ADAPTIVE BANDIT ALGORITHMS FOR CONTEXTUAL MATCHING MARKETS

Shiyun Lin[*] **Simon Mauras**[†] **Vianney Perchet**[‡] **Nadav Merlis**[§]

ABSTRACT

We study bandit learning in matching markets, where players and arms constitute the two market sides, and the players' utilities are linear in the arm contexts. In each round, new arms arrive with observable contexts. Then, the algorithm matches them to players, aiming to minimize each player's regret against a *stable matching benchmark*. This contextual structure creates significant complexity: subtle context shifts can slightly alter one player's utility while completely reconfiguring the underlying benchmark, causing large regret spikes for others. We address this in two settings: *stochastic* contexts, drawn from a latent distribution, and *adversarial* contexts, which may be arbitrary. For the stochastic case, we introduce a novel minimum preference gap to capture learning difficulty and provide a fully adaptive algorithm with an instance-dependent poly-logarithmic regret upper bound. We also establish matching instance-independent regret upper and lower bounds under a mild distributional assumption. For the adversarial setting, we propose a tractable regret notion that remains valid under arbitrary contexts and achieves an instance-independent sublinear regret bound via an adaptive algorithm.

1 INTRODUCTION

Two-sided matching markets provide a foundational theoretical framework for modeling and analyzing coordination problems across a wide range of economic and social domains. Classic applications include college admissions (matching students with schools) (Gale & Shapley, 1962), labor markets (assigning jobs to workers) (Kelso Jr & Crawford, 1982; Mine et al., 2013), medical residency placement (pairing doctors with hospitals) (Roth, 1984), and ride-sharing (matching passengers with drivers) (Shi et al., 2023). In these markets, participants on each side hold preferences over the other. For example, jobs rank workers by skill, while workers rank jobs based on compensation or location. The central objective is to achieve a *stable matching*, where no participant pair would mutually prefer to match with each other over their current assignments. Given full preference lists, the Gale-Shapley Deferred Acceptance algorithm (Gale & Shapley, 1962) efficiently finds a stable matching that is optimal for one side of the market.

However, in online marketplaces like crowdsourcing and gig platforms, worker preferences over jobs are often not known *a priori* and must instead be learned through repeated matchings during task delegation. Recent work has thus modeled this scenario using multi-player multi-armed bandits (Das & Kamenica, 2005; Liu et al., 2020; Basu et al., 2021; Li et al., 2022; Kong & Li, 2023; Zhang & Fang, 2024; Lin et al., 2025), casting the two sides of the market as players and arms. This framework enables the development of learning algorithms that integrate Gale-Shapley dynamics with bandit-feedback techniques.

Most existing work models the market as a stochastic multi-armed bandit, assuming noisy feedback from a single, *static* preference profile. However, this assumption is often unrealistic in real-world applications. In labor markets, for example, player preferences shift with external factors such as project cycles, difficulty, and reward structures. To capture such *dynamic preferences*, a canonical approach uses contextual bandits (Chu et al., 2011; Abbasi-Yadkori et al., 2011), where the matching utility for a player-arm pair is modeled as the inner product of the player's latent preference vector and the arm's observed feature (or context) vector. The platform observes noisy utility feedback after each match and must learn the underlying preference vectors to rank arms effectively with the contextual information in each round, aiming for per-step (approximate) stability in the long run.

[*]Center for Statistical Science, School of Mathematical Sciences, Peking University, shiyunlin@stu.pku.edu.cn

[†]INRIA, FairPlay Joint Team, simon.mauras@inria.fr

[‡]CREST, ENSAE, IP Paris; Criteo AI lab, FairPlay Joint Team, Vianney.perchet@normalesup.org

[§]Technion - Israel Institute of Technology, nmerlis@technion.ac.il

Consider an online freelance platform. While a worker's preference type – defined by stable valuations for attributes like pay and complexity – remains fixed, the realized features of a specific job (e.g., its budget, deadline, and required skills) are naturally stochastic. In practice, these features can be modeled as independent draws from a fixed but unknown distribution associated with the job's category (such as graphic design or copywriting). This motivates our first setting, in which the contextual feature vector of each arm is generated *stochastically* from a category-specific distribution in every matching round. In contrast, other realistic matching environments involve non-stationary or even strategically manipulated contexts. On a competitive procurement platform, for instance, clients may adaptively adjust a project's described terms, complexity, or evaluation metrics – either due to market shifts or to strategically exploit contractors' learned bidding behaviors. To capture such adversarial dynamics, we also analyze a setting in which the sequence of context vectors can be *arbitrary and potentially adversarial*.

Adaptive algorithms in bandits (Wu et al., 2018; Hao et al., 2020) dynamically balance exploration and exploitation based on observed data, improving both statistical efficiency and practical deployment. In two-sided matching markets, this adaptivity is crucial for automatically adjusting to changing user distributions, enhancing performance in non-stationary environments. We therefore seek to develop *adaptive* bandit algorithms for contextual matching markets.

1.1 Main Contributions

Novel Metric for Characterizing Problem Difficulty To characterize problem difficulty, we introduce novel metrics tailored to each setting. For the stochastic environment, we propose a *minimum preference gap*, a relaxed version of prior gap definitions that offers better realism and practicality. For the adversarial environment, we define a tractable form of *regret* measured against a player-optimal stable matching benchmark – employing its exact form in the large-gap regime and an approximation in the small-gap regime. Together, these metrics form a unified analytical framework that enables us to derive meaningful theoretical guarantees for our proposed algorithms.

Fully Adaptive Algorithms We develop separate algorithms for the stochastic and adversarial settings. Both employ an *adaptive phase transition* scheme, alternating between refining parameter estimates and performing (approximately) stable matchings. This adaptive design allows handling *arbitrary context covariance structures* in the stochastic case and *arbitrary context sequences* in the adversarial case. For the stochastic setting, we add a batched design that maintains and shrinks a candidate minimum preference gap toward the true value, enabling operation without prior knowledge of the true gap with respect to the distribution.

Theoretical Regret Guarantees We establish non-asymptotic regret upper bounds for our proposed algorithms under minimal assumptions, achieving instance-dependent poly-logarithmic rate in the stochastic setting and instance-independent sublinear rate in the adversarial setting. For the stochastic setting, under the additional condition of a strictly positive minimum eigenvalue for the context covariance matrix, we provide a logarithmic regret guarantee. In the asymptotic regime of the stochastic setting, we also derive instance-independent regret bounds, proving tight $\tilde{\Theta}\left(T^{2/3}\right)$ upper and lower bounds, where T is the time horizon. Notably, this result contrasts with prior work on stochastic bandit learning in matching markets with static preferences, where instance-independent guarantees are only possible for approximation benchmark.

1.2 Related Work

Bandit Learning in Matching Markets Das & Kamenica (2005) initiated the study of bandit learning in matching markets by introducing the problem and proposing empirical solutions. Liu et al. (2020) later provided a formal theoretical formulation and established initial guarantees on stable regret. Subsequent research has expanded this line of work, aiming to improve regret bounds (Sankararaman et al., 2021; Kong & Li, 2023) and explore various problem variants, including decentralized matching (Liu et al., 2021; Basu et al., 2021), many-to-one markets (Wang et al., 2022), settings with preference indifference (Kong et al., 2025; Lin et al., 2025), and models with transferable utilities (Jagadeesan et al., 2023). For a comprehensive survey, please refer to Li et al. (2025).

Despite extensive work on static preferences, research on dynamic preference structures remains limited. Recent approaches model non-stationarity via constraints on the total number of preference changes (Muthirayan et al., 2023) or on per-round utility drift (Ghosh et al., 2024). In contextual matching markets, previous work assumes prior knowledge of the minimum preference gap and covariance structure (Li et al., 2022) or studies decentralized settings with fixed and known number of environments (Parikh et al., 2024).

In contrast, we develop *fully adaptive*, centralized algorithms for both stochastic and adversarial settings, requiring *no prior knowledge* of the environment.

Contextual Combinatorial Bandits Contextual combinatorial bandits (CCB) extend the multi-armed bandit framework to the setting where an agent, given contextual information, must select a *set of items* (a "super arm") to maximize an aggregate reward. Under semi-bandit feedback, Qin et al. (2014), Wen et al. (2015), and Takemura et al. (2021) incorporate linear reward assumption into the model, while Chen et al. (2018) studies volatile arms with submodular rewards. In contrast, Li et al. (2024) examines the full-bandit setting, and Zierahn et al. (2023) considers nonstochastic contexts.

We study contextual bandits in matching markets, where the focus is on forming bilateral matches between players and arms. We receive reward feedback for every matched pair and consider both stochastic and adversarial contexts. A key innovation is our *player-optimal stable regret* metric, which ensures fairness for each player rather than optimizing for a single aggregate benchmark as in CCB, introducing unique algorithmic challenges.

2 Problem Formulation

In this section, we formulate the problem setting of contextual bandits in matching markets over a finite horizon T. For a positive integer n, we use $[n]$ to represent $\{1, 2, \cdots, n\}$.

Consider a market with N players and K arms. We define the set of players as $\mathcal{N} = \{p_1, p_2, \cdots, p_N\}$ and the set of arms as $\mathcal{K} = \{a_1, a_2, \cdots, a_K\}$. Following previous work (Liu et al., 2020; 2021; Sankararaman et al., 2021; Kong & Li, 2023; Lin et al., 2025), we assume $N \leq K$ to ensure that each player has a chance of being matched.

2.1 Preference Structure

Fixed and known preferences of arms over players We assume each arm a_j has a fixed, strict, and known preference ordering P_{a_j} over players. That is, if $p_i \succ_{a_j} p_{i'}$ in P_{a_j}, arm a_j strictly prefers player p_i to player $p_{i'}$. This knowledge of arm-side preferences is a mild and common assumption, consistent with prior work in bandit learning in matching markets (Liu et al., 2020; Kong & Li, 2023; Lin et al., 2025), and reflects settings where arms (e.g., jobs or schools) have well-established rankings over players (e.g., candidates or applicants) through prior evaluations such as recruitment interviews and entrance examinations.

Dynamic and unknown preferences of players over arms In contrast, players' preferences over arms are dynamic, uncertain, and must be learned over time. We model these preferences via a linear contextual utility model. At each time step t, the utility of player p_i for arm a_j is given by $\boldsymbol{U}_{i,j}(t) = \theta_i^\top \mathrm{x}_j(t)$, where $\theta_i \in \mathbb{R}^d$ is an unknown *preference parameter* intrinsic to player p_i and fixed across all T rounds, and $\mathrm{x}_j(t) \in \mathbb{R}^d$ is the *contextual feature vector* of arm a_j at time t, revealed before the matching decision. The true utility is not directly observed; instead, upon matching player p_i to arm a_j at time t, the platform receives a noisy reward $y_{i,j}(t) = \boldsymbol{U}_{i,j}(t) + \epsilon_{i,j}(t)$, with $\epsilon_{i,j}(t)$ being i.i.d. noise with mean 0. And if a player is not matched at time t, the received reward $y_i(t)$ is set to be 0. With all player-arm pairs, we could construct a utility matrix $\boldsymbol{U}(t)$ encoding the preferences of players over arms. This formulation captures settings such as online labor markets, where workers (players) have latent preferences over jobs (arms) that can only be inferred through repeated interactions. In such markets, the ranking of a worker over available jobs is revealed only when job tasks are posted, while the platform can access the fixed preferences of jobs over workers through prior screening or interviews.

For the valuation model, we make the following assumptions throughout the paper.

Assumption 1. Contexts are bounded: $\|\mathrm{x}_j(t)\|_2 \leq B_x, \forall j \in [K]$ almost surely.

Assumption 2. The noises are bounded: $|\epsilon_{i,j}(t)| \leq B_\epsilon, \forall i \in [N], \forall j \in [K]$ almost surely.

Assumption 3. The preference parameters are bounded: $\|\theta_i\|_2 \leq B_\theta, \forall i \in [N]$.

Under Assumptions 1, 2, and 3, the linear valuation model and the additive noise implies $|y_{i,j}(t)| \leq B_y := B_\theta B_x + B_\epsilon$ almost surely. Following conventional linear contextual bandits literature, we assume $B_y \leq 1$.

2.2 MATCHING AND STABILITY

Stability is a key concept to characterize the matching in the market, which ensures there is no *justified envy* in the market. Formally, a matching μ_t is stable at time t if no player and arm has an incentive to abandon their current partner according to the utility matrix $\boldsymbol{U}(t)$ and the preference profile $(P_a)_{a\in\mathcal{K}}$.

Definition 1 (Stability)**.** A matching μ_t is stable at time t if there is no *blocking pair* (p_i, a_j) such that $p_i \succ_{a_j} \mu_t(a_j)$ and $\boldsymbol{U}_{i,j}(t) > \boldsymbol{U}_{i,\mu_t(p_i)}(t)$.

Stable matchings may not be unique. Let $\mathcal{S}_t := \{\mu : \mu \text{ is a stable matching w.r.t. } \boldsymbol{U}(t) \text{ and } (P_a)_{a\in\mathcal{K}}\}$ be the set of all stable matchings with respect to $\boldsymbol{U}(t)$ and $(P_a)_{a\in\mathcal{K}}$. Given complete preferences, the Gale-Shapley (GS) algorithm (Gale & Shapley, 1962) finds a stable matching after repeated proposals from one side of the market to the other. When the preferences for both sides are strict, due to the lattice structure of the problem, the matching returned by the GS algorithm is always optimal within $\mathcal{S}_t$ for each member of the proposing side (Knuth, 1976). Let $\mu_t^* = \{(i, \mu_t^*(i)) : i \in [N]\} \in \mathcal{S}_t$ be the players' most preferred stable matching. That is to say, $\boldsymbol{U}_{i,\mu_t^*(i)}(t) \geq \boldsymbol{U}_{i,\mu(i)}(t)$ for any $\mu \in \mathcal{S}_t, i \in [N]$. We define $\boldsymbol{U}_i^*(t) := \boldsymbol{U}_{i,\mu_t^*(i)}(t) = \max_{\mu\in\mathcal{S}_t} \boldsymbol{U}_{i,\mu_t(i)}(t)$ as the player's *optimal stable share* at time t.

2.3 REGRET DEFINITION

The performance of an algorithm is measured by the *player-optimal stable regret*. For a player $p_i \in \mathcal{N}$, this regret is defined at time T as the sum, over all rounds, of the difference in reward between being matched by the time-varying player-optimal stable policy μ_t^* and by the algorithm's implemented policy μ_t:

$$Reg_i(T) = \sum_{t=1}^{T} \boldsymbol{U}_i^*(t) - \mathbb{E}\left[\sum_{t=1}^{T} y_i(t)\right], \tag{1}$$

where $y_i(t) = y_{i,\mu_t(i)}(t)$ is the observed reward of player p_i under the matching μ_t at time t, with the expectation over algorithmic randomness and noise.

The goal of the problem is to design an online matching algorithm that minimizes the regret for every player by learning the unknown preference parameters $\{\theta_i\}$ while ensuring stable or near-stable matchings over time.

3 CONTEXTUAL MATCHING MARKETS WITH STOCHASTIC CONTEXTS

In this section, we consider the setting where the context feature vector $\mathrm{x}_j(t)$ of each arm a_j is drawn i.i.d. from a fixed but unknown distribution $\mathcal{D}_{\mathrm{x}_j}$ at every round t.

Given the realized contexts x_j, for $j \in [K]$, define the *minimum difference* between any two utilities for player p_i as[1] $\delta_{\min}^{(i)} = \min_{j,j'\in[K]} |(\theta_i)^\top(\mathrm{x}_j - \mathrm{x}_{j'})|$, and the minimum difference for the utility matrix as $\delta_{\min} = \min_{i\in[N]} \delta_{\min}^{(i)}$.

The difficulty of the learning and matching problem is directly tied to the minimum difference: a larger difference simplifies the task by making rankings more easily distinguishable. To formalize this measure of problem hardness, we define the following minimum preference gap.

Definition 2 (Minimum Preference Gap)**.** Given the ground-truth context distribution $(\mathcal{D}_{\mathrm{x}_j})_{j\in[K]}$ and preference parameters $(\theta_i)_{i\in[N]}$, the *minimum preference gap* $\Delta_{\min}$ is defined as the supremum of all gaps for which the event $\{\delta_{\min} \geq \Delta\}$ is sufficiently likely – specifically, the probability of the event is greater than $1 - \frac{\log T}{T\Delta^2}$. That is,

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

The definition of the minimum preference gap is motivated by the exploration-exploitation trade-off. To estimate a utility within a precision Δ and with confidence $1 - \frac{1}{T}$, we require $\mathcal{O}\left(\log T/\Delta^2\right)$ exploration samples. Conversely, for a given Δ, if $\mathbb{P}\left(\delta_{\min} \geq \Delta\right) > 1 - \zeta$ for some threshold $\zeta > 0$, then during exploitation there is at most an $\mathcal{O}(\zeta)$ probability of making a ranking error due to indistinguishability, leading to an $\mathcal{O}(\zeta T)$ regret contribution. Balancing the exploration regret, $\mathcal{O}\left(\log T/\Delta^2\right)$, against the exploitation regret, $\mathcal{O}(\zeta T)$, by setting these terms equal yields the optimal threshold $\zeta = \frac{\log T}{T\Delta^2}$, which directly informs our definition.

[1] Without causing ambiguity, we omit the time index in the following.

Remark 3.1. Our minimum preference gap $\Delta_{\min}$ generalizes the stricter sub-optimality gap $\tilde{\Delta}$ used in prior work (Li et al., 2022), which relies on a deterministic condition: a uniform lower bound for $\delta_{\min}$ over all T rounds. This could be viewed as $\mathbb{P}\left(\delta_{\min} \geq \tilde{\Delta}\right) = 1$. In contrast, our framework adopts a probabilistic soft threshold. It is clear that $\Delta_{\min} \geq \tilde{\Delta}$, and thus any regret upper bound established using our condition $\Delta_{\min}$ immediately provides a valid guarantee for the more restrictive $\tilde{\Delta}$.

3.1 Batched Adaptive Regret-Balancing Algorithm

We present Batched Adaptive Regret-Balancing (BARB, Algorithm 1) for contextual bandits in matching markets with stochastic contexts. BARB operates in batches. In batch k, it maintains a candidate gap Δ_k – initialized with the input Δ_1 and shrunk across batches (Line 21) – to iteratively approximate the true minimum preference gap $\Delta_{\min}$.

Each batch executes two adaptive phases: an *exploration phase* (Line 8 - 10), where we use maximum cardinality matchings (Line 9) to collect samples to refine parameter estimates, and an *exploitation phase*, where the Gale-Shapley algorithm is deployed when rankings are distinguishable (Line 13); otherwise, evidence is gathered to advance to the next batch (Line 14 - 16).

Take $\mathcal{G}_k^{(i)}$ as the set of rounds in which player p_i explores, i.e., p_i is involved in the maximum cardinality matching in those rounds in $\mathcal{G}_k^{(i)}$. Let $i_s = \mu_s(i)$ be the arm matched to player p_i under the implemented matching μ_s at time s. In time t_k of the exploration phase of the k-th batch, we use the following update rule for the Gram matrix $\boldsymbol{V}_i^{(t_k)}$ and the preference parameter $\hat{\theta}_i^{(t_k)}$:[2]

$$\boldsymbol{V}_i^{(t_k)} = \lambda \boldsymbol{I} + \sum_{s<t_k, s\in\mathcal{G}_k^{(i)}} \mathrm{x}_{i_s}\mathrm{x}_{i_s}^\top, \quad \hat{\theta}_i^{(t_k)} = \left(\boldsymbol{V}_i^{(t_k)}\right)^{-1} \sum_{s<t_k, s\in\mathcal{G}_k^{(i)}} \mathrm{x}_{i_s} y_{i,i_s}. \tag{3}$$

Given the ridge parameter λ, let $\delta = T^{-2}$, let the input parameter $\eta = B_y\sqrt{d\log\left(\frac{1+TB_x^2/\lambda}{\delta}\right)} + \lambda^{1/2}B_\theta$. Classical result on ridge regression in the bandit framework (Abbasi-Yadkori et al., 2011) shows that $\|\theta_i - \hat{\theta}_i^{(t_k)}\|_{\boldsymbol{V}_i^{(t_k)}} \leq \eta$ with high probability. On the other hand, the relationship $|\boldsymbol{U}_{i,j} - \hat{\boldsymbol{U}}_{i,j}| \leq \|\theta_i - \hat{\theta}_i^{(t_k)}\|_{\boldsymbol{V}_i^{(t_k)}}\|\mathrm{x}_j(t_k)\|_{\left(\boldsymbol{V}_i^{(t_k)}\right)^{-1}}$ suggests that having $\|\mathrm{x}_j(t_k)\|_{\left(\boldsymbol{V}_i^{(t_k)}\right)^{-1}} \leq \frac{\Delta_k}{\eta} := \xi_k$ suffices to ensure the estimated utility $\boldsymbol{U}$ is of precision Δ_k with high probability. This observation motivates the design of the adaptive criterion (Line 7) to transit between exploration and exploitation within each batch.

During the exploitation phase of batch k, for every player p_i, we construct confidence intervals with radius Δ_k for the estimated utility of each arm and track overlaps between intervals of distinct arms using a counter N_k. An excessive overlap indicates that Δ_k may be too large relative to $\Delta_{\min}$. Consequently, when N_k exceeds a specified threshold (Line 20), the algorithm moves to batch $k+1$.

3.2 Theoretical Analysis

We now present the regret upper bound for each player by following the BARB algorithm.

Theorem 3.1 (Upper Bound for Stochastic Contexts)**.** *Assume that the environment follows Assumption 1, 2 and 3, and the contexts are stochastic. Following the BARB algorithm, the player-optimal stable regret for $p_i \in \mathcal{N}$ is*

$$Reg_i(T) = \mathcal{O}\left(\frac{\log^2 T}{\Delta_{\min}^2} + \frac{\log T}{\Delta_{\min}^2}\right) = \mathcal{O}\left(\frac{\log^2 T}{\Delta_{\min}^2}\right). \tag{4}$$

We provide the proof sketch for Theorem 3.1 as follows, while the detailed proof is deferred to Appendix B.

Proof Sketch. In the exploration phase of the k-th batch, from the contruction of graph G_{t_k}, we know that $\|\mathrm{x}_{i_{t_k}}(t_k)\|_{\left(\boldsymbol{V}_i^{(t_k)}\right)^{-1}} > \xi_k$, while by elliptical potential lemma (Carpentier et al., 2020), we have $\sum_{t_k\in\mathcal{G}_k^{(i)}}\|\mathrm{x}_{i_{t_k}}(t_k)\|_{\left(\boldsymbol{V}_i^{(t_k)}\right)^{-1}} \leq \sqrt{|\mathcal{G}_k^{(i)}| d\log\left(\frac{|\mathcal{G}_k^{(i)}|+d\lambda}{d\lambda}\right)}$ and hence $|\mathcal{G}_k^{(i)}| \leq \frac{d\log\left(\frac{T+d\lambda}{d\lambda}\right)}{\xi_k^2}$. Let $\mathcal{G}_k$ be the

[2]Without ambiguity, we omit the time index s for the observed context and the reward.

Algorithm 1 Batched Adaptive Regret-Balancing (BARB), abstract version

Input: Time horizon T, preference profile $(P_a)_{a\in\mathcal{K}}$, context dimension d, parameter η, candidate gap Δ_1, ridge penalty parameter λ.

1: $t \leftarrow 1$.
2: **for** $k = 1, 2, \cdots$ **do**
3: $\quad\xi_k \leftarrow \frac{\Delta_k}{\eta}$; $\boldsymbol{V}_i^{(1)} \leftarrow \lambda \boldsymbol{I}_{d\times d}, \hat{\theta}_i^{(1)} \leftarrow \mathbf{0}_d, \forall i \in [N]$.
4: $\quad N_k \leftarrow 0$. ▷ Exploitation rounds with overlap CIs.
5: $\quad$**for** $t_k = 1, 2, \cdots$ **do**
6: $\quad\quad$Observe context $\mathrm{x}_j(t_k)$ for arm $a_j, \forall j \in [K]$.
7: $\quad\quad$**if** $\exists i, j$, s.t. $\|x_j(t_k)\|_{\left(\boldsymbol{V}_i^{(t_k)}\right)^{-1}} > \xi_k$ **then**
8: $\quad\quad\quad$Form a bipartite graph G_{t_k} such that $\forall (i,j) \in G_{t_k}$ has $\|\mathrm{x}_j(t_k)\|_{\left(\boldsymbol{V}_i^{(t_k)}\right)^{-1}} > \xi_k$.
9: $\quad\quad\quad\mu_{t_k} \leftarrow$ **Maximum Cardinality Matching** on G_{t_k}.
10: $\quad\quad\quad$Update $\boldsymbol{V}_i^{(t_k+1)}$ and $\hat{\theta}^{(t_k+1)}$ with Eq.(3).
11: $\quad\quad$**else**
12: $\quad\quad\quad\hat{\boldsymbol{U}}_{i,j}(t_k) \leftarrow \left(\hat{\theta}_i^{(t_k)}\right)^\top \mathrm{x}_j(t_k)$.
13: $\quad\quad\quad\mu_{t_k} \leftarrow$ **Deferred Acceptance** with $\hat{\boldsymbol{U}}(t_k)$.
14: $\quad\quad\quad$**if** $\exists i$, overlapping utility CIs exist **then**
15: $\quad\quad\quad\quad N_k \leftarrow N_k + 1$.
16: $\quad\quad\quad$**end if**
17: $\quad\quad$**end if**
18: $\quad\quad t \leftarrow t + 1$.
19: $\quad\quad$**if** $t > T$ **then** Stop and return. **end if**
20: $\quad\quad$**if** $N_k > \frac{3\log T}{16\Delta_k^2}$ **then**
21: $\quad\quad\quad\Delta_{k+1} \leftarrow \frac{\Delta_k}{\sqrt{2}}$, enter the next batch.
22: $\quad\quad$**end if**
23: $\quad$**end for**
24: **end for**

set of rounds that any player explores, we have $\mathcal{G}_k = \cup_i \mathcal{G}_k^{(i)}$ and hence $|\mathcal{G}_k| \le \sum_i |\mathcal{G}_k^{(i)}| \le \frac{Nd\log\left(\frac{T+d\lambda}{d\lambda}\right)}{\xi_k^2}$. Since $\xi_k = \frac{\Delta_k}{\eta}$ with $\eta = \mathcal{O}\left(\sqrt{\log T}\right)$, we have that the exploration regret is no larger than $B_y|\mathcal{G}_k| = \mathcal{O}\left(\frac{\log^2 T}{\Delta_k^2}\right)$.

In the exploitation phase of the k-th batch, we incur no regret when the confidence intervals for the utilities are non-overlapped by calling the Gale-Shapley algorithm. On the other hand, by the stopping threshold, we have at most N_k rounds with overlapped confidence intervals in the k-th exploitation phase, by the specified threshold for N_k (Line 20), the exploitation regret for the k-th batch is $\mathcal{O}(\frac{\log T}{\Delta_k^2})$.

Finally, with the multiplicative Chernoff bound (Mitzenmacher & Upfal, 2017), we could stop exploration at the first batch when $\Delta_k < \Delta_{\min}/4$ with high probability. By the doubling trick applied in the algorithm design, the regret of the k-th batch dominate the sum of the regret of the first $k-1$ batches, and hence the final regret is as shown in Eq.(4). □

The regret upper bound in Theorem 3.1 holds for a fully adaptive algorithm. The algorithm's strength is its minimal assumption: contexts need only be stochastic, with no further requirements on the underlying distribution. Under the additional structural assumption that the context covariance matrix has a uniformly bounded minimum eigenvalue, we can eliminate a $\log T$ factor from the regret bound. This refined bound then explicitly reflects the problem's hardness by inversely relating to this eigenvalue lower bound.

Assumption 4. Let $\mathcal{D}_{\mathrm{x}}$ be the joint distribution of contexts for all K arms. Define the context covariance matrix as $V = \mathbb{E}\left[XX^\top \mid X \in \mathcal{D}_{\mathrm{x}}\right]$. Then there exists a deterministic constant $\tilde{\lambda}$ such that for all $X \in \mathcal{D}_{\mathrm{x}}$ we have the minimum eigenvalue of the covariance matrix $\lambda_{\min}(V) \ge \tilde{\lambda} > 0$.

When Assumption 4 holds, it is possible to design a batched explore-then-commit (batched-ETC) algorithm that retains the regret-balancing principle of BARB (Algorithm 1) without requiring adaptive exploration. The resulting algorithm achieves an alternative regret upper bound of $\mathcal{O}(\frac{\log T}{\tilde{\lambda}^2\Delta_{\min}^2})$. The algorithmic details and the regret analysis are deferred to Appendix C.

3.3 MATCHING ASYMPTOTIC REGRET BOUNDS

In this subsection, we demonstrate that when the distribution of the minimum difference $\delta_{\min}$ satisfies a mild regularity condition, the minimum preference gap definition (Definition 2) yields an asymptotic regret upper bound of $\tilde{\mathcal{O}}(T^{2/3})$ as $T \to \infty$. Furthermore, we construct an instance that satisfies this regularity condition and attains a matching lower bound of $\Omega(T^{2/3})$, establishing the tightness of the asymptotic regret rate. These matching regret bounds also justify the rationality of the minimum preference gap definition.

We first present the following assumption that ensures the cumulative distribution function (CDF) of $\delta_{\min}$ is dominated by a linear function when it is close enough to 0.

Assumption 5. Under the definition of minimum difference $\delta_{\min}$, denote the CDF of $\delta_{\min}$ as $F(\Delta)$, for some constant Δ_0 sufficiently close to 0, and for some constant $c > 0$, we have, $\forall \Delta \leq \Delta_0, F(\Delta) \leq c\Delta$.

Assumption 5 is typically satisfied when the minimum difference $\delta_{\min}$ has a bounded probability density function (PDF) in a neighbourhood of zero.

The following theorem shows the asymptotic regret upper bound under Assumption 5.

Theorem 3.2. *Under Assumption 5, when T is sufficiently large, the regret of the BARB algorithm satisfies $Reg_i(T) = \tilde{\mathcal{O}}(T^{2/3})$, $\forall i \in [N]$.*

The proof of Theorem 3.2 can be found in Appendix D. We now provide a matching minimax regret lower bound in the following theorem.

Theorem 3.3. *Let $N = K = 3$. For any policy π, there exists an instance satisfying Assumption 5 such that for some constants $c, \Delta_0 > 0$, when $\Delta \leq \Delta_0$, the CDF of $\delta_{\min}$ obeys $F_T(\Delta) \leq c\Delta$ for all $T \geq 1$.*[3] *Then at least one player suffers regret $\Omega(T^{2/3})$, i.e., there exists $i \in [N]$ such that $Reg_i(T) \geq f(c)T^{2/3}$, where $f(c)$ depends only on c.*

The proof idea is to construct two instances with carefully designed preference parameters and context distributions, such that their resulting utility matrices differ in only a single entry. Insufficient sampling of this critical entry leads to high regret for one player in the first instance, while sampling it too frequently incurs high *social regret* (the sum of regrets across all players) in the second instance. This inherent trade-off yields the stated regret lower bound. The details can be found in Appendix E.

4 CONTEXTUAL MATCHING MARKETS WITH ADVERSARIAL CONTEXTS

In contrast to the stochastic setting, we now analyze worst-case scenarios in which no assumptions are placed on context generation, i.e., contexts may even be chosen by an adaptive adversary. This adversarial model offers robustness guarantees against non-stationary environments and strategic manipulation, ensuring algorithm performance even when stochastic assumptions fail to hold.

In the stochastic setting (Section 3), the regret bound scales as $1/\Delta_{\min}^2$, implying the problem becomes harder as $\Delta_{\min}$ shrinks. However, $\Delta_{\min}$ is defined with respect to a stationary context distribution, which no long exists in the adversarial case. Here, an adversary can select arbitrary contexts, potentially forcing the minimum utility difference $\delta_{\min}$ arbitrarily small for arbitrarily many rounds. This makes the problem intractable under the usual regret measure. To restore tractability, we first introduce a refined, tractable regret definition suitable for adversarial contexts. We then develop an adaptive algorithm for this setting and establish a corresponding regret upper bound.

4.1 TRACTABLE REGRET WITH ADVERSARIAL CONTEXTS

The inverse scaling of regret with the minimum preference gap represents a fundamental limit in bandit learning for matching markets, an unavoidable phenomenon demonstrated in Example 2 in Liu et al. (2020). Consequently, when the minimum gap is small or the preference profile admits ties, standard regret guarantees become vacuous. The underlying issue is that a market with ties lacks a single, universally utility-maximizing stable matching, creating a dilemma where uncertainty in one player-arm pair can propagate regret to other players. To address this challenge, Lin et al. (2025) introduced an approximation regret framework in stochastic bandits for static markets. In our setting – contextual bandits for dynamic

[3] The context distribution and hence the distribution of $\delta_{\min}$ may depend on T. Yet, the condition holds uniformly over T, and so the guarantees of Theorem 3.2 holds starting from some $T_0 > 0$.

markets with adversarial contexts – we adapt this idea and propose a regret definition that smoothly interpolates between rounds with large and small utility differences.

Before presenting the refined regret definition, we first introduce some notation about ε-stable matching and approximation oracle.

Definition 3 (ε-Stability)**.** Given $\varepsilon \geq 0$, a matching μ_t is ε-stable at time t if there is no *ε-blocking pair* (p_i, a_j) such that $p_i \succ_{a_j} \mu_t(a_j)$ and $\boldsymbol{U}_{i,j}(t) > \boldsymbol{U}_{i,\mu_t(p_i)}(t) + \varepsilon$.

The notion of ε-stability is a relaxation of stability, where setting $\varepsilon = 0$ makes it equivalent to stability (Definition 1). In general, ε-stable matching is not unique. We define

$$\mathcal{S}_t^{\varepsilon} := \{\mu : \mu \text{ is an } \varepsilon\text{-stable matching w.r.t. } \boldsymbol{U}(t) \text{ and } (P_a)_{a\in\mathcal{K}}\},$$

and we call a_j a *valid ε-stable match* for player p_i at time t if there exists an ε-stable matching matches p_i with a_j, and it is the *optimal ε-stable match* of player p_i at time t if it is the most preferred valid ε-stable match, i.e., there exists a matching $\mu_t^{\varepsilon} \in \mathcal{S}_t^{\varepsilon}$ such that $\mu_t^{\varepsilon}(i) = a_j$ and $\boldsymbol{U}_{i,\mu_t^{\varepsilon}(i)}(t) = \max_{\mu\in\mathcal{S}_t^{\varepsilon}} \boldsymbol{U}_{i,\mu(i)}(t)$. We say $\boldsymbol{U}_{i,\mu_t^{\varepsilon}(i)}(t)$ is the *ε-optimal stable share* for player p_i at time t, denoted as $\boldsymbol{U}_i^{\varepsilon}(t)$.

Since no single ε-stable matching maximizes all players' utilities simultaneously, Lin et al. (2025) proposed a utility approximation under bounded instability, ensuring each player achieves a certain satisfaction level in the presence of (statistical) ties in a static market. We adapt this to the contextual setting: when the minimum utility difference $\delta_{\min}$ is large – making preference rankings and the unique player-optimal stable matching easy to identify – we use the optimal stable share $\boldsymbol{U}_i^*(t)$ as the benchmark. Conversely, when $\delta_{\min}$ is small, we adopt an α-approximation of $\boldsymbol{U}_i^{\varepsilon}(t)$, where $\alpha \in (0, 1]$ is an approximation ratio. Formally, given predetermined parameters Δ and ε with $\varepsilon < \Delta$ – typically set as $\varepsilon = \Delta/2$ for simplicity – we define the player's *α-approximate Δ-optimal stable regret* as follows:

$$Reg_i^{\alpha,\Delta}(T) = \sum_{t=1}^{T}\left[\boldsymbol{U}_i^*(t)\mathbb{1}\{\delta_{\min}(t) > \Delta\} + \alpha\boldsymbol{U}_i^{\varepsilon}(t)\mathbb{1}\{\delta_{\min}(t) \leq \Delta\}\right] - \mathbb{E}\left[\sum_{t=1}^{T} y_i(t)\right]. \tag{5}$$

We assume access to an offline approximation oracle that, when given players' utility matrix $\boldsymbol{U}$, the uncertainty set with radius γ, this oracle returns a matching whose expected utility is guaranteed to be at least an α-fraction of $\boldsymbol{U}_i^{\varepsilon}$, with an additive error bounded by $2\gamma + \varepsilon$. For instance, Lin et al. (2025) constructed an oracle with an approximation ratio of $1/\lfloor \log_2 N + 2\rfloor$. For more details, please refer to Appendix F. In our online learning algorithm, we call this offline oracle with estimated utilities and constructed confidence intervals to approach the benchmark utility when $\delta_{\min}$ is small.

4.2 Adaptive Explore-Choose Oracle Algorithm

We propose the Adaptive Explore-Choose Oracle Algorithm (AdECO) for contextual bandits in matching markets with adversarial contexts. AdECO's exploration scheme follows the same adaptive principle as BARB: When $\|\mathrm{x}_j(t)\|_{\left(\boldsymbol{V}_i^{(t)}\right)^{-1}}$ is greater than the threshold ξ, we perform exploration through maximum cardinality matching on a bipartite graph $\mathcal{G}_t$, and update the estimation with the contextual information and the collected rewards (Line 4 - 7). Otherwise, the algorithm exploits by choosing between excuting the Gale-Shapley algorithm (Line 11) or calling the α-approximation oracle (Line 13), based on whether the confidence intervals for the utilities are sufficiently separated.

The setting of the parameter η follows the same idea as that employed in the stochastic contexts case. Take $\mathcal{G}^{(i)}$ as the set of rounds in which player p_i explores, i.e., p_i is involved in the maximum cardinality matching in those rounds $s \in \mathcal{G}^{(i)}$. In time t of the exploration phase, we use the following update rule for the Gram matrix $\boldsymbol{V}_i^{(t)}$ and the preference parameter $\hat{\theta}_i^{(t)}$:

$$\boldsymbol{V}_i^{(t)} = \lambda\boldsymbol{I} + \sum_{s<t, s\in\mathcal{G}^{(i)}} x_{i_s}x_{i_s}^{\top}, \quad \hat{\theta}_i^{(t)} = \left(\boldsymbol{V}_i^{(t)}\right)^{-1} \sum_{s<t, s\in\mathcal{G}^{(i)}} x_{i_s}y_{i,i_s}. \tag{6}$$

4.3 Theoretical Analysis

We now present the regret upper bound for each player by following the AdECO algorithm.

Theorem 4.1 (Upper Bound for Adversarial Contexts)**.** *Following the AdECO algorithm with parameter Δ and ε, the α-approximate Δ-optimal stable regret for $p_i \in \mathcal{N}$ is*

$$Reg_i^{\alpha,\Delta}(T) = \mathcal{O}\left(\frac{Nd\log^2 T}{(\Delta-\varepsilon)^2} + \frac{\Delta+\varepsilon}{2}\cdot T\right). \tag{7}$$

Algorithm 2 Adaptive Explore-Choose Oracle (AdECO), abstract version

Input: Time horizon T, preference profile $(P_a)_{a\in\mathcal{K}}$, context dimension d, parameter Δ, ε, and η, ridge penalty parameter λ.

$\xi \leftarrow \frac{\Delta-\varepsilon}{4\eta}$; $\boldsymbol{V}_i^{(1)} \leftarrow \lambda \boldsymbol{I}_{d\times d}, \hat{\theta}_i^{(1)} \leftarrow \mathbf{0}_d, \forall i \in [N]$.
for $t = 1, 2, \cdots, T$ **do**
  Observe context $\mathrm{x}_j(t)$ for arm a_j, $\forall j \in [K]$.
  if $\exists i, j$, s.t. $\|\mathrm{x}_j(t)\|_{\left(\boldsymbol{V}_i^{(t)}\right)^{-1}} > \xi$ **then**
    Form a bipartite graph G_t such that $\forall (i,j) \in G_t$ has $\|\mathrm{x}_j(t)\|_{\left(\boldsymbol{V}_i^{(t)}\right)^{-1}} > \xi$.
    $\mu_t \leftarrow$ **Maximum Cardinality Matching** on G_t.
    Update $\boldsymbol{V}_i^{(t+1)}$ and $\hat{\theta}^{(t+1)}$ according to Eq.(6).
  else
    $\hat{\boldsymbol{U}}_{i,j}(t) \leftarrow \left(\hat{\theta}_i^{(t)}\right)^\top \mathrm{x}_j(t)$.
    if Distance between any two CIs $\geq \varepsilon$ **then**
      $\mu_t \leftarrow$ **Deferred Acceptance** with $\hat{\boldsymbol{U}}(t)$.
    else
      $\mu_t \leftarrow \alpha$**-Approximation Oracle** with $\hat{\boldsymbol{U}}(t)$ and instability tolerance $\frac{\Delta+\varepsilon}{2}$.
    end if
  end if
end for

By setting $\Delta = \mathcal{O}(T^{-1/3})$, $\varepsilon = \Delta/2$, we have the regret upper bound as $\mathcal{O}(T^{2/3})$.

We provide the proof sketch for Theorem 4.1 as follows, while the detailed proof is deferred to Appendix G.

Proof Sketch. Similar to the proof in Theorem 3.1, in the exploration phase, by elliptical potential lemma (Carpentier et al., 2020), we have that $|\mathcal{G}| \leq \frac{Nd\log\left(\frac{T+d\lambda}{d\lambda}\right)}{\xi^2} = \frac{16\eta^2 Nd\log\left(\frac{T+d}{d}\right)}{(\Delta-\varepsilon)^2} = \mathcal{O}\left(\frac{Nd\log^2 T}{(\Delta-\varepsilon)^2}\right)$.

In every exploitation round, if we call the Gale-Shapley algorithm, we incur 0 or negative regret since it returns the unique player-optimal stable matching when we have correct estimation of the preference ranking, while if we call the α-approximation oracle with $\hat{\boldsymbol{U}}(t)$ and instability tolerance $\frac{\Delta+\varepsilon}{2}$, we are in the scenario where $\delta_{\min}$ is small and we incur $\frac{\Delta+\varepsilon}{2}$ approximation regret.

Therefore, the final regret is $\mathcal{O}\left(\frac{Nd\log^2 T}{(\Delta-\varepsilon)^2} + \frac{\Delta+\varepsilon}{2}\cdot T\right)$. □

Remark 4.1. Inspired by the adversarial context results, we could call back the stochastic context case, if the distribution of $\delta_{\min}$ satisfies Assumption 5, then by the asymptotic regret upper bound (Theorem 3.2), we know that by setting a maximum exploration time of $\mathcal{O}(T^{\frac{2}{3}+\zeta})$ for some small ζ, it suffices to finish exploration. And if the exploration could not stop until $\mathcal{O}(T^{\frac{2}{3}+\zeta})$ rounds, it is probably that the distribution of $\delta_{\min}$ is of poor condition, i.e., the probability of having small enough $\delta_{\min}$ is rather high. In this case, we could switch between the GS oracle and the approximation oracle as we propose in the adversarial context case, and gain a meaningful regret guarantee with respect to the α-approximate Δ-optimal stable regret.

5 Numerical Experiments

In this section, we empirically validate the convergence of BARB (stochastic contexts) and AdECO (adversarial contexts). For the stochastic setting, we also benchmark performance against Batched-ETC (Appendix C) and ETC (Li et al., 2022).

Each experiment runs for $T = 200k$ rounds, and results are averaged over 20 independent trials. Besides the averaged cumulative regret, we also plot the standard error, defined as the standard deviation divided by $\sqrt{20}$.

Due to space constraints, we present results for the stochastic setting with a large minimum preference gap $\Delta_{\min}$ and a small minimum eigenvalue of the context covariance matrix; additional comparisons are provided in Appendix H.

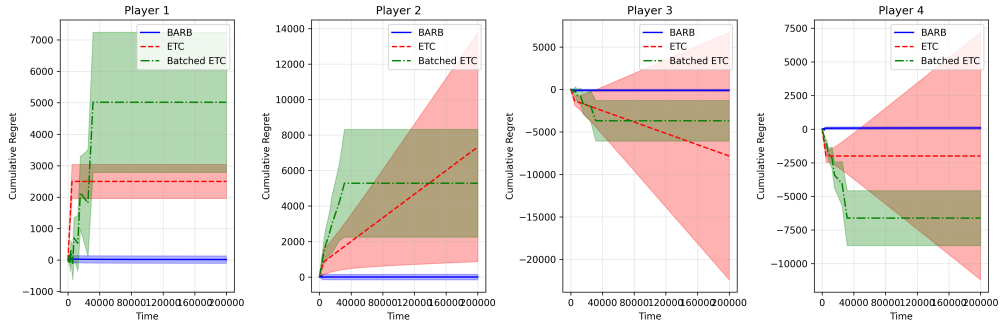

Figure 1: Experimental Comparison of BARB, Batched-ETC, and ETC in a market with 4 players and 4 arms.

We evaluate player-optimal stable regret in a market with $N = 4$ players and $K = 4$ arms. Contexts have dimension $d = 3$. Each arm's preference over players is a random permutation of $\{1, \cdots, N\}$. For preference and context generation, we first sample two orthonormal vectors $v, z \in \mathbb{R}^d$. We set $\theta_i = v$ for all $i \in [N]$, ensuring $\|\theta_i\| = 1$. Next, we define equally spaced values $\gamma_j = \frac{j-1}{K-1}$ for $j \in [K]$ and set $\tilde{\mathrm{x}}_j = \gamma_j v + \sqrt{1-\gamma_j^2} z$. At each time t, the k-th entry of the context vector $\mathrm{x}_j(t)$ is independently drawn from a Gaussian distribution with mean $\tilde{\mathrm{x}}_{j,k}$ and variance 0.01. The ground-truth utility for player p_i and arm a_j at time t is $\boldsymbol{U}_{i,j}(t) = \theta_i^\top \mathrm{x}_j(t)$. When a match occurs, a reward is sampled from the Gaussian distribution with mean $\boldsymbol{U}_{i,j}(t)$ and variance 0.05. Then we would truncate it so that the absolute value of the observed reward is bounded by $B_y = 1$.

The ridge penalty parameter is set to $\lambda = 1$ for all algorithms. For ETC, the exploration length is $h = 5000$; for Batched-ETC, the initial exploration length is $T_1 = 100$; and for BARB, the initial candidate gap is $\Delta_1 = 0.5$.

We present the player-optimal stable regret for each player in Figure 1. The results demonstrate that BARB outperforms both ETC and Batched-ETC in settings where the minimum eigenvalue $\lambda_{\min}$ of the context covariance matrix is small. In Appendix H, we provide a complementary example where $\lambda_{\min}$ is large, using the same algorithmic parameters. There, BARB achieves performance comparable to ETC. Since the covariance structure of contexts is typically unknown in practice, the fully adaptive design of BARB and the fact that its regret upper bound does not depend on the covariance structure of the context distribution offer more robust performance across varying problem instances than the parameter-dependent ETC.

6 Conclusions and Discussions

This paper studies contextual bandits in matching markets under a linear utility model, with the objective of minimizing individual *player-optimal stable regret*. The problem is complex: minor context shifts can drastically reconfigure the stable matching benchmark, causing disproportionate regret. We address this for both *stochastic and adversarial contexts*, introducing a novel *minimum preference gap* measure for the former and a tractable regret notion for the latter. We develop fully adaptive algorithms with theoretical guarantees for both settings. For the stochastic case, we also provide matched instance-independent regret bounds. While the current framework is centralized, extending it to decentralized markets presents a natural direction for future work.

Acknowledgements

This research was supported by Israel Science Foundation research grant (ISF's No. 4118/25) and the Maimonides Fund's Future Scientists Center. Vianney Perchet's research was supported in part by the French National Research Agency (ANR) in the framework of the PEPR IA FOUNDRY project (ANR-23-PEIA-0003) and through the grant DOOM ANR-23-CE23-0002. It was also funded by the European Union (ERC, Ocean, 101071601). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

REFERENCES

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

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.

Alexandra Carpentier, Claire Vernade, and Yasin Abbasi-Yadkori. The elliptical potential lemma revisited. *arXiv preprint arXiv:2010.10182*, 2020.

Lixing Chen, Jie Xu, and Zhuo Lu. Contextual combinatorial multi-armed bandits with volatile arms and submodular reward. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 208–214. JMLR Workshop and Conference Proceedings, 2011.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

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

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. Competing bandits in non-stationary matching markets. *IEEE Transactions on Information Theory*, 70(4): 2831–2850, 2024.

Botao Hao, Tor Lattimore, and Csaba Szepesvari. Adaptive exploration in linear contextual bandit. In *International Conference on Artificial Intelligence and Statistics*, pp. 3536–3545. PMLR, 2020.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

Meena Jagadeesan, Alexander Wei, Yixin Wang, Michael I Jordan, and Jacob Steinhardt. Learning equilibria in matching markets with bandit feedback. *Journal of the ACM*, 70(3):1–46, 2023.

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.

Donald E Knuth. Mariages stables et leurs relations avec d'autres problèmes combinatoires (stable marriage and its relation to other combinatorial problems). In *CRM Proceedings and Lecture Notes*, volume 10. Les Presses de l'Université de Montréal, 1976.

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, Jingqi Tang, Mingzhu Li, Pinyan Lu, John C.S. Lui, and Shuai Li. Bandit learning in matching markets with indifference. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=7ENakslm9J`.

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

Shuai Li, Zilong Wang, and Fang Kong. A survey on bandit learning in matching markets. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 10546–10554, 2025.

Yexin Li, Zhancun Mu, and Siyuan Qi. A contextual combinatorial bandit approach to negotiation. In *International Conference on Machine Learning*, pp. 28448–28465. PMLR, 2024.

Yuantong Li, Chi-hua Wang, Guang Cheng, and Will Wei Sun. Dynamic matching bandit for two-sided online markets. *arXiv preprint arXiv:2205.03699*, 2022.

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*, 2025. URL https://openreview.net/forum?id=HPmDgKOlM3.

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.

Tsunenori Mine, Tomoyuki Kakuta, and Akira Ono. Reciprocal recommendation for job matching with bidirectional feedback. In *2013 Second IIAI International Conference on Advanced Applied Informatics*, pp. 39–44. IEEE, 2013.

Michael Mitzenmacher and Eli Upfal. *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis*. Cambridge university press, 2017.

Deepan Muthirayan, Chinmay Maheshwari, Pramod Khargonekar, and Shankar Sastry. Competing bandits in time varying matching markets. In *Learning for Dynamics and Control Conference*, pp. 1020–1031. PMLR, 2023.

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

Lijing Qin, Shouyuan Chen, and Xiaoyan Zhu. Contextual combinatorial bandit and its application on diversified online recommendation. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pp. 461–469. SIAM, 2014.

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.

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.

Chengchun Shi, Runzhe Wan, Ge Song, Shikai Luo, Hongtu Zhu, and Rui Song. A multiagent reinforcement learning framework for off-policy evaluation in two-sided markets. *The Annals of Applied Statistics*, 17(4):2701–2722, 2023.

Kei Takemura, Shinji Ito, Daisuke Hatano, Hanna Sumita, Takuro Fukunaga, Naonori Kakimura, and Ken-ichi Kawarabayashi. Near-optimal regret bounds for contextual combinatorial semi-bandits with linear payoff functions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 9791–9798, 2021.

Joel A Tropp. An introduction to matrix concentration inequalities. *Foundations and trends® in machine learning*, 8(1-2):1–230, 2015.

Matilde Tullii, Solenne Gaucher, Nadav Merlis, and Vianney Perchet. Improved algorithms for contextual dynamic pricing. *Advances in Neural Information Processing Systems*, 37:126088–126117, 2024.

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.

Zheng Wen, Branislav Kveton, and Azin Ashkan. Efficient learning in large-scale combinatorial semi-bandits. In *International Conference on Machine Learning*, pp. 1113–1122. PMLR, 2015.

Huasen Wu, Xueying Guo, and Xin Liu. Adaptive exploration-exploitation tradeoff for opportunistic bandits. In *International Conference on Machine Learning*, pp. 5306–5314. PMLR, 2018.

YiRui Zhang and Zhixuan Fang. Decentralized two-sided bandit learning in matching market. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024.

Lukas Zierahn, Dirk van der Hoeven, Nicolo Cesa-Bianchi, and Gergely Neu. Nonstochastic contextual combinatorial bandits. In *International conference on artificial intelligence and statistics*, pp. 8771–8813. PMLR, 2023.

A BATCHED ADAPTIVE REGRET-BALANCING ALGORITHM

The full version of Batched Adaptive Regret-Balancing (BARB) algorithm is presented in Algorithm 3.

Algorithm 3 Batched Adaptive Regret-Balancing (BARB), full version

Input: Time horizon T, preference profile $(P_a)_{a\in\mathcal{K}}$, context dimension d, parameter η, initial candidate gap Δ_1, ridge penalty parameter λ.

$t \leftarrow 1$.
for $k = 1, 2, \cdots$ **do**
  $\xi_k \leftarrow \frac{\Delta_k}{\eta}$.
  $\boldsymbol{V}_i^{(1)} \leftarrow \lambda \boldsymbol{I}_{d\times d}$, $\hat{\theta}_i^{(1)} \leftarrow \mathbf{0}_d$, $\forall i \in [N]$.
  $N_k \leftarrow 0$. ▷ Number of rounds with overlapping CIs during exploitation.
  for $t_k = 1, 2, \cdots$ **do**
    Observe context $\mathrm{x}_j(t_k)$ for every arm a_j, $\forall j \in [K]$.
    if $\exists i \in [N], j \in [K]$, s.t. $\|\mathrm{x}_j(t_k)\|_{(\boldsymbol{V}_i^{(t_k)})^{-1}} > \xi_k$ **then** ▷ Exploration
      Collect all (i, j) such that $\|\mathrm{x}_j(t_k)\|_{(\boldsymbol{V}_i^{(t_k)})^{-1}} > \xi_k$, form a bipartite graph G_{t_k}.
      Find a **maximum cardinality matching** μ^c on G_{t_k}, let $M_{t_k}^c := \{i \in [N] : i \text{ is matched in } \mu^c\}$ match

$$\mu_{t_k}(i) \leftarrow \begin{cases} \mu^c(i) & \text{if } i \in M_{t_k}^c, \\ \emptyset & \text{otherwise.} \end{cases}$$

      Update $\boldsymbol{V}_i^{(t_k+1)}$ and $\hat{\theta}_i^{(t_k+1)}$, $\forall i \in M_{t_k}^c$ according to Eq.(3).
      $\boldsymbol{V}_i^{(t_k)} \leftarrow \boldsymbol{V}_i^{(t_k-1)}$ and $\hat{\theta}_i^{(t_k)} \leftarrow \hat{\theta}_i^{(t_k-1)}$, $\forall i \in [N] \setminus M_{t_k}^c$.
    else ▷ Exploitation
      $\hat{\theta}_i^{(t_k)} \leftarrow \hat{\theta}_i^{(t_k-1)}$, $\forall i \in [N]$.
      $\hat{\boldsymbol{U}}_{i,j}(t_k) = (\hat{\theta}_i^{(t_k)})^\top \mathrm{x}_j(t_k)$, $F_i \leftarrow$ False.
      $\mu_{t_k}(i) \leftarrow \mu^{GS}(i)$, $\forall i \in [N]$. (Input $\hat{\boldsymbol{U}}(t_k)$ and $(P_a)_{a\in\mathcal{K}}$ for **Gale-Shapley**)
      for $i = 1, 2, \cdots, N$ **do**
        $\hat{\boldsymbol{U}}_{i,\cdot}^{\text{sort}}(t_k) \leftarrow \text{Sort}\left(\hat{\boldsymbol{U}}_{i,\cdot}(t_k), \text{decreasing}\right)$
        $\Delta_{i,\min} \leftarrow \min\left\{\hat{\boldsymbol{U}}_{i,j}^{\text{sort}}(t_k) - \hat{\boldsymbol{U}}_{i,j+1}^{\text{sort}}(t_k)\right\}, j \in [N]$
        if $\Delta_{i,\min} > 2\Delta_k$ **then**
          $F_i \leftarrow$ True
        end if
      end for
      if $\exists i \in [N]$ s.t. $F_i ==$ False **then** ▷ CIs are overlapped
        $N_k \leftarrow N_k + 1$.
      end if
    end if
    $t \leftarrow t + 1$.
    if $t > T$ **then**
      Stop and return.
    end if
    if $N_k > \frac{3\log T}{16\Delta_k^2}$ **then** ▷ Exploitation regret exceeds exploration regret
      Break the loop and enter the next batch, set $\Delta_{k+1} \leftarrow \frac{\Delta_k}{\sqrt{2}}$,
    end if
  end for
end for

B PROOF OF THEOREM 3.1

Proof. Denote k^* as the largest batch number of Algorithm 1. That is to say, the counter N_k would not meet the pre-defined threshold in batch k^*. In batch k, given Δ_k, for the exploration phase, denote the estimation indices for player p_i as $\mathcal{G}_k^{(i)} = \left\{ t_{k_1}^{(i)}, \cdots, t_{k_{|\mathcal{G}_k^{(i)}|}}^{(i)} \right\}$. Notice that for $t_k \in \mathcal{G}_k^{(i)}$, we have $\|\mathrm{x}_j(t_k)\|_{\left(\boldsymbol{V}_i^{(t_k)}\right)^{-1}} > \xi_k = \frac{\Delta_k}{\eta}$ for $j = \mu_{t_k}(i)$, and we only use the contexts and matching reward from $\mathcal{G}_k^{(i)}$ to update the Gram matrix $\boldsymbol{V}_i$ and the estimation of the preference parameter $\hat{\theta}_i$. Take $\mathcal{G}_k$ as the set of rounds in which the market explores, we have $\mathcal{G}_k = \cup_{i\in[N]}\mathcal{G}_k^{(i)}$. Define $\mathcal{F}_k = \left\{ \exists t_k \notin \mathcal{G}_k, i \in [N], j \in [K] : |\hat{\boldsymbol{U}}_{i,j}(t) - \boldsymbol{U}_{i,j}(t)| > \Delta_k \right\}$ as the bad event that some preference of players over arms is not estimated well during the exploitation phase. When $k > k^*$, we define $\mathbb{P}(\mathcal{F}_k) = 0$. For the exploitation phase, using Δ_k as the width for the confidence interval constructed for the estimated utility, let $\mathcal{F}_d^{(t)} := \left\{ \hat{\boldsymbol{U}}_{i,j}^{\text{sort}}(t) - \hat{\boldsymbol{U}}_{i,j+1}^{\text{sort}}(t) > 2\Delta_k, \forall j \in [N], \forall i \in [N] \right\} \cap \left\{ \hat{\boldsymbol{U}}_{i,N}^{\text{sort}}(t) - \hat{\boldsymbol{U}}_{i,j}^{\text{sort}}(t) > 2\Delta_k, \forall j \in [N+1, K], \forall i \in [N] \right\}$ as the event that the first $(N+1)$-ranked arms have disjoint confidence intervals for any player p_i at time t. Denote T'_k as the total length of the k-th batch. Denote $\mathcal{F}_{\Delta_{k^*}} := \left\{ \Delta_{k^*} \geq \frac{\Delta_{\min}}{4\sqrt{2}} \right\}$ as the event that the algorithm stops at an early batch.

The optimal stable regret for each player p_i by following BARB (Algorithm 1) satisfies

$$
\begin{aligned}
Reg_i(T) =& \mathbb{E}\left[\sum_{t=1}^{T}\left(\boldsymbol{U}_i^*(t)-\boldsymbol{U}_{i,\mu_t(i)}(t)\right)\right] && (8)\\
=& \mathbb{E}\left[\sum_{t=1}^{T}\left(\boldsymbol{U}_i^*(t)-\boldsymbol{U}_{i,\mu_t(i)}(t)\right)\mathbb{1}\{\forall k\in[T]:\neg\mathcal{F}_k\}\right]+\mathbb{E}\left[\sum_{t=1}^{T}\left(\boldsymbol{U}_i^*(t)-\boldsymbol{U}_{i,\mu_t(i)}(t)\right)\mathbb{1}\{\exists k\in[T]:\mathcal{F}_k\}\right]\\
\leq& \mathbb{E}\left[\sum_{t=1}^{T}\left(\boldsymbol{U}_i^*(t)-\boldsymbol{U}_{i,\mu_t(i)}(t)\right)\mathbb{1}\{\forall k\in[T]:\neg\mathcal{F}_k\}\right]+B_yT\mathbb{P}\left(\exists k\in[T]:\mathcal{F}_k\right) && (9)\\
\leq& \mathbb{E}\left[\sum_{t=1}^{T}\left(\boldsymbol{U}_i^*(t)-\boldsymbol{U}_{i,\mu_t(i)}(t)\right)\mathbb{1}\{\forall k\in[T]:\neg\mathcal{F}_k\}\right]+B_yT\cdot\frac{NK}{T} && (10)\\
=& \mathbb{E}\left[\sum_{t=1}^{T}\left(\boldsymbol{U}_i^*(t)-\boldsymbol{U}_{i,\mu_t(i)}(t)\right)\mathbb{1}\{\{\forall k\in[k^*]:\neg\mathcal{F}_k\}\cap\mathcal{F}_{\Delta_{k^*}}\}\right]\\
&+\mathbb{E}\left[\sum_{t=1}^{T}\left(\boldsymbol{U}_i^*(t)-\boldsymbol{U}_{i,\mu_t(i)}(t)\right)\mathbb{1}\{\{\forall k\in[k^*]:\neg\mathcal{F}_k\}\cap\neg\mathcal{F}_{\Delta_{k^*}}\}\right]+B_yNK\\
\leq& \mathbb{E}\left[\sum_{t=1}^{T}\left(\boldsymbol{U}_i^*(t)-\boldsymbol{U}_{i,\mu_t(i)}(t)\right)\mathbb{1}\{\{\forall k\in[k^*]:\neg\mathcal{F}_k\}\cap\mathcal{F}_{\Delta_{k^*}}\}\right]\\
&+B_yT\mathbb{P}\left(\{\forall k\in[k^*]:\neg\mathcal{F}_k\}\cap\neg\mathcal{F}_{\Delta_{k^*}}\right)+B_yNK && (11)\\
\leq& \mathbb{E}\left[\sum_{k=1}^{k^*}\left[\sum_{t_k\in\mathcal{G}_k}\left(\boldsymbol{U}_i^*(t_k)-\boldsymbol{U}_{i,\mu_t(i)}(t_k)\right)+\sum_{t_k\notin\mathcal{G}_k}\left(\boldsymbol{U}_i^*(t_k)-\boldsymbol{U}_{i,\mu_t(i)}(t_k)\right)\right]\mathbb{1}\{\{\forall k\in[k^*]:\neg\mathcal{F}_k\}\cap\mathcal{F}_{\Delta_{k^*}}\}\right]\\
&+B_yT\cdot\frac{1}{T}+B_yNK && (12)\\
\leq& \mathbb{E}\Bigg[\sum_{k=1}^{k^*}\Bigg[B_y|\mathcal{G}_k|\mathbb{1}\{\{\forall k\in[k^*]:\neg\mathcal{F}_k\}\cap\mathcal{F}_{\Delta_{k^*}}\} && (13)\\
&+\sum_{t_k\notin\mathcal{G}_k}\Bigg[\left(\boldsymbol{U}_i^*(t_k)-\boldsymbol{U}_{i,\mu_t(i)}(t_k)\right)\mathbb{1}\Big\{\{\forall k\in[k^*]:\neg\mathcal{F}_k\}\cap\mathcal{F}_{\Delta_{k^*}}\cap\mathcal{F}_d^{(t_k)}\Big\}\\
&+\left(\boldsymbol{U}_i^*(t_k)-\boldsymbol{U}_{i,\mu_t(i)}(t_k)\right)\mathbb{1}\Big\{\{\forall k\in[k^*]:\neg\mathcal{F}_k\}\cap\mathcal{F}_{\Delta_{k^*}}\cap\neg\mathcal{F}_d^{(t_k)}\Big\}\Bigg]\Bigg]\Bigg]+B_y(1+NK)\\
\leq& \mathbb{E}\Bigg[\sum_{k=1}^{k^*}\Bigg[B_y|\mathcal{G}_k|\mathbb{1}\{\{\forall k\in[k^*]:\neg\mathcal{F}_k\}\cap\mathcal{F}_{\Delta_{k^*}}\}\\
&+\sum_{t_k\notin\mathcal{G}_k}\left(\boldsymbol{U}_i^*(t_k)-\boldsymbol{U}_{i,\mu_t(i)}(t_k)\right)\mathbb{1}\Big\{\{\forall k\in[k^*]:\neg\mathcal{F}_k\}\cap\mathcal{F}_{\Delta_{k^*}}\cap\neg\mathcal{F}_d^{(t_k)}\Big\}\Bigg]\Bigg]+B_y(1+NK) && (14)\\
\leq& B_y\mathbb{E}\left[\sum_{k=1}^{k^*}\left[|\mathcal{G}_k|+N_k\right]\mathbb{1}\{\{\forall k\in[k^*]:\neg\mathcal{F}_k\}\cap\mathcal{F}_{\Delta_{k^*}}\}\right]+B_y(1+NK) && (15)\\
\leq& B_y\mathbb{E}\left[\sum_{k=1}^{k^*}\left(\frac{\eta^2Nd\log\left(\frac{T+d\lambda}{d\lambda}\right)}{\Delta_k^2}+\frac{3\log T}{16\Delta_k^2}\right)\mathbb{1}\{\{\forall k\in[k^*]:\neg\mathcal{F}_k\}\cap\mathcal{F}_{\Delta_{k^*}}\}\right]+B_y(1+NK) && (16)\\
\leq& 2B_y\mathbb{E}\left[\left(\frac{\eta^2Nd\log\left(\frac{T+d\lambda}{d\lambda}\right)}{\Delta_{k^*}^2}+\frac{3\log T}{16\Delta_{k^*}^2}\right)\mathbb{1}\{\{\forall k\in[k^*]:\neg\mathcal{F}_k\}\cap\mathcal{F}_{\Delta_{k^*}}\}\right]+B_y(1+NK) && (17)\\
\leq& 64B_y\left(\frac{\eta^2Nd\log\left(\frac{T+d\lambda}{d\lambda}\right)}{\Delta_{\min}^2}+\frac{3\log T}{\Delta_{\min}^2}\right)+B_y(1+NK)=\mathcal{O}\left(\frac{\log^2T}{\Delta_{\min}^2}+\frac{\log T}{\Delta_{\min}^2}\right) && (18)
\end{aligned}
$$

Eq.(8) holds according to the assumption that the noises are i.i.d. with mean 0. Eq.(9), (11), and (13) comes from the fact that under Assumptions 1, 2 and 3, both the OSS and the expected reward are in $[0, B_y]$, and hence the instantaneous regret $\boldsymbol{U}_i^*(t) - \boldsymbol{U}_{i,\mu_t(i)}(t)$ is bounded by B_y. Eq.(10) relies on Lemma B.1. Eq.(12) holds according to Lemma B.2. Eq. (14) holds because, with Lemma B.3, under the good event $\neg\mathcal{F}_k$ (ensuring all utilities are well-estimated) and the event $\mathcal{F}_d^{(t_k)}$ (ensuring the top $(N+1)$-ranked confidence intervals are disjoint at time t_k), the Gale-Shapley subroutine invoked by Algorithm 1 is guaranteed to output the OSS for that round. Eq.(15) follows from the algorithm setting that N_k counts the number of times during exploitation that the top $(N+1)$-ranked CIs are overlapped. Eq.(16) holds according to Lemma B.4 and the algorithm setting that when $N_k > \frac{3\log T}{16\Delta_k^2}$, we would move to the next batch. Eq.(17) holds based on Lemma B.5. Eq.(18) follows from the parameter setting that $\eta = \mathcal{O}\left(\sqrt{\log T}\right)$.

□

Lemma B.1. *Let $\delta = T^{-2}$, $\eta = B_y\sqrt{d\log\left(\frac{1+TB_x^2/\lambda}{\delta}\right)} + \lambda^{1/2}B_\theta$, we have $\mathbb{P}\left(\exists k \in [T] : \mathcal{F}_k\right) \leq \frac{NK}{T}$.*

Proof.

$$\begin{aligned}
\mathbb{P}\left(\exists k \in [T] : \mathcal{F}_k\right) =& \mathbb{P}\left(\exists k \in [T], \exists t_k \notin \mathcal{G}_k, i \in [N], j \in [K] : |\hat{\boldsymbol{U}}_{i,j}(t_k) - \boldsymbol{U}_{i,j}(t_k)| > \Delta_k\right) \\
\leq& \sum_{k=1}^{T}\sum_{i=1}^{N}\sum_{j=1}^{K}\mathbb{P}\left(\exists t_k \notin \mathcal{G}_k : |\hat{\boldsymbol{U}}_{i,j}(t_k) - \boldsymbol{U}_{i,j}(t_k)| > \Delta_k\right) && (19)\\
\leq& \sum_{k=1}^{T}\sum_{i=1}^{N}\sum_{j=1}^{K}\mathbb{P}\left(\exists t_k \notin \mathcal{G}_k : \|\hat{\theta}_i - \theta_i\|_{\boldsymbol{V}_i^{(t_k)}}\|\mathrm{x}_j(t)\|_{\left(\boldsymbol{V}_i^{(t_k)}\right)^{-1}} > \Delta_k\right) && (20)\\
\leq& \sum_{k=1}^{T}\sum_{i=1}^{N}\sum_{j=1}^{K}\mathbb{P}\left(\exists t_k \notin \mathcal{G}_k : \|\hat{\theta}_i - \theta_i\|_{\boldsymbol{V}_i^{(t_k)}} > \eta\right) && (21)\\
\leq& TNK\delta = \frac{NK}{T}. && (22)
\end{aligned}$$

Eq.(19) follows from the union bound. Eq.(20) holds according to Cauchy-Schwarz inquality. Eq.(21) comes from the fact that for any $t_k \notin \mathcal{G}_k$, we have $\|\mathrm{x}_j(t_k)\|_{\left(\boldsymbol{V}_i^{(t_k)}\right)^{-1}} \leq \xi_k = \frac{\Delta_k}{\eta}$ in Algorithm 1. Eq.(22) holds according to Lemma J.2 (and notice it trivially holds if $k > k^*$). □

Lemma B.2. *It holds that $\mathbb{P}\left(\{\forall k \in [k^*] : \neg\mathcal{F}_k\} \cap \neg\mathcal{F}_{\Delta_{k^*}}\right) \leq \frac{1}{T}$.*

Proof. Let k_0 be the first k for which $\Delta_k < \Delta_{\min}/4$, i.e., $\Delta_{k_0} \geq \Delta_{\min}/4\sqrt{2}$ since $\Delta_{k+1} = \Delta_k/\sqrt{2}$. Since in batch k_0, the threshold for determining moving to the next batch is $N_{k_0} \geq \frac{3\log T}{16\Delta_{k_0}^2}$, we have

$$\neg\mathcal{F}_{\Delta_{k^*}} = \left\{\Delta_{k^*} < \frac{\Delta_{\min}}{4\sqrt{2}}\right\} \subseteq \left\{N_{k_0} \geq \frac{3\log T}{16\Delta_{k_0}^2}\right\}.$$

Therefore, to prove the claim, it suffices to show that

$$\mathbb{P}\left(\{\forall k \in [k^*] : \neg\mathcal{F}_k\} \cap \left\{N_{k_0} \geq \frac{3\log T}{16\Delta_{k_0}^2}\right\}\right) \leq \frac{1}{T}.$$

Let $\tilde{N}_k$ be the number of rounds of the k-th batch such that $\delta_{\min}^{(t_k)} \leq 4\Delta_k$. Under the event $\neg\mathcal{F}_{k_0}$, when there exists overlap CIs with CI radius Δ_{k_0} at time t_{k_0}, there must exist a player p_i and two arms a_j and $a_{j'}$ with $\hat{\boldsymbol{U}}_{i,j}(t_{k_0}) \geq \hat{\boldsymbol{U}}_{i,j'}(t_{k_0})$, such that $\hat{\boldsymbol{U}}_{i,j}(t_{k_0}) - \Delta_{k_0} \leq \hat{\boldsymbol{U}}_{i,j'}(t_{k_0}) + \Delta_{k_0}$, and hence $\boldsymbol{U}_{i,j}(t_{k_0}) \leq \hat{\boldsymbol{U}}_{i,j}(t_{k_0}) + \Delta_{k_0} \leq \hat{\boldsymbol{U}}_{i,j'}(t_{k_0}) + 3\Delta_{k_0} \leq \boldsymbol{U}_{i,j'}(t_{k_0}) + 4\Delta_{k_0}$, which implies $\delta_{\min}^{(t_{k_0})} \leq 4\Delta_{k_0}$. Therefore, under the event $\neg\mathcal{F}_{k_0}$, we have $N_{k_0} \leq \tilde{N}_{k_0}$, and hence it suffices to show that

$$\mathbb{P}\left(\{\forall k \in [k^*] : \neg\mathcal{F}_k\} \cap \left\{\tilde{N}_{k_0} \geq \frac{3\log T}{16\Delta_{k_0}^2}\right\}\right) \leq \mathbb{P}\left(\tilde{N}_{k_0} \geq \frac{3\log T}{16\Delta_{k_0}^2}\right) \leq \frac{1}{T}.$$

As $\Delta_{k_0} < \Delta_{\min}$, and recall the definition of the minimum preference gap $\Delta_{\min}$ (Definition 2), for every round t_{k_0} in batch k_0, we have

$$\mathbb{P}\left(\delta_{\min}^{(t_{k_0})} \leq 4\Delta_{k_0}\right) \leq \mathbb{P}\left(\delta_{\min}^{(t_{k_0})} \leq \Delta_{\min}\right) \leq \frac{\log T}{T\Delta_{\min}^2}.$$

Denote the length of the k-th batch as T'_k, we have

$$\mathbb{E}\left[\tilde{N}_{k_0}\right] = \mathbb{E}\left[\sum_{t_{k_0}=1}^{T'_{k_0}} \mathbb{1}\left\{\delta_{\min}^{(t_{k_0})} \leq 4\Delta_{k_0}\right\}\right] \leq T \cdot \frac{\log T}{T\Delta_{\min}^2} = \frac{\log T}{\Delta_{\min}^2}.$$

Let ρ satisfy $(1+\rho)\mathbb{E}\left[\tilde{N}_{k_0}\right] = \frac{3\log T}{\Delta_{\min}^2}$, then $\rho\mathbb{E}\left[\tilde{N}_{k_0}\right] \geq \frac{2\log T}{\Delta_{\min}^2}$ and $\rho \geq 2$. Therefore,

$$\begin{aligned}
\mathbb{P}\left(\tilde{N}_{k_0} \geq \frac{3\log T}{16\Delta_{k_0}^2}\right) &\leq \mathbb{P}\left(\tilde{N}_{k_0} \geq \frac{3\log T}{\Delta_{\min}^2}\right) \\
&= \mathbb{P}\left(\tilde{N}_{k_0} \geq (1+\rho)\,\mathbb{E}\left[\tilde{N}_{k_0}\right]\right) \\
&\leq \exp\left(-\frac{\rho^2\mathbb{E}\left[\tilde{N}_{k_0}\right]}{2+\rho}\right) \qquad (23) \\
&\leq \exp\left(-\frac{\rho\mathbb{E}\left[\tilde{N}_{k_0}\right]}{2}\right) \\
&\leq \exp\left(-\frac{\log T}{\Delta_{\min}^2}\right) \\
&\leq \frac{1}{T}. \qquad (24)
\end{aligned}$$

Eq.(23) holds according to the multiplicative Chernoff bound (Lemma J.3) by noticing that $\tilde{N}_{k_0}$ is the sum of independent Bernoulli trials. Eq.(24) follows from the assumptions that $B_y \leq 1$ so that $\Delta_{\min} \leq 1$. □

Lemma B.3. *In the Gale-Shapley algorithm, at most N proposals are made per player before termination. Consequently, each player's optimal stable match must be among its top N preferred arms.*

Proof. Within the Gale-Shapley algorithm, a proposal from a player results in the arm being tentatively matched to that player. To prove the claim, assume by contradiction that more than N distinct arms are proposed to. This would imply that at least $N+1$ players are involved in proposals. However, since there are exactly N players, once N proposals have been made, every player must be tentatively matched. At this point, the algorithm terminates. Therefore, the maximum number of distinct arms that can be proposed to is N. Since each player proposes to arms strictly in descending order of preference, it follows that a player's optimal stable match must be among its top N ranked arms. □

Lemma B.4. *In batch k, given η and the ridge penalty parameter λ, for a given Δ_k, let $\mathcal{G}_k$ denote the set of rounds in which the system explores. Almost surely, the length of the exploration phase $\mathcal{G}_k$ is bounded as*

$$|\mathcal{G}_k| \leq \frac{\eta^2 N d \log\left(\frac{T+d\lambda}{d\lambda}\right)}{\Delta_k^2}. \qquad (25)$$

Proof. The following proof is adapted from the proof in Tullii et al. (2024, Lemma 1).

Let $\mathcal{G}_k^{(i)} = \left\{t_k : p_i \in M_{t_k}^c\right\}$ denote the set of rounds in which player p_i is involved in the maximum cardinality matching, and updates its estimation for the Gram matrix and the preference parameter. For player p_i, for all $t_k \in \mathcal{G}_k^{(i)}$, we can write the Gram matrix as

$$\boldsymbol{V}_i^{(t_k)} = \lambda \boldsymbol{I}_d + \sum_{s<t_k, s\in\mathcal{G}_k^{(i)}} \mathrm{x}_{i_s}\mathrm{x}_{i_s}^\top,$$

where $i_s = \mu_s(i)$. By the elliptical potential lemma (Lemma J.1), we have that

$$\sum_{t_k \in \mathcal{G}_k^{(i)}} \|\mathrm{x}_{i_{t_k}}\|_{\left(\boldsymbol{V}_i^{(t_k)}\right)^{-1}} \leq \sqrt{|\mathcal{G}_k^{(i)}| d \log\left(\frac{|\mathcal{G}_k^{(i)}| + d\lambda}{d\lambda}\right)}.$$

Since for all $t_k \in \mathcal{G}_k^{(i)}$, we have that $\|\mathrm{x}_{i_{t_k}}\|_{\left(\boldsymbol{V}_i^{(t_k)}\right)^{-1}} \geq \xi_k = \frac{\Delta_k}{\eta}$, this implies that

$$|\mathcal{G}_k^{(i)}| \cdot \frac{\Delta_k}{\eta} \leq \sqrt{|\mathcal{G}_k^{(i)}| d \log\left(\frac{|\mathcal{G}_k^{(i)}| + d\lambda}{d\lambda}\right)}.$$

Now, almost surely, $|\mathcal{G}_k^{(i)}| \leq T$. Using this bound and reorganizing the inequality leads to

$$|\mathcal{G}_k^{(i)}| \leq \frac{\eta^2 d \log\left(\frac{T+d\lambda}{d\lambda}\right)}{\Delta_k^2}.$$

When the system explores at time t_k, at least one player p_i and one arm a_j triggers the condition that $\|\mathrm{x}_j(t_k)\|_{\left(\boldsymbol{V}_i^{(t_k)}\right)^{-1}} > \xi_k$ and is involved in the maximum cardinality matching. Therefore, we have $\mathcal{G}_k = \cup_{i \in [N]} \mathcal{G}_k^{(i)}$, and hence

$$|\mathcal{G}_k| \leq \sum_{i \in [N]} |\mathcal{G}_k^{(i)}| \leq \frac{\eta^2 N d \log\left(\frac{T+d\lambda}{d\lambda}\right)}{\Delta_k^2}.$$

□

Lemma B.5. *For any $n > 1$, $\sum_{k=1}^{n-1} \frac{1}{\Delta_k^2} \leq \frac{1}{\Delta_n^2}$.*

Proof. From the setting of Algorithm 1, we know that $\Delta_{k+1} = \frac{\Delta_k}{\sqrt{2}}$. Therefore,

$$\sum_{k=1}^{n-1} \frac{1}{\Delta_k^2} = \sum_{i=1}^{n-1} \frac{1}{2^i} \cdot \frac{1}{\Delta_n^2} \leq \frac{1}{\Delta_n^2}.$$

□

C BATCHED EXPLORE-THEN-COMMIT ALGORITHM

In this section, we propose the Batched Explore-then-Commit (Batched-ETC) algorithm. Following a similar regret-balancing idea as BARB, Algorithm 4 operates in batches. Starting with an initial exploration length T_1 that is the algorithm's input, in each batch k we keep a counter N_k counting the number of overlapping confidence intervals during exploitation. When N_k exceeds a certain threshold, it triggers a fresh batch $k+1$ where the exploration length $T_{k+1} = 2T_k$.

Algorithm 4 Batched Explore-then-Commit (Batched-ETC)

Input: Time horizon T, preference profile $(P_a)_{a\in\mathcal{K}}$, context dimension d, initial exploration length T_1, ridge penalty parameter λ.

$t \leftarrow 1$.
for $k = 1, 2, \cdots$ **do**
  $\boldsymbol{V}_i^{(0)} \leftarrow \lambda \boldsymbol{I}_{d\times d}$, $S_i^{(0)} \leftarrow \mathbf{0}_d$, $\hat{\theta}_i^{(0)} \leftarrow \mathbf{0}_d$, $\forall i \in [N]$.
  $N_k \leftarrow 0$. ▷ Number of rounds with overlapping CIs during exploitation.
  for $t_k = 1, 2, \cdots, T_k$ **do** ▷ Learning Step
    Observe context $\mathrm{x}_j(t_k)$ for every arm a_j, $\forall j \in [K]$.
    Choose a matching μ_{t_k}, match player p_i to arm $a_{i_{t_k}}$ where $i_{t_k} = \mu_{t_k}(i)$.
    For every player p_i, collect the reward $y_i(t_k)$.
    $\boldsymbol{V}_i^{(t_k)} \leftarrow \boldsymbol{V}_i^{(t_k-1)} + \mathrm{x}_{i_{t_k}}\mathrm{x}_{i_{t_k}}^\top$, $S_i^{(t_k)} \leftarrow S_i^{(t_k-1)} + \mathrm{x}_{i_{t_k}} y_i(t_k)$
  end for
  $\hat{\theta}_i^{(T_k)} \leftarrow \left(\boldsymbol{V}_i^{(T_k)}\right)^{-1} S_i^{(T_k)}$.
  for $t_k = T_k + 1, \cdots$ **do** ▷ Exploitation Step
    $\hat{\boldsymbol{U}}_{i,j}(t_k) = (\hat{\theta}_i^{(t_k)})^\top \mathrm{x}_j(t_k)$, $F_i \leftarrow$ False.
    $\mu_{t_k}(i) \leftarrow \mu^{GS}(i), \forall i \in [N]$. (Input $\hat{\boldsymbol{U}}(t_k)$ and $(P_a)_{a\in\mathcal{K}}$ for **Gale-Shapley**)
    $\Delta_k \leftarrow \sqrt{\frac{\log T}{T_k}}$
    for $i = 1, 2, \cdots, N$ **do**
      $\hat{\boldsymbol{U}}_{i,\cdot}^{\text{sort}}(t_k) \leftarrow \text{Sort}\left(\hat{\boldsymbol{U}}_{i,\cdot}(t_k), \text{decreasing}\right)$
      $\Delta_{i,\min} \leftarrow \min\left\{\hat{\boldsymbol{U}}_{i,j}^{\text{sort}}(t_k) - \hat{\boldsymbol{U}}_{i,j+1}^{\text{sort}}(t_k)\right\}, j \in [N]$
      if $\Delta_{i,\min} > 2\Delta_k$ **then**
        $F_i \leftarrow$ True
      end if
    end for
    $t \leftarrow t + 1$.
    if $t > T$ **then**
      Stop and return.
    end if
    if $\exists i \in [N]$, s.t. $F_i ==$ False **then** ▷ CIs are overlapped
      $N_k \leftarrow N_k + 1$.
      if $N_k > \frac{3\log T}{16\Delta_k^2}$ **then** ▷ Exploitation regret exceeds exploration regret
        Break the loop and enter the next batch, set $T_{k+1} \leftarrow 2T_k$,
      end if
    end if
  end for
end for

C.1 THEORETICAL ANALYSIS OF BATCHED-ETC (ALGORITHM 4)

Theorem C.1 (Regret Upper Bound for Batched-ETC with stochastic contexts)**.** *Assume that the environment follows Assumption 1, 2 and 3. Further, assume that the context distribution follows Assumption 4, following the Batched-ETC algorithm, the player-optimal stable regret for $p_i \in \mathcal{N}$ is*

$$Reg_i(T) = \mathcal{O}\left(\frac{\log T}{\tilde{\lambda}^2 \Delta_{\min}^2} + \frac{\log T}{\Delta_{\min}^2}\right). \tag{26}$$

Proof. Denote k^* as the largest batch number of Algorithm 4. That is to say, the counter N_k would not meet the pre-defined threshold in batch k^*. In batch k, given the exploration length T_k, define $\Delta_k = C\frac{\sqrt{\log T}}{\tilde{\lambda}\sqrt{T_k}}$ as the width for the confidence interval constructed for the estimated utility in the exploitation phase, where C is a constant specified as in Lemma C.1. Define $\mathcal{F}_k = \left\{\exists t_k > T_k, i \in [N], j \in [K] : |\hat{\boldsymbol{U}}_{i,j}(t) - \boldsymbol{U}_{i,j}(t)| > \Delta_k\right\}$ as the bad event that some preference of players over arms is not estimated well in the exploitation phase of batch k. When $k > k^*$, we define $\mathbb{P}(\mathcal{F}_k) = 0$. Let $\mathcal{F}_d^{(t)} := \left\{\hat{\boldsymbol{U}}_{i,j}^{\text{sort}}(t) - \hat{\boldsymbol{U}}_{i,j+1}(t) > 2\Delta_k, \forall j \in [N], \forall i \in [N]\right\} \cap \left\{\hat{\boldsymbol{U}}_{i,N}^{\text{sort}}(t) - \hat{\boldsymbol{U}}_{i,j}^{\text{sort}}(t) > 2\Delta_k, \forall j \in [N+1, K], \forall i \in [N]\right\}$ as the event that the first $(N+1)$-ranked arms have disjoint confidence intervals for any player p_i at time t. Denote T_k' as the total length of the k-th batch. Define $\mathcal{F}_{\Delta_{k^*}} := \left\{\Delta_{k^*} \geq \frac{\Delta_{\min}}{4\sqrt{2}}\right\}$ as the event that the algorithm stops at an early batch.

The optimal stable regret for each player p_i by following Batched-ETC (Algorithm 4) satisfies

$$
\begin{aligned}
Reg_i(T) =& \mathbb{E}\left[\sum_{t=1}^{T}\left(\boldsymbol{U}_i^*(t) - \boldsymbol{U}_{i,\mu_t(i)}(t)\right)\right] && (27)\\
=& \mathbb{E}\left[\sum_{t=1}^{T}\left(\boldsymbol{U}_i^*(t) - \boldsymbol{U}_{i,\mu_t(i)}(t)\right)\mathbb{1}\{\forall k\in[T]:\neg\mathcal{F}_k\}\right] + \mathbb{E}\left[\sum_{t=1}^{T}\left(\boldsymbol{U}_i^*(t) - \boldsymbol{U}_{i,\mu_t(i)}(t)\right)\mathbb{1}\{\exists k\in[T]:\mathcal{F}_k\}\right]\\
\leq& \mathbb{E}\left[\sum_{t=1}^{T}\left(\boldsymbol{U}_i^*(t) - \boldsymbol{U}_{i,\mu_t(i)}(t)\right)\mathbb{1}\{\forall k\in[T]:\neg\mathcal{F}_k\}\right] + B_yT\mathbb{P}\left(\exists k\in[T]:\mathcal{F}_k\right) && (28)\\
\leq& \mathbb{E}\left[\sum_{t=1}^{T}\left(\boldsymbol{U}_i^*(t) - \boldsymbol{U}_{i,\mu_t(i)}(t)\right)\mathbb{1}\{\forall k\in[T]:\neg\mathcal{F}_k\}\right] + B_yT\cdot\frac{NK}{T} && (29)\\
=& \mathbb{E}\left[\sum_{t=1}^{T}\left(\boldsymbol{U}_i^*(t) - \boldsymbol{U}_{i,\mu_t(i)}(t)\right)\mathbb{1}\{\{\forall k\in[k^*]:\neg\mathcal{F}_k\}\cap\mathcal{F}_{\Delta_{k^*}}\}\right]\\
&+\mathbb{E}\left[\sum_{t=1}^{T}\left(\boldsymbol{U}_i^*(t) - \boldsymbol{U}_{i,\mu_t(i)}(t)\right)\mathbb{1}\{\{\forall k\in[k^*]:\neg\mathcal{F}_k\}\cap\neg\mathcal{F}_{\Delta_{k^*}}\}\right] + B_yNK\\
\leq& \mathbb{E}\left[\sum_{t=1}^{T}\left(\boldsymbol{U}_i^*(t) - \boldsymbol{U}_{i,\mu_t(i)}(t)\right)\mathbb{1}\{\{\forall k\in[k^*]:\neg\mathcal{F}_k\}\cap\mathcal{F}_{\Delta_{k^*}}\}\right] + B_yT\mathbb{P}\left(\neg\mathcal{F}_{\Delta_{k^*}}\right) + B_yNK && (30)\\
\leq& \mathbb{E}\left[\sum_{k=1}^{k^*}\left[B_yT_k + \sum_{t_k=T_k+1}^{T_k'}\left(\boldsymbol{U}_i^*(t_k) - \boldsymbol{U}_{i,\mu_t(i)}(t_k)\right)\right]\mathbb{1}\{\{\forall k\in[k^*]:\neg\mathcal{F}_k\}\cap\mathcal{F}_{\Delta_{k^*}}\}\right] + B_yT\cdot\frac{1}{T} + B_yNK && (31)\\
\leq& \mathbb{E}\Bigg[\sum_{k=1}^{k^*}\Bigg[B_yT_k\mathbb{1}\{\{\forall k\in[k^*]:\neg\mathcal{F}_k\}\cap\mathcal{F}_{\Delta_{k^*}}\}\\
&+\sum_{t_k=T_k+1}^{T_k'}\Bigg[\left(\boldsymbol{U}_i^*(t_k) - \boldsymbol{U}_{i,\mu_t(i)}(t_k)\right)\mathbb{1}\Big\{\{\forall k\in[k^*]:\neg\mathcal{F}_k\}\cap\mathcal{F}_{\Delta_{k^*}}\cap F_d^{(t_k)}\Big\}\\
&+\left(\boldsymbol{U}_i^*(t_k) - \boldsymbol{U}_{i,\mu_t(i)}(t_k)\right)\mathbb{1}\Big\{\{\forall k\in[k^*]:\neg\mathcal{F}_k\}\cap\mathcal{F}_{\Delta_{k^*}}\cap\neg F_d^{(t_k)}\Big\}\Bigg]\Bigg]\Bigg] + B_y(1+NK)\\
\leq& \mathbb{E}\Bigg[\sum_{k=1}^{k^*}\Bigg[B_yT_k\mathbb{1}\{\{\forall k\in[k^*]:\neg\mathcal{F}_k\}\cap\mathcal{F}_{\Delta_{k^*}}\}\\
&+\sum_{t_k=T_k+1}^{T_k'}\left(\boldsymbol{U}_i^*(t_k) - \boldsymbol{U}_{i,\mu_t(i)}(t_k)\right)\mathbb{1}\Big\{\{\forall k\in[k^*]:\neg\mathcal{F}_k\}\cap\mathcal{F}_{\Delta_{k^*}}\cap\neg F_d^{(t_k)}\Big\}\Bigg]\Bigg] + B_y(1+NK) && (32)\\
\leq& B_y\mathbb{E}\left[\sum_{k=1}^{k^*}\left[T_k+N_k\right]\mathbb{1}\{\{\forall k\in[k^*]:\neg\mathcal{F}_k\}\cap\mathcal{F}_{\Delta_{k^*}}\}\right] + B_y(1+NK) && (33)\\
\leq& B_y\mathbb{E}\left[\sum_{k=1}^{k^*}\left(\frac{C^2\log T}{\tilde{\lambda}^2\Delta_k^2} + \frac{3\log T}{16\Delta_k^2}\right)\mathbb{1}\{\{\forall k\in[k^*]:\neg\mathcal{F}_k\}\cap\mathcal{F}_{\Delta_{k^*}}\}\right] + B_y(1+NK) && (34)\\
\leq& 2B_y\mathbb{E}\left[\left(\frac{C^2\log T}{\tilde{\lambda}^2\Delta_{k^*}^2} + \frac{3\log T}{16\Delta_{k^*}^2}\right)\mathbb{1}\{\{\forall k\in[k^*]:\neg\mathcal{F}_k\}\cap\mathcal{F}_{\Delta_{k^*}}\}\right] + B_y(1+NK) && (35)\\
\leq& 64B_y\left(\frac{C^2\log T}{\tilde{\lambda}^2\Delta_{\min}^2} + \frac{3\log T}{16\Delta_{\min}^2}\right) + B_y(1+NK) = \mathcal{O}\left(\frac{\log T}{\tilde{\lambda}^2\Delta_{\min}^2} + \frac{\log T}{\Delta_{\min}^2}\right).
\end{aligned}
$$

Eq.(27) holds according to the assumption that the noises are i.i.d. with mean 0. Eq.(28), (30) comes from the fact that under Assumptions 1, 2, 3, both the OSS and the expected reward are in $[0, B_y]$, and hence the instantaneous regret $\boldsymbol{U}_i^*(t) - \boldsymbol{U}_{i,\mu_t(i)}(t)$ is bounded by B_y. Eq.(29) relies on Lemma C.1. Eq.(31) follows from the same reasoning as we used in Lemma B.2. Eq.(32) holds because, with Lemma B.3, under the good event $\neg\mathcal{F}_k$ (ensuring all utilities are well-estimated) and the event $\mathcal{F}_d^{(t_k)}$ (ensuring the top $(N+1)$-ranked confidence intervals are disjoint at time t_k), the Gale-Shapley subroutine invoked by Algorithm 4 is guaranteed to output the OSS for that round. Eq.(33) follows from the algorithm setting that N_k counts the number of times during exploitation that the top $(N+1)$-ranked CIs are overlapped. Eq.(34) holds according to the definition of Δ_k and that the algorithm would move to the next batch whenever $N_k > \frac{3\log T}{16\Delta_k^2}$. Eq.(35) follows from the definition of Δ_k and that $T_{k+1} = 2T_k$, we have $\Delta_{k+1} = \frac{\Delta_k}{\sqrt{2}}$, and hence the regret from the k-th batch would dominate the cumulative regret from the first $(k-1)$-th batches as shown by Lemma B.5.

□

Lemma C.1. *Under Assumptions 1, 3 and 4, we have that for any given time t_k in batch k, for any fixed player p_i, arm a_j,*

$$\mathbb{P}\left(|\hat{\boldsymbol{U}}_{i,j}(t) - \boldsymbol{U}_{i,j}(t)| > \frac{B_x}{\tilde{\lambda} + \frac{T_k\lambda}{2}}\left(B_x B_\epsilon\sqrt{2T_k\log(4T^3)} + \lambda B_\theta\right)\right) \leq \frac{1}{T^3}.$$

Furthermore, for some constant C such that $C\frac{\sqrt{\log T}}{\tilde{\lambda}\sqrt{T_k}} \geq \frac{B_x}{\tilde{\lambda}+\frac{T_k\lambda}{2}}\left(B_x B_\epsilon\sqrt{2T_k\log(4T^3)} + \lambda B_\theta\right)$, we have

$$\mathbb{P}\left(|\hat{\boldsymbol{U}}_{i,j}(t) - \boldsymbol{U}_{i,j}(t)| > C\frac{\sqrt{\log T}}{\tilde{\lambda}\sqrt{T_k}}\right) \leq \frac{1}{T^3}.$$

And hence,

$$\mathbb{P}\left(\exists k \in [T] : \mathcal{F}_k\right) \leq \frac{NK}{T}.$$

Proof. Let $X = \left(X_1^\top, \cdots, X_K^\top\right)^\top$ be the contexts for all K arms, let $\Sigma = \mathbb{E}\left[XX^\top \mid X \in \mathcal{D}_x\right]$, $\Sigma_i = \mathbb{E}\left[X_i X_i^\top \mid X_i \in \mathcal{D}_{x_i}\right]$. Under Assumption 4, $\lambda_{\min}(\Sigma) \geq \tilde{\lambda}$. By Cauchy's eigenvalue interlacing theorem (Lemma J.4), we have that $\lambda_{\min}(\Sigma_i) \geq \lambda_{\min}(\Sigma) \geq \tilde{\lambda} > 0$. Therefore, under Assumption 4, it suffices for any player p_i to sample a fixed arm to learn the preference parameter θ_i. Without loss of generality, we assume that the algorithm match p_i to a_i during the exploration phase.

In the k-th batch, let $X_i^{(k)} \in \mathbb{R}^{T_k \times d}$ be the contexts of arm a_i during the exploration phase, without ambiguity, we use X_i to replace $X_i^{(k)}$. For player p_i, in the exploitation phase, we have

$$\begin{aligned}
\|\hat{\theta}_i^{(k)} - \theta_i\|_2 &= \|\left(X_i^\top X_i + \lambda \boldsymbol{I}_d\right)^{-1} X_i^\top \left(X_i\theta_i + \epsilon\right) - \theta_i\|_2 \\
&= \|\left(X_i^\top X_i + \lambda \boldsymbol{I}_d\right)^{-1}\left(X_i^\top X_i + \lambda \boldsymbol{I}_d - \lambda \boldsymbol{I}_d\right)\theta_i + \left(X_i^\top X_i + \lambda \boldsymbol{I}_d\right)^{-1} X_i^\top \epsilon - \theta\|_2 \\
&= \|\theta_i - \lambda\left(X_i^\top X_i + \lambda \boldsymbol{I}_d\right)^{-1}\theta_i + \left(X_i^\top X_i + \lambda \boldsymbol{I}_d\right)^{-1} X_i^\top \epsilon - \theta\|_2 \\
&= \|\left(X_i^\top X_i + \lambda \boldsymbol{I}_d\right)^{-1}\left(X_i^\top \epsilon - \lambda\theta_i\right)\|_2 \\
&\leq \frac{1}{\lambda_{\min}\left(X_i^\top X_i + \lambda \boldsymbol{I}_d\right)}\|X_i^\top \epsilon - \lambda\theta_i\|_2 \\
&\leq \frac{1}{\lambda_{\min}\left(X_i^\top X_i + \lambda \boldsymbol{I}_d\right)}\left(\|X_i^\top \epsilon\| + \lambda B_\theta\right).
\end{aligned}$$

By matrix Bernstein inequality (Lemma J.5), we have that with probability at least $1 - o(1/T^3)$, $\lambda_{\min}\left(X_i^\top X_i\right) > \frac{T_k}{2}\lambda_{\min}(\Sigma_i) \geq \frac{T_k\tilde{\lambda}}{2}$, and hence with probability at least $1 - o(1/T^3)$, we have

$$\frac{1}{\lambda_{\min}\left(X_i^\top X_i + \lambda \boldsymbol{I}_d\right)} \leq \frac{1}{\lambda + \frac{T_k\tilde{\lambda}}{2}}. \tag{36}$$

On the other hand, $X_i^\top \epsilon = \sum_{t=1}^{T_k} \epsilon_t X_{i,t}$ is the sum of independent random vectors. By Hoeffding's inequality, for any $s > 0$

$$\mathbb{P}\left(\|\sum_{t=1}^{T_k} \epsilon_t X_{i,t}\|_2 \geq s \right) \leq 2\exp\left(-\frac{s^2}{2B_\epsilon^2 \sum_{t=1}^{T_k} \|X_{i,t}\|_2^2} \right),$$

where $\sum_{t=1}^{T_k} \|X_{i,t}\|_2^2 \leq T_k B_x^2$. Let the RHS of the above inequality be $\frac{1}{2T^3}$, then we have

$$\mathbb{P}\left(\|X_i^\top \epsilon\| \geq B_x B_\epsilon \sqrt{2T_k \log(4T^3)} \right) \leq \frac{1}{2T^3}. \tag{37}$$

Combining Eq.(36) and Eq.(37) with union bound, we have that

$$\mathbb{P}\left(\|\hat{\theta}_i^{(k)} - \theta_i\|_2 > \frac{1}{\lambda + \frac{T_k \tilde{\lambda}}{2}} \left(B_x B_\epsilon \sqrt{2T_k \log(4T^3)} + \lambda B_\theta \right) \right) \leq \frac{1}{T^3}. \tag{38}$$

Finally, as $|\hat{\boldsymbol{U}}_{i,j}(t) - \boldsymbol{U}_{i,j}(t)| = |\left(\hat{\theta}_i - \theta_i\right)^\top \mathrm{x}_j(t)| \leq \|\hat{\theta}_i - \theta_i\|_2 \|\mathrm{x}_j(t)\|_2 \leq B_x \|\hat{\theta}_i - \theta_i\|_2$, we reach the first part of the conclusion. While for the last part, we have

$$\begin{aligned}
\mathbb{P}\left(\exists k \in [T] : \mathcal{F}_k\right) &= \mathbb{P}\left(\exists k \in [T], \exists t_k \in [T_k + 1, T_k'], i \in [N], j \in [K] : |\hat{\boldsymbol{U}}_{i,j}(t_k) - \boldsymbol{U}_{i,j}(t_k)| > \Delta_k \right) \\
&\leq \sum_{k=1}^{T} \sum_{t_k = T_k + 1}^{T_k'} \sum_{i=1}^{N} \sum_{j=1}^{K} \mathbb{P}\left(|\hat{\boldsymbol{U}}_{i,j}(t_k) - \boldsymbol{U}_{i,j}(t_k)| > \Delta_k \right) \\
&\leq T^2 N K \frac{1}{T^3} = \frac{NK}{T}.
\end{aligned}$$

□

D PROOF OF THEOREM 3.2

Proof. When the context distribution satisfies Assumption 5, there exists some threshold Δ_0 and some constant c such that the CDF of $\delta_{\min}$: $F(\Delta)$ satisfies $F(\Delta) \leq c\Delta$ for any $\Delta \leq \Delta_0$. Therefore,

$$\mathbb{P}\left(\delta_{\min} \geq \Delta\right) = 1 - F(\Delta) \geq 1 - c\Delta, \forall \Delta \leq \Delta_0$$

For the constant c and the threshold Δ_0 given above, there exists a $T_0 \geq e$ such that $1 - c\Delta_0 = 1 - \frac{\log T_0}{T_0 \Delta_0^2}$, and hence for $T \geq T_0$ sufficiently large, we have

$$1 - c\Delta_0 \leq 1 - \frac{\log T}{T\Delta_0^2}.$$

When $T \geq T_0$ is fixed, the LHS of the above inequality is a decreasing function of Δ_0 while the RHS is an increasing function of Δ_0. Furthermore, when $\Delta_0 \to 0$, $1 - c\Delta_0 \to 1$ while $1 - \frac{\log T}{T\Delta_0^2} \to -\infty$. Therefore, there exists a $\bar{\Delta} \leq \Delta_0$ such that

$$1 - F\left(\bar{\Delta}\right) \geq 1 - c\bar{\Delta} = 1 - \frac{\log T}{T\bar{\Delta}^2}. \tag{39}$$

Solve Eq.(39), we have $\bar{\Delta} = \left(\frac{\log T}{cT}\right)^{1/3} = \tilde{\Theta}\left(T^{-1/3}\right)$. And

$$\mathbb{P}\left(\delta_{\min} \geq \Delta\right) = 1 - F(\Delta) \geq 1 - \frac{\log T}{T\Delta^2}, \quad \forall \Delta \leq \bar{\Delta}.$$

From the minimum preference gap definition (Definition 2), we know that $\Delta_{\min}$ is the maximum Δ such that $\mathbb{P}\left(\delta_{\min} \geq \Delta\right) \geq 1 - \frac{\log T}{T\Delta^2}$. Therefore,

$$\Delta_{\min} \geq \bar{\Delta} \Longrightarrow \Delta_{\min} = \tilde{\Omega}\left(T^{-1/3}\right).$$

Finally, as the regret upper bound for Algorithm 1 is $\mathcal{O}\left(\frac{\log^2 T}{\Delta_{\min}^2}\right)$ as shown in Theorem 3.1, we have the conclusion that

$$Reg_i(T) = \tilde{\mathcal{O}}\left(T^{2/3}\right), \quad \forall i \in [N] \quad \text{as} \quad T \to \infty.$$

□

E PROOF OF THEOREM 3.3

Proof. We prove Theorem 3.3 by contradiction. Fix a policy π, assume that this policy achieves $o\left(T^{2/3}\right)$ regret for each player in the market.

Let $\mathcal{N} = \{p_1, p_2, p_3\}$, $\mathcal{K} = \{a_1, a_2, a_3\}$, and $p_1 \succ p_2 \succ p_3$ for all arms. Throughout the proof, we assume that given the contexts and the preference parameters, all observations are Gaussian of unit variance, that is, when matching p_i to a_j at round t, we observe $y_i(t) \sim \mathcal{N}\left(\theta_i^\top x_j(t), 1\right)$, where θ_i is the preference parameter for player p_i and $x_j(t)$ is the observed context for arm a_j at time t. Let $\tau = T^{-1/3}, \varphi = \frac{1}{2}, \psi = \frac{1}{8}$. Consider two instances $\boldsymbol{\nu}$ and $\boldsymbol{\nu'}$ with the following preference parameters and context distributions, where the distribution of x_2 depends on the distribution of x_1, and the two instances only differ in the first entry of θ_1:

$$\theta_1 = \begin{bmatrix}\beta\\1\\0\\0\end{bmatrix}, \theta_2 = \begin{bmatrix}0\\0\\1\\0\end{bmatrix}, \theta_3 = \begin{bmatrix}0\\0\\0\\1\end{bmatrix}$$

$$x_1 = \begin{bmatrix}U\\0\\1\\\psi\end{bmatrix}, x_2 = \begin{bmatrix}0\\1\\0\\f(U)\end{bmatrix}, x_3 = \begin{bmatrix}0\\0\\\psi\\0\end{bmatrix},$$

where U is a uniform distribution on $[0,1]$, $f(U) = 1 \cdot \mathbb{1}\left\{U > \frac{1}{1+\tau}\right\} + \varphi \cdot \mathbb{1}\left\{U \leq \frac{1}{1+\tau}\right\}$; $\beta = 1$ in instance $\boldsymbol{\nu}$ and $\beta = 1+\tau$ in instance $\boldsymbol{\nu'}$. Therefore, we could write the utility matrix $\boldsymbol{U}$ for $\boldsymbol{\nu}$ and $\boldsymbol{U}'$ for $\boldsymbol{\nu'}$ as follows:

$$\boldsymbol{U} = \begin{bmatrix}U & 1 & 0\\1 & 0 & \psi\\\psi & f(U) & 0\end{bmatrix}, \quad \boldsymbol{U}' = \begin{bmatrix}(1+\tau)\cdot U & 1 & 0\\1 & 0 & \psi\\\psi & f(U) & 0\end{bmatrix}.$$

In Lemma E.2, we show that the above two instances are valid for proving a matching regret lower bound corresponding to the regret upper bound in Theorem 3.2, in the sense that the two instances satisfy Assumption 5.

Furthermore, since there are 3 players and 3 arms in the market, and the payoffs are non-negative for every player at every time, we can, without loss of generality, assume that every player gets matched to one arm every time. Denote the instantaneous regret for player p_i as $\boldsymbol{U}_i^*(t) - \boldsymbol{U}_{p_i,\mu_t(p_i)}(t)$.

Lemma E.1 (Properties of Instances $\boldsymbol{\nu}$ and $\boldsymbol{\nu'}$)**.** *Based on the utility matrices $\boldsymbol{U}$ and $\boldsymbol{U}'$, we have the following properties of $\boldsymbol{\nu}$ and $\boldsymbol{\nu'}$:*

1. Benchmark Utilities

- *The benchmark utilities for the three players in instance $\boldsymbol{\nu}$ are $\boldsymbol{U}^* = (1,1,0)^\top$.*
- *The benchmark utilities for the three players in instance $\boldsymbol{\nu'}$ are $\boldsymbol{U}^* = (1,1,0)^\top$ when $U \leq \frac{1}{1+\tau}$, and $\boldsymbol{U}^* = ((1+\tau)U, \psi, 1)^\top$ when $U > \frac{1}{1+\tau}$.*

2. Regret under Consideration

- *In instance $\boldsymbol{\nu}$, the regret for p_2, i.e., $Reg^{p_2}_{\boldsymbol{\nu},\pi}(T)$, under any context realizations would be non-negative.*
- *In instance $\boldsymbol{\nu'}$, the social regret for $p_1 + p_2 + p_3$, i.e., $Reg^{p_1+p_2+p_3}_{\boldsymbol{\nu'},\pi}(T)$, under any context realizations would be non-negative.*

The proof of Lemma E.1 is provided at the end of this section.

Let $\mathbb{P}_{\boldsymbol{\nu},\pi}$ be the joint probability measure over the history, and $\mathbb{E}_{\boldsymbol{\nu},\pi}$ be the expectation induced by instance $\boldsymbol{\nu}$ and policy π, and $\mathbb{P}_{\boldsymbol{\nu'},\pi}$, $\mathbb{E}_{\boldsymbol{\nu'},\pi}$ be defined similarly. The two instances $\boldsymbol{\nu}$ and $\boldsymbol{\nu'}$ only differ in one entry

of the utility matrix, i.e., the entry corresponds to the pair (p_1, a_1). From the divergence decomposition theorem (Lemma J.6), we know that

$$\begin{aligned} D\left(\mathbb{P}_{\boldsymbol{\nu},\pi}, \mathbb{P}_{\boldsymbol{\nu}',\pi}\right) &= \sum_{i=1}^{N}\sum_{j=1}^{K} \mathbb{E}_{\boldsymbol{\nu},\pi} N_{ij}(T) \cdot D(\boldsymbol{\nu}_{ij}, \boldsymbol{\nu}'_{ij}) \\ &= \mathbb{E}_{\boldsymbol{\nu},\pi}\left[N_{11}(T)\right] \cdot D(\boldsymbol{\nu}_{11}, \boldsymbol{\nu}'_{11}), \end{aligned}$$

where $\boldsymbol{\nu}_{ij}$ is the distribution of utilities obtained when player p_i is matched to arm a_j in the environment $\boldsymbol{\nu}$, $N_{ij}(T) \in \mathbb{N} \cup 0$ is the number of times player p_i is matched to arm a_j, up to and including time T. By the chain rule of KL divergence (Lemma J.7), we have

$$\begin{aligned} D(\boldsymbol{\nu}_{11}, \boldsymbol{\nu}'_{11}) &= D(\mathbb{P}_{\boldsymbol{\nu},\pi}(U), \mathbb{P}_{\boldsymbol{\nu}',\pi}(U)) + D\left(\mathbb{P}_{\boldsymbol{\nu},\pi}(\boldsymbol{U}_{11}|U), \mathbb{P}_{\boldsymbol{\nu},\pi}(\boldsymbol{U}'_{11}|U)\right) \\ &= 0 + \mathbb{E}_{u\sim U[0,1]} \frac{(\tau u)^2}{2} \\ &\leq \frac{\tau^2}{2}. \end{aligned}$$

Therefore, $D\left(\mathbb{P}_{\boldsymbol{\nu},\pi}, \mathbb{P}_{\boldsymbol{\nu}',\pi}\right) \leq \frac{\tau^2}{2} \cdot \mathbb{E}_{\boldsymbol{\nu},\pi}[N_{11}(T)]$. Now consider the event $A = \left\{N_{11}(T) \geq \frac{\tau T}{4}\right\}$ and its complement $A^c = \left\{N_{11}(T) < \frac{\tau T}{4}\right\}$, by Bretagnolle-Huber inequality (Lemma J.8), we have

$$\mathbb{P}_{\boldsymbol{\nu},\pi}(A) + \mathbb{P}_{\boldsymbol{\nu}',\pi}(A^c) \geq \frac{1}{2}\exp\left(-D\left(\mathbb{P}_{\boldsymbol{\nu},\pi}, \mathbb{P}_{\boldsymbol{\nu}',\pi}\right)\right) \geq \frac{1}{2}\exp\left(-\frac{\tau^2}{2} \cdot \mathbb{E}_{\boldsymbol{\nu},\pi}[N_{11}(T)]\right).$$

In instance $\boldsymbol{\nu}$, the stable matching benchmark for player p_2 is arm a_1 and it is the arm providing maximum utility for p_2, therefore, every time when we match (p_1, a_1), p_2 could only choose between a_2 and a_3, which would incur positive regret and the regret could not be compensate by the rounds that p_2 is matched to a_1, this leads to $(1-\psi) \cdot \mathbb{E}_{\boldsymbol{\nu},\pi}[N_{11}(T)] \leq Reg^{p_2}_{\boldsymbol{\nu},\pi}(T)$. From the assumption, we know that $\tau = T^{-1/3}$ and $Reg^{p_2}_{\boldsymbol{\nu},\pi}(T) = o(T^{2/3})$, i.e., $\limsup_{T\to\infty} \frac{Reg^{p_2}_{\boldsymbol{\nu},\pi}(T)}{T^{2/3}} = 0$ and hence there exist a constant c_1 such that $\mathbb{P}_{\boldsymbol{\nu},\pi}(A) + \mathbb{P}_{\boldsymbol{\nu}',\pi}(A^c) \geq c_1$, which in turns implies either $\mathbb{P}_{\boldsymbol{\nu},\pi}(A) \geq \frac{c_1}{2} := c_2$ or $\mathbb{P}_{\boldsymbol{\nu}',\pi}(A^c) \geq c_2$.

Case I: $\mathbb{P}_{\boldsymbol{\nu},\pi}(A) \geq c_2$.

In instance $\boldsymbol{\nu}$, when p_1 is matched to a_1, the best possible case for p_2 is that $p_2 - a_3$ such that p_2 collects ψ expected reward in the corresponding round, and hence the regret in this round is $1-\psi$. While a_1 is the arm with the maximum utility for p_2, this regret cannot be compensated in the rounds that p_2 is matched to a_1. Hence, we have

$$Reg^{p_2}_{\boldsymbol{\nu},\pi}(T) \geq (1-\psi) \cdot \frac{\tau T}{4} \cdot \mathbb{P}_{\boldsymbol{\nu},\pi}(A).$$

If $\mathbb{P}_{\boldsymbol{\nu},\pi}(A) \geq c_2$, we have $Reg^{p_2}_{\boldsymbol{\nu},\pi}(T) \geq c_3 \tau T = \Omega\left(T^{2/3}\right)$ for some constant c_3, which is a contradiction.

Case II: $\mathbb{P}_{\boldsymbol{\nu}',\pi}(A^c) \geq c_2$.

In instance $\boldsymbol{\nu}'$, let $N_u(T)$ be the number of times such that $U > \frac{1}{1+\tau}$, as U follows a uniform distribution on $[0,1]$, we have $N_u(T) \sim Binomial\left(T, \frac{\tau}{1+\tau}\right)$. Define the event $B = \left\{N_u(T) \geq \frac{\tau T}{2}\right\}$ and its complement $B^c = \left\{N_u(T) < \frac{\tau T}{2}\right\}$.

$$\begin{aligned} \mathbb{P}_{\boldsymbol{\nu},\pi}(A^c) &= \mathbb{P}_{\boldsymbol{\nu}',\pi}(A^c \cap B) + \mathbb{P}_{\boldsymbol{\nu}',\pi}(A^c \cap B^c) \\ &\leq \mathbb{P}_{\boldsymbol{\nu}',\pi}(A^c \cap B) + \mathbb{P}_{\boldsymbol{\nu}',\pi}\left(N_u < \frac{\tau T}{2}\right) \\ &\leq \mathbb{P}_{\boldsymbol{\nu}',\pi}(A^c \cap B) + \exp\left(-\frac{T\tau(1-\tau)^2}{8(1+\tau)}\right), \end{aligned}$$

where the last inequality follows from Lemma E.3. When $\tau = T^{-1/3}$, we have that the second term of the last inequality being $o\left(\exp\left(-T^{1/3}\right)\right) \leq c_4$ for some constant $c_4 < c_2$ when T is sufficiently large, leading to the requirement that $\mathbb{P}_{\boldsymbol{\nu}',\pi}\left(A^c \cap B\right) \geq c_2 - c_4 := c_5 > 0$.

Define $N^+_{ij}(T)$ and $N^-_{ij}(T)$ as the number of times such that player p_i is matched to arm a_j in the rounds that $U > \frac{1}{1+\tau}$ and $U \leq \frac{1}{1+\tau}$, respectively. Therefore, we have $N_{ij}(T) = N^+_{ij}(T) + N^-_{ij}(T)$, and

$\left\{N_{11}(T) < \frac{\tau T}{4}\right\} \subseteq \left\{N_{11}^{+}(T) < \frac{\tau T}{4}\right\}$, which further implies

$$\mathbb{P}_{\boldsymbol{\nu}',\pi}\left(N_{11}^{+}(T) < \frac{\tau T}{4}, N_u(T) \geq \frac{\tau T}{2}\right) \geq \mathbb{P}_{\boldsymbol{\nu}',\pi}(A^c \cap B) \geq c_5.$$

We then consider the event $\left\{N_{11}^{+}(T) < \frac{\tau T}{4}, N_u(T) \geq \frac{\tau T}{2}\right\}$. We further decompose the event as

$$\begin{aligned}
&\left\{N_{11}^{+}(T) < \frac{\tau T}{4}, N_u(T) \geq \frac{\tau T}{2}\right\} \\
=&\left\{N_{11}^{+}(T) < \frac{\tau T}{4}, N_u(T) \geq \frac{\tau T}{2}, N_{13}^{+}(T) > \frac{\tau T}{8}\right\} \cup \left\{N_{11}^{+}(T) < \frac{\tau T}{4}, N_u(T) \geq \frac{\tau T}{2}, N_{13}^{+}(T) \leq \frac{\tau T}{8}\right\},
\end{aligned}$$

then we have

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{\nu}',\pi}\left(N_{11}^{+}(T) < \frac{\tau T}{4}, N_u(T) \geq \frac{\tau T}{2}, N_{13}^{+}(T) > \frac{\tau T}{8}\right) + \mathbb{P}_{\boldsymbol{\nu}',\pi}\left(N_{11}^{+}(T) < \frac{\tau T}{4}, N_u(T) \geq \frac{\tau T}{2}, N_{13}^{+}(T) \leq \frac{\tau T}{8}\right) \\
\geq&\mathbb{P}_{\boldsymbol{\nu}',\pi}\left(N_{11}^{+}(T) < \frac{\tau T}{4}, N_u(T) \geq \frac{\tau T}{2}\right) \geq \mathbb{P}_{\boldsymbol{\nu}',\pi}(A^c \cap B) \\
\geq&c_5,
\end{aligned}$$

which implies either $\mathbb{P}_{\boldsymbol{\nu}',\pi}\left(N_{11}^{+}(T) < \frac{\tau T}{4}, N_u(T) \geq \frac{\tau T}{2}, N_{13}^{+}(T) > \frac{\tau T}{8}\right) \geq \frac{c_5}{2} := c_6$ or $\mathbb{P}_{\boldsymbol{\nu}',\pi}\left(N_{11}^{+}(T) < \frac{\tau T}{4}, N_u(T) \geq \frac{\tau T}{2}, N_{13}^{+}(T) \leq \frac{\tau T}{8}\right) \geq c_6$.

Case II.a: $\mathbb{P}_{\boldsymbol{\nu}',\pi}\left(N_{11}^{+}(T) < \frac{\tau T}{4}, N_u(T) \geq \frac{\tau T}{2}, N_{13}^{+}(T) > \frac{\tau T}{8}\right) \geq c_6$**.**

As the benchmark for player p_1 is always the one with the largest expected utility, the instantaneous regret for p_1 is always non-negative. Furthermore, every time when we match p_1 to a_3, the instantaneous regret for p_1 is no smaller than 1. Combining the above two points, we have

$$\begin{aligned}
Reg_{\boldsymbol{\nu}',\pi}^{p_1} &\geq \frac{\tau T}{8}\mathbb{P}\left(N_{13}^{+}(T) > \frac{\tau T}{8}\right) \\
&\geq \frac{\tau T}{8}\mathbb{P}_{\boldsymbol{\nu}',\pi}\left(N_{11}^{+}(T) < \frac{\tau T}{4}, N_u(T) \geq \frac{\tau T}{2}, N_{13}^{+}(T) > \frac{\tau T}{8}\right) \\
&\geq \frac{\tau T}{8} \cdot c_6 = \Omega\left(T^{2/3}\right),
\end{aligned}$$

since $\tau = T^{-1/3}$, which is a contradiction.

Case II.b: $\mathbb{P}_{\boldsymbol{\nu}',\pi}\left(N_{11}^{+}(T) < \frac{\tau T}{4}, N_u(T) \geq \frac{\tau T}{2}, N_{13}^{+}(T) \leq \frac{\tau T}{8}\right) \geq c_6$**.**

If $N_{11}^{+}(T) < \frac{\tau T}{4}$ and $N_{13}^{+}(T) \leq \frac{\tau T}{8}$, we have $N_{32}^{+}(T) \leq N_{11}^{+}(T) + N_{13}^{+}(T) \leq \frac{3\tau T}{8}$. This is because, without loss of generality, we assume that every player gets matched to one arm every time, and so when arm a_2 is occupied by player p_3, player p_1 is matched to either a_1 or a_3. Therefore,

$$\mathbb{P}_{\boldsymbol{\nu}',\pi}\left(N_{32}^{+}(T) \leq \frac{3\tau T}{8}, N_u(T) \geq \frac{\tau T}{2}\right) \geq c_6.$$

When $U > \frac{1}{1+\tau}$, the benchmark utility for p_3 is 1, which is the largest utility provided by arm a_2 and the hence instantaneous regret for p_3 is always non-negative during these rounds. We have

$$\begin{aligned}
Reg_{\boldsymbol{\nu}',\pi}^{p_3}(T) &= \mathbb{E}\left[\sum_{t=1}^{T}\left[\boldsymbol{U}_3^*(t) - \boldsymbol{U}_{p_3,\mu_t(p_3)}(t)\right]\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\left(\boldsymbol{U}_3^*(t) - \boldsymbol{U}_{p_3,\mu_t(p_3)}(t)\right)\mathbb{1}\left\{U_t > \frac{1}{1+\tau}\right\}\right] + \mathbb{E}\left[\sum_{t=1}^{T}\left(\boldsymbol{U}_3^*(t) - \boldsymbol{U}_{p_3,\mu_t(p_3)}(t)\right)\mathbb{1}\left\{U_t \leq \frac{1}{1+\tau}\right\}\right],
\end{aligned}$$

where

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\left(\boldsymbol{U}_3^*(t)-\boldsymbol{U}_{p_3,\mu_t(p_3)}(t)\right)\mathbb{1}\left\{U_t>\frac{1}{1+\tau}\right\}\right]\\
\geq&\mathbb{E}\left[\sum_{t=1}^{T}\left(\boldsymbol{U}_3^*(t)-\boldsymbol{U}_{p_3,\mu_t(p_3)}(t)\right)\mathbb{1}\left\{\left\{U_t>\frac{1}{1+\tau}\right\}\cap\left\{N_{32}^{+}(T)\leq\frac{3\tau T}{8}\right\}\cap\left\{N_u(T)\geq\frac{\tau T}{2}\right\}\right\}\right],\\
\geq&(1-\psi)\frac{\tau T}{8}\cdot c_6\\
=&\Omega\left(T^{\frac{2}{3}}\right),
\end{aligned}
\tag{40}
$$

where the last inequality comes from the fact that every time when $U_t > \frac{1}{1+\tau}$ while p_3 is not matched to a_2, the instantaneous regret for p_3 is at least $1-\psi$. And under the event $\left\{N_{32}^{+}(T)\leq\frac{3\tau T}{8}\right\}\cap\left\{N_u(T)\geq\frac{\tau T}{2}\right\}$, the number of times that p_3 is not matched to a_2 is at least $\frac{\tau T}{2}-\frac{3\tau T}{8}=\frac{\tau T}{8}$.

To ensure that $Reg^{p_3}_{\boldsymbol{\nu'},\pi}(T)=o\left(T^{2/3}\right)$, we need to compensate p_3 during the rounds when $U\leq\frac{1}{1+\tau}$ so that the regret incurred from Eq.(40) does not cause a contradiction to the assumption. While the benchmark utility for p_3 is 0, provided by arm a_3 in these rounds, the instantaneous regret for p_3 during these rounds is always non-positive, and hence there exists some constant c_7, say $c_7 := c_6/2$, such that

$$
\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}\left[\left(\boldsymbol{U}_3^*(t)-\boldsymbol{U}_{p_3,\mu_t(p_3)}(t)\right)\mathbb{1}\left\{U_t\leq\frac{1}{1+\tau}\right\}\right]&=\mathbb{E}_{\boldsymbol{\nu'},\pi}\left[-\psi\cdot N_{31}^{-}(T)-\varphi\cdot N_{32}^{-}(T)\right]\\
&\leq -c_7\frac{(1-\psi)\tau T}{8}.
\end{aligned}
$$

As $\psi<\varphi$ in our assumption, we have that under the policy π,

$$
\mathbb{E}_{\boldsymbol{\nu'},\pi}\left[\varphi(N_{31}^{-}(T)+N_{32}^{-}(T))\right]\geq c_7\frac{(1-\psi)\tau T}{8}\quad\Longrightarrow\quad\mathbb{E}_{\boldsymbol{\nu'},\pi}\left[N_{31}^{-}(T)+N_{32}^{-}(T)\right]\geq c_7\frac{(1-\psi)\tau T}{8\varphi}.
$$

When $U\leq\frac{1}{1+\tau}$, the benchmark stable matching is $\{p_1-a_2,p_2-a_1,p_3-a_3\}$ with the social benchmark utility of 2. While by matching p_3 to a_1 or a_2, the total collected rewards for the three players would be at most $\max\{(1+\tau)\cdot U+\psi+\varphi,1+\varphi,\psi,1+2\psi\}\leq 1+\psi+\varphi$. Therefore, when $U\leq\frac{1}{1+\tau}$, matching p_3 to a_1 or a_2 incurs the instantaneous social regret for the three players no less than $1-\varphi-\psi$.

Finally, from Lemma E.1, we know that the instananeous social regret for $p_1+p_2+p_3$ is always non-negative, regardless of the value of $\boldsymbol{U}$ or the chosen matching, we have

$$
\begin{aligned}
Reg^{p_1+p_2+p_3}_{\boldsymbol{\nu'},\pi}(T)=&\mathbb{E}\left[\sum_{t=1}^{T}\left[\sum_{i=1}^{3}\left(\boldsymbol{U}_i^*(t)-\boldsymbol{U}_{p_i,\mu_t(p_i)}(t)\right)\right]\right]\\
\geq&\mathbb{E}\left[\sum_{t=1}^{T}\left[\sum_{i=1}^{3}\left(\boldsymbol{U}_i^*(t)-\boldsymbol{U}_{p_i,\mu_t(p_i)}(t)\right)\mathbb{1}\left\{U_t\leq\frac{1}{1+\tau},p_3\text{ matches to }a_1\right\}\right]\right]\\
&+\mathbb{E}\left[\sum_{t=1}^{T}\left[\sum_{i=1}^{3}\left(\boldsymbol{U}_i^*(t)-\boldsymbol{U}_{p_i,\mu_t(p_i)}(t)\right)\mathbb{1}\left\{U_t\leq\frac{1}{1+\tau},p_3\text{ matches to }a_2\right\}\right]\right]\\
\geq&(1-\varphi-\psi)\mathbb{E}_{\boldsymbol{\nu'},\pi}\left[N_{31}^{-}(T)+N_{32}^{-}(T)\right]\\
\geq&(1-\varphi-\psi)\cdot c_7\frac{(1-\psi)\tau T}{8\varphi}\\
:=&c_8\tau T=\Omega\left(T^{\frac{2}{3}}\right)
\end{aligned}
$$

for some constant c_8, which is a contradiction.

Therefore, there does not exist a policy π such that the regret for every player is $o(T^{2/3})$, which confirms the theorem. □

Lemma E.2. *For instances $\boldsymbol{\nu}$ and $\boldsymbol{\nu'}$, there exists a constant $c=3$ such that $\mathbb{P}\left(\delta_{\min}\leq\Delta\right)=F(\Delta)\leq c\Delta$ for any $\Delta\leq\frac{1}{16}$.*

Proof. **In instance $\boldsymbol{\nu}$:** For the three players, we have

$$\delta_{\min}^{(p_1)} = \min\{1-U, U\}, \quad \delta_{\min}^{(p_2)} = \frac{1}{2}, \quad \delta_{\min}^{(p_3)} = \frac{1}{8}.$$

Therefore, when $\Delta \leq \frac{1}{16}$, we have

$$\mathbb{P}(\delta_{\min} \leq \Delta) \leq \mathbb{P}(U \geq 1-\Delta) + \mathbb{P}(U \leq \Delta) = 2\Delta.$$

In instance $\boldsymbol{\nu}'$: For the three players, we have

$$\delta_{\min}^{(p_1)} = \min\{(1+\tau)U, |(1+\tau)U-1|\}, \quad \delta_{\min}^{(p_2)} = \frac{1}{2}, \quad \delta_{\min}^{(p_3)} = \frac{1}{8}.$$

Therefore, when $\Delta \leq \frac{1}{16}$, we have

$$\mathbb{P}(\delta_{\min} \leq \Delta) \leq \mathbb{P}((1+\tau)U \leq \Delta) + \mathbb{P}(1-\Delta \leq (1+\tau)U \leq 1+\Delta) \leq \frac{3}{1+\tau}\Delta.$$

Combining the two instances, we have that for any $\Delta \leq \frac{1}{16}$, we have

$$\mathbb{P}(\delta_{\min} \leq \Delta) \leq \max\left\{\frac{3}{1+\tau}, 2\right\}\Delta \leq 3\Delta.$$

□

Lemma E.3. *If $N_u \sim Binomial\left(T, \frac{\tau}{1+\tau}\right)$, we have*

$$\mathbb{P}\left(N_u < \frac{\tau T}{2}\right) \leq \exp\left(-\frac{T\tau(1-\tau)^2}{8(1+\tau)}\right).$$

Proof. N_u is the sum of independent Bernoulli trials, and $\mathbb{E}[N_u] = \frac{\tau T}{1+\tau}$. Let ρ solving the equality $(1-\rho)\mathbb{E}[N_u] = \frac{\tau T}{2}$, we have $\rho = \frac{1-\tau}{2}$. From Lemma J.9, we know that

$$\begin{aligned}\mathbb{P}\left(N_u < \frac{\tau T}{2}\right) &= \mathbb{P}(N_u \leq (1-\rho)\mu) \\ &\leq \exp\left(-\frac{\mathbb{E}[N_u]\rho^2}{2}\right) \\ &= \exp\left(-\frac{T\tau(1-\tau)^2}{8(1+\tau)}\right).\end{aligned}$$

□

Proof of Lemma E.1. We prove the properties as follows.

1. Benchmark Utilities

- Under instance $\boldsymbol{\nu}$, the preference ranking of player p_1 over the arms is $a_2 \succ a_1 \succ a_3$ almost surely, while the preference ranking of player p_2 is $a_1 \succ a_3 \succ a_2$, and for p_3 it is $a_2 \succ a_1 \succ a_3$, and hence the stable matching benchmark is $\mu^* = \{p_1 - a_2, p_2 - a_1, p_3 - a_3\}$, which corresponds to $\boldsymbol{U}^* = (1,1,0)^\top$.
- Under instance $\boldsymbol{\nu}'$, when $U \leq \frac{1}{1+\tau}$, we have that $\boldsymbol{U}'_{11} \leq \boldsymbol{U}'_{12}$ almost surely, so that the preference ranking of players over arms would be the same as that in instance $\boldsymbol{\nu}$ and $\boldsymbol{U}^* = (1,1,0)^\top$. When $U > \frac{1}{1+\tau}$, the preference ranking of player p_1 over the arms would be altered to $a_1 \succ a_2 \succ a_3$ and hence the stable matching benchmark shifts to $\mu^* = \{p_1 - a_1, p_2 - a_3, p_3 - a_2\}$, which corresponds to $\boldsymbol{U}^* = ((1+\tau)U, \psi, 1)^\top$.

2. Regret under Consideration

- In instance $\boldsymbol{\nu}$, for player p_2, the maximum utility is from the matching (p_2, a_1), which has a utility of 1. This pair is also the stable matching pair, and hence the regret for p_2 is always non-negative.

- In instance $\boldsymbol{\nu}'$, when $U \leq \frac{1}{1+\tau}$, the maximum weight matching is $\{(p_1 - a_2), (p_2 - a_1), (p_3 - a_3)\}$, providing a social welfare of 2; when $U > \frac{1}{1+\tau}$, the maximum weight matching is $\{(p_1 - a_1), (p_2 - a_3), (p_3 - a_2)\}$, providing a social welfare of $\frac{9}{8} + (1+\tau)U > 2$. Therefore, in either case, the stable matching is the same as the maximum weight matching, and hence the social regret for $p_1 + p_2 + p_3$ would be non-negative.

□

F ADAPTIVE EXPLORE-CHOOSE ORACLE ALGORITHM

The full version of Adative Explore-Choose Oracle (AdECO) algorithm is presented in Algorithm 5.

Algorithm 5 Adaptive Explore-Choose Oracle (AdECO)

Input: Time horizon T, preference profile $(P_a)_{a\in\mathcal{K}}$, context dimension d, parameter Δ, ε, and η, ridge penalty parameter λ.

$\xi \leftarrow \frac{\Delta-\varepsilon}{4\eta}$.
$\boldsymbol{V}_i^{(1)} \leftarrow \lambda \boldsymbol{I}_{d\times d}$, $\hat{\theta}_i^{(1)} \leftarrow \mathbf{0}_d$, $\forall i \in [N]$.
for $t = 1, 2, \cdots, T$ **do**
  Observe context $\mathrm{x}_j(t)$ for every arm a_j, $\forall j \in [K]$.
  if $\exists i \in [N], j \in [K]$, s.t. $\|\mathrm{x}_j(t)\|_{(\boldsymbol{V}_i^{(t)})^{-1}} > \xi$ **then** ▷ Exploration
    Collect all (i, j) such that $\|\mathrm{x}_j(t)\|_{(\boldsymbol{V}_i^{(t)})^{-1}} > \xi$, form a bipartite graph G_t.
    Find a **maximum cardinality matching** μ_c on G_t, let $M^c := \{i \in [N] : i \text{ is matched in } \mu^c\}$, match

$$\mu_t(i) \leftarrow \begin{cases} \mu^c(i) & \text{if } i \in M^c, \\ \emptyset & \text{otherwise.} \end{cases}$$

    Update $\boldsymbol{V}_i^{(t+1)}$ and $\hat{\theta}_i^{(t+1)}$, $\forall i \in M^c$ according to Eq.(6).
    $\boldsymbol{V}_i^{(t)} \leftarrow \boldsymbol{V}_i^{(t-1)}$ and $\hat{\theta}_i^{(t)} \leftarrow \hat{\theta}_i^{(t-1)}$, $\forall i \in [N] \setminus M^c$.
  else ▷ Exploitation
    $\hat{\theta}_i^{(t)} \leftarrow \hat{\theta}_i^{(t-1)}$, $\forall i \in [N]$.
    $\hat{\boldsymbol{U}}_{i,j}(t) = \left(\hat{\theta}_i^{(t)}\right)^\top \mathrm{x}_j(t)$, $F_i \leftarrow$ False.
    for $i = 1, 2, \cdots, N$ **do**
      $\hat{\boldsymbol{U}}_{i,\cdot}^{\text{sort}}(t) \leftarrow \text{Sort}\left(\hat{\boldsymbol{U}}_{i,\cdot}(t), \text{decreasing}\right)$
      $\Delta_{i,\min} \leftarrow \min\left\{\hat{\boldsymbol{U}}_{i,j}^{\text{sort}}(t) - \hat{\boldsymbol{U}}_{i,j+1}^{\text{sort}}(t)\right\}, j \in [K-1]$
      if $\Delta_{i,\min} > \frac{\Delta+\varepsilon}{2}$ **then**
        $F_i \leftarrow$ True
      end if
    end for
    if $\forall i \in [N]$, $F_i ==$ True **then** ▷ CIs are well-separated
      $\mu_t(i) \leftarrow \mu^{GS}(i)$, $\forall i \in [N]$. (Input $\hat{\boldsymbol{U}}(t)$ and $(P_a)_{a\in\mathcal{K}}$ for **Gale-Shapley**)
    else
      $\mu_t(i) \leftarrow \mu^{approx}(i)$, $\forall i \in [N]$. (Input $\hat{\boldsymbol{U}}(t)$ and $(P_a)_{a\in\mathcal{K}}$ for α-**approximation oracle** with instability tolerance $\frac{\Delta+\varepsilon}{2}$)
    end if
  end if
end for

When the confidence intervals for the utilities overlap, we cannot distinguish the strict preference ranking for players over arms. Moreover, when there are ties in the preference profile, the optimal stable share of each player might be achieved from different stable matchings. In this scenario, outputting a deterministic matching with the estimated utility matrix may benefit some players while incurring large regret for others. Therefore, we should consider distributions over matchings and approximation. Here, we assume access to an offline oracle that, given a utility matrix $\boldsymbol{U}$ and the uncertainty set with confidence radius γ, outputs a randomized matching guaranteeing each player p_i an α of $\boldsymbol{U}_i^\varepsilon(p_i)$ in expectation, with additional error $2\gamma + \varepsilon$.

Before we formally present the definition of the approximation oracle, we first define the ε-optimal stable share within an uncertainty set. Given a utility matrix $\boldsymbol{U}$ and the arms' preference profile $(P_a)_{a\in\mathcal{K}}$, define the set of ε-stable matchings as

$$\mathcal{S}_{\boldsymbol{U}}^\varepsilon := \{\mu : \mu \text{ is an } \varepsilon - \text{stable matching w.r.t. } \boldsymbol{U} \text{ and } (P_a)_{a\in\mathcal{K}}\},$$

and the ε-optimal stable share with respect to $\boldsymbol{U}$ is

$$\boldsymbol{U}_\varepsilon^*(p_i) := \max_{\mu \in \mathcal{S}_{\boldsymbol{U}}^\varepsilon} \boldsymbol{U}(p_i, \mu(p_i)), \quad \forall i \in [N].$$

Given an uncertainty set $\mathcal{U}$ of utility matrices, we define the ε-optimal stable share within $\mathcal{U}$ as

$$\mathcal{U}_\varepsilon^*(p_i) := \sup_{\boldsymbol{U} \in \mathcal{U}} \max_{\mu \in \mathcal{S}_{\boldsymbol{U}}^\varepsilon} \boldsymbol{U}(p_i, \mu(p_i)), \quad \forall i \in [N]. \tag{41}$$

The $(\boldsymbol{\alpha}, \gamma, \varepsilon)$-approximation oracle is defined as follows.

Definition 4 ($(\boldsymbol{\alpha}, \gamma, \varepsilon)$-Approximation Oracle)**.** An $(\boldsymbol{\alpha}, \gamma, \varepsilon)$-approximation oracle takes a rectangular uncertainty set $\mathcal{U}$ with center $\boldsymbol{U}$, confidence radius γ as input and returns a (randomized) matching $\tilde{\mu}$ satisfying: $\mathbb{E}\left[\boldsymbol{U}_{\tilde{\mu}}(p_i)\right] \geq \boldsymbol{\alpha}^{\mathcal{U}}(p_i) \cdot \mathcal{U}_\varepsilon^*(p_i) - (2\gamma + \varepsilon)$ for every player p_i, where $\boldsymbol{\alpha}^{\mathcal{U}} \in (0,1]^N$ is a player-specific approximation ratio vector (often simplified to $\boldsymbol{\alpha}$). If $\boldsymbol{\alpha}^{\mathcal{U}}(p_i) = \alpha$ is uniform across players and independent of $\mathcal{U}$, we call it an $(\alpha, \gamma, \varepsilon)$-approximation oracle, and without ambiguity, we simply denote it as α-approximation oracle.

For example, Algorithm 6 (Lin et al., 2025) guarantees that for any input utility matrix $\boldsymbol{U}$ with instability tolerance $2\gamma + \varepsilon$, let $\mathcal{U}$ be the rectangular uncertainty set with center $\boldsymbol{U}$ and confidence radius γ, we have $\boldsymbol{\alpha}^{\boldsymbol{U}}(p_i) \geq 1/\lfloor \log_2 N + 2 \rfloor$. We restate the algorithm as follows, while omitted the detailed analysis.

Algorithm 6 ε-Oracle for Approximated Player Optimal Stable Matching

Input: N players, K arms, Utility matrix $\boldsymbol{U}$ that encodes the preference of players over arms, strict preference profile P_a of arms, an integer $m \geq 1$, and the instability tolerance $\varepsilon \geq 0$.

1: For each arm $a \in \mathcal{A}$, duplicate it m times and denote the i-th copy as $a^{(i)}$.
2: Each replica $a^{(i)}$ shares the same preference P_a as the original arm a.
3: For every player p and job $a^{(i)}$, define the utility

$$\boldsymbol{U}(p, a^{(i)}) := \boldsymbol{U}(p, a) - (i-1)\varepsilon$$

and use it to generate the players' preference profile P_w (breaking ties in favor of lower indices).
4: Run Gale-Shapley algorithm on P_w and P_a to compute a player-optimal stable matching $\tilde{\mu}$.
5: For each $i \in [m]$, build a matching $\tilde{\mu}_i$, which matches each arm a with $\tilde{\mu}_i(a) := \tilde{\mu}(a^{(i)})$.

Output: The distribution D which selects each matching $\tilde{\mu}_i$ with probability $1/m$.

G PROOF FOR THEOREM 4.1

Proof. Given Δ and ε, we define $\gamma = \frac{\Delta-\varepsilon}{4}$, $\xi = \frac{\Delta-\varepsilon}{4\eta}$, for the exploration phase, denote the estimation indices for player p_i as $\mathcal{G}^{(i)} = \left\{t_1^{(i)}, \cdots, t_{|\mathcal{G}^{(i)}|}^{(i)}\right\}$. Notice that for $t \in \mathcal{G}^{(i)}$, we have $\|x_j(t)\|_{\left(\boldsymbol{V}_i^{(t)}\right)^{-1}} > \xi$ for $j = \mu_t(i)$, and we only use the contexts and matching rewards from $\mathcal{G}^{(i)}$ to update the Gram matrix $\boldsymbol{V}_i$ and the estimation of the preference parameter $\hat{\theta}_i$. Take $\mathcal{G}$ as the set of rounds in which the market explores, we have $\mathcal{G} = \cup_{i\in[N]}\mathcal{G}^{(i)}$. Define $\mathcal{F} = \left\{\exists t \notin \mathcal{G}, i \in [N], j \in [K] : |\hat{\boldsymbol{U}}_{i,j}(t) - \boldsymbol{U}_{i,j}(t)| > \gamma\right\}$ as the bad event that some preference of players over arms is not estimated well during the exploitation phase. Using γ as the width for the confidence interval constructed for the estimated utility, let $\mathcal{F}_d^{(t)} := \left\{\hat{\boldsymbol{U}}_{i,j}^{\text{sort}}(t) - \hat{\boldsymbol{U}}_{i,j+1}^{\text{sort}}(t) > 2\gamma + \varepsilon = \frac{\Delta+\varepsilon}{2}, \forall j \in [K-1], \forall i \in [N]\right\}$ as the event that all arms have well-separated confidence intervals for any player p_i at time t in the sense that the distance between any two confidence intervals for the arms is at least ε. Define the instantaneous regret for a player p_i at time t as $(\boldsymbol{U}_i^*(t)\mathbb{1}\{\delta_{\min}(t) > \Delta\} + \alpha\boldsymbol{U}_i^{\varepsilon}(t)\mathbb{1}\{\delta_{\min}(t) \leq \Delta\}) - \boldsymbol{U}_{i,\mu_t(i)}(t)$, where $\varepsilon < \Delta$ and is typically chosen as $\Delta/2$.

The optimal stable regret for each player p_i by following AdECO (Algorithm 2) satisfies

$$
\begin{aligned}
Reg_i^{\alpha,\Delta}(T) =& \mathbb{E}\left[\sum_{t=1}^{T}\left[(\boldsymbol{U}_i^*(t)\mathbb{1}\{\delta_{\min}(t)>\Delta\}+\alpha\boldsymbol{U}_i^{\varepsilon}(t)\mathbb{1}\{\delta_{\min}(t)\le\Delta\})-\boldsymbol{U}_{i,\mu_t(i)}(t)\right]\right] && (42)\\
=& \mathbb{E}\left[\sum_{t=1}^{T}\left[(\boldsymbol{U}_i^*(t)\mathbb{1}\{\delta_{\min}(t)>\Delta\}+\alpha\boldsymbol{U}_i^{\varepsilon}(t)\mathbb{1}\{\delta_{\min}(t)\le\Delta\})-\boldsymbol{U}_{i,\mu_t(i)}(t)\right]\mathbb{1}\{\neg\mathcal{F}\}\right]\\
&+\mathbb{E}\left[\sum_{t=1}^{T}\left[(\boldsymbol{U}_i^*(t)\mathbb{1}\{\delta_{\min}(t)>\Delta\}+\alpha\boldsymbol{U}_i^{\varepsilon}(t)\mathbb{1}\{\delta_{\min}(t)\le\Delta\})-\boldsymbol{U}_{i,\mu_t(i)}(t)\right]\mathbb{1}\{\mathcal{F}\}\right]\\
\le& \mathbb{E}\left[\sum_{t=1}^{T}\left[(\boldsymbol{U}_i^*(t)\mathbb{1}\{\delta_{\min}(t)>\Delta\}+\alpha\boldsymbol{U}_i^{\varepsilon}(t)\mathbb{1}\{\delta_{\min}(t)\le\Delta\})-\boldsymbol{U}_{i,\mu_t(i)}(t)\right]\mathbb{1}\{\neg\mathcal{F}\}\right]+B_yT\mathbb{P}(\mathcal{F}) && (43)\\
\le& \mathbb{E}\left[\sum_{t=1}^{T}\left[(\boldsymbol{U}_i^*(t)\mathbb{1}\{\delta_{\min}(t)>\Delta\}+\alpha\boldsymbol{U}_i^{\varepsilon}(t)\mathbb{1}\{\delta_{\min}(t)\le\Delta\})-\boldsymbol{U}_{i,\mu_t(i)}(t)\right]\mathbb{1}\{\neg\mathcal{F}\}\right]+B_yT\cdot\frac{NK}{T} && (44)\\
\le& \mathbb{E}\left[\sum_{t\in\mathcal{G}}\left[(\boldsymbol{U}_i^*(t)\mathbb{1}\{\delta_{\min}(t)>\Delta\}+\alpha\boldsymbol{U}_i^{\varepsilon}(t)\mathbb{1}\{\delta_{\min}(t)\le\Delta\})-\boldsymbol{U}_{i,\mu_t(i)}(t)\right]\mathbb{1}\{\neg\mathcal{F}\}\right]\\
&+\mathbb{E}\left[\sum_{t\notin\mathcal{G}}\left[(\boldsymbol{U}_i^*(t)\mathbb{1}\{\delta_{\min}(t)>\Delta\}+\alpha\boldsymbol{U}_i^{\varepsilon}(t)\mathbb{1}\{\delta_{\min}(t)\le\Delta\})-\boldsymbol{U}_{i,\mu_t(i)}(t)\right]\mathbb{1}\{\neg\mathcal{F}\}\right]+B_yNK\\
\le& \mathbb{E}\left[B_y|\mathcal{G}|\mathbb{1}\{\neg\mathcal{F}\}\right] && (45)\\
&+\mathbb{E}\left[\sum_{t\notin\mathcal{G}}\left[(\boldsymbol{U}_i^*(t)\mathbb{1}\{\delta_{\min}(t)>\Delta\}+\alpha\boldsymbol{U}_i^{\varepsilon}(t)\mathbb{1}\{\delta_{\min}(t)\le\Delta\})-\boldsymbol{U}_{i,\mu_t(i)}(t)\right]\mathbb{1}\left\{\neg\mathcal{F}\cap\mathcal{F}_d^{(t)}\right\}\right]\\
&+\mathbb{E}\left[\sum_{t\notin\mathcal{G}}\left[(\boldsymbol{U}_i^*(t)\mathbb{1}\{\delta_{\min}(t)>\Delta\}+\alpha\boldsymbol{U}_i^{\varepsilon}(t)\mathbb{1}\{\delta_{\min}(t)\le\Delta\})-\boldsymbol{U}_{i,\mu_t(i)}(t)\right]\mathbb{1}\left\{\neg\mathcal{F}\cap\neg\mathcal{F}_d^{(t)}\right\}\right]+B_yNK\\
\le& \mathbb{E}\left[B_y|\mathcal{G}|\mathbb{1}\{\neg\mathcal{F}\}\right] && (46)\\
&+\mathbb{E}\left[\sum_{t\notin\mathcal{G}}\left[(\boldsymbol{U}_i^*(t)\mathbb{1}\{\delta_{\min}(t)>\Delta\}+\alpha\boldsymbol{U}_i^{\varepsilon}(t)\mathbb{1}\{\delta_{\min}(t)\le\Delta\})-\boldsymbol{U}_{i,\mu_t(i)}(t)\right]\mathbb{1}\left\{\neg\mathcal{F}\cap\neg\mathcal{F}_d^{(t)}\right\}\right]+B_yNK\\
\le& \mathbb{E}\left[\left(B_y|G|+\frac{\Delta+\varepsilon}{2}\cdot T\right)\mathbb{1}\{\neg\mathcal{F}\}\right]+B_yNK && (47)\\
\le& B_y\frac{\eta^2Nd\log\left(\frac{T+d\lambda}{\lambda}\right)}{\gamma^2}+\frac{\Delta+\varepsilon}{2}\cdot T+B_yNK && (48)\\
=& \mathcal{O}\left(\frac{Nd\log^2T}{(\Delta-\varepsilon)^2}+\frac{\Delta+\varepsilon}{2}\cdot T\right).
\end{aligned}
$$

Eq.(42) according to the assumption that the noises are i.i.d. with mean 0. Eq.(43) and (45) comes from the fact that under Assumptions 1, 2, 3, both the OSS and the expected reward are in $[0, B_y]$, and $\alpha \in (0, 1]$, hence the instantaneous regret $(\boldsymbol{U}_i^*(t)\mathbb{1}\{\delta_{\min}(t) > \Delta\} + \alpha\boldsymbol{U}_i^{\varepsilon}(t)\mathbb{1}\{\delta_{\min}(t) \le \Delta\}) - \boldsymbol{U}_{i,\mu_t(i)}(t)$ is bounded by B_y. Eq.(44) holds based on Lemma G.1. Eq.(46) holds according to Lemma G.2. Eq.(47) follows from Lemma G.3. Eq.(48) holds based on Lemma B.4 by setting the parameters accordingly.

□

Lemma G.1. *Let* $\delta = T^{-1}, \eta = B_y\sqrt{d\log\left(\frac{1+TB_x^2/\lambda}{\delta}\right)} + \lambda^{1/2}B_\theta$, *we have* $\mathbb{P}(\mathcal{F}) \le \frac{NK}{T}$.

Proof.

$$\begin{aligned}
\mathbb{P}(\mathcal{F}) =& \mathbb{P}\left(\exists t \notin \mathcal{G}, i \in [N], j \in [K] : |\hat{\boldsymbol{U}}_{i,j}(t) - \boldsymbol{U}_{i,j}(t)| > \gamma\right) \\
\leq& \sum_{i=1}^{N}\sum_{j=1}^{K} \mathbb{P}\left(\exists t \notin \mathcal{G} : |\hat{\boldsymbol{U}}_{i,j}(t) - \boldsymbol{U}_{i,j}(t)| > \gamma\right) && (49)\\
\leq& \sum_{i=1}^{N}\sum_{j=1}^{K} \mathbb{P}\left(\exists t \notin \mathcal{G} : \|\hat{\theta}_i - \theta_i\|_{\boldsymbol{V}_i^{(t)}} \|\mathrm{x}_j(t)\|_{\left(\boldsymbol{V}_i^{(t)}\right)^{-1}} > \gamma\right) && (50)\\
\leq& \sum_{i=1}^{N}\sum_{j=1}^{K} \mathbb{P}\left(\exists t \notin \mathcal{G} : \|\hat{\theta}_i - \theta_i\|_{\boldsymbol{V}_i^{(t)}} > \eta\right) && (51)\\
\leq& NK\delta = \frac{NK}{T}. && (52)
\end{aligned}$$

Eq.(49) follows from the union bound. Eq.(50) holds according to Cauchy-Schwarz inequality. Eq.(51) comes from the fact that for any $t \notin \mathcal{G}$, we have $\|\mathrm{x}_j(t)\|_{\left(\boldsymbol{V}_i^{(t)}\right)^{-1}} \leq \xi = \frac{\gamma}{\eta}$ in Algorithm 2. Eq.(52) holds according to Lemma J.2. □

Lemma G.2. *At time t, under the event $\neg\mathcal{F} \cap \mathcal{F}_d^{(t)}$, the instantaneous regret of Algorithm 2 is non-positive, i.e.,*

$$\mathbb{E}\left[\left((\boldsymbol{U}_i^*(t)\mathbb{1}\{\delta_{\min}(t) > \Delta\} + \alpha\boldsymbol{U}_i^{\varepsilon}(t)\mathbb{1}\{\delta_{\min}(t) \leq \Delta\}) - \boldsymbol{U}_{i,\mu_t(i)}(t)\right)\mathbb{1}\left\{\neg\mathcal{F} \cap \mathcal{F}_d^{(t)}\right\}\right] \leq 0.$$

Proof. When $\delta_{\min}(t) > \Delta$, the Gale-Shapley subroutine called by Algorithm 2 outputs the unique player-optimal stable matching with respect to the ground-truth utility matrix $\boldsymbol{U}$, as the preference rankings of players over arms induced from $\hat{\boldsymbol{U}}$ would be the same as those from $\boldsymbol{U}$ under the event $\neg\mathcal{F} \cap \mathcal{F}_d^{(t)}$. Specifically, for any player p_i, any two arms a_j and $a_{j'}$ such that $\hat{\boldsymbol{U}}_{i,j}(t) > \hat{\boldsymbol{U}}_{i,j'}(t)$, under the event $\mathcal{F}_d^{(t)}$, we have $\hat{\boldsymbol{U}}_{i,j}(t) - \hat{\boldsymbol{U}}_{i,j'}(t) \geq 2\gamma + \varepsilon$, so that $\boldsymbol{U}_{i,j}(t) - \boldsymbol{U}_{i,j'}(t) \geq \hat{\boldsymbol{U}}_{i,j}(t) - \gamma - (\hat{\boldsymbol{U}}_{i,j'}(t) + \gamma) \geq \varepsilon$. Therefore, every player received the expected reward equal to their corresponding OSS in time t, leading to the instantaneous regret being 0.

When $\delta_{\min}(t) \leq \Delta$, from Lemma G.4, we know that if we call the Gale-Shapley oracle with the estimated utility matrix $\hat{\boldsymbol{U}}$ and the preference profile P_a, the resulting matching gives every player the expected reward equal to $\boldsymbol{U}_i^{\varepsilon}(t)$ with respect to the ground-truth utility matrix $\boldsymbol{U}$ and P_a. This is because under the event $\neg\mathcal{F} \cap \mathcal{F}_d^{(t)}$, the minimum difference between any two arms is no less than ε. And since $\alpha \in (0, 1]$, we have that $\boldsymbol{U}_i^*(t) = \boldsymbol{U}_i^{\varepsilon}(t) \geq \alpha\boldsymbol{U}_i^{\varepsilon}(t)$, and hence the instantaneous regret is non-positive. □

Lemma G.3. *At time t, under the event $\neg\mathcal{F}\cap\neg\mathcal{F}_d^{(t)}$, the instantaneous regret of Algorithm 2 is no larger than $\frac{\Delta+\varepsilon}{2}$, i.e.,*

$$\mathbb{E}\left[\left[(\boldsymbol{U}_i^*(t)\mathbb{1}\{\delta_{\min}(t) > \Delta\} + \alpha\boldsymbol{U}_i^{\varepsilon}(t)\mathbb{1}\{\delta_{\min}(t) \leq \Delta\}) - \boldsymbol{U}_{i,\mu_t(i)}(t)\right]\mathbb{1}\left\{\neg\mathcal{F}\cap\neg\mathcal{F}_d^{(t)}\right\}\right] \leq \frac{\Delta+\varepsilon}{2}.$$

Proof. At time t, under the event $\neg\mathcal{F}\cap\neg\mathcal{F}_d^{(t)}$, there exists a player p_i, two arms a_j and $a_{j'}$ such that $\hat{\boldsymbol{U}}_{i,j}(t) - \hat{\boldsymbol{U}}_{i,j'}(t) \leq 2\gamma + \varepsilon$ while $|\hat{\boldsymbol{U}}_{i,j}(t) - \boldsymbol{U}_{i,j}(t)| \leq \gamma$ and $|\hat{\boldsymbol{U}}_{i,j'}(t) - \boldsymbol{U}_{i,j'}(t)| \leq \gamma$. Therefore, $|\boldsymbol{U}_{i,j}(t) - \boldsymbol{U}_{i,j'}(t)| \leq 4\gamma + \varepsilon$. Since $\gamma = \frac{\Delta-\varepsilon}{4}$, we have $\delta_{\min}(t) \leq 4\gamma + \varepsilon = \Delta$, which implies that the benchmark utility would be $\alpha\boldsymbol{U}_i^{\varepsilon}(t)$ in this case.

Under the event $\neg\mathcal{F}$, the ground-truth utility matrix $\boldsymbol{U}$ lies in the uncertainty set $\mathcal{U}$ with center $\hat{\boldsymbol{U}}$ and radius γ, and hence $\boldsymbol{U}_i^{\varepsilon} \leq \mathcal{U}_{\varepsilon}^*(p_i)$ for any $i \in [N]$ according to the definition of $\mathcal{U}_{\varepsilon}^*(p_i)$ (Eq.(41)). From Definition 4, we know that when calling the $(\alpha, \gamma, \varepsilon)$-approximation oracle with input $\hat{\boldsymbol{U}}$ and confidence radius γ, the receiving expected utility satisfies

$$\mathbb{E}[\boldsymbol{U}_{i,\mu_t(i)}(t)] \geq (\alpha\mathcal{U}_{\varepsilon}^*(p_i))(t) - (2\gamma + \varepsilon) \geq \alpha\boldsymbol{U}_i^{\varepsilon}(t) - (2\gamma + \varepsilon)$$

Specifically, we prove the corresponding property for Algorithm 6 in Lemma G.5.

Therefore, when setting the approximation ratio $\alpha = 1/\lfloor \log_2 N + 2 \rfloor$, we have the instantaneous regret

$$\mathbb{E}\left[\left(\alpha \boldsymbol{U}_i^{\varepsilon}(t) - \boldsymbol{U}_{i,\mu_t(i)}(t)\right) \mathbb{1}\left\{\neg\mathcal{F} \cap \neg\mathcal{F}_d^{(t)}\right\}\right] \leq 2\gamma + \varepsilon = \frac{\Delta + \varepsilon}{2}.$$

□

Lemma G.4. *Given a utility matrix $\boldsymbol{U}$ and the preference profile P_a with N players and K arms, define the minimum difference between any two utilities for player p_i as*

$$\delta_{\min}^{(i)} = \min_{j,j' \in [K]} |\boldsymbol{U}_{i,j} - \boldsymbol{U}_{i,j'}|,$$

and $\delta_{\min} = \min_{i \in [N]} \delta_{\min}^{(i)}$. If $\delta_{\min} \geq \varepsilon$, the set of ε-stable matchings $\mathcal{S}^{\varepsilon}$ is the same as the set of stable matching $\mathcal{S}$ with respect to $\boldsymbol{U}$ and P_a, and hence for any player p_i, $\boldsymbol{U}_i^{\varepsilon} = \boldsymbol{U}_i^*$.

Proof. From the definition of the set of stable matchings $\mathcal{S}^{\varepsilon}$, we know that $\mathcal{S} \subseteq \mathcal{S}^{\varepsilon}$.If there is a matching $\mu \in \mathcal{S}^{\varepsilon}$ such that $\mu \notin \mathcal{S}$, there exists a blocking pair (p_i, a_j) such that $p_i \succ_{a_j} \mu(a_j)$ and $\boldsymbol{U}_{i,j} > \boldsymbol{U}_{i,\mu(p_i)}$. Since $\delta_{\min} \geq \varepsilon$, we also have $\boldsymbol{U}_{i,j} \geq \boldsymbol{U}_{i,\mu(p_i)} + \varepsilon$, which implies (p_i, a_j) is also an ε-blocking pair, violating the ε-stability of μ, which is a contradition. Therefore, $\mathcal{S} = \mathcal{S}^{\varepsilon}$ and hence $\boldsymbol{U}_i^{\varepsilon} = \boldsymbol{U}_i^*$. □

Lemma G.5. *Given an uncertainty set $\mathcal{U}$, define the center $\hat{\boldsymbol{U}}$ of the set $\mathcal{U}$, and the uncertainty parameter as*

$$\hat{\boldsymbol{U}}(p,a) = \frac{\inf_{\boldsymbol{U} \in \mathcal{U}} \boldsymbol{U}(p,a) + \sup_{\boldsymbol{U} \in \mathcal{U}} \boldsymbol{U}(p,a)}{2}, \quad \gamma = \frac{1}{2} \sup_{\boldsymbol{U}_1, \boldsymbol{U}_2 \in \mathcal{U}} \|\boldsymbol{U}_1 - \boldsymbol{U}_2\|_{\max}.$$

Run the α-approximation oracle defined in Algorithm 6 with input $\hat{\boldsymbol{U}}$, $m = \lfloor \log_2 N + 2 \rfloor$ and instability tolerance $2\gamma + \varepsilon$, we could generate a distribution $D \in \Delta(\mathcal{I})$ such that

$$\boldsymbol{U}_D(p_i) \geq \frac{\mathcal{U}_{\varepsilon}^*(p_i)}{m} - (2\gamma + \epsilon), \quad \forall i \in [N].$$

Proof. By running the α-approximation oracle with $\hat{\boldsymbol{U}}$, $2\gamma + \varepsilon$, $\lfloor \log_2 N + 2 \rfloor$, from Lemma J.10, we have

$$\boldsymbol{U}_D(p_i) \geq \frac{\hat{\boldsymbol{U}}_{2\gamma+\varepsilon}^*(p_i)}{m} - (2\gamma + \varepsilon), \quad \forall i \in [N]. \tag{53}$$

By construction, any utility matrix $\boldsymbol{U} \in \mathcal{U}$ satisfies $\|\boldsymbol{U} - \hat{\boldsymbol{U}}\|_{\max} \leq \gamma$. From Lemma G.6, we know that for any matching $\mu \in \mathcal{S}_{\boldsymbol{U}}^{\varepsilon}$, $\forall \boldsymbol{U} \in \mathcal{U}$, we have $\mu \in \mathcal{S}_{\hat{\boldsymbol{U}}}^{2\gamma+\varepsilon}$, that is $\bigcup_{\boldsymbol{U} \in \mathcal{U}} \mathcal{S}_{\boldsymbol{U}}^{\varepsilon} \subseteq \mathcal{S}_{\hat{\boldsymbol{U}}}^{2\gamma+\varepsilon}$. Therefore, for the ε-optimal stable share, we have

$$\mathcal{U}_{\varepsilon}^*(p_i) \leq \hat{\boldsymbol{U}}_{2\gamma+\varepsilon}^*(p_i), \quad \forall i \in [N]. \tag{54}$$

Combining Eq.(53) and (54), the conclusion holds. □

Lemma G.6. *Fix the preferences of arms over players. Given two utility matrix $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ such that $\|\boldsymbol{U}_1 - \boldsymbol{U}_2\|_{\max} < \gamma$, then any ε-stable matching μ for $\boldsymbol{U}_1$, is also $(2\gamma + \varepsilon)$-stable with respect to $\boldsymbol{U}_2$.*

Proof. If a matching μ is ε-stable with respect to $\boldsymbol{U}_1$, then for any (p, a) pair such that $p \succ_a \mu(a)$, we must have

$$\boldsymbol{U}_1(p,a) \leq \boldsymbol{U}(p, \mu(p)) + \varepsilon, \tag{55}$$

from the definition of ε-stable matching.

Since $\|\boldsymbol{U}_1 - \boldsymbol{U}_2\|_{\max} \leq \gamma$, we have that for any (p, a) pair, $|\boldsymbol{U}_1(p,a) - \boldsymbol{U}(p,a)| \leq \gamma$. Therefore,

$$\boldsymbol{U}_2(w,a) \leq \boldsymbol{U}_1(w,a) + \gamma \leq \boldsymbol{U}_1(w, \mu(w)) + \gamma + \epsilon \leq \boldsymbol{U}_2(w, \mu(w)) + 2\gamma + \epsilon, \tag{56}$$

where the first and the last inequality come from $\|\boldsymbol{U}_1 - \boldsymbol{U}_2\|_{\max} \leq \gamma$, while the second inequality holds according to Eq.(55). Therefore, combining $p \succ_a \mu(a)$ and Eq.(56), we can conlude that matching μ is $(2\gamma + \varepsilon)$-stable with respect to $\boldsymbol{U}_2$. □

H ADDITIONAL NUMERICAL EXPERIMENTS

In this section, we provide additional numerical experiment results.

For the stochastic contexts case, we evaluate player-optimal stable regret in a market with $N = 4$ players and $K = 4$ arms. Contexts have dimension $d = 3$. Each arm's preference over players is a random permutation of $\{1, \cdots, N\}$. Each entry of the preference parameter θ_i is uniformly randomly sampled from $[0, 1]$; while for the contexts, we firstly sample a Gaussian distribution with mean 10 and variance 1 for each entry and then normalize the vector so that $\|x_j(t)\|_2 = 1$.

The ridge penalty parameter is set to $\lambda = 1$ for all algorithms under comparison. For ETC, the exploration length is $h = 5000$; for Batched-ETC, the initial exploration length is $T_1 = 100$; and for BARB, the initial candidate gap is $\Delta_1 = 0.5$. In short, the algorithm parameter settings are the same as those in the example presented in the main text, while this example is in good condition in the sense that the minimum eigenvalue of the context covariance matrix would be large compared to the first example.

Figure 2 shows the player-optimal stable regret for each player. Due to batched explorations, the performances of BARB and Batched-ETC are slightly worse than that of ETC. However, when the minimum eigenvalue of the context covariance matrix vanishes, the performance of ETC and Batched-ETC may significantly deteriorate, as shown in Figure 1, which implies BARB should be a more robust choice.

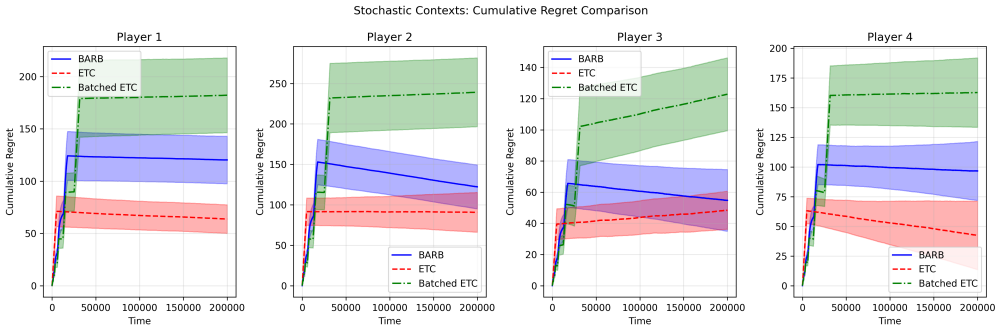

Figure 2: Experimental Comparison of BARB, Batched-ETC, and ETC in a market with 4 players and 4 arms.

For generality, we vary the market size $N = K \in \{3, 6, 9, 12\}$ to show the performance of the algorithm BARB. In Figure 3, the result shows that the maximum cumulative regret over players increases as the market size becomes larger. This dependency is consistent with the theoretical results.

Furthermore, to assess the robustness of the algorithm's performance with respect to the choice of the initial candidate gap Δ_1, we conduct a sensitivity analysis by varying Δ_1 over the set $0.4, 0.6, 0.8, 1.0$. The results, presented in Figure 4, show that the maximum cumulative regrets over players under different choices exhibit little variation, indicating that the algorithm is robust to this parameter selection.

For the adversarial contexts, we consider a market with $N = 4$ players and $K = 4$ arms. Contexts have dimension $d = 3$. Each arm's preference over players is a random permutation of $\{1, \cdots, N\}$. To confuse the algorithm, we intentionally alternate between generating contexts, thereby producing a utility matrix characterized by a mix of large and small gaps.

Given that the ε-optimal stable share is typically NP-hard to compute, we compare the rewards collected by each player under two conditions: (i) using the Gale-Shapley or approximation oracle with a known utility matrix, and (ii) following the AdECO algorithm (Algorithm 2).

Figure 5 shows the numerical results for each player in the adversarial setting. From Figure 5, we find that the collected reward curve of AdECO closely follows that of the oracle, demonstrating that its adaptive exploration strategy provides a good estimation.

Furthermore, we consider a more challenging scenario where, in each round, the adversary randomly selects the market to be in either the large-gap or small-gap region. We vary the probability of choosing the small-gap region as $p \in \{0.1, 0.5, 0.9\}$ and report the difference between the reward collected by the oracle—an algorithm that exactly knows the ground-truth utility and selects the offline oracle accordingly—and that collected by AdECO. The maximum difference across players is shown in Figure 6. As the probability of

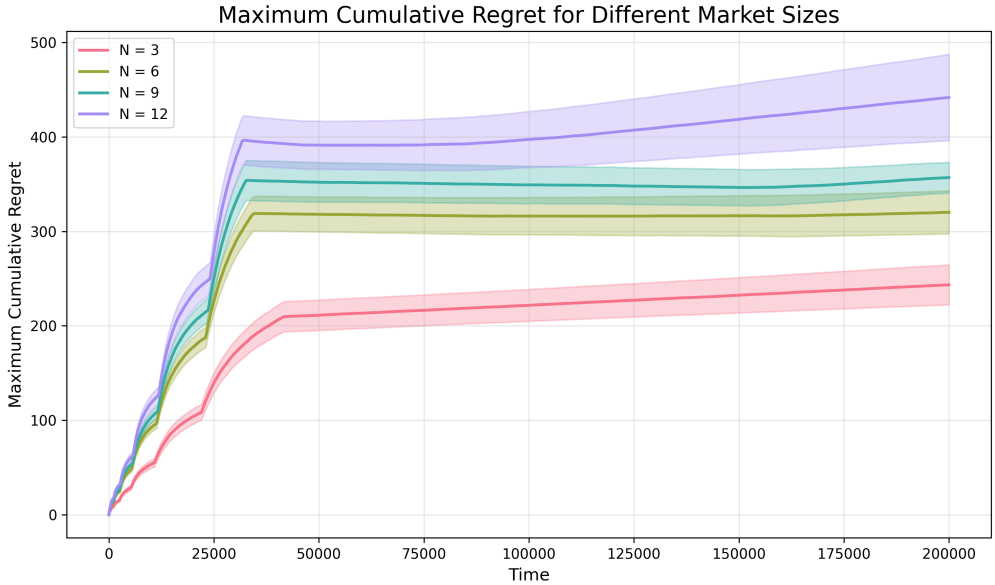

Figure 3: Experimental comparison of BARB with different market sizes.

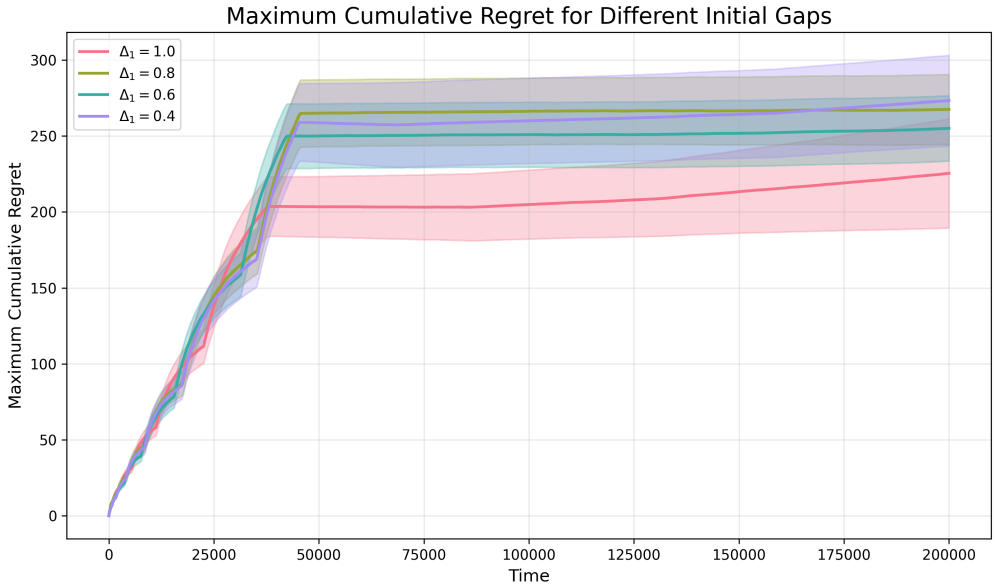

Figure 4: Experimental comparison of BARB with different initial candidate gaps Δ_1 in a market with $N = 4$ players and $K = 4$ arms.

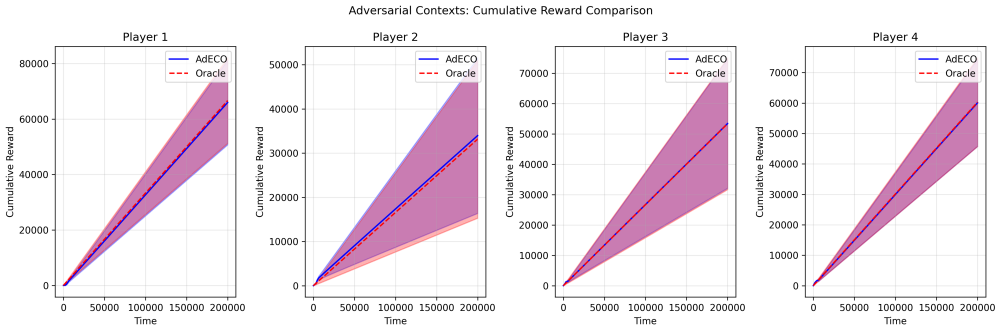

Figure 5: Experimental Comparison of AdECO and the offline oracle with 4 players and 4 arms.

the small-gap regime increases, the reward difference between the two algorithms decreases. This trend may be attributed to the choice of performance metric in each regime: in the small-gap regime, we use approximation regret as the metric, whereas in the large-gap regime, we use standard regret. The difference in scale between these two metrics likely accounts for the observed results.

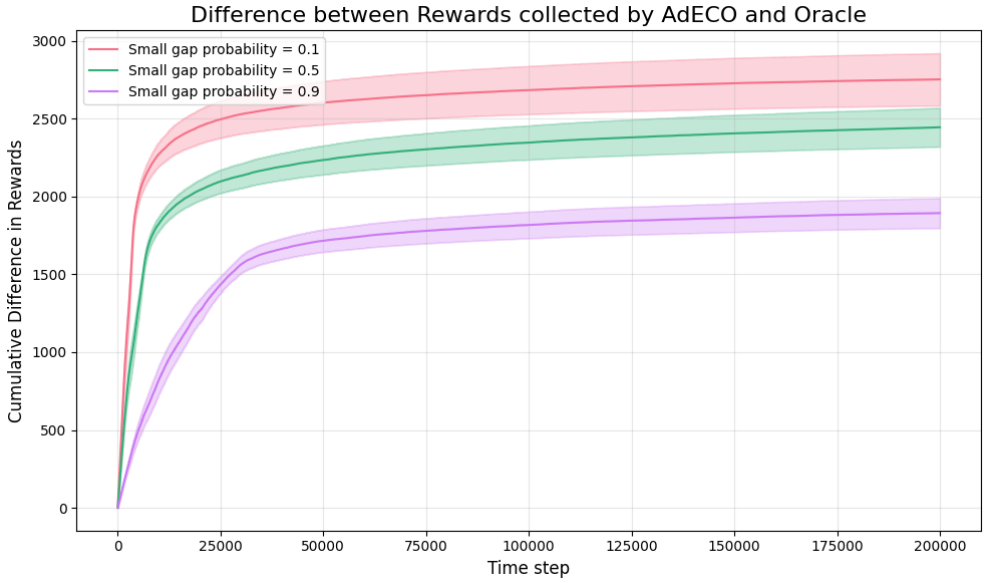

Figure 6: Adversarial contexts with varying probability of choosing the small-gap regime.

I ILLUSTRATION OF THE MINIMUM DIFFERENCE $\delta_{\min}$

In this section, we provide an illustrative example on the minimum difference $\delta_{\min}$.

For simplicity, we consider a stochastic contextual market with one player and three arms and the dimension of the context is $d = 1$. The latent preference parameter $\theta_1 = 1$, and the context distribution for the three arms are $Unif[0, \frac{1}{2}]$, $Unif[\frac{1}{4}, \frac{3}{4}]$ and $Unif[\frac{1}{2}, 1]$, respectively. Therefore, the distributions of the matching utilities are $\boldsymbol{U}_1 \sim Unif[0, \frac{1}{2}]$, $\boldsymbol{U}_2 \sim Unif[\frac{1}{4}, \frac{3}{4}]$ and $\boldsymbol{U}_3 \sim Unif[\frac{1}{2}, 1]$, respectively [4].

According to the definition of minimum preference, we have

$$\delta_{\min} = \min\left\{|\boldsymbol{U}_1 - \boldsymbol{U}_2|, |\boldsymbol{U}_2 - \boldsymbol{U}_3|, |\boldsymbol{U}_3 - \boldsymbol{U}_1|\right\}.$$

To derive the CDF $F_{\delta_{\min}}(x)$ of the minimum difference, we have that

$$F_{\delta_{\min}}(x) = \mathbb{P}\left(\delta_{\min} \leq x\right) = 1 - \mathbb{P}\left(\delta_{\min} > x\right).$$

Let $S(x) := \mathbb{P}\left(\delta_{\min} > x\right)$, given that the utilities all follow uniform distribution, we have that the joint density is constant on the cube

$$\Omega = \left[0, \frac{1}{2}\right] \times \left[\frac{1}{4}, \frac{3}{4}\right] \times \left[\frac{1}{2}, 1\right], \quad Vol\left(\Omega\right) = \frac{1}{8}.$$

Therefore,

$$S(x) = 8 \cdot Vol\left(\{(\boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{U}_3) \in \Omega : |\boldsymbol{U}_i - \boldsymbol{U}_j| > x, \forall i < j\}\right).$$

We partition the cube Ω into regions where the order of $(\boldsymbol{U}_1, \boldsymbol{U}_2, \boldsymbol{U}_3)$ is fixed. Only the following orderings have positive volume:

(A) $\boldsymbol{U}_1 \leq \boldsymbol{U}_2 \leq \boldsymbol{U}_3$.
(B) $\boldsymbol{U}_2 \leq \boldsymbol{U}_1 \leq \boldsymbol{U}_3$.
(C) $\boldsymbol{U}_1 \leq \boldsymbol{U}_3 \leq \boldsymbol{U}_2$.

Denote the corresponding volumes as V_A, V_B and V_C, we have that

$$S(x) = 8 \cdot \left(V_A(x) + V_B(x) + V_C(x)\right).$$

We work with different regions separately.

Region A : $\boldsymbol{U}_1 \leq \boldsymbol{U}_2 \leq \mu_3$.

Here, $\delta_{\min} = \min\{\boldsymbol{U}_2 - \boldsymbol{U}_1, \boldsymbol{U}_3 - \boldsymbol{U}_2\}$, thus we have $\boldsymbol{U}_2 - \boldsymbol{U}_1 > x$ and $\boldsymbol{U}_3 - \boldsymbol{U}_2 > x$ in region A.

Therefore, given $\boldsymbol{U}_2$, we have

$$\boldsymbol{U}_1 \in \left[0, \min\left\{\frac{1}{2}, \boldsymbol{U}_2 - x\right\}\right], \quad \boldsymbol{U}_3 \in \left[\max\left\{\frac{1}{2}, \boldsymbol{U}_2 + x\right\}, 1\right],$$

and hence

$$V_A(x) = \int_{\boldsymbol{U}_2} \max\left\{0, \min\left\{\frac{1}{2}, \boldsymbol{U}_2 - x\right\}\right\} \cdot \max\left\{0, 1 - \max\left\{\frac{1}{2}, \boldsymbol{U}_2 + x\right\}\right\} d\boldsymbol{U}_2,$$

which makes

$$V_A(x) = \begin{cases} \frac{4}{3}x^3 - \frac{1}{2}x^2 - \frac{1}{4}x + \frac{3}{32}, & 0 \leq x \leq \frac{1}{4}, \\ -\frac{4}{3}x^3 + 2x^2 - x + \frac{1}{6}, & \frac{1}{4} \leq x \leq \frac{1}{2}. \end{cases}$$

Region B : $\boldsymbol{U}_2 \leq \boldsymbol{U}_1 \leq \boldsymbol{U}_3$.

Here, $\delta_{\min} = \min\{\boldsymbol{U}_1 - \boldsymbol{U}_2, \boldsymbol{U}_3 - \boldsymbol{U}_1\}$, then given $\boldsymbol{U}_1$, we have

$$\boldsymbol{U}_2 \in \left[\frac{1}{4}, \min\left\{\frac{3}{4}, \boldsymbol{U}_1 - x\right\}\right], \quad \boldsymbol{U}_3 \in \left[\max\left\{\frac{1}{2}, \boldsymbol{U}_1 + x\right\}, 1\right].$$

With a similar derivation to that in region A, we have that

$$V_B(x) = \begin{cases} \frac{2}{3}x^3 + \frac{1}{8}x^2 - \frac{1}{8}x + \frac{1}{64}, & 0 \leq x \leq \frac{1}{8} \\ -\frac{2}{3}x^3 + \frac{5}{8}x^2 - \frac{3}{16}x + \frac{7}{384}, & \frac{1}{8} \leq x \leq \frac{1}{4}. \end{cases}$$

[4] Without causing ambiguity, we drop the index of the player and focus on the index of the arm.

Region C $: \boldsymbol{U}_1 \leq \boldsymbol{U}_3 \leq \boldsymbol{U}_2$.

Here, $\delta_{\min} = \min\{\boldsymbol{U}_3 - \boldsymbol{U}_1, \boldsymbol{U}_2 - \boldsymbol{U}_3\}$, then given $\boldsymbol{U}_3$, we have that

$$\boldsymbol{U}_1 \in \left[0, \min\left\{\frac{1}{2}, \boldsymbol{U}_3 - x\right\}\right], \quad \boldsymbol{U}_2 \in \left[\max\left\{\frac{1}{4}, \boldsymbol{U}_3 + x\right\}, \frac{3}{4}\right].$$

Similarly, we have

$$V_C(x) = \begin{cases} \frac{2}{3}x^3 + \frac{1}{8}x^2 - \frac{1}{8}x + \frac{1}{64}, & 0 \leq x \leq \frac{1}{8} \\ -\frac{2}{3}x^3 + \frac{5}{8}x^2 - \frac{3}{16}x + \frac{7}{384}, & \frac{1}{8} \leq x \leq \frac{1}{4}. \end{cases}$$

Combining the above three regions, we have the final CDF of $\delta_{\min}$ as $F_{\delta_{\min}}(x) = 1 - S(x)$, and

$$F_{\delta_{\min}}(x) = \begin{cases} 0, & x < 0 \\ 1 - 8 \cdot \left(\frac{8}{3}x^3 - \frac{1}{4}x^2 - \frac{1}{2}x + \frac{1}{8}\right), & 0 \leq x \leq \frac{1}{8}, \\ 1 - 8 \cdot \left(\frac{3}{4}x^2 - \frac{5}{8}x + \frac{25}{192}\right), & \frac{1}{8} \leq x \leq \frac{1}{4} \\ 1 - 8 \cdot \left(-\frac{4}{3}x^3 + 2x^2 - x + \frac{1}{6}\right), & \frac{1}{4} \leq x \leq \frac{1}{2} \\ 1, & x \geq \frac{1}{2}. \end{cases}$$

The cumulative distribution function (CDF) of $\delta_{\min}$ for the above example is depicted in Figure 7. We also plot the function $\frac{\log T}{Tx^2}$ for three distinct time horizons: 1000, 10000, and 100000. In this plot, the crossing points correspond to the minimum preference gap $\Delta_{\min}$ (Definition 2) for each respective T.

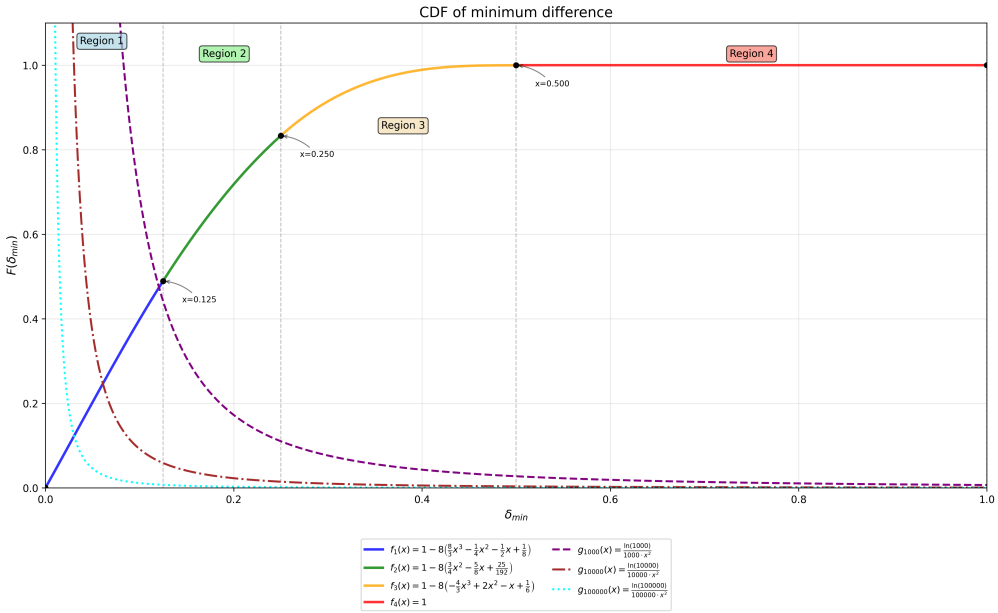

Figure 7: CDF of $\delta_{\min}$ and the minimum preference gap.

J TECHNICAL LEMMAS

Lemma J.1 (Elliptical Potential Lemma, Proposition 1 (Carpentier et al., 2020)). *Let $x_1, \cdots, x_T$ be a sequence of arbitrary vectors in $\mathbb{R}^d$ such that $\|x_i\|_2 \le 1$. For any $1 \le t \le T$, define $\boldsymbol{V}_t = \sum_{s=1}^{t-1} x_s x_s^\top + \lambda \boldsymbol{I}$. The elliptical potential satisfies*

$$\sum_{t=1}^{T} \|x_t\|_{\boldsymbol{V}_{t+1}^{-1}} = \sum_{t=1}^{T} \sqrt{x_t^\top \boldsymbol{V}_{t+1}^{-1} x_t} \le \sqrt{Td \log\left(\frac{T + d\lambda}{d\lambda}\right)},$$

and for any $p \in \mathbb{R}^+, \lambda \ge 1$, we have

$$\sum_{t=1}^{T} \|x_t\|_{\boldsymbol{V}_{t+1}^{-p}} \le \sum_{t=1}^{T} \|x_t\|_{\boldsymbol{V}_t^{-p}} \le 2^{\frac{p}{2}} \sum_{t=1}^{T} \|x_t\|_{\boldsymbol{V}_{t+1}^{-p}}.$$

Lemma J.2 (Confidence Ellipsoid, Theorem 2 (Abbasi-Yadkori et al., 2011)). *Let $\mathrm{x}_1, \cdots, \mathrm{x}_T$ be a sequence of vectors in $\mathbb{R}^d$ such that for all $t \ge 1$, $\|\mathrm{x}_t\|_2 \le B_x$. Let $\theta \in \mathbb{R}^d$ satisfy $\|\theta\|_2 \le B_\theta$ and $y_t = \theta^\top \mathrm{x}_t + \epsilon_t$, where ϵ_t is a centered, i.i.d. noise term independent of x_t and of $(\mathrm{x}_s, \epsilon_s)_{s<t}$, and $|\epsilon_t| \le B_\epsilon$. Therefore, we have $|y_t| \le B_y = B_\theta \cdot B_x + B_\epsilon$. Let $\boldsymbol{V} = \lambda \boldsymbol{I}$ for $\lambda > 0$, define*

$$\boldsymbol{V}_t = \boldsymbol{V} + \sum_{s=1}^{t} \mathrm{x}_s \mathrm{x}_s^\top, \quad \boldsymbol{S}_t = \sum_{s=1}^{t} y_s \mathrm{x}_s,$$

and let $\hat{\theta}_t$ be the ℓ^2-regularized least square estimate of θ with regularization parameter λ, i.e.,

$$\hat{\theta}_t = \boldsymbol{V}_t^{-1} S_t.$$

Then, for any $\delta > 0$, with probability at least $1 - \delta$, for all $t \ge 0$, θ lies in the set

$$\mathcal{C}_t = \left\{ \theta \in \mathbb{R}^d : \|\hat{\theta}_t - \theta\|_{\boldsymbol{V}_t} \le B_y \sqrt{d \log\left(\frac{1 + tB_x^2/\lambda}{\delta}\right)} + \lambda^{1/2} B_\theta \right\}.$$

Lemma J.3 (Multiplicative Chernoff Bound, Deviation Above the Mean, Theorem 4.4 (Mitzenmacher & Upfal, 2017)). *Let $X_1, \cdots, X_n$ be independent Bernoulli trials such that $\mathbb{P}(X_i) = p_i$, Let $X = \sum_{i=1}^{n} X_i$ and $\mu = \mathbb{E}[X]$. Then for any $\rho \ge 0$, we have*

$$\mathbb{P}(X \ge (1+\rho)\mu) \le \exp\left(-\frac{\rho^2 \mu}{2 + \rho}\right).$$

Lemma J.4 (Cauchy's Eigenvalue Interlacing Theorem, Theorem 4.3.17 (Horn & Johnson, 2012)). *Suppose $A \in \mathbb{R}^{n \times n}$ is symmetric. Let $B \in \mathbb{R}^{m \times m}$ with $m < n$ be a principal submatrix (obtained by deleting both i-th row and i-th column for some values of i. Suppose A has eigenvalues $\lambda_1 \le \cdots \le \lambda_n$ and B has eigenvalues $\beta_1 \le \cdots \le \beta_m$. Then*

$$\lambda_k \le \beta_k \le \lambda_{k+n-m} \quad \textit{for} \quad k = 1, \cdots, m.$$

Lemma J.5 (Matrix Bernstein Inequality, Theorem 1.6.2 (Tropp, 2015)). *Let $S_1, \cdots, S_n$ be independent, random matrices with common dimensions $d \times d$, and assume that $\|S_i\|_2 \le M$. Let $\mathbb{E}[S_i] = \Sigma$, $V = \|\sum_{i=1}^{n} \mathbb{E}\left[(S_i - \Sigma)^2\right]\| \le n\|\Sigma\|_2^2$. Then for any $t \ge 0$,*

$$\mathbb{P}\left(\|\frac{1}{n}\sum_{i=1}^{n} S_i - \Sigma\|_2 \ge t\right) \le 2d \cdot \exp\left(-\frac{nt^2/2}{V/n + Mt/3}\right) \lesssim 2d \exp\left(-C_1 n\right),$$

where C_1 is a constant.

Lemma J.6 (Divergence Decomposition, Lemma 15.1 (Lattimore & Szepesvári, 2020)). *For two bandit instances $\boldsymbol{\nu} = \{\nu_{ij} : i \in [N], j \in [K]\}$, and $\boldsymbol{\nu'} = \{\nu'_{ij} : i \in [N], j \in [K]\}$, fix some policy π and let $\mathbb{P}_{\boldsymbol{\nu},\pi}$ and $\mathbb{P}_{\boldsymbol{\nu'},\pi}$ be the probability measures induced by the T-round interconnection of π and $\boldsymbol{\nu}$ (respectively, π and $\boldsymbol{\nu'}$), the following divergence decomposition holds,*

$$D\left(\mathbb{P}_{\boldsymbol{\nu},\pi}, \mathbb{P}_{\boldsymbol{\nu'},\pi}\right) = \sum_{i=1}^{N} \sum_{j=1}^{K} \mathbb{E}_{\boldsymbol{\nu},\pi}\left[N_{ij}(T)\right] \cdot D\left(\nu_{ij}, \nu'_{ij}\right).$$

Lemma J.7 (Chain Rule of KL Divergence, Theorem 2.5.3 (Cover, 1999))**.** *Consider two random variables X and Y with two possible joint probability mass functions $p(x, y)$ and $q(x, y)$. The KL divergence between $p(x, y)$ and $q(x, y)$ is*

$$D\left(p(x,y)\|q(x,y)\right) = D\left(p(x)\|q(x)\right) + D\left(p(y|x)\|q(y|x)\right).$$

Lemma J.8 (Bretagnolle-Huber Inequality, Theorem 14.2 (Lattimore & Szepesvári, 2020))**.** *Let P and Q be probability measures on the same measurable space $(\Omega, \mathcal{F})$, and let $A \in \mathcal{F}$ be an arbitrary event. Then,*

$$P(A) + Q(A^c) \geq \frac{1}{2}\exp\left(-D(P,Q)\right),$$

where $A^c = \Omega \backslash A$ is the complement of A, $D(\cdot,\cdot)$ is the KL divergence.

Lemma J.9 (Multiplicative Chernoff Bound, Deviation Below the Mean, Theorem 4.5 (Mitzenmacher & Upfal, 2017))**.** *Let $X_1, \cdots, X_n$ be independent Bernoulli trials such that $\mathbb{P}(X_i) = p_i$. Let $X = \sum_{i=1}^n X_i$ and $\mu = \mathbb{E}[X]$. Then for any $0 < \rho < 1$, we have*

$$\mathbb{P}\left(X \leq (1-\rho)\mu\right) \leq \exp\left(-\frac{\mu\rho^2}{2}\right).$$

Lemma J.10 (Approximate Optimal Stable Share Guarantee of Algorithm 6, Theorem 5 (Lin et al., 2025))**.** *Given any utility matrix $\boldsymbol{U}$, parameter $m = \lfloor \log_2 N + 2 \rfloor$, and the instability tolerance $\varepsilon \geq 0$, Algorithm 6 computes a distribution $D \in \Delta(\mathcal{I})$, where $\mathcal{I}$ is the set of internally stable matchings, such that $\boldsymbol{U}_D(p) \geq \frac{\boldsymbol{U}_\varepsilon^*(p)}{m} - \varepsilon, \forall p \in \mathcal{N}$.*