2020; Sankararaman et al., 2021; Basu et al., 2021; Zhang et al., 2022; Kong & Li, 2023; Kong et al., 2024) assume preference values over different arms are distinct, i.e., $\mu_{i,j} \neq \mu_{i,j'}$ for any player p_i and arms $a_j, a_{j'}$. However, this assumption is often unrealistic in practical applications, as multiple arms may yield similar expected rewards, making it difficult for players to explicitly differentiate their preferences. Following recent advancements (Kong et al., 2025; Lin et al., a), we relax this assumption by allowing for indifference, where a player may receive identical expected rewards from different arms, i.e., $\mu_{i,j} = \mu_{i,j'}$ for $j \neq j'$. On the arm side, each arm a_j has a preference ranking over players. Denote $\pi_{j,i}$ as the rank of p_i in a_j's preference rankings. In matching market with indifference, arms may also have indifferent preferences. We use $\pi_{j,i} \prec \pi_{j,i'}$ to represent that a_j favors p_i more than $p_{i'}$, and $\pi_{j,i} = \pi_{j,i'}$ denotes that a_j is indifferent between the two players. We assume known arm-side preferences, a common setting in existing bandit-based matching frameworks (Liu et al., 2020; Sankararaman et al., 2021; Basu et al., 2021; Kong & Li, 2023; Kong et al., 2024).

At each time slot $t = 1, 2, 3...$, each player p_i proposes to an arm $A_i(t) \in \mathcal{K} \cup \{-1\}$, where -1 represents that player p_i does not propose to any arm in this round. Correspondingly, each arm a_j receives proposals from $A_j^{-1}(t) = \{p_i : A_i(t) = a_j\}$ and accepts $\bar{A}_j^{-1}(t) \in \arg\min_{p_i \in A_j^{-1}(t)} \pi_{j,i}$. When there are multiple choices, the agents break the tie by choosing the agent with the smallest index on the other side. Player p_i receives a reward: $X_{i,A_i(t)}(t) = \langle \theta_i^*, C_{i,A_i(t)} \rangle + \eta_t$ if its proposal is accepted, and η_t is a noise term following a 1-subgaussian distribution with zero mean (Lattimore & Szepesvári, 2020). If p_i's proposal is rejected by $A_i(t)$, it receives $X_{i,A_i(t)}(t) = 0$. We use $\bar{A}(t) = (p_i, \bar{A}_i(t))_{i \in [N]}$ to denote the matching outcome at time t. $\bar{A}_i(t) = A_i(t)$ if p_i is successfully matched and $\bar{A}_i(t) = -1$ otherwise. Players also receive context vectors, while unmatched players are assigned the zero vector.

We consider agents seeking a stable matching. For a matching outcome $\bar{A} := (\bar{A}_i)_{i \in [N]}$, if there exist two matching pairs $(p_i, \bar{A}_i)$, $(p_j, \bar{A}_j)$ such that $\mu_{i,\bar{A}_j} > \mu_{i,\bar{A}_i}$ and $\pi_{\bar{A}_j,i} \prec \pi_{\bar{A}_j,j}$, then p_i and $\bar{A}_j$ have the incentive to abandon their current partner and match with each other. Such pairs of $(p_i, \bar{A}_j)$ are referred to as blocking pairs. We define a stable matching as a matching without any blocking pair. Let $\mathcal{M} := \{m := (m_i)_{i \in [N]} : m \text{ is stable}\}$ denote the set of all stable matchings.

Under the setting of strict preferences, we define Δ as the minimum preference gap between the top N arms among all players and focus on the player-optimal regret defined as the difference between $\mu_{i,m_i^*} = \max_{m^* \in \mathcal{M}} \mu_{i,m_i^*}$ (the player's maximum reward that can be obtained in all stable matchings) and the received reward, i.e., $\text{Reg}_i(T) = \mathbb{E}[\sum_{t=1}^{T} (\mu_{i,m_i^*} - X_{i,A_i(t)})]$, where the expectation is with respect to randomness of players' actions and the reward distribution.

For markets with indifference, we define Δ as the minimum non-zero preference gap between all arms among all players. Notice that in matching markets with indifference, there is no concept of a player-optimal or player-pessimal stable matching (Kong et al., 2025). Thus, we focus on the stable regret defined as the difference between $\mu_{i,m_i} = \min_{m \in \mathcal{M}} \mu_{i,m_i}$ (the player's minimum reward that can be obtained in all stable matchings) and the received reward, i.e., $\text{Reg}_i(T) = \mathbb{E}[\sum_{t=1}^{T} (\mu_{i,m_i} - X_{i,A_i(t)})]$.

4 Algorithm for Markets with Strict Preferences

In this section, we investigate the setting with strict preference. Since all contexts reveal information of the preference vector in some direction of the vector space $\mathbb{R}^d$, the challenge lies in finding a set of informative arms that allow players to efficiently estimate their preference vector in every dimension.

Inspired by the G-optimal design problem (Atkinson, 2014), we propose a linear contextual bandit algorithm, named Contextual Explore-then-Gale-Shapley (CETGS), which efficiently learns the players' preference vectors through sampling only a small subset of arms and achieves a player-optimal stable matching. We present our algorithm for decentralized markets requiring inter-agent communication. In the absence of a central server that records the preferences on both sides, we follow existing works (Kong et al., 2024; 2025) by leveraging arm-proposing communication for agents.

Algorithm 1 demonstrates the decentralized CETGS algorithm from the view of player p_i. At the beginning, players initialize $\text{Stopping_Flag}_i = \text{False}$, where $\text{Stopping_Flag}_i = \text{True}$ means the

Algorithm 1 Decentralized contextual explore-then-Gale-Shapley(CETGS) from the view of p_i

Input: Player set $\mathcal{N}$, arm set $\mathcal{K}$, horizon T, confidence parameter δ
1: **Initialize:** $\text{Stopping_Flag}_i = \text{False}$, $\pi_{j,i} = -1, \text{index}_i = -1, \forall j \in [K],\ \forall i \in [N]$
2: **do:** $(\text{index}_i, \pi_{j,i}, C_j)_{i\in[N],\ j\in[K]} = \text{Index}\ (\mathcal{N}, \mathcal{K})$ as Algorithm 5
3: Find the G-optimal design by calculating $\pi_i = \arg\max_{\pi\in\mathcal{P}(\mathcal{K})} \log\det \sum_{j=1}^{K} \pi(j) C_j C_j^\top$.
4: **Set** $l = 0,\ t_l = 0$
5: **while** $\exists n \in [N]$ s.t. $\text{Stopping_Flag}_n == \text{False}$ **do**
6: Let $\xi_l = 2^{-l}, \quad (T_{i,j,l})_{i\in[N],\ j\in[K]} = \left\lceil \frac{2d\pi_i(j)}{\xi_l^2} \log\left(\frac{K}{\delta}\right) \right\rceil_{i\in[N],\ j\in[K]}$
7: $(T_l, \{C_{i,A_i(t)}, X_{i,A_i(t)}\}_{t=t_l}^{t_l+T_l}) = \text{Routing}\ (\pi_{j,i}, T_{i,j,l})_{i\in[N],\ j\in[K]}$
8: **Calculate:**
$V_{i,l} = \sum_{j\in[K]} T_{i,j,l} C_j C_j^\top, \quad \hat{\theta}_{i,l} = V_{i,l}^{-1} \sum_{t=t_l}^{t_l+T_l} C_{i,A_i(t)} X_{i,A_i(t)}(t)$
9: **for** any $j, j' \in [K]$ **do**
10: **if** $\langle \hat{\theta}_{i,l}, C_j - C_{j'} \rangle > 2\xi_l$ **then**
11: Set $\mu_{i,j} > \mu_{i,j'}$
12: **end if**
13: **end for**
14: **if** the first N ranking of player is determined **then**
15: **Set** $\text{Stopping_Flag}_i = \text{True}$
16: Select arm $= a_{\text{index}_i}$
17: **else**
18: Select arm $= -1$
19: **end if**
20: $t_{l+1} = t_l + T_l + 1,\ l = l + 1$
21: **end while**
22: **for** $p = 1, 2, \ldots, N$ **do**
23: **for** $j = 1, \ldots, N$ **do**
24: Let $\sigma_i = (\sigma_{i,1}, \ldots, \sigma_{i,N})$ be player i's top N arms, sorted s.t. $\mu_{i,\sigma_{i,1}} > \cdots > \mu_{i,\sigma_{i,N}}$
25: **If** $p = \text{index}_\text{i}$, select $\sigma_{i,j}$
26: **else** select arm $= -1$
27: **end for**
28: **end for**
29: $\bar{A} = \text{Offline GS}\left((\pi_{j,i})_{i\in[N],\ j\in[K]}, (\sigma_i)_{i\in[N]}\right)$
30: **for** $t = 1, 2, \ldots, T - NK - N^2 - t_l$ **do**
31: Select $A_i(t) = \bar{A}_i$
32: **end for**

ranking of the top N arms of player i is determined. Then players proceed to the index phase (Line 2) according to Algorithm 5 in Appendix C to accomodate the decentralized setting, followed by the exploration phase starting from Line 3. Since G-optimal and D-optimal designs are equivalent by Lemma J.3 in Appendix J, these two terms are considered synonymous in the following passages. Line 3 describes the process of searching for a G-optimal design over the arm set, where $\pi_i \in \mathcal{P}(\mathcal{K})$ denotes the arm sampling distribution, whose entries are the sampling proportions of each arm. Finding a G-optimal design enables us to construct a high-probability bound for the players' preference vector θ^* by sampling at most $d(d+1)/2$ arms, instead of the entire arm set. Subsequent to Line 4, the exploration phase can be further divided into super-phases with exponentially growing lengths. At the beginning of each super-phase, players compute the number of samples for each arm as specified in Line 6, where δ is a confidence parameter chosen as $1/T$. Then, they perform local routing (Line 7) following Algorithm 2 to ensure that all players can sample the exact amount of arms calculated in Line 6. During the routing process, players iteratively receive rewards and contexts from sampling different arms, which are then used to compute the estimated preference vector $\hat{\theta}$ (Line 8), and further determine their preference according to Line 9–13. At the end of each super-phase, players will communicate whether their top-N rankings are determined (Line 14-19). If player p_i has determined its top-N ranking, it sets $\text{Stopping_Flag}_i = \text{True}$ and selects the arm with index index_i obtained in the index phase to send this message (Line 16), while players that have not determined their top-N ranking choose -1 (Line 18). The super-phase index l is then updated (Line 20). If all of the

players have determined their top-N preference rankings denoted by $\sigma_i = (\sigma_{i,1} \dots \sigma_{i,N})$, i.e., every player selects the arm corresponding to their index during communication at the end of a super-phase, they share their top-N preferences by sequentially choosing arms from their preference lists over N^2 time slots (Line 22-28). After sharing their preference rankings, players enter the final local Gale-Shapley (GS) phase, where each player determines their proposed arm as specified in Algorithm 6 (Line 30–32), and a player-optimal stable matching is guaranteed at each time slot. Notably, since G-optimal design solutions may not be unique, we employ a tie-breaking rule to ensure coordination: lower-indexed players are prioritized for lower-indexed arms whenever feasible. Algorithm 6 is an offline version of the Gale-Shapley algorithm (Gale & Shapley, 1962) (see Appendix D). Lemma F.7 ensures stable matching optimality by showing that each player's optimal match necessarily falls within their top-N preferences.

Algorithm 2 Routing

Input: Required sample counts $(T_{i,j,l})_{i\in[N],\, j\in[K]}$
1: **Initialize:** $T_l = 0$, $t_{i,j,l} = T_{i,j,l}$, $\forall i \in [N],\ \forall j \in [K]$
2: **while** $\exists\, i \in [N], j \in [K]$ s.t. $t_{i,j,l} > 0$ **do**
3: Construct a bipartite multigraph $G_l(T_l)$, where p_i and a_j are connected by $t_{i,j,l}$ edges
4: Set $d_{p_i} = \sum_{j\in[K]} t_{i,j,l}$, $d_{a_j} = \sum_{i\in[N]} t_{i,j,l}$, and $\Delta_l = \max_{v\in\mathcal{N}\cup\mathcal{K}} d_v$
5: Find a matching M_l in $G_l(T_l)$ that matches all vertices with the maximum degree Δ_l
6: **for** $p_i \in \mathcal{N}$ **do**
7: **if** $(p_i, a_j) \in M_l$ for some $a_j \in \mathcal{K}$ **then**
8: Select $A_i(T_l) = a_j$ and receive $X_{i,A_i(T_l)}(T_l)$ and $C_{i,A_i(T_l)}$
9: **else**
10: Select $A_i(T_l) = -1$
11: **end if**
12: **end for**
13: Set $t_{i,j,l} = t_{i,j,l} - 1$ for each $(p_i, a_j) \in M_l$, $\quad T_l = T_l + 1$
14: **end while**
Output: $\Big(T_l, \{C_{i,A_i(t)}, X_{i,A_i(t)}\}_{t=t_l}^{t_l+T_l}\Big), \forall i \in [N]$

Algorithm 2 demonstrates the routing process. Specifically, we formulate this routing problem as an edge-coloring problem on a bipartite multigraph. This formulation enables us to construct a routing algorithm that completes the required sampling of all players using the minimum possible number of global time slots with the help of graph theory (Gabow & Kariv, 1978; König, 1916). This construction is formalized in Lemma F.5. For every exploration phase l, at each time slot T_l, we form a bipartite multi-graph $G_l(T_l)$ where player i and arm j are treated as vertices on opposite sides connected by $t_{i,j,l}$ parallel edges (Line 3). Then we find a matching M_l in $G_l(T_l)$ that matches all vertices with the maximum degree (Line 4-5). If player i matches arm j then we reduce $t_{i,j,l}$ by one, the algorithm then proceeds the next time slot (Line 13). The routing procedure will stop when all players have selected the required number of arm (Line 2).

4.1 Theoretical Results

Theorem 4.1. *Following the CETGS algorithm with $\delta = 1/T$, the player-optimal stable regret of each player p_i satisfies* $\mathrm{Reg}_i(T) \le O\big(\max\{d\log T/\Delta^2,\ Nd\ \max_j \pi_i(j)\log T/\Delta^2\}\big)$.

The proof of Theorem 4.1 is deferred to Appendix F. Existing algorithms (Liu et al., 2020; Kong & Li, 2023; Zhang et al., 2022; Kong et al., 2024) suffer from a regret dependency on K, as they learn the reward of each arm independently. Despite the state-of-the-art bounds of $O\left(K\log T/\Delta^2\right)$ (Zhang et al., 2022; Kong & Li, 2023) and $O\left(N^2\log T/\Delta^2 + K\log T/\Delta\right)$ (Kong et al., 2024), such a dependence on the arm-set size hinders scalability in massive markets. The core of CETGS lies in learning a unified preference vector rather than individual arm rewards. This allows each matching to inform the estimation of all arms, significantly enhancing information efficiency.

Theorem 4.2. *Under serial dictatorship, there exists an OSB instance with unit normal noise such that the regret of the lowest-ranked player p_n satisfies* $\mathrm{Reg}_n(T) \ge \Omega\big(\max\left\{N\log T/\Delta^2,\ d\log T/\Delta\right\}\big)$.

Due to space limits, the proof is deferred to Appendix G. The term $\Omega\big(N \log T/\Delta^2\big)$ captures the cost induced by matching collisions due to serial dictatorship while the term $\Omega(d \log T/\Delta)$ captures the statistical cost of identifying the optimal arm under a d-dimensional linear preference model.

Discussion of CETGS under the tabular case. With contexts reduced to orthogonal one-hot vectors in $\mathbb{R}^K$, this tabular case is equivalent to the stochastic bandit setting. As each context informs only one dimension of θ^*, players must sample all arms for accurate estimation, causing the G-optimal design to yield a uniform distribution with $\pi_i(j) = 1/K$. Hence, by replacing d with K, the upper bound degenerates into $O\big(K \log T/\Delta^2\big)$, matching the order of ETGS (Zhang et al., 2022; Kong & Li, 2023).

Extensions to the indifference setting. CETGS algorithm produces an optimal-stable matching after a stopping condition in Line 5 of Algorithm 1 is met, which requires that all players have distinguished their top N preferred arms. However, under indifference, the stopping condition may never be satisfied (e.g., the top $N + 1$ rewards of a player are the same), and the algorithm may suffer $O(T)$ regret. Recent works by Kong et al. (2025) propose an adaptive arm-proposing strategy for handling indifference where arms adaptively propose to players and players accept the optimal one from the proposed arms. However, this strategy cannot be directly integrated with the G-optimal design. One critical dilemma is that the available arm set for each player varies with time, causing arms to follow different G-optimal designs in different time slots. Therefore, whenever a player follows a new design, they need to resample other arms excluded from the previous design and the previously collected information is wasted to some extent. To address these challenges, we first present an extension of CETGS named Contextual Exploration with Indifference (CEWI). This approach is implemented by enumerating all matchings and only involves arms that result in a semi-stable matching to the available arm set to calculate G-optimal design at the beginning of each phase. CEWI achieves $O(Nd \log T/\Delta^2)$ regret, improving the $O(NK \log T/\Delta^2)$ bound of Kong et al. (2025) when $d \ll K$. The technical details of CEWI are deferred to Appendix H.

Theorem 4.3. *Following CEWI algorithm, the stable regret of each player p_i satisfies* $\mathrm{Reg}_i(T) \leq O\big(Nd \log T/\Delta^2\big)$. *If Δ = 0, then* $\mathrm{Reg}_i(T) \leq O(\log T)$ *for all players.*

5 Algorithm under Indifference with Computational Efficiency

For the CEWI algorithm, exhaustive enumeration of all matchings leads to exponential time complexity, which limits its practical scalability as the market size increases. The core challenge of removing such enumeration is that since players have no information of which arms are tied in advance, they cannot determine an appropriate available arm set at the beginning of each super-phase that will guarantee a stable matching when the rankings of all players are determined.

To address this challenge, we move beyond the G-optimal exploration and consider alternative exploration designs. Inspired by arm-proposing GS exploration by Kong et al. (2025), we designate the arms as the proposing side, where players select from those not yet identified as suboptimal by Linear-UCB. Integrating the arm-proposing mechanism into contextual bandits within matching markets presents further challenges, specifically in designing a sampling policy that bypasses G-optimal design while establishing efficient communication. While players in the tabular case choose arms with the fewest samples to target those with the most uncertainty, this heuristic fails in contextual linear bandits. Here, arms are structurally correlated via a shared feature space, whereas sampling counts alone are insufficient to capture the true uncertainty in arm estimation. We address this by selecting the arm with the largest estimated reward variance $\|C_j\|_{V_{i,t}^{-1}}$ to capture the estimation uncertainty under the Linear-UCB criterion.

Furthermore, in the tabular setting they considered, players explore in phases of exponentially increasing length. Throughout a given phase, players rely exclusively on information from preceding phases and decide to communicate at the end of the phase. Directly extending this tabular strategy to a contextual setting is problematic. This is because, in the tabular case, pulling the arm with the minimum sample count maintains a maximum sample difference of one among available arms, thereby incurring negligible regret even if certain arms are identified as suboptimal during the phase. In contrast, the variance-based selection in contextual linear bandits does not guarantee negligible regret, as players may repeatedly choose high-variance arms even after they are identified as suboptimal, thereby incurring large regret that scales poorly with the phase length. Alternatively, we develop an

instant-communication mechanism that enables real-time sharing of updated preference estimates, significantly reducing the overhead of phase exploration and improving the overall regret bound.

Drawing upon the aforementioned insights, we propose C-EAGS, a decentralized enumeration-free algorithm for markets with indifference that achieves computational efficiency at the modest cost of an additional $d \log T$ regret term compared to CEWI. Algorithm 3 details the C-EAGS procedure. We first define a preference matrix Better for each player p_i to characterize their pairwise preference relations over arms. $\mathrm{Better}(i, j, j') = 1$ indicates that player p_i has identified arm a_j as superior to $a_{j'}$, whereas a value of 0 denotes an undecided pairwise preference. The linear UCB and LCB of a_j for p_i at time slot t are defined as $\mathrm{UCB}_{i,t}(a_j) = \langle \hat{\theta}_{i,t}, C_j \rangle + \sqrt{\beta_t} \|C_j\|_{V_{i,t}^{-1}}$ and $\mathrm{LCB}_{i,t}(a_j) = \langle \hat{\theta}_{i,t}, C_j \rangle - \sqrt{\beta_t} \|C_j\|_{V_{i,t}^{-1}}$ by taking $\sqrt{\beta_t} = \sqrt{\lambda} M + \sqrt{2 \log T + d \log((d\lambda + tL^2)/d\lambda)}$. We initialize the $\mathbb{R}^{K \times K}$ matrix Better to zero, $\mathrm{Update_Flag}_i = \mathrm{False}$ and $\mathrm{Update_Pairs}_i = \emptyset$. Let $V_{i,0} = \lambda I_d$ with $\lambda > 0$, and $\mathrm{communicate} = \mathrm{False}$ indicates the players skip communication phase in the next time step. All players first enter the index phase (Line 2) following Algorithm 5, then determine whether to explore or communicate based on the previous time slot (Line 4-19). If no player updates its matrix Better, players run a local arm-guided Gale-Shapley (LAGS) algorithm based on the arms' preference $\pi_{j,i}$ and each player's matrix Better (Line 10) following Algorithm 4 to explore arms without collisions. Following each matching, players update $V_{i,t}$, $\hat{\theta}_{i,t}$ (Line 12-13), and the matrix Better using linear UCB and LCB (Line 14-18). If a new update occurs in $\mathrm{Better}(i, j, j')$, $\mathrm{Update_Flag}_i$ is set to True. If p_i updates its $\mathrm{Better}(i, j, j')$, then p_i will broadcast this information to other players (Line 4-9). Specifically, to enter a communication phase (Line 4-6), p_i selects $A_i(t) = -1$ to indicate it has made an update of its preference ranking. Other players can observe that p_i is unmatched and thereby recognize that they will enter a communication phase. The communication process follows Algorithm 7 (detailed in Appendix E). After that, players continue to explore and this process persists throughout the horizon.

Algorithm 3 Decentralize contextual exploration with arm-guided GS(C-EAGS) from the view of p_i

Input: Player set $\mathcal{N}$, arm set $\mathcal{K}$, horizon T.
1: Initialize: $\mathrm{Better}(i, j, j') = 0$, $V_{i,0} = \lambda I$, $\mathrm{communicate} = \mathrm{False}$, $\mathrm{Update_Flag}_i = \mathrm{False}$, $\mathrm{Update_Pairs}_i = \emptyset$, $\pi_{j,i} = -1$, $\mathrm{index}_i = -1$ for $\forall i \in [N], j, j' \in [K]$
2: **do:** $(\mathrm{index}_i, \pi_{j,i}, C_j)_{i \in [N],\ j \in [K]} = \mathrm{Index}\ (\mathcal{N}, \mathcal{K})$ as Algorithm 5
3: **for** $t = 1, 2, \ldots, T$ **do**
4: **if** $\exists i \in \mathcal{N}$ s.t. $\bar{A}_i(t-1) = -1 \wedge \mathrm{communication} = \mathrm{False}$ **then**
5: $\mathrm{communicate} = \mathrm{True}$
6: $\mathrm{Communication}(\mathrm{Update_Flag}_i, \mathrm{Update_Pairs}_i, \mathrm{Better}(i, j, j'))$ as Algorithm 7
7: **else if** $\mathrm{Update_Flag}_i = \mathrm{True}$ **then**
8: $A_i(t) = -1$
9: **else**
10: $m = \mathrm{LAGS}(\pi_{j,i}, V_{i,t-1}^{-1}, \mathrm{Better}(i, j, j'))_{i \in [N],\ j, j' \in [K]}$ as Algorithm 4
11: **Select** $A_i(t) = m_i$, receives $X_{i, A_i(t)}$
12: $V_{i,t} = V_{i,t-1} + C_{i,A_i(t)} C_{i,A_i(t)}^{\top}$
13: $\hat{\theta}_{i,t} = V_{i,t}^{-1} \sum_{s=1}^{t} C_{i,A_i(s)} X_{i,A_i(s)}$
14: **if** $\exists\ j, j' \in \mathcal{K}$ s.t. $\mathrm{UCB}_{i,t}(a_j) < \mathrm{LCB}_{i,t}(a_{j'})$, $\mathrm{Better}(i, j', j) = 0$ **then**
15: $\mathrm{Better}(i, j', j) = 1$
16: $\mathrm{Update_Flag}_i = \mathrm{True}$
17: $\mathrm{Update_Pairs}_i.\mathrm{add}((j', j))$
18: **end if**
19: **end if**
20: **end for**

Algorithm 4 implements the LAGS algorithm. Each arm a_j maintains two state variables: a proposing rank r_j, and the variable t_j^{-1} for its current temporary matching. Specifically, r_j is initialized to 1 and adaptively increased for exploration, while t_j^{-1} starts at -1 to indicate the arm is initially unmatched (Line 1-2). Correspondingly, at each round players will update their available arm set $\mathcal{A}_i$ based on arm proposals and t_i, the player temporarily matched to p_i. As long as an arm a_j remains unmatched and has not exhausted its top N preferred players (Line 3), it proposes to the player situated at the r_j-th position in its preference list (Line 4). The proposed player will receive this proposal, update its

Algorithm 4 Local arm-guided Gale-Shapley (LAGS) from the view of p_i

Input: $(\pi_{j,i})$, $V_{i,t}$, Better$(i, j, j'), \forall i \in [N],\ j, j' \in [K]$.
1: **Initialize:** $\forall i \in [N]$, set its available arm set $\mathcal{A}_i = \emptyset$, temporarily matched arm $t_i = -1$
2: **Initialize:** $\forall j \in [K]$, set current proposing ranking $r_j = 1$, temporarily matched player $t_j^{-1} = -1$
3: **while** $\exists a_j$ such that $t_j^{-1} = -1$ and $r_j \leq N$ **do**
4: Let p_i be the player ranked at position r_j, i.e., $p_i = \pi_{j,r_j}$
5: Update the available arm set: $\mathcal{A}_i = \mathcal{A}_i \cup \{a_j\}$
6: $\mathcal{E}_i = \{a_j \in \mathcal{A}_i : \exists j' \in \mathcal{A}_i \text{ s.t. Better}(i, j', j) = 1\}$
7: Update the temporarily matched arm as: $m_i = \arg\max_{j \in \mathcal{A}_i \setminus \mathcal{E}_i} \|C_j\|_{V_{i,t-1}^{-1}}$
8: Let a_p be the arm temporarily matched with p_i, i.e., $a_p = t_i$. Update $t_p^{-1} = p_i$
9: **for** all arms $a_{j'} \in \mathcal{A}_i$ s.t. $a_{j'} \neq t_i$ **do**
10: $r_{j'} = r_{j'} + 1$
11: $t_{j'}^{-1} = -1$
12: **end for**
13: **end while**
14: **Output:** matching outcome $m = (m_i)_{i \in [N]}$

available arm set and eliminate the suboptimal arms denoted by $\mathcal{E}_i$ based on its Better matrix (Line 5-6). The player will then select the arm with the largest $\|C_j\|_{V_{i,t-1}^{-1}}$ from the arms that have not yet been distinguished in $\mathcal{A}_i$ (denoted as $\mathcal{A}_i \setminus \mathcal{E}_i$), and both the player and the arm will update their temporary matches (Line 7-8). Subsequently, the arms that are rejected will increase their current proposing ranking by 1 and need to propose to other players in the following rounds (Line 9-12). The condition $N \leq K$ guarantees a full matching of all players (Line 14).

5.1 Theoretical Results

Theorem 5.1. *Following C-EAGS algorithm, the stable regret of each player p_i satisfies* $\mathrm{Reg}_i(T) \leq O\big(Nd^2 \log^2 T/\Delta^2\big)$. *If Δ = 0, then* $\mathrm{Reg}_i(T) \leq O(NK)$ *for all players.*

The proof of Theorem 5.1 is in Appendix I. This upper bound is independent of K in the main term, improving the state-of-the-art regret bound of $O(NK \log T/\Delta^2)$ by Kong et al. (2025).

Discussion of C-EAGS under the tabular case. In the tabular case where the context vectors are all orthogonal one-hot vectors in $\mathbb{R}^K$, $V_{i,t}$ degenerates into a diagonal matrix in $\mathbb{R}^{K \times K}$ where the j-th entry of the diagonal equals to the number of times a_j has been sampled by p_i. Thus, the arm-selection criterion of choosing $m_i = \arg\max_{j \in \mathcal{A}_i \setminus \mathcal{E}_i} \|C_j\|_{V_{i,t-1}^{-1}}$ is equivalent to choosing $m_i = \arg\min_{j \in \mathcal{A}_i \setminus \mathcal{E}_i} T_{i,j}$, which reduces our algorithm to that of Kong et al. (2025). Here $T_{i,j}$ denotes the number of time a_j has been selected by p_i.

Extensive experiments confirming our algorithm's efficacy are deferred to Appendix K.

6 Conclusion

In this paper, we extend bandit learning in two-sided matching to the linear contextual bandit framework, proposing upper bounds of lower orders in markets with both strict and indifferent preferences. For markets with strict preferences, we propose a decentralized algorithm integrating G-optimal exploration with Gale-Shapley, which facilitates the efficiency of the learning process, providing an optimal stable regret upper bound of $O\big(\max\{d \log T/\Delta^2, Nd \max_j \pi_i(j) \log T/\Delta^2\}\big)$ which eliminates the term K from the main term. Extending this to markets with indifference, we use enumeration to ensure long-term stability, achieving a stable regret of $O(Nd \log T/\Delta^2)$. To avoid exponential time complexity due to enumerating all matchings, we provide an alternative contextual arm-guided GS algorithm that achieves $O(Nd^2 \log^2 T/\Delta^2)$ regret. Both algorithms improve state-of-the-art results (Kong et al., 2025) by eliminating the K factor from the regret bound. We also establish a lower bound of $\Omega\big(\max\{N \log T/\Delta^2,\ d \log T/\Delta\}\big)$ to characterize the difficulty of the problem. In future work, we aim to design a computationally efficient algorithm achieving $O(Nd \log T/\Delta^2)$ regret in markets with indifference, a bound currently only theoretical for CEWI.

AI USE STATEMENT

In this work, we used generative AI tools for linguistic refinement of the manuscript, such as improving wording, grammatical correctness, and overall readability. We also use generative AI tools to organize the code repositories. They were not involved in the development of research ideas, methodological design, theoretical analysis, proof construction, experimental setup, or interpretation of the results. All AI-assisted revisions were examined by the authors, who remain fully responsible for the accuracy and final presentation of the manuscript.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Anthony C Atkinson. Optimal design. *Wiley StatsRef: Statistics Reference Online*, pp. 1–17, 2014.

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

Sanmay Das and Emir Kamenica. Two-sided bandits and the dating market. In *IJCAI*, volume 5, pp. 19, 2005.

Tamás Fleiner. A fixed-point approach to stable matchings and some applications. *Mathematics of Operations research*, 28(1):103–126, 2003.

Harold N Gabow and Oded Kariv. Algorithms for edge coloring bipartite graphs. In *Proceedings of the tenth annual ACM symposium on Theory of computing*, pp. 184–192, 1978.

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

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

Meena Jagadeesan, Alexander Wei, Yixin Wang, Michael Jordan, and Jacob Steinhardt. Learning equilibria in matching markets from bandit feedback. *Advances in Neural Information Processing Systems*, 34:3323–3335, 2021.

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

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

Fang Kong, Junming Yin, and Shuai Li. Thompson sampling for bandit learning in matching markets. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 3164–3170. International Joint Conferences on Artificial Intelligence Organization, 2022.

Fang Kong, Zilong Wang, and Shuai Li. Improved analysis for bandit learning in matching markets. *Advances in Neural Information Processing Systems*, 37:91904–91929, 2024.

Fang Kong, Jingqi Tang, Mingzhu Li, Pinyan Lu, John CS Lui, and Shuai Li. Bandit learning in matching markets with indifference. In *The Thirteenth International Conference on Learning Representations*, 2025.

Dénes König. Über graphen und ihre anwendung auf determinantentheorie und mengenlehre. *Mathematische Annalen*, 77:453–465, 1916. doi: 10.1007/BF01456961.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Shiyun Lin, Simon Mauras, Nadav Merlis, and Vianney Perchet. Stable matching with ties: Approximation ratios and learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, a.

Shiyun Lin, Simon Mauras, Vianney Perchet, and Nadav Merlis. Adaptive bandit algorithms for contextual matching markets. b.

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

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.

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

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

Satush Parikh, Soumya Basu, Avishek Ghosh, and Abishek Sankararaman. Competing bandits in decentralized contextual matching markets. *arXiv preprint arXiv:2411.11794*, 2024.

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.

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

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

Aleksandrs Slivkins. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

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

Yirui Zhang, Siwei Wang, and Zhixuan Fang. Matching in multi-arm bandit with collision. *Advances in Neural Information Processing Systems*, 35:9552–9563, 2022.

A TABLE OF COMPARISONS OF RELATED STATE-OF-THE-ART RESULTS

References	Regret bound	Setting	Holds under indifference?
Liu et al. (2020)	$O\left(\frac{K\log T}{\Delta^2}\right)*\#$	known Δ, gap_1	✓
	$O\left(\frac{NK\log T}{\Delta^2}\right)\#$	gap_2	***X***
Liu et al. (2021)	$O\left(\frac{N^5K^2\log^2 T}{\varepsilon N^4\Delta^2}\right)$	gap_2	***X***
Sankararaman et al. (2021)	$O\left(\frac{NK\log T}{\Delta^2}\right)$ $\Omega\left(\max\left\{\frac{N\log T}{\Delta^2}, \frac{K\log T}{\Delta}\right\}\right)$	serial dictatorship, gap_1	***X***
Basu et al. (2021)	$O\left(K\log^{1+\varepsilon}T + 2^{(\frac{1}{\Delta^2})^{\frac{1}{\varepsilon}}}\right)*$	gap_2	✓
Maheshwari et al. (2022)	$O\left(\frac{CNK\log T}{\Delta^2}\right)$	α-reducible condition, gap_1	***X***
Kong et al. (2022)	$O\left(\frac{N^5K^2\log^2 T}{\varepsilon N^4\Delta^2}\right)$	gap_2	***X***
Zhang et al. (2022)	$O\left(\frac{K\log T}{\Delta^2}\right)*$	gap_2	***X***
Kong & Li (2023)	$O\left(\frac{K\log T}{\Delta^2}\right)*$	gap_3	***X***
Kong et al. (2024)	$O\left(\frac{N^2\log T}{\Delta^2}+\frac{K\log T}{\Delta}\right)*$	gap_4	***X***
	$O\left(\frac{N\log T}{\Delta^2}+\frac{K\log T}{\Delta}\right)\#$	α-condition, gap_3	***X***
Parikh et al. (2024)	$O\left(\frac{d^2L^2\log T}{\kappa\Delta^2}\right)*$	gap_3, spectrum assumption	***X***
Lin et al. (b)	$O\left(\frac{Nd^2\log^2 T}{\Delta^2}\right)*$	gap_2, shifting contexts	***X***
	$O\left(\frac{Nd^2\log^2 T}{\Delta^2}\right)$	adversarial contexts, α-approximate regret	✓
Kong et al. (2025)	$O\left(\frac{NK\log T}{\Delta^2}\right)$	gap_2(non-zero)	✓
Ours	$O\left(\max\left\{\frac{d\log T}{\Delta^2}, \max_j \frac{Nd\,\pi_i(j)\log T}{\Delta^2}\right\}\right)$	gap_3	***X***
	$\Omega\left(\max\left\{\frac{N\log T}{\Delta^2}, \frac{d\log T}{\Delta}\right\}\right)$	serial dictatorship, gap_1	
	$O\left(\frac{Nd\log T}{\Delta^2}\right)$	gap_2(non-zero)	✓
	$O\left(\frac{Nd^2\log^2 T}{\Delta^2}\right)$	gap_2(non-zero)	✓

Table 1: Comparisons of related state-of-the-art results. Regret bounds marked with $*$ denote the player-optimal stable regret, whereas those without $*$ denote the player-pessimal stable regret in markets with strict preference and stable regret in markets with indifference; bounds marked with **#** are derived under the centralized setting. Throughout, N denotes the number of players, K is the number of arms, d denotes the dimension of the context vector satisfying $d \ll K$, and T is the time horizon. Additionally, L denotes the maximum norm of all context vectors, Δ is the minimum preference gap between arms, ε denotes an algorithm-dependent parameter, κ denotes the parameter corresponding to the context vector constraint, $\pi_i(j)$ is as sampling distribution $\in \left[\frac{2}{d^2+d}, 1\right]$ and C is a parameter associated with the unique stable matching condition. Notice that the choice of Δ varies across different works. Specifically gap_1 denotes the minimum preference gap between the player-optimal stable arm and the next arm in the preference ranking over all players; gap_2 denotes the minimum preference gap between any two distinct arms across all players; gap_3 denotes the minimum preference gap among the top $N{+}1$ ranked arms across all players; and gap_4 denotes the minimum preference gap among arms ranked above the next arm of the player-optimal stable arm in the preference ranking across all players. One can verify that $\text{gap}_1 \geq \text{gap}_4 \geq \text{gap}_3 \geq \text{gap}_2$ holds. Notice in Kong et al. (2025) and our work, due to the appearance of tied arms, in the indifference setting we consider gap_2 as the minimum non-zero preference gap between all arms among all players.

B Comparison With Previous Contextual Matching Literature

In this section, we provide detailed comparisons with the two closely related works proposed by Lin et al. (b) and Parikh et al. (2024).

Comparison with Lin et al. (b). Lin et al. (b) and our work differ along four fundamental dimensions: the contextual model, the information structure, the resulting regret guarantee, and the focus on the indifference setting.

First, Lin et al. (b) consider a stochastic-context model in which contexts are repeatedly sampled from a fixed distribution. In contrast, we study a stationary contextual matching market, where each arm is associated with a fixed feature vector throughout the learning horizon. Second, Lin et al. (b) study a centralized setting in which a central server coordinates the players' actions and aggregates the information collected across the market. Our algorithms, by contrast, operate in a decentralized setting. Each player makes decisions without centralized coordination. Third, the two works adopt different notions of the preference gap and achieve different theoretical guarantees. Lin et al. (b) define

$$\delta_{\min}^{(i)} = \min_{\substack{j,j' \in [K] \\ j \neq j'}} \left| \theta_i^\top (C_j - C_{j'}) \right|, \qquad \delta_{\min} = \min_{i \in [N]} \delta_{\min}^{(i)},$$

and introduce the probabilistic minimum preference gap

$$\Delta_{\min} := \sup \left\{ \Delta : \mathbb{P}(\delta_{\min} \geq \Delta) \geq 1 - \frac{\log T}{T\Delta^2} \right\}.$$

When their stochastic-context model is specialized to the stationary setting considered in our paper, this quantity reduces to the smallest pairwise preference gap, denoted by Δ_2 in Appendix A. In contrast, our regret analysis depends on the less restrictive gap Δ_3, where $\Delta_3 > \Delta_2$.

These differences are reflected in the resulting regret guarantees. Lin et al. (b) establish a regret bound of $O\left(\frac{Nd^2 \log^2 T}{\Delta_2^2}\right)$, whereas CETGS achieves $O\Big(\max\left\{\frac{d \log T}{\Delta^2}, \max_j \frac{Nd\,\pi_i(j) \log T}{\Delta^2}\right\}\Big)$. Accordingly, in terms of the dependence on d and T, our bound improves upon that of Lin et al. (b) by a factor of at least $d \log T$. Moreover, our guarantee depends on the more favorable gap Δ_3 rather than the minimum pairwise gap Δ_2. Importantly, these improvements are achieved under a decentralized information structure, rather than the centralized setting assumed by Lin et al. (b).

Finally, for the setting with indifference, Lin et al. (b) adopt an α-approximate optimal stable regret, under which each player is guaranteed an α fraction of her optimal stable utility. We study exact stable regret relative to the player's least preferred stable matching without introducing an approximation factor.

Comparison with Parikh et al. (2024) Parikh et al. (2024) study contextual matching markets with a finite collection of unknown environments. Their framework imposes substantially stronger structural assumptions on the arm features and does not directly extend to markets with indifference.

First, Assumption 2 of Parikh et al. (2024) requires the K arms to be partitioned into multiple groups, with each group containing d linearly independent feature vectors. This condition, known as the spectrum assumption, effectively requires the arm set to contain multiple disjoint full-rank subsets. Such a requirement can be restrictive when the arm features are highly correlated, redundant, or concentrated in only a small number of informative directions.

Our formulation avoids this requirement: it only imposes a mild global spanning condition on the arm features. Moreover, the algorithms in Parikh et al. (2024) fail to hold under the indifference setting where repetitive exploration may incur linear regret.

Since our work and Parikh et al. (2024) consider substantially different environmental models, structural assumptions, and learning objectives, we do not regard their regret guarantees as directly comparable.

C INDEX PHASE

Algorithm 5 aims to learn a unique index for all players, acquire the context vector of each arm, and broadcast the preference of each arm to all players. In the first time slot of phase j, all players propose to arm a_j (Line 1-4) and only the most preferred player will be accepted and receives context vector C_j. This particular player will then receive index 1 and receive context C_j (Line 5) and all players infer from the observed matching outcome of a_j that a_j favors this player the most. During phase j, players who have not matched with a_j will re-propose to a_j and this procedure terminates after N rounds, once all players have been matched with a_j. After executing Algorithm 5, players will record locally the preference ranking of all arms and receive an index for later communications.

Algorithm 5 Index phase from the view of p_i

Input: Player set $\mathcal{N}$, arm set $\mathcal{K}$
1: **for** $j \in [K]$ **do**
2: Arm $= a_j$
3: **for** round $\tau = 1, 2, \ldots, N$ **do**
4: $A_i(\tau) =$ Arm
5: Set Arm $= -1$ and $\text{index}_i = \tau$ **if** accepted by $A_i(\tau)$ and receive context C_j
6: Update $\pi_{j,\bar{A}_j^{-1}(\tau)} = \tau$
7: **end for**
8: **end for**
9: **Output**: $\text{index}_i, (\pi_{j,i})_{j\in[K]}, (C_j)_{j\in[K]}$

D OFFLINE GALE-SHAPLEY

Algorithm 6 Offline GS

Input: $(\pi_{j,i})_{j\in[K],i\in[N]}, (\sigma_i)_{i\in[N]}$
1: **Initialize:** $s_i = 1,\ \bar{A}_i = -1, \forall i \in [N]$
2: **while** $\exists n \in [N]$ s.t. $\bar{A}_n = -1$ **do**
3: **for** $\forall i \in [N]$ **do**
4: $A_i = a_{\sigma_{i,s_i}}$
5: $a_{\sigma_{i,s_i}}$ selects $p \in \arg\min_{p_i \in A^{-1}_{\sigma_{i,s_i}}} \pi_{\sigma_{i,s_i},i}$
6: $s_i = s_i + 1$ **if** $\bar{A}_i = -1$
7: **end for**
8: **end while**
9: **return** $\bar{A} = (\bar{A}_1, \bar{A}_2, ..., \bar{A}_N)$

E COMMUNICATION

The communication process is shown in Algorithm 7 where players broadcast their updates of their Better matrix one-by-one based on the order of their index. If the player with the current index has no updates to broadcast, the index counter is increased by one to allow the next player to broadcast as shown in Line 3-5. If a player has information to update, for all $\mathrm{Update_Pairs}(j, j')$, to signal a preference for a_j over $a_{j'}$, the player sequentially selects a_j and $a_{j'}$ across two rounds as shown in Line 8-11. Other players observe the matching outcomes of both slots and record the results accordingly (Line 12). After completing its broadcast, the player selects no arm, as shown in 14 and set $\mathrm{Update_Flag}_i = \mathrm{False}$, $\mathrm{Update_Pairs}_i = \{\ \}$ as shown in Line 16-17, and the next player starts to broadcast in the following rounds. After communicating, players set $\mathrm{communicate} = \mathrm{False}$ (Line 22).

Algorithm 7 Communication

Input: Update_Flag$_i$, Update_Pairs$_i$, Better(i, j, j') $\forall i \in [N], j, j' \in [K]$.
Initialize: $P = 1$
while $P \leq N$ **do**
  if for p_i with index$_i = P$, Update_Flag$_P$ = False **then**
    $P = P + 1$
  else
    for player p_i with index$_i = P$, $T = 0$ **do**
      for $(j, j') \in$ Update_Pairs$_i$ **do**
        $A_i(T) = a_j$
        $T = T + 1$
        $A_i(T) = a_{j'}$
        $T = T + 1$
        Update Better$(i, j, j') = 1$
      end for
      $A_i(T) = -1$
      $T = T + 1$
      Update_Flag$_i$ = False
      Update_Pairs$_i$ = { }
      $P = P + 1$
    end for
  end if
end while
communicate = False

F Proof of Theorem 4.1

For convenience, with a slight abuse of notation, let $\hat{\theta}_{i,l}$ denote the value of $\hat{\theta}_i$ at the end of phase l. We define the failure event

$$\mathcal{F} = \left\{ \exists i \in [N], j \in [K], |\langle \hat{\theta}_{i,l} - \theta_i^*, C_j \rangle| > \xi_l \right\} \tag{1}$$

which represents the event that some player's estimated preference value for some arm deviates significantly from its true value at the end of phase l. Let l_{max} correspond to the super-phase when the top-N rankings of all players are determined. Let $\Delta_{\min} > 0$ denote the minimum preference gap among the top $N + 1$ arms for all players (referred to as gap$_3$), where the positivity of $\Delta_{\min}$ is guaranteed by the absence of indifference. The stable regret can be decomposed as follows:

$$\begin{aligned}
\text{Reg}_i(T) &= \mathbb{E}\left[\sum_{t=1}^{T}(u_{i,m_i^*} - X_{i,A_i(t)}(t))\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\{\{\bar{A}(t) \neq m^\star\}\}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\{\{\bar{A}(t) \neq m^\star\}\} \mid \neg\mathcal{F}\right] + \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\{\mathcal{F}\}\right] \\
&\leq NK + 256\max\left\{\frac{d\log T}{\Delta_{\min}^2},\ \max_j \frac{Nd\,\pi_i(j)\log T}{\Delta_{\min}^2}\right\} + \frac{256Nd\log K}{\Delta_{\min}^2} \\
&\quad + \frac{Nd(d+1)\log\frac{8}{\Delta_{\min}}}{2} + \log T + N^2 + \frac{N\log T}{T},
\end{aligned}$$

where the last inequality holds from Lemma F.1 and F.2.

It is worth noting that, as summarized in Appendix A, Table 1, the terms $\log T/\Delta^2$ and $\log T/\Delta$ are typically regarded as the leading terms in the matching market literature, as T is assumed sufficiently large while Δ may be sufficiently small. Therefore, despite the presence of lower-order K-dependent terms such as NK, our result can be upper bounded by $O\big(\max\{d\log T/\Delta^2, Nd\ \max_j \pi_i(j)\log T/\Delta^2\}\big)$, which is independent of K.

Lemma F.1. *By taking $\delta = 1/T$, we have*

$$\mathbb{E}\left[\sum_{l=0}^{l_{\max}} \mathbb{I}(\mathcal{F})\right] \leq \frac{N \log T}{T}.$$

Proof. Recall that $\mathcal{F}$ is defined as Eq. 1. Then,

$$\begin{aligned}
\mathbb{E}\left[\sum_{l=0}^{l_{\max}} \mathbb{I}(\mathcal{F})\right] &= \mathbb{E}\left[\sum_{l=0}^{l_{\max}} \mathbb{I}\left\{\exists i \in [N], j \in [K], |\langle \hat{\theta}_{i,l} - \theta_i^*, C_j\rangle| > \xi_l\right\}\right] \\
&= \mathbb{E}\left[\sum_{l=0}^{l_{\max}} \sum_{i=1}^{N} \sum_{j=1}^{K} \mathbb{I}\left\{|\langle \hat{\theta}_{i,l} - \theta_i^*, C_j\rangle| > \xi_l\right\}\right] \\
&\leq \log T \cdot NK \cdot \frac{\delta}{K} \qquad (2) \\
&= \log T \cdot NK \cdot \frac{1}{TK} \qquad (3) \\
&= \frac{N \log T}{T},
\end{aligned}$$

where Eq.2 follows by Lemma J.1, J.2, J.4, and the fact that there exist at most $\log T$ super-phases which will be discussed in Lemma F.3. Eq.3 follows by taking $\delta = 1/T$.

□

Lemma F.2.

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\{\bar{A}(t) \neq m^\star\} \,\middle|\, \neg\mathcal{F}\right] \leq NK + \log T + N^2 + 256 \max\left\{\frac{d \log T}{\Delta_{\min}^2}, \max_j \frac{Nd\,\pi_i(j) \log T}{\Delta_{\min}^2}\right\} \\
+ \frac{256 Nd \log K}{\Delta_{\min}^2} + \frac{Nd(d+1)\log\left(\frac{8}{\Delta_{\min}}\right)}{2}.
\end{aligned}$$

Proof. The regret can be further decomposed into the regret accumulated during the index phase (Algorithm 1 Line 2), communication (Algorithm 1 Line 14-19) and exploration (Algorithm 1 Line 4-13). Denote t_l(communication) as the number of time slots running on communication after super-phase l. τ_t(Index phase) is an indicator indicating whether the current time slot is running the index phase, where τ_t(Index phase) $= 1$ if time slot t is running the index phase and 0 otherwise.

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\{\{\bar{A}(t) \neq m^\star\}\} \mid \neg\mathcal{F}\}\right] \\
\leq &\mathbb{E}\left[\sum_{t=1}^{T} \tau_t(\text{Index phase}) \mid \neg\mathcal{F}\right] + \mathbb{E}\left[\sum_{l=0}^{l_{\max}} t_l(\text{communication}) \mid \neg\mathcal{F}\right] \\
&+ \mathbb{E}\left[\sum_{l=0}^{l_{\max}} \sum_{\tau=1}^{T_l} \mathbb{I}\{\{\bar{A}(t) \neq m^\star\}\} \mid \neg\mathcal{F}\right] \\
\leq &\mathbb{E}\left[\sum_{t=1}^{T} \tau_t(\text{Index phase}) \mid \neg\mathcal{F})\right] + \mathbb{E}\left[\sum_{l=0}^{l_{\max}} t_l(\text{communication}) \mid \neg\mathcal{F}\right] \\
&+ \mathbb{E}\left[\sum_{l=0}^{l_{\max}} \sum_{\tau=1}^{T_l} \mathbb{I}\Big(\exists i \in [N], j \in [K] \text{ s.t. } u_{i,j} > u_{i,A_i(t)} \text{ and } \pi_{j,i} \prec \pi_{j,A_j(t)} \mid \neg\mathcal{F}\Big)\right] \\
\leq &NK + \log T + N^2 + 256 \max\left\{\frac{d \log T}{\Delta_{\min}^2},\ \max_j \frac{Nd\,\pi_i(j) \log T}{\Delta_{\min}^2}\right\}
\end{aligned}$$

$$+\frac{256Nd\log K}{\Delta_{\min}^2}+\frac{Nd(d+1)\log\frac{8}{\Delta_{\min}}}{2}. \tag{4}$$

For Eq.equation 4, the regret for the index phase can be bounded by NK since during this process, N players sequentially match with K arms occupying NK rounds. The regret bound for communication and exploration follows by Lemmas F.3 and F.4. □

Lemma F.3.

$$\mathbb{E}\left[\sum_{l=0}^{l_{\max}} t_l(\text{communication}) \mid \neg\mathcal{F}\right] \le \log T + N^2.$$

Proof. Since communication happens only for one time slot at the end of each super-phase as shown in Line 14-19 of Algorithm 1, and the super-phases grow exponentially by the choice of $\xi_l = 2^{-l}$, it holds that there are at most $\log T$ phases and the communication time is bounded by $\log T$. Furthermore, since N players will communicate their top-N preference rankings one-by-one as shown in Line 22-28 of Algorithm 1, the regret for this procedure can be bounded by N^2. □

Lemma F.4.

$$\mathbb{E}\left[\sum_{l=0}^{l_{\max}}\sum_{\tau=1}^{T_l}\mathbb{I}\left\{\{\bar{A}(t)\neq m^\star\}\right\}\,\middle|\,\neg\mathcal{F}\right] \le 256\max\left\{\frac{d\log T}{\Delta_{\min}^2},\ \max_j\frac{Nd\,\pi_i(j)\log T}{\Delta_{\min}^2}\right\} + \frac{256Nd\log K}{\Delta_{\min}^2}+\frac{Nd(d+1)\log\frac{8}{\Delta_{\min}}}{2}.$$

Proof.

$$\begin{aligned}
&\mathbb{E}\left[\sum_{l=0}^{l_{\max}}\sum_{\tau=1}^{T_l}\mathbb{I}\left\{\{\bar{A}(t)\neq m^\star\}\right\}\,\middle|\,\neg\mathcal{F}\right]\\
&\le \mathbb{E}\left[\sum_{l=0}^{l_{\max}}\sum_{\tau=1}^{T_l}1\,\middle|\,\neg\mathcal{F}\right]\\
&\le \mathbb{E}\left[\sum_{l=0}^{l_{\max}}\max\left\{\frac{2d}{\xi_l^2}\log\left(\frac{K}{\delta}\right),\ \max_j\frac{2Nd\,\pi(j)}{\xi_l^2}\log\left(\frac{K}{\delta}\right)\right\}+\frac{Nd(d+1)}{2}\,\middle|\,\neg\mathcal{F}\right] \qquad (5)\\
&= \mathbb{E}\left[\sum_{l=0}^{l_{\max}}\max\left\{\frac{2d}{\xi_l^2}\log\left(\frac{K}{\delta}\right),\ \max_j\frac{2Nd\,\pi(j)}{\xi_l^2}\log\left(\frac{K}{\delta}\right)\right\}\,\middle|\,\neg\mathcal{F}\right]+\mathbb{E}\left[\sum_{l=0}^{l_{\max}}\frac{Nd(d+1)}{2}\,\middle|\,\neg\mathcal{F}\right].
\end{aligned}$$

For Eq. 5, the term $Nd(d+1)/2$ derives from Lemma J.3, while the first term of Eq. 5 comes from Lemma F.5. For convenience, we abbreviate $\pi_i(j)$ as $\pi(j)$ since each player shares the same sampling distribution.

By Lemma F.6, under the event $\neg\mathcal{F}$, the stopping condition for the algorithm to enter GS is that $\xi_l < \frac{\Delta_{\min}}{4}$. Notice by the choice of $\xi_l = 2^{-l}$, it holds that $\xi_{l_{\max}-1} > \frac{\Delta_{\min}}{4}$ then $\xi_{l_{\max}} > \frac{\Delta_{\min}}{8}$. Then we have $2^{-l_{\max}} > \frac{\Delta_{\min}}{8}$. Hence

$$\begin{aligned}
&\sum_{l=0}^{l_{\max}}\max\left\{\frac{2d}{\xi_l^2}\log\left(\frac{K}{\delta}\right),\ \max_j\frac{2Nd\,\pi(j)}{\xi_l^2}\log\left(\frac{K}{\delta}\right)\right\}\\
&\quad< 2\max\left\{\frac{2d}{\xi_{l_{\max}}^2}\log\left(\frac{K}{\delta}\right),\ \max_j\frac{2Nd\,\pi(j)}{\xi_{l_{\max}}^2}\log\left(\frac{K}{\delta}\right)\right\}\\
&\quad< 2\max\left\{\frac{2d}{\frac{\Delta_{\min}^2}{64}}\log\left(\frac{K}{\delta}\right),\ \max_j\frac{2Nd\,\pi(j)}{\frac{\Delta_{\min}^2}{64}}\log\left(\frac{K}{\delta}\right)\right\}\\
&\quad= \max\left\{\frac{256d}{\Delta_{\min}^2}\log\left(\frac{K}{\delta}\right),\ \max_j\frac{256Nd\,\pi(j)}{\Delta_{\min}^2}\log\left(\frac{K}{\delta}\right)\right\}.
\end{aligned}$$

By taking δ as $\frac{1}{T}$ and the fact that $l_{\max} < \log \frac{8}{\Delta_{\min}}$ by $2^{-l_{\max}} > \frac{\Delta_{\min}}{8}$, we obtain the desired result. □

Lemma F.5. *The number of time slots for each phase l is:*

$$\max\left\{\left\lceil \frac{2d}{\xi_l^2}\log\left(\frac{K}{\delta}\right)\right\rceil,\ \max_j N\left\lceil \frac{2d\,\pi(j)}{\xi_l^2}\log\left(\frac{K}{\delta}\right)\right\rceil.\right\}. \tag{6}$$

Proof. We represent the matching procedure in each phase l by a bipartite multigraph G_l, with players on one side and arms on the other. Each player-arm pair (i, j) is connected by $T_{i,j,l}$ parallel edges, where each edge represents one required match.

One can see that scheduling all required collision-free matches is equivalent to partitioning the edge set of G_l into different colors where there does not exist two edges sharing the same color that are connected to the same vertex (each player and arm can only match with one agent on the other side per time slot). Figure 1 provides an intuitive sketch of this idea.

By Lemma J.6, the minimum number of colors $\chi'(G_l)$ needed is:

$$\chi'(G_l) = \Delta(G_l) = \max\left\{\max_i \sum_j T_{i,j,l},\ \max_j \sum_i T_{i,j,l}\right\}. \tag{7}$$

which equals the maximum vertex degree in the bipartite graph $\Delta(G_l)$. Since $\sum_j \pi(j) = 1$ and all players use the same distribution, we have

$$\sum_j T_{i,j,l} = \left\lceil \frac{2d}{\xi_l^2}\log\left(\frac{K}{\delta}\right)\right\rceil, \qquad \sum_i T_{i,j,l} = N\left\lceil \frac{2d\,\pi(j)}{\xi_l^2}\log\left(\frac{K}{\delta}\right)\right\rceil.$$

Substituting these identities into Eq. 7 establishes the result. The optimality of Algorithm 2 follows directly from the proof and results of Gabow & Kariv (1978).

□

Lemma F.6. *Under $\neg\mathcal{F}$, $\xi_l < \frac{\Delta_{\min}}{4}$ implies that the top N arms of all players are ranked.*

Proof. For any player i and any two arms j and j' with $\mu_{i,j} > \mu_{i,j'}$ where j and j' are arms in player i's top N arms, let $\Delta_{i,j,j'}$ be $\mu_{i,j} - \mu_{i,j'}$.

$$\begin{aligned}
\langle \hat{\theta}_{i,l}, (C_j - C_{j'})\rangle &= \langle (\hat{\theta}_{i,l} - \theta_i^*), (C_j - C_{j'})\rangle + \Delta_{i,j,j'} \\
&= \langle (\hat{\theta}_{i,l} - \theta_i^*), C_j\rangle - \langle (\hat{\theta}_{i,l} - \theta_i^*), C_{j'}\rangle + \Delta_{i,j,j'} \\
&\geq -2\xi_l + \Delta_{i,j,j'} \\
&> 2\xi_l, \quad \text{since } \Delta_{i,j,j'} \geq \Delta_{\min} \text{ and } \xi_l < \frac{\Delta_{\min}}{4}.
\end{aligned}$$

By Algorithm 1 Line 9-13, player p_i ranks a_j above $a_{j'}$. □

Lemma F.7. *In the optimal stable matching generated by the GS algorithm, each player is guaranteed to match with one of its top-N preferred arm.*

Proof. Under the Offline GS in Algorithm 6, a player will only propose to the arm ranked R-th in its preference ranking if it is rejected by all of the top R arms which are all occupied by other players. By contradiction, if in the final matching the matched arm of a player is not its top N preferred arm, then all of the top N arms are occupied by other players. However, there are only $N - 1$ players, and we meet a contradiction. □

G Proof of Theorem 4.2

In this section, we first give the definitions of serial dictatorship and optimally stable bandits (OSB).

Definition 1 (Serial Dictatorship)**.** *A matching market satisfies* serial dictatorship *if all arms have an identical strict preference ranking over players, i.e.,*

$$\pi_{j,1} \prec \pi_{j,2} \prec \cdots \prec \pi_{j,N}, \qquad \forall a_j \in \mathcal{K},$$

after reindexing the players according to their rankings without loss of generality.

Definition 2 (Optimally Stable Bandits (OSB))**.** *A bandit instance is an optimally stable bandit (OSB) instance if each player is matched with its optimal arm under the unique stable matching, i.e.,*

$$m_i^* = \arg\max_{a_j \in \mathcal{K}} \mu_{i,j}, \qquad \forall p_i \in \mathcal{N}.$$

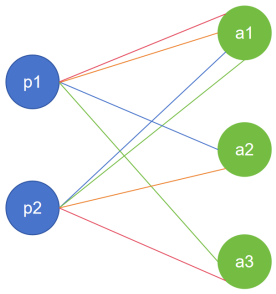


Figure 1: An example of a 2×3 bipartite graph with $\chi'(\mathrm{G}_l) = \Delta(\mathrm{G}_l) = 4$.

Note that the serial dictatorship and OSB conditions are not newly introduced in this work; both have been studied in previous matching market literature (Sankararaman et al., 2021).

We consider the player with the **lowest** ranking among the serial dictatorship. Without loss of generality, denote this player by p_N and its optimal arm as a_K. Let $N_j^{(i)}(T)$ denote the number of times player i is matched with arm j throughout the horizon. $\Delta_{\min}^{(N)}$ corresponds to the preference gap between a_K (the optimal arm of p_N) and the second-most-preferred arm of p_N and $\Delta_j^{(N)}$ refers to the preference gap between a_K and arbitrary arm a_j.

Having defined the above notation, we further decompose the regret of p_N into two parts in the next lemma.

Lemma G.1. *Under the serial dictatorship and OSB conditions, the cumulative regret of the lowest-ranked player p_N satisfies*

$$\mathrm{Reg}_N(T) \geq \max \left\{ \underbrace{\sum_{k=1}^{K} \Delta_k^{(N)} \mathbb{E}\left[N_k^{(N)}(T)\right]}_{\textit{Exploration}}, \underbrace{\sum_{j'=1}^{N-1} \Delta_{\min}^{(N)} \mathbb{E}\left[N_K^{(j')}(T)\right]}_{\textit{Blocked}} \right\}.$$

Proof. The first term corresponds to the regret incurred when player p_N is matched with suboptimal arms due to exploration. In particular, whenever p_N is matched with arm a_j, it incurs an expected regret of $\Delta_j^{(N)}$. The second term accounts for the regret incurred when a_K, the optimal arm of p_N, is occupied by a higher-ranked player due to exploration under serial dictatorship. Consequently, p_N must either match with a suboptimal arm or remain unmatched, resulting in a per-round regret of at least $\Delta_{\min}^{(N)}$. Therefore, the cumulative regret of p_N is lower bounded by the maximum of the two terms. □

Having defined this notation, consider the bandit instance where the first $K-1$ arm contexts satisfy $\mathrm{span}\{C_1, C_2, ..., C_{k-1}\} = \mathbb{R}^{d-1}$ and $\mathrm{span}\{C_1, C_2, ..., C_k\} = \mathbb{R}^d$, were C_K is a one-hot vector. Thus, for every player, their preference for a_K can be learned only by directly matching with a_K.

For p_N, let a_K have mean $1/2$ and its second-preferred-arm have mean $1/4$. Let the other arms have mean $1/2 - 2\Delta$, where Δ is between $1/8$ and $1/4$. Then by definition $\Delta_{\min}^{(N)} = 1/4$. For each higher-ranked player p_i, let its optimal arm have mean $1/2$, while all other arms have mean $1/2 - \Delta$. Following the definition of minimum preference gap in Appendix A, one can verify that Δ corresponds to gap_1.

We first lower bound the blocking regret in Lemma G.2.

Lemma G.2. *Under serial dictatorship and OSB conditions, the regret incurred by p_N due to higher-ranked players exploring its optimal arm a_K satisfies:*

$$\sum_{j'=1}^{N-1} \Delta_{\min}^{(N)} \mathbb{E}\Big[N_K^{(j')}(T)\Big] = \Omega\left(\frac{N \log T}{\Delta^2}\right).$$

Proof. Since C_K is one-hot, Lemma J.7 can be directly applied. Therefore, we have:

$$\liminf_{T\to\infty} \frac{\sum_{j'=1}^{N-1} \Delta_{\min}^{(N)} \mathbb{E}\Big[N_K^{(j')}(T)\Big]}{\log T} \geq \sum_{j'=1}^{N-1} \frac{\Delta_{\min}^{(N)}}{D_{\inf}\Big(\nu_{j'K}, \mu_{j' m_{j'}^*}, \mathcal{P}\Big)}. \tag{8}$$

Since the arm reward distributions are normal, we have $D_{\inf}(\mathcal{N}(1/2-\Delta,1), 1/2, \mathcal{P}) = \Delta^2/2$, where $\mathcal{N}$ denotes normal distribution. Substituting it and $\Delta_{\min}^{(N)} = 1/4$ into the right hand side of Eq. 8, we obtain:

$$\sum_{j'=1}^{N-1} \frac{\Delta_{\min}^{(N)}}{D_{\inf}\Big(\nu_{j'K}, \mu_{j' m_{j'}^*}, \mathcal{P}\Big)} = \sum_{j'=1}^{N-1} \frac{1/4}{\Delta^2/2} = \frac{N-1}{2\Delta^2}.$$

Rearranging Eq. 8, we obtain:

$$\sum_{j'=1}^{N-1} \Delta_{\min}^{(N)} \mathbb{E}\Big[N_K^{(j')}(T)\Big] \geq \left(\frac{N-1}{2\Delta^2}\right) \log T$$

as $T \to \infty$, which completes the proof. □

Next, we lower bound the regret incurred due to exploration in Lemma G.3.

Lemma G.3. *Under the serial dictatorship and OSB conditions, the regret incurred by the lowest-ranked player p_N due to exploration satisfies:*

$$\sum_{k=1}^{K} \Delta_k^{(N)} \mathbb{E}\Big[N_k^{(N)}(T)\Big] = \Omega\left(\frac{d \log T}{\Delta}\right).$$

Definition 3 (Consistent Policy)**.** *A policy π is called* consistent *over a class of bandit instances $\mathcal{E}$ if, for every $\nu \in \mathcal{E}$ and every $p > 0$,*

$$\lim_{T\to\infty} \frac{\mathrm{Reg}(\pi, \nu, T)}{T^p} = 0.$$

This definition is stated for the single player setting where $\mathrm{Reg}(\pi, \nu, T)$ corresponds to the regret accumulated over horizon T under policy π on instance $\nu \in \mathcal{E}$. In our multi-player setting, by Theorem 4.1, it holds that there exists such policies that are consistent for all players. We focus on these policies in the remaining proof of this section.

Proof. By Lemma J.8,

$$\liminf_{T\to\infty} \frac{\sum_{k=1}^{K} \Delta_k^{(N)} \mathbb{E}\Big[N_k^{(N)}(T)\Big]}{\log T} \geq c(\mathcal{K}, \theta_N^*). \tag{9}$$

where

$$c(\mathcal{K}, \theta_N^*) = \inf_{\alpha \in [0,\infty)^{\mathcal{K}}} \sum_{k\in\mathcal{K}} \alpha(a_k) \Delta_k^{(N)}$$
$$\text{s.t.} \quad \|C_k\|_{H_\alpha^{-1}}^2 \leq \frac{\Delta_k^{(N)}}{2}, \quad \forall k \in \mathcal{K} \text{ with } \Delta_k^{(N)} > 0,$$

and

$$H_\alpha = \sum_{k\in\mathcal{K}} \alpha(a_k) C_k C_k^\top.$$

By the linearity and cyclic property of trace,

$$d = \sum_{k\in\mathcal{K}} \alpha(a_k)\,\mathrm{tr}\left(C_k C_k^\top H_\alpha^{-1}\right) = \sum_{k\in\mathcal{K}} \alpha(a_k) C_k^\top H_\alpha^{-1} C_k = \sum_{k\in\mathcal{K}} \alpha(a_k)\|C_k\|^2_{H_\alpha^{-1}}. \tag{10}$$

Also we have

$$\alpha(a_K) C_K C_K^\top \preceq H_\alpha,$$

which implies

$$\alpha(a_K)\|C_K\|^2_{H_\alpha^{-1}} \le 1. \tag{11}$$

Since the suboptimal arms have gaps at most 2Δ, this implies

$$\|C_k\|^2_{H_\alpha^{-1}} \le \frac{\left(\Delta_k^{(N)}\right)^2}{2} \le 2\Delta^2.$$

Define $\mathcal{S} = \mathcal{K}/\{K\}$, by Eq. 10 and Eq. 11

$$\sum_{k\in\mathcal{S}} \alpha(a_k) \ge \frac{d-1}{2\Delta^2}.$$

Since every suboptimal arm in $\mathcal{K}$ has a regret gap of at least Δ, we have

$$c(\mathcal{K}, \theta_N^*) \ge \sum_{k\in\mathcal{S}} \alpha(a_k)\Delta_k^{(N)} \ge \Delta \sum_{k\in\mathcal{S}} \alpha(a_k) \ge \frac{d-1}{2\Delta}.$$

Combining the above results with Eq. 9, we obtain

$$\sum_{k=1}^{K} \Delta_k^{(N)} \mathbb{E}\left[N_k^{(N)}(T)\right] \ge \frac{(d-1)\log T}{2\Delta}.$$

as $T \to \infty$, which completes the proof.

Finally, combining Lemmas G.1,G.2 and G.3, the proof of Theorem 4.2 is complete. □

H CONTEXTUAL EXPLORATION WITH INDIFFERENCE (CEWI)

Similar to CETGS, CEWI will perform an index phase as Algorithm 5 at the beginning of the exploration phase. After the index phase, the following time slots are still partitioned into numerous super-phases with exponentially growing of length, and the number of sampling counts is derived from G-optimal design as Algorithm 1 Line 3. For each player p_i, a Better matrix initialized as $0 \in \mathbb{R}^{K\times K}$ is used is to record the player's preference between arms where $\mathrm{Better}(i,j,j') = 1$ implies p_i has determined that a_j is better than $a_{j'}$. $\mathrm{Update_Flag}_i = \mathrm{False}$ means p_i has not updated its Better matrix in this super-phase, while $\mathrm{Update_Pairs}_i$ initialized as $\{\ \}$ is the set of arm pairs (j,j') corresponding to the update of $\mathrm{Better}(i,j,j') = 1$. The key modification of CEWI yields that in CETGS the available set for all players to compute G-optimal design is the entire arm set $\mathcal{K}$ as Algorithm 1 Line 3, but in CEWI, each player's available set is formulated by an enumeration of all matchings according to each player's Better matrix. A matching $\hat{A} := ((p_i, \bar{A}_i))_{i\in[N]}$ is defined as a semi-stable matching if there do not exist two matching pairs $(p_i, \bar{A}_i), (p_j, \bar{A}_j)$, such that $\mathrm{Better}(i, \bar{A}_j, \bar{A}_i) = 1$ and $\pi_{\bar{A}_j,i} \prec \pi_{\bar{A}_j,j}$. Denote the set of all semi-stable matchings as phase as $\mathcal{S} := \{\hat{A} = ((p_i, \bar{A}_i))_{i\in[N]}, \hat{A} \text{ is semi-stable}\}$. Then the available arm set for p_i is defined as $\{a : a \in (p_i, a)\ s.t. \exists\ \hat{A} \in \mathcal{S} \text{ satisfying } (p_i, a) \in \hat{A}\}$.

After formulating the sampling time of each arm from G-optimal design, players start to route with a modified routing strategy different from Algorithm 2. In this new routing strategy, starting from player p with index = 1, it will start sampling arms with the smallest index a_j and moves on to

the next arm when the sampling of a_j is finished. Since each sampled arm of p is associated with a semi-stable matching $\hat{A}$, other players will select the arms matched with them in that particular semi-stable matching $\hat{A}$ s.t. $(p, a_j) \in \hat{A}$. Notice that (p, a_j) may be involved in numerous semi-stable matchings outcomes, we choose the semi-stable matching that assigns lower-indexed players to lower-indexed arms whenever feasible. After p finishes sampling, the player with index $= 2$ will start sampling, and the rest of the players will continue the aforementioned process until the sampling procedure of the player with index $= N$ is finished. When players' finished sampling in the next time step, players who updated their Better matrix based on the criterion in Algorithm 1 Line 9-13 will choose the arms with their unique index and other players will choose -1. If there exists player who need to update their Better matrix, they will sequentially broadcast their updates of their Better matrix following the decentralized communication procedure as Algorithm 7 which will be introduced in the next paragraph, and players will move on to the next super-phase after communication.

The intuition of CEWI is that with the decrease of ξ_l, all semi-stable matchings will eventually be equivalent to a stable matching, and the matching at each time slot in the exploration phase will be a stable-matching. However, the optimality of these stable matchings are not guaranteed for that different stable matching may appear in different time slots due to sampling needs. The stable regret for CEWI is $O\big(Nd\log T/\Delta^2\big)$ which is better than the state-of-the-art upper bound Kong et al. (2025) of AE-AGS which is $O\big(NK\log T/\Delta^2\big)$ since $d \ll K$. Here Δ is the minimum non-zero preference gap between all arms among all players. In a special case where Δ = 0 meaning all players have the same preference towards all arms, the regret of CEWI will degenerate into $O(\log T)$.

Theorem 4.3. *Following CEWI algorithm, the stable regret satisfies*

$$\mathrm{Reg}_i(T) \le O\big(Nd\log T/\Delta^2\big)\,.$$

If Δ = 0, then $\mathrm{Reg}_i(T) \le O(\log T)$ for all players.

Proof. Define Δ as the non-zero minimum preference gap between all arms among all players. The expected regret for player i running CEWI is:

$$\begin{aligned}
\mathrm{Reg}_i(T) &= \mathbb{E}\left[\sum_{t=1}^{T}(u_{i,m_i} - X_{i,A_i(t)}(t))\right] \\
&\le \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\{\bar{A}(t) \notin \mathcal{M}\}\right] \\
&\le \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\{\bar{A}(t) \notin \mathcal{M}\} \mid \neg\mathcal{F}\right] + \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\{\mathcal{F}\}\right] \\
&\le NK + \frac{256Nd\log T}{\Delta^2} + \frac{256Nd\log K}{\Delta^2} + \frac{Nd(d+1)\log\frac{8}{\Delta}}{2} \\
&\quad + \log T + 2NK^2 + N, \qquad (12)
\end{aligned}$$

where Eq. 12 holds following Lemma F.1, F.2, F.4 and I.2 by replacing $\Delta_{\min}$ with Δ and $\mathbb{I}\{\bar{A}(t) \neq m^\star\}$ with $\mathbb{I}\{\bar{A}(t) \notin \mathcal{M}\}$.

Notice when $\Delta = 0$, which means that all players favor all arms to the same extend, the expected regret then degenerates into:

$$\begin{aligned}
\mathrm{Reg}_i(T) &= \mathbb{E}\left[\sum_{t=1}^{T}(u_{i,m_i} - X_{i,A_i(t)}(t))\right] \\
&\le NK + \log T + N, \qquad (13)
\end{aligned}$$

where Eq. 13 holds since under the assumption of $\Delta = 0$, all matchings will be a stable matching during the exploration phase, hence suffering from 0 regret, and the communication in Algorithm 7 will never occur under $\neg\mathcal{F}$ since all arms will be treated equally. Notice that players are unaware of $\Delta = 0$ and still have to broadcast their update status by choosing $A_i(t) = -1$ at the end of each super-phase, which will cause regret bounded by $\log T$ demonstrated in Lemma F.3. □

I PROOF OF THEOREM 5.1

Let $\mathcal{G}$ denote the failure event of the estimated $\hat{\theta}$ is out of bounds of the high-probability elliptical bound:

$$\mathcal{G} = \left\{ \exists i \in [N], t \in [T] \text{ s.t. } \|\theta_i^* - \hat{\theta}_{i,t}\|^2_{V_{i,t-1}} > \beta_t \right\}.$$

Where $\sqrt{\beta_t} = \sqrt{\lambda}M + \sqrt{2\log\frac{1}{\delta} + d\log\left(\frac{d\lambda+tL^2}{d\lambda}\right)}$. The expected regret for player i can be decomposed as:

$$\begin{aligned}
\text{Reg}_i(T) &= \mathbb{E}\left[\sum_{t=1}^{T}(u_{i,m_i} - X_{i,A_i(t)})\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\{\bar{A}(t) \notin \mathcal{M}\}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\{\bar{A}(t) \notin \mathcal{M} \mid \neg\mathcal{G}\}\right] + \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\{\mathcal{G}\}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T}\tau_t(\text{Index phase}) \mid \neg\mathcal{G}\right] \\
&\quad + \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\Big(\exists i \in [N], j \in [K] \text{ s.t. } u_{i,j} > u_{i,A_i(t)} \textit{ and } \pi_{j,i} \prec \pi_{j,A_j(t)}\Big) \mid \neg\mathcal{G}\right] \\
&\quad + \mathbb{E}\left[\sum_{t=1}^{T}\omega_t(\text{communication}^*) \mid \neg\mathcal{G}\right] + \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\{\mathcal{G}\}\right] \\
&\leq NK + \frac{32\lambda M^2 Nd}{\Delta^2}\log\left(1 + \frac{TL^2}{\lambda d}\right) \\
&\quad + \frac{64MNd\sqrt{2\lambda\log T + \lambda d\log\left(1+\frac{TL^2}{d\lambda}\right)}}{\Delta^2}\log\left(1 + \frac{TL^2}{d\lambda}\right) \\
&\quad + \frac{64Nd\log T}{\Delta^2}\log\left(1 + \frac{TL^2}{d\lambda}\right) + \frac{32Nd^2}{\Delta^2}\log^2\left(1 + \frac{TL^2}{d\lambda}\right) + 2NK^2 + N. \qquad (14)
\end{aligned}$$

Here communication* denotes the communication procedure in Algorithm 7, $\omega_t(\text{communication}^*)$ = 1 if the current time slot t is running communication* and 0 otherwise. The detailed proof of Eq.14 are shown in Lemma I.1, I.2, I.3.

When $\Delta = 0$, the regret bound degenerates into:

$$\begin{aligned}
\text{Reg}_i(T) &= \mathbb{E}\left[\sum_{t=1}^{T}(u_{i,m_i} - X_{i,A_i(t)})\right] \\
&\leq NK. \qquad (15)
\end{aligned}$$

Here Eq.15 holds for that when $\Delta = 0$, under $\neg\mathcal{G}$, players will never update their Better matrix and only the index phase produces a regret bounded by NK by Lemma F.1. The term N follows from the expectation of the failure event over the horizon by Lemma I.1.

Lemma I.1. *By setting the confidence parameter $\delta = 1/T$, we have*

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\{\mathcal{G}\}\right] \leq N.$$

Proof.

$$\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{I}\{\mathcal{G}\}\right] \leq T \cdot \mathbb{E}\left[\sum_{i=1}^{N}\mathbb{I}\left\{\|\theta_i^* - \hat{\theta}_{i,t-1}\|^2_{V_{i,t-1}} > \beta_t\right\}\right]$$

$$\leq TN \cdot \delta \tag{16}$$
$$= TN \cdot \frac{1}{T} \tag{17}$$
$$= N,$$

where Eq.16 follows from Lemma J.1 and J.5, and Eq.17 holds by setting the confidence parameter $\delta = 1/T$. □

Lemma I.2. *For the communication phase, we have the following bounds:*

$$\mathbb{E}\left[\sum_{t=1}^{T} \omega_t(\text{communication}^*) \mid \neg\mathcal{G}\right] \leq N\left(\frac{K(K-1)}{2} + \frac{2K(K-1)}{2} + \frac{K(K-1)}{2}\right) \tag{18}$$
$$\leq 2NK^2.$$

Proof. Since there are K arms, the total update pairs is $\frac{K(K-1)}{2}$. The first term of Eq.18 is due to updates when Update_Flag is True and choosing $A_{i,t} = -1$ as Algorithm 3 Line 8. The second term is the total transmission steps for $\frac{K(K-1)}{2}$ pairwise updates as Algorithm 7 Line 7-12, and the third term is due to a single player marking the end of its transmission by choosing $A_{i,t} = -1$ as Algorithm 7 Line 14. □

Lemma I.3.

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\Big\{\exists i \in [N], j \in [K] \text{ s.t. } u_{i,j} > u_{i,A_i(t)} \text{ and } \pi_{j,i} \prec \pi_{j,A_j(t)} \mid \neg\mathcal{G}\Big\}\right]$$
$$\leq \frac{32\lambda M^2 Nd}{\Delta^2}\log\left(1 + \frac{TL^2}{\lambda d}\right) + \frac{64MNd\sqrt{2\lambda\log T + \lambda d\log\left(1+\frac{TL^2}{d\lambda}\right)}}{\Delta^2} \cdot \log\left(1 + \frac{TL^2}{d\lambda}\right)$$
$$+ \frac{64Nd\log T}{\Delta^2} \cdot \log\left(1 + \frac{TL^2}{d\lambda}\right) + \frac{32Nd^2}{\Delta^2}\log^2\left(1 + \frac{TL^2}{d\lambda}\right). \tag{19}$$

Proof. By Lemma I.5, we have

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\Big\{\exists i \in [N], j \in [K] \text{ s.t. } u_{i,j} > u_{i,A_i(t)} \text{ and } \pi_{j,i} \prec \pi_{j,A_j(t)} \mid \neg\mathcal{G}\Big\}\right]$$
$$\leq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N} \mathbb{I}(a_{i,t} \text{ is not the best arm in } A_{i,t} \mid \neg\mathcal{G})\right]$$
$$\leq \mathbb{E}\left[\sum_{i=1}^{N}\sum_{t=1}^{T} \mathbb{I}(\text{the rank of } A_{i,t} \text{ is not determined} \mid \neg\mathcal{G})\right]$$
$$\leq \mathbb{E}\left[\sum_{i=1}^{N}\sum_{t=1}^{T} \mathbb{I}\left(\max_{j\in\mathcal{A}_i\setminus\mathcal{E}_i} \|C_j\|_{V_{i,t-1}^{-1}} > \frac{\Delta}{4\sqrt{\beta_t}} \mid \neg\mathcal{G}\right)\right].$$

Since

$$\sum_{t=1}^{T} \mathbb{I}\left(\max_{j\in\mathcal{A}_i\setminus\mathcal{E}_i} \|C_j\|_{V_{i,t-1}^{-1}} > \frac{\Delta}{4\sqrt{\beta_t}}\right) = \sum_{t=1}^{T} \mathbb{I}\left(\max_{j\in\mathcal{A}_i\setminus\mathcal{E}_i} \|C_j\|^2_{V_{i,t-1}^{-1}} > \frac{\Delta^2}{16\beta_t}\right).$$

Note that $\frac{\Delta}{4\sqrt{\beta_t}} \leq 1$ by the choice of β_t. Thus, applying Lemma I.6, we have

$$\sum_{t=1}^{T} \mathbb{I}\left(\max_{j\in\mathcal{A}_i\setminus\mathcal{E}_i} \|C_j\|_{V_{i,t-1}^{-1}} > \frac{\Delta}{4\sqrt{\beta_t}}\right) \cdot \left(\frac{\Delta}{4\sqrt{\beta_t}}\right)^2 \leq \sum_{t=1}^{T}\left(1 \wedge \max_{a_j\in A_{i,t}} \|C_{i,j}\|^2_{V_{i,t-1}^{-1}}\right)$$
$$\leq 2\log\frac{\det(V_{i,T})}{\det(\lambda I)} \tag{20}$$

$$\leq 2d \log\left(1 + \frac{TL^2}{\lambda d}\right), \tag{21}$$

we obtain:

$$\sum_{t=1}^{T} \mathbb{I}\left(\max_{j \in \mathcal{A}_i \setminus \mathcal{E}_i} \|C_j\|_{V_{i,t-1}^{-1}} > \frac{\Delta}{4\sqrt{\beta_t}}\right) \leq \frac{16\beta_t}{\Delta^2} 2d \log\left(1 + \frac{TL^2}{\lambda d}\right).$$

Substituting

$$\sqrt{\beta_t} = \sqrt{\lambda} M + \sqrt{2 \log \frac{1}{\delta} + d \log\left(\frac{d\lambda + tL^2}{d\lambda}\right)} \leq \sqrt{\lambda} M + \sqrt{2 \log \frac{1}{\delta} + d \log\left(\frac{d\lambda + TL^2}{d\lambda}\right)},$$

and taking δ as 1/T, we obtain the final result in Eq.19. □

Lemma I.4. *For player i and arm j, UCB$_{i,t}(a_j)$ and LCB$_{i,t}(a_j)$ are defined as:*

$$\textit{UCB}_{i,t}(a_j) = \langle \hat{\theta}_{i,t}, C_j \rangle + \sqrt{\beta_t} \|C_j\|_{V_{i,t}^{-1}},$$
$$\textit{LCB}_{i,t}(a_j) = \langle \hat{\theta}_{i,t}, C_j \rangle - \sqrt{\beta_t} \|C_j\|_{V_{i,t}^{-1}},$$

where

$$\sqrt{\beta_t} = \sqrt{\lambda} M + \sqrt{2 \log \frac{1}{\delta} + d \log\left(\frac{d\lambda + tL^2}{d\lambda}\right)}.$$

With probability at least $1 - \delta$,

$$\textit{LCB}_{i,t}(a_j) - \sqrt{\beta_t} \|C_j\|_{V_{i,t}^{-1}} \leq \langle \theta_i^*, C_j \rangle \leq \textit{UCB}_{i,t}(a_j) + \sqrt{\beta_t} \|C_j\|_{V_{i,t}^{-1}}.$$

Proof. Under event $\neg\mathcal{G}$, by Lemma J.5, we have:

$$\|\hat{\theta}_{i,t} - \theta_i^*\|_{V_{i,t}} \leq \sqrt{\beta_t}.$$

For any arm a_j with context feature C_j, we consider the estimation error of the expected reward:

$$\left|\langle \hat{\theta}_{i,t}, C_j \rangle - \langle \theta_i^*, C_j \rangle\right| = \left|\langle \hat{\theta}_{i,t} - \theta_i^*, C_j \rangle\right|.$$

Applying the generalized Cauchy-Schwarz inequality for matrices $V_{i,t}$ and its inverse $V_{i,t}^{-1}$, we obtain:

$$\begin{aligned}
\left|\langle \hat{\theta}_{i,t} - \theta_i^*, C_j \rangle\right| &= \left|(\hat{\theta}_{i,t} - \theta_i^*)^\top V_{i,t}^{\frac{1}{2}} V_{i,t}^{-\frac{1}{2}} C_j\right| \\
&\leq \left\|V_{i,t}^{\frac{1}{2}}(\hat{\theta}_{i,t} - \theta_i^*)\right\|_2 \cdot \left\|V_{i,t}^{-\frac{1}{2}} C_j\right\|_2 \\
&= \|\hat{\theta}_{i,t} - \theta_i^*\|_{V_{i,t}} \cdot \|C_j\|_{V_{i,t}^{-1}} \\
&\leq \sqrt{\beta_t} \|C_j\|_{V_{i,t}^{-1}}.
\end{aligned}$$

Expanding the absolute value directly yields the desired upper and lower bounds:

$$-\sqrt{\beta_t} \|C_j\|_{V_{i,t}^{-1}} \leq \langle \theta_i^*, C_j \rangle - \langle \hat{\theta}_{i,t}, C_j \rangle \leq \sqrt{\beta_t} \|C_j\|_{V_{i,t}^{-1}},$$

which rigorously holds with probability at least $1 - \delta$ by Lemma J.5. □

Lemma I.5. *Under $\neg\mathcal{G}$, for player i and and arm j, j', let $\Delta_{i,j,j'} > 0$ be the preference gap between arm j' and j with $u_{i,j'} > u_{i,j}$. Then if $\max_{j,j'} \|C_j\|_{V_{i,t-1}^{-1}} < \frac{\Delta_{i,j,j'}}{4\sqrt{\beta_t}}$, then UCB$_{i,t}(a_j) <$ LCB$_{i,t}(a_j')$.*

Proof. The proof follows by the same idea in Lemma F.6,

$$\begin{aligned}
&\text{LCB}_{i,t_{i,t}}(a_j') - \text{UCB}_{i,t_{i,t}}(a_j) \\
\geq &\langle \theta_i^*, (C_{j'} - C_j) \rangle - 2\sqrt{\beta_t} \|C_{j'}\|_{V_{i,t}^{-1}} - 2\sqrt{\beta_t} \|C_j\|_{V_{i,t}^{-1}}
\end{aligned} \tag{22}$$

$$
\begin{aligned}
&= \Delta_{i,j,j'} - 2\sqrt{\beta_t}\|C_{j'}\|_{V_{i,t}^{-1}} - 2\sqrt{\beta_t}\|C_j\|_{V_{i,t}^{-1}} \\
&> 0, \qquad \text{since } \max_{j,j'} \|C_j\|_{V_{i,t-1}^{-1}} < \frac{\Delta_{i,j,j'}}{4\sqrt{\beta_t}}.
\end{aligned}
$$

where Eq.22 comes from Lemma I.4. □

Lemma I.6. *Let $V_{i,0} = \lambda I \in \mathbb{R}^{d\times d}$ be a positive definite initialization matrix. For player i, let $C_{i,A_i(1)}, \ldots, C_{i,A_i(T)} \in \mathbb{R}^d$ be a sequence of observed context vectors such that $\|C_{i,A_i(t)}\|^2 \leq L^2 < \infty$ for all $t \in [T]$. Assume the covariance matrix updates via $V_{i,t} = V_{i,t-1} + C_{i,A_i(t)}C_{i,A_i(t)}^\top$. Then, we have:*

$$
\sum_{t=1}^{T} \left(1 \wedge \|C_{i,A_i(t)}\|^2_{V_{i,t-1}^{-1}}\right) \leq 2d \log\left(1 + \frac{TL^2}{\lambda d}\right).
$$

Proof. Notice that $C_{i,A_i(t)} = \max_{j\in\mathcal{A}_i\setminus\mathcal{E}_i} \|C_j\|_{V_{i,t-1}^{-1}}$ due to Algorithm 4 Line 7. Using the identity that for any scalar $u \geq 0$, $u \wedge 1 \leq 2\ln(1+u)$, we obtain :

$$
\sum_{t=1}^{T} \left(1 \wedge \|C_{i,A_i(t)}\|^2_{V_{i,t-1}^{-1}}\right) \leq 2\sum_{t=1}^{T} \log\left(1 + \|C_{i,A_i(t)}\|^2_{V_{i,t-1}^{-1}}\right). \tag{23}
$$

For any step $t \geq 1$, we can rewrite the covariance matrix by the sum of the previous matrix and the current update:

$$
V_{i,t} = V_{i,t-1} + C_{i,A_i(t)}C_{i,A_i(t)}^\top = V_{i,t-1}^{1/2}\left(I + V_{i,t-1}^{-1/2}C_{i,A_i(t)}C_{i,A_i(t)}^\top V_{i,t-1}^{-1/2}\right)V_{i,t-1}^{1/2}.
$$

Since the determinant is a multiplicative map, and the matrix $(I + yy^\top)$ has determinant of $(1 + \|y\|^2)$ and 1. By taking determinants on both sides, we yield:

$$
\det(V_{i,t}) = \det(V_{i,t-1})\cdot\det\left(I + V_{i,t-1}^{-1/2}C_{i,A_i(t)}C_{i,A_i(t)}^\top V_{i,t-1}^{-1/2}\right) = \det(V_{i,t-1})\left(1 + \|C_{i,A_i(t)}\|^2_{V_{i,t-1}^{-1}}\right).
$$

and by recursive decomposing $V_{i,T}$, we have:

$$
\det(V_{i,T}) = \det(V_{i,0})\prod_{t=1}^{T}\left(1 + \|C_{i,A_i(t)}\|^2_{V_{i,t-1}^{-1}}\right). \tag{24}
$$

From Eq.24, we divide both sides by $\det(V_{i,0})$, and take the logarithm of both sides. Combining this with Eq.23, we arrive at

$$
\sum_{t=1}^{T} \left(1 \wedge \|C_{i,A_i(t)}\|^2_{V_{i,t-1}^{-1}}\right) \leq 2\log\frac{\det(V_{i,T})}{\det(V_{i,0})}. \tag{25}
$$

Using the inequality of arithmetic and geometric means, since the trace of a matrix equals to the sum of its eigenvalues $\lambda_1 \ldots\ldots \lambda_d$, the determinant of $V_{i,T}$ is bounded by:

$$
\det(V_{i,T}) = \prod_{k=1}^{d}\lambda_k \leq \left(\frac{\mathrm{trace}(V_{i,T})}{d}\right)^d.
$$

Since $\mathrm{trace}(V_{i,T}) = \mathrm{trace}(\lambda I) + \sum_{t=1}^{T}\mathrm{trace}(C_{i,A_i(t)}C_{i,A_i(t)}^\top) \leq d\lambda + TL^2$, and $\det(V_{i,0}) = \lambda^d$, we derive:

$$
\frac{\det(V_{i,T})}{\det(V_{i,0})} \leq \frac{((d\lambda + TL^2)/d)^d}{\lambda^d} = \left(1 + \frac{TL^2}{\lambda d}\right)^d.
$$

It follows from Eq. 25 and the above inequality that the desired result holds. □

J TECHNICAL LEMMAS

Lemma J.1 (Corollary 5.5 in Lattimore & Szepesvári (2020)). *Assume that $X_1, X_2, \ldots, X_n$ are independent, σ-sub-Gaussian random variables centered around μ. Then for any $\varepsilon > 0$,*

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

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

Lemma J.2 (Equation 20.2 in Lattimore & Szepesvári (2020)). *Assume the following:*

(a) No regularization: $\lambda = 0$ and V_t is invertible.

(b) Independent sub-gaussian noise: $(\eta_s)_s$ are independent and 1-sub-gaussian.

(c) Fixed design: $A_1, \ldots, A_t$ are deterministically chosen without the knowledge of $X_1, \ldots, X_t$.

Then,

$$\mathbb{P}\left(\langle \hat{\theta}_t - \theta^*, x\rangle \geq \sqrt{2\|x\|^2_{V_t^{-1}} \log\left(\frac{1}{\delta}\right)}\right) \leq \delta.$$

Taking x as $-x$, it immediately follows that:

$$\mathbb{P}\left(|\langle \theta^* - \hat{\theta}_t, x\rangle| \geq \sqrt{2\|x\|^2_{V_t^{-1}} \log\left(\frac{1}{\delta}\right)}\right) \leq \delta.$$

Lemma J.3 (Theorem 21.1 in Lattimore & Szepesvári (2020)). *Assume that $\mathcal{A} \subset \mathbb{R}^d$ is compact and $\mathrm{span}(\mathcal{A}) = \mathbb{R}^d$. $g(\pi) = \max_{a\in\mathcal{A}} \|a\|^2_{V(\pi)^{-1}}$. The following are equivalent:*

(a) π^* is a minimiser of g,

(b) π^* is a maximiser of $f(\pi) = \log\det V(\pi)$,

(c) $g(\pi^*) = d$.

Furthermore, there exists a minimiser π^* of g such that

$$|\mathrm{Supp}(\pi^*)| \leq \frac{d(d+1)}{2}.$$

Lemma J.4 (Chapter 21 in Lattimore & Szepesvári (2020)). *Assume that π is a G-optimal design and $a \in \mathrm{Supp}(\pi)$. Let*

$$n_a = \left\lceil \frac{2\pi(a)g(\pi)}{\varepsilon^2} \log\left(\frac{1}{\delta}\right)\right\rceil.$$

Then, choosing each action $a \in \mathrm{Supp}(\pi)$ exactly n_a times ensures that

$$V = \sum_{a\in\mathrm{Supp}(\pi)} n_a a a^\top \succeq \frac{2g(\pi)}{\varepsilon^2} \log\left(\frac{1}{\delta}\right) V(\pi),$$

and for any $a \in \mathcal{A}$, with probability $1-\delta$,

$$\langle \hat{\theta} - \theta^*, a\rangle \leq \sqrt{2\|a\|^2_{V^{-1}} \log\left(\frac{1}{\delta}\right)} \leq \varepsilon.$$

Taking a as $-a$, it immediately follows that:

$$|\langle \theta^* - \hat{\theta}, a\rangle| \leq \sqrt{2\|a\|^2_{V^{-1}} \log\left(\frac{1}{\delta}\right)} \leq \varepsilon.$$

Lemma J.5 (Assumption 19.1 & Theorem 20.5 in Lattimore & Szepesvári (2020)). *Let $\hat{\theta}_t$ be an estimate of θ^* given by*

$$\hat{\theta}_t = V_t^{-1} \sum_{s=1}^{t} A_s X_s,$$

where V_t is a $d \times d$ matrices given by

$$V_0 = \lambda I \quad \textit{and} \quad V_t = V_0 + \sum_{s=1}^{t} A_s A_s^\top.$$

Assume that the confidence set $\mathcal{C}_t$ is closed, centered at $\hat{\theta}_{t-1}$, and satisfies

$$\mathcal{C}_t \subseteq \mathcal{E}_t = \left\{ \theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_{t-1}\|^2_{V_{t-1}} \leq \beta_t \right\}.$$

For any $\delta \in (0,1)$, if β_n is chosen to be

$$\sqrt{\beta_t} = \sqrt{\lambda} m + \sqrt{2 \log\left(\frac{1}{\delta}\right) + d \log\left(\frac{d\lambda + tL^2}{d\lambda}\right)},$$

then with probability $1-\delta$, for all $t \in [T]$, it holds that $\theta_* \in \mathcal{C}_t$.

Lemma J.6 (Kőnig's Edge-Coloring Theorem (König, 1916)). *If G is a bipartite multi-graph (parallel edges are allowed), then its edge-chromatic number equals its maximum degree, i.e., $\chi'(G) = \Delta(G)$.*

$$\chi'(G) = \Delta(G)$$

Lemma J.7 (Appendix D.3 in Sankararaman et al. (2021)). *For any arm $a_{j'}$, its expected matched time with a_K satisfies:*

$$\liminf_{T\to\infty} \frac{\mathbb{E}\left[N_K^{(j')}(T)\right]}{\log T} \geq \frac{1}{D_{\inf}\left(\nu_{j'K}, \mu_{j'm_*^{(j')}}, \mathcal{P}\right)}.$$

where $\nu_{j'K}$ denotes the reward distribution of arm a_K for player $p_{j'}$, $\mathcal{P}$ is the class of normal distributions, and $D_{\inf}\left(\nu_{j'K}, \mu_{j'm_*^{(j')}}, \mathcal{P}\right) = \inf_{\nu' \in \mathcal{P}} \{D_{\mathrm{KL}}(\nu_{j'K} \| \nu') : \mu(\nu') > \mu_{j'm_*^{(j')}}\}$, where $\mu(\nu')$ denotes the mean of ν'.

Lemma J.8 (Corollary 25.2 in Lattimore & Szepesvári (2020)). *Let $\mathcal{K} \subset \mathbb{R}^d$ be a finite set that spans $\mathbb{R}^d$, and let $\theta_N^* \in \mathbb{R}^d$ be such preference vector that there is a unique optimal arm. Then, for any consistent policy and player p_N,*

$$\liminf_{T\to\infty} \frac{\sum_{k=1}^{K} \Delta_k^{(N)} \mathbb{E}\left[N_k^{(N)}(T)\right]}{\log n} \geq c(\mathcal{K}, \theta_N^*),$$

where

$$c(\mathcal{K}, \theta_N^*) = \inf_{\alpha \in [0,\infty)^{\mathcal{K}}} \sum_{k \in \mathcal{K}} \alpha(a_k) \Delta_k^{(N)}$$

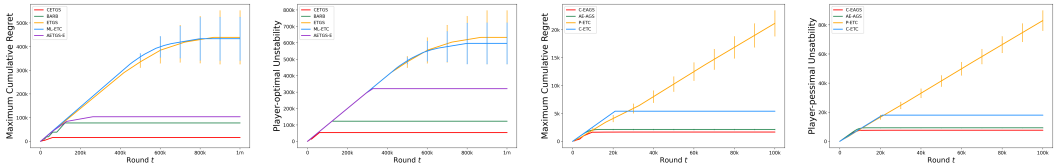
$$\textit{s.t.} \quad \|C_k\|^2_{H_\alpha^{-1}} \leq \frac{(\Delta_k^{(N)})^2}{2}, \quad \forall k \in \mathcal{K} \textit{ with } \Delta_k^{(N)} > 0,$$

and

$$H_\alpha = \sum_{k \in \mathcal{K}} \alpha(a_k) C_k C_k^\top.$$

K EXPERIMENTS

In this section, we compare our Algorithm 1 (CETGS) with AETGS-E (Kong et al., 2024), ETGS (Kong & Li, 2023), ML-ETC (Zhang et al., 2022) and BARB Lin et al. (b) in markets with strict preferences. Under the indifference setting, we compare our C-EAGS with AE-AGS (Kong et al., 2025), C-ETC (Liu et al., 2020), and P-ETC (Basu et al., 2021). We report the maximum cumulative regret defined as the largest cumulative regret among all players and the cumulative market unsuitability. In each experiment, we run all algorithms for $T = 10^6$ and $T = 10^5$ rounds in markets with strict preferences and indifference, respectively. We report the averaged results of 20 independent trials. The error bars, representing the standard errors calculated as the standard deviations divided by $\sqrt{20}$, are plotted. The player's preference vector θ^* and arm context C_j are randomly generated following Δ. We use unit-variance Gaussian noise and generate arm-side preferences through random permutations.



(a) Optimal regret (b) Market unstability (c) Stable regret (d) Market unstability

Figure 2: Experimental comparisons. (a)-(b) show CETGS with $N = 3, K = 300, \Delta = 0.15$. (c)-(d) show C-EAGS with $N = 7, K = 7, \Delta = 0.3$. In all cases $d = 2$.

As shown in Figure 2, our CETGS and C-EAGS algorithms, which utilize contextual linear bandits to learn a unified preference vector, exhibit the least cumulative regret and market uncertainty compared to baselines in matching markets with strict preferences and indifference, respectively. In contrast, the stochastic bandit-based baselines cannot avoid exploring all K arms to determine their preferences, which generates additional regret compared to our proposed methods. Additionally, CETGS utilizes the G-optimal design to efficiently learn preference information, therefore consistently outperforms the contextual baseline BARB Lin et al. (b) in the strict preference setting

We further provide extended experiments to justify our results. Table 2 compares the running times of CEWI and C-EAGS, highlighting the inefficiency of CEWI in large matching markets and demonstrating the necessity of C-EAGS. Furthermore, we conduct extensive experiments on CETGS and C-EAGS under varying parameters. For CETGS, we vary $N \in \{2, 4, 5, 6\}$, $\Delta \in \{0.1, 0.2, 0.25, 0.3\}$, $K \in \{100, 200, 400, 500\}$, and $d \in \{3, 4, 5, 6\}$, as shown in Figures 3–6. For C-EAGS, we vary $N = K \in \{5, 6, 8, 9\}$ and $\Delta \in \{0.2, 0.25, 0.35, 0.4\}$, as shown in Figures 7 and 8. Each experiment is conducted for 20 trials; the mean and standard deviation (or standard error) are reported.

	$N = K = 8$	$N = K = 9$	$N = K = 10$	$N = K = 11$
CEWI	4.490± 0.003s	13.393± 0.047s	116.102±0.224s	1465.748±5.804s
C-EAGS	33.049±0.018	38.885±0.038	45.736±0.039	53.289±0.235s

Table 2: Running time comparison between CEWI and C-EAGS with different market size

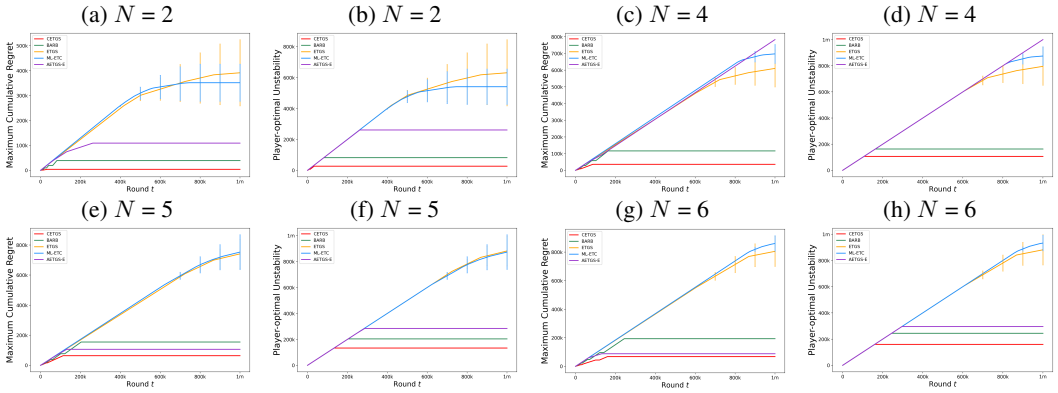


Figure 3: Experimental comparisons of our CETGS with AETGS-E, ETGS, ML-ETC, Phased ETC and BARB in one-to-one decentralized markets with $K = 300$ arms, $d = 2$ and $\Delta = 0.15$ and different N.

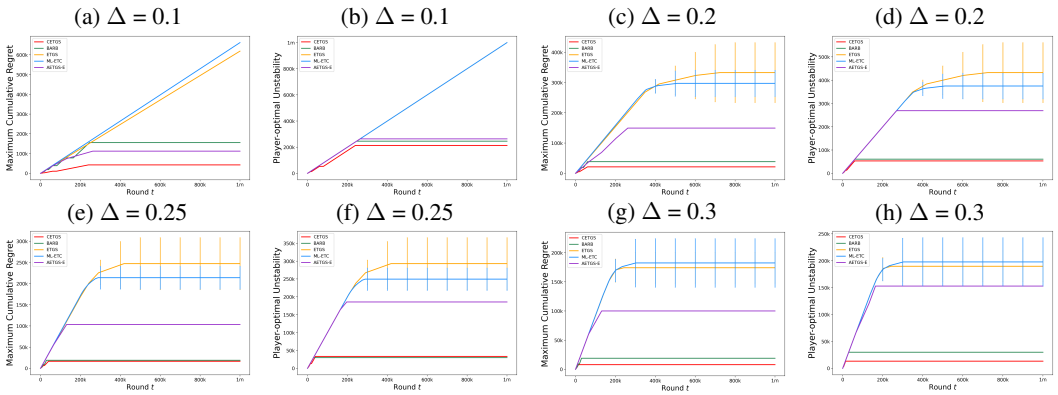


Figure 4: Experimental comparisons of our CETGS with AETGS-E, ETGS, ML-ETC, Phased ETC and BARB in one-to-one decentralized markets with $N = 3$ players, $K = 300$ arms, $d = 2$ and different Δ.

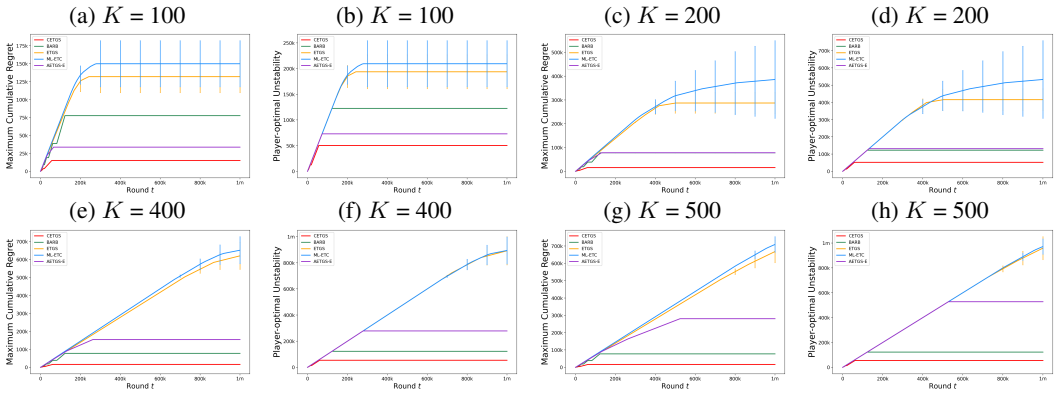


Figure 5: Experimental comparisons of our CETGS with AETGS-E, ETGS, ML-ETC, Phased ETC and BARB in one-to-one decentralized markets with $N = 3$ players, $d = 2$, $\Delta = 0.15$ and different K.

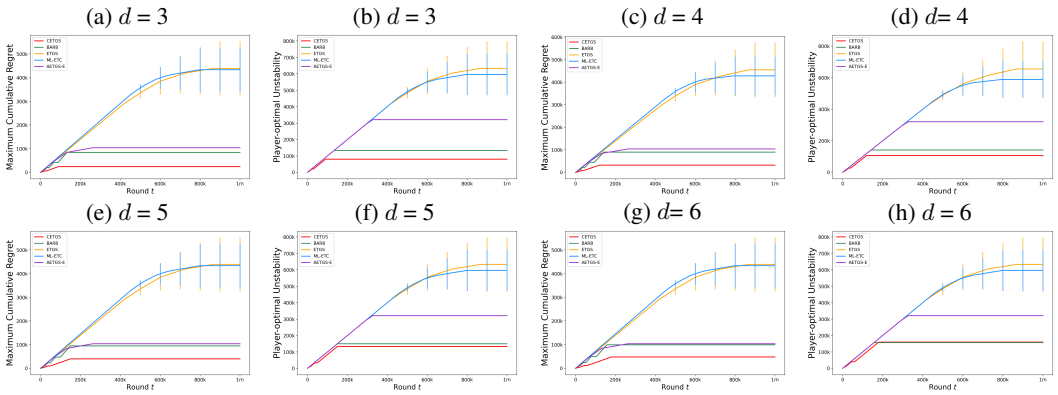


Figure 6: Experimental comparisons of our CETGS with AETGS-E, ETGS, ML-ETC, Phased ETC and BARB in one-to-one decentralized markets with $N = 3$ players, $K = 300$ arms, $\Delta = 0.15$ and different d.

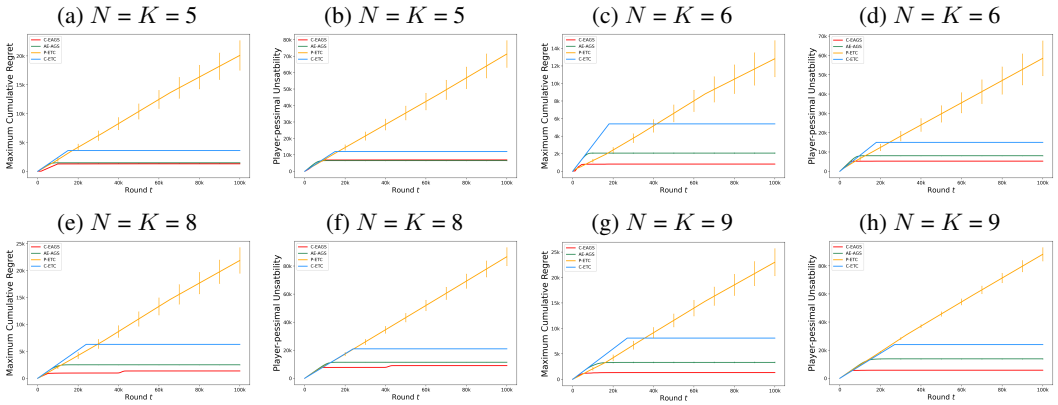


Figure 7: Experimental comparisons of our C-EAGS with AE-AGS, C-ETC and P-ETC in one-to-one decentralized markets with $d = 2$, $\Delta = 0.3$, different N and K.

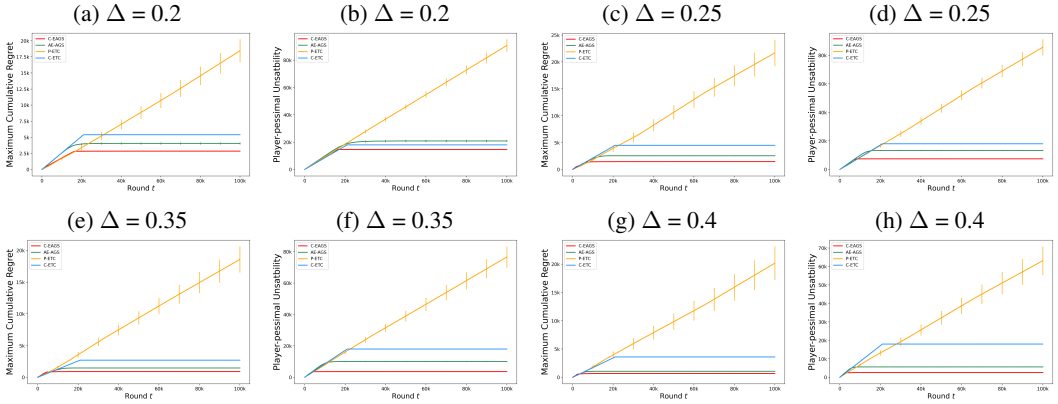


Figure 8: Experimental comparisons of our C-EAGS with AE-AGS, C-ETC and P-ETC in one-to-one decentralized markets with $N = 7$ players and $K = 7$ arms, $d = 2$ and different Δ.