
Probably Correct Optimal Stable Matching under Two-Sided Uncertainty

Andreas Athanasopoulos¹

Anne-Marie George²

Christos Dimitrakakis¹

¹University of Neuchâtel, Neuchâtel, Switzerland

²University of Oslo, Oslo, Norway

Abstract

We study a sequential learning problem for stable matchings in two-sided markets where preferences on both sides are initially unknown. We focus on a centralized setting where an algorithm matches agents at each time step and receives noisy rewards that reflect the preferences of the matched agents, following a semi-bandit feedback structure. We adopt a pure exploration perspective, aiming to efficiently identify the optimal stable matching with high probability. Our work extends prior results by handling *two-sided uncertainty* and by exploiting *partial preference* information. A central ingredient is the notion of **pervasive stable matching**, which enables the identification of optimal stable matchings under partial preferences. We propose elimination-based algorithms whose stopping criteria exploit the structure of the learned partial preferences, and provide a refined sample-complexity analysis. Beyond pure exploration, we extend our approach to regret minimization and establish regret bounds with respect to the *optimal* stable matching that avoid dependence on the minimum reward gap $\Delta_{\min}$.

1 INTRODUCTION

Two-sided matching markets model situations in which two distinct sets of agents aim to match with each other based on their preferences [Gusfield and Irving, 1989]. Classic examples include students applying to universities, workers matching with firms, and organ donors paired with patients [Roth, 1984]. A fundamental solution concept in this setting is *stable matching* [Gale and Shapley, 1962], a notion of equilibrium in the market, where no pair of agents has an incentive to deviate from the proposed assignment.

Classical models typically assume that preferences are

known in advance. In contrast, modern applications such as ride-sharing systems and online labor markets motivate sequential learning formulations, where agents’ preferences are initially unknown and must be learned over time through repeated interactions.

A growing body of literature casts this learning scenario within the Multi-Armed Bandit (MAB) framework, a direction initiated by Das and Kamenica [2005]. In the centralized version of the problem, a platform matches the agents and receives noisy feedback reflecting their underlying preferences. This setting creates an inherent tension between exploration and exploitation: the platform must decide whether to exploit current estimates of a stable matching or continue exploring to learn the agents’ preferences.

The problem is typically studied within a regret-minimization framework. In this work, we instead adopt a pure exploration perspective, aiming to efficiently identify an *optimal stable matching* with high probability.

1.1 RELATED WORK

In their foundational work, Das and Kamenica introduce this learning problem and empirically evaluate the performance of MAB algorithms across several preference structures. A follow-up by Liu et al. [2020] introduced the notions of *Player-Optimal* and *Player-Pessimal* stable regret. Focusing on a setting where only one side of the market has unknown preferences, they analyzed an Explore-Then-Commit algorithm, proving sublinear bounds for both regret notions. They also examined an Upper Confidence Bound approach, establishing sublinear bounds for Player-Pessimal stable regret; however, they proved an impossibility result for bounding Player-Optimal regret.

Later studies closed this gap by achieving sublinear player-optimal stable regret [Sankararaman et al., 2021, Basu et al., 2021, Kong and Li, 2023], though they primarily focused on one-sided uncertainty and specific market structures. More recent studies have begun to consider settings where both

sides of the market have uncertain preferences [Pagare and Ghosh, 2024, Zhang and Fang, 2024]. However, these works employ Explore-Then-Commit type algorithms. In contrast, this work adopts a pure exploration framework that is more naturally suited to the analysis of such algorithms and enables refined results on regret with respect to the *optimal* stable matching.

The pure exploration objective of identifying an *optimal stable matching* with high probability was introduced by Athanopoulos et al. [2025], while the work of Hosseini et al. [2024] focuses on identifying *any* stable matching. Both of these studies propose elimination-based algorithms restricted to one-sided uncertainty. We extend this literature by tackling a two-sided learning environment, providing refined arguments for the analysis of elimination algorithms.

A core component of our approach is utilizing the agents’ *partial preference* rankings to identify an optimal stable matching, thereby circumventing the need to learn their full preference orderings. This approach is motivated by the concept of *pervasive stable matching*, introduced by Rastegari et al. [2014] to reason about optimal stable matchings under partial information in *non-learning* scenarios. Specifically, a partial preference profile of the agents admits a pervasive stable matching if that matching remains optimally stable under every possible full refinement of the given preferences. A closely related notion is that of a *super stable matching*, initially introduced to address stability in instances with indifferent preferences [Irving, 1994]. While super stability was recently leveraged by Basu [2025] to bound regret in sequential learning scenarios, our work specifically relies on pervasive stable matchings. This distinction is crucial, as our objective is to identify the *optimal* stable matching, rather than *any* stable matching.

Appendix A contains an extended review of related literature, along with a comprehensive discussion on metrics.

1.2 CONTRIBUTIONS

In this work, we demonstrate how partial preferences and the concepts of *pervasive stable matching* can be applied in a learning scenario to identify the optimal stable matching with high probability under *two-sided market uncertainty*.

Our specific contributions are structured as follows:

- **Uniform Strategies:** We present two variants of uniform exploration strategies that assume additional information about the reward structure. We show that reasoning about partial preferences significantly reduces the overall sample complexity compared to approaches that learn the full rankings, thereby avoiding the traditional dependence on the minimum reward gap.
- **Elimination Algorithm & Technical Contribution:** We propose an adaptive elimination algorithm that re-

quires no prior knowledge of reward gaps, utilizing a stopping rule based on partial preferences. Furthermore, we provide a refined complexity bound based on edge-coloring arguments on bipartite graphs.

- **Extended Elimination Rule:** We also present an extension of the elimination algorithm that further eliminates pairs upon identifying a super stable matching.
- **Regret Minimization:** Beyond our primary objective of pure exploration, we propose a meta-algorithm for the regret minimization setting. To the best of our knowledge, this provides the first regret bound with respect to the *optimal* stable matching under *two-sided uncertainty* that avoids a dependence on the minimum reward gap $\Delta_{\min}$.
- **Theoretical Lower Bounds:** Finally, we conclude our analysis by establishing a lower bound on the sample complexity, utilizing known results from the combinatorial MAB literature [Jourdan et al., 2021].

The remainder of the paper is organized as follows. Section 2 introduces the necessary notation and background regarding the stable matching problem and partial preferences. Section 3 then formally defines our learning problem and explains how we utilize the concept of pervasive stable matching. In Section 4, we analyze uniform sampling strategies, discussing the impact of utilizing partial preferences. Section 5 presents our elimination-based algorithm, followed by an extension of the elimination procedure and a discussion on a meta-algorithm for the regret minimization setting. Next, Section 6 establishes a lower bound on the sample complexity. In Section 7, we evaluate our algorithms through simulation on random instances. Further discussion and omitted proofs are provided in the Appendix.

2 BACKGROUND

We consider a learning problem in a two-sided matching market, consisting of two distinct, non-empty sets of agents, $\mathcal{P}$ and $\mathcal{A}$, with $K = |\mathcal{P}|$ and $N = |\mathcal{A}|$ denoting the number of agents on each side, with arbitrary K, N ¹.

The agents $\mathfrak{A} = \mathcal{P} \cup \mathcal{A}$ are matched one-to-one according to a matching $\mathbb{M} \subseteq \mathcal{P} \times \mathcal{A}$, i.e., pairs in $\mathbb{M}$ are pairwise disjoint. For convenience, we use the functional representation of a matching, where $m(p) = a$ and $m(a) = p$ for every $(p, a) \in \mathbb{M}$, and $m(a) = \perp$ for any unmatched agent $a \in \mathfrak{A}$. We define $OS(a)$ as the opposite side of agent a , i.e., $OS(a) = \mathcal{A}$ if $a \in \mathcal{P}$, and $OS(a) = \mathcal{P}$ if $a \in \mathcal{A}$.

Each agent $a \in \mathfrak{A}$ has an underlying true preference over the agents on the opposite side, represented by a strict preference ranking F_a . These preferences are derived from utility functions $\mu_a : OS(a) \cup \{\perp\} \rightarrow \mathbb{R}$,

¹Prior work usually assume $K \leq N$ [Liu et al., 2021, Basu, 2025, Kong and Li, 2023, Athanopoulos et al., 2025].

where $\mu_{a,m(a)}$ denotes the utility that agent a obtains when matched with $m(a) \in OS(a)$, and $\mu_{a,\perp}$ is the utility of being unmatched. We assume that for every agent $a \in \mathcal{A}$, the utility values $\{\mu_{a,i}\}_{i \in OS(a) \cup \{\perp\}}$ are distinct, thereby ensuring no ties in preferences. The preference ranking is then uniquely constructed by sorting these utilities: $F_a = \arg \text{sort}_{i \in OS(a) \cup \{\perp\}} \mu_{a,i}$. We denote by $\succ_{F_a}$ the corresponding strict order relation over $OS(a) \cup \{\perp\}$, where $i \succ_{F_a} j$ if and only if $\mu_{a,i} > \mu_{a,j}$ for $i, j \in OS(a) \cup \{\perp\}$.

In order for a matching m to align with the agents' preferences, Gale and Shapley [1962] introduced the concept of stable matching, a notion of equilibrium in which no two agents have an incentive to deviate. Formally:

Definition 2.1 (Stable Matching). A matching m is stable under the true preference profile $\{F_a\}_{a \in \mathcal{A}}$ if:

1. **Individual rationality:** For every agent $a \in \mathcal{A}$, it holds that $m(a) \succeq_{F_a} \perp$.
2. **No blocking pairs:** There exists no pair $(p, a) \in \mathcal{P} \times \mathcal{A}$ such that agent p prefers a over $m(p)$ and agent a prefers p over $m(a)$, i.e., $a \succ_{F_p} m(p)$ and $p \succ_{F_a} m(a)$.

Gale and Shapley [1962] proved that the set of stable matchings $\mathcal{S}$ is always non-empty by introducing the Gale–Shapley (GS) algorithm, which is based on sequential proposals from one side of the market to the other. The algorithm returns the (unique) *optimal stable matching* m^* for the proposing side, in the sense that every proposer receives the best partner they can obtain in any stable matching. Conversely, the resulting matching is pessimal for the side receiving the proposals, providing each agent on that side with their least-preferred partner among all stable matchings. In the remainder of the paper we concentrate on $\mathcal{P}$ -optimal stable matchings m^* , simply referred to as *optimal*.

2.1 STABLE MATCHING FROM PARTIAL PREFERENCES

If we correctly estimate the full preference rankings of the agents, we can identify the optimal stable matching using the GS algorithm. However, learning the complete preference orders is not necessary: the optimal stable matching can be identified using only a correct partial preference order, as described by Rastegari et al. [2014].

Stable matchings under certain types of partial preferences form the basis of our proposed algorithms. We begin with the definition of a partial ranking.

Definition 2.2 (Partial ranking). A partial ranking P_a of an agent $a \in \mathcal{A}$ over the set of agents $OS(a)$ is a partial order $\succ_{P_a}$, i.e., asymmetric and transitive, that can be represented as directed acyclic graph (DAG) with edges $(i, j) \subseteq OS(a) \times OS(a)$ for all $i \neq j$ with $i \succ_{P_a} j$ (meaning $i \not\preceq_{P_a} j$).

Note that all partial rankings are asymmetric and transitive, and that a full ranking is a special case of a partial ranking.

Definition 2.3 (Full ranking). A full ranking F_a of an agent a over the set of agents $OS(a)$ is a partial ranking such that, for every distinct pair $(i, j) \in OS(a) \times OS(a)$, either $i \succ_{F_a} j$ or $j \succ_{F_a} i$.

We now define when a partial ranking P_a is compatible with P'_a , in the sense that P'_a is a refinement of P_a .

Definition 2.4 (Compatible Partial Rankings). A partial ranking P_a is said to be *compatible* with another partial ranking P'_a , denoted by $P_a \mapsto P'_a$, if for any $i, j \in OS(a)$ such that $i \succ_{P'_a} j$, it is not the case that $j \succ_{P_a} i$. We overload this definition to partial *preference profiles* $P_A = \{P_a\}_{a \in \mathcal{A}}$ and $P'_A = \{P'_a\}_{a \in \mathcal{A}}$ of the agents, writing $P_A \mapsto P'_A$ if $P_a \mapsto P'_a$ for all $a \in \mathcal{A}$.

Given a partial preference profile P_A , we can describe the set of full ranking profiles $F_A = \{F_a\}_{a \in \mathcal{A}}$ that refine P_A as $\mathcal{F}(P_A) = \{F_A \mid P_A \mapsto F_A\}$. This motivates the following definition of a super stable matching [Irving, 1994], a concept originally proposed for preferences with indifference: we say that a matching m is *super stable* w.r.t. the preference profile P_A if it is stable w.r.t. every full ranking profile $F_A \in \mathcal{F}(P_A)$.

The work of Rastegari et al. [2014] introduce a similar definition of *pervasive stable matching* (PSM), for a matching that is stable and optimal with respect to every full refinement of the partial preferences $F_A \in \mathcal{F}(P_A)$.

Definition 2.5 (Pervasive Stable Matching). We say that a matching m is a *pervasive stable matching* for the preference profile P_A if m is the optimal stable matching for every full ranking profile $F_A \in \mathcal{F}(P_A)$. We also say that P_A *admits a pervasive stable matching* if such a matching exists.

Previous work shows that the set of super stable matchings may be empty, and whenever it is non-empty, it forms a distributive lattice [Spieker, 1995]. In this case, there exists a unique optimal super stable matching for each side. In the following example we highlight that an optimal super stable matching is not necessarily the same as the pervasive stable matching [Rastegari et al., 2014].

Example 1. Consider an instance with two agents on each side, $\mathcal{P} = \{p_1, p_2\}$ and $\mathcal{A} = \{a_1, a_2\}$, respectively. We assume that the agents in $\mathcal{A}$ have correctly estimated their preferences, while agents in $\mathcal{P}$ have no knowledge about their preferences yet, i.e. $P_p = \emptyset \forall p \in \mathcal{P}$, making them indifferent between a_1 and a_2 , denoted by $a_1 \sim a_2$:

p_1	p_2	a_1	a_2
$a_1 \sim a_2$	$a_1 \sim a_2$	$p_1 \succ p_2$	$p_2 \succ p_1$

In this setting, there are four possible full ranking profiles in $\mathcal{F}(\{P_a\}_{a \in \mathcal{A}})$ that can refine the partial preferences of the agents in $\mathcal{P}$ that we illustrate in the next table.

Instance	p_1	p_2	a_1	a_2
I_1	$a_1 \succ a_2$	$a_1 \succ a_2$	$p_1 \succ p_2$	$p_2 \succ p_1$
I_2	$a_1 \succ a_2$	$a_2 \succ a_1$	$p_1 \succ p_2$	$p_2 \succ p_1$
I_3	$a_2 \succ a_1$	$a_1 \succ a_2$	$p_1 \succ p_2$	$p_2 \succ p_1$
I_4	$a_2 \succ a_1$	$a_2 \succ a_1$	$p_1 \succ p_2$	$p_2 \succ p_1$

The only super stable matching, and thus the optimal, is $M = \{(p_1, a_1), (p_2, a_2)\}$ as the matching is stable with respect to every possible profile I_j with $j = 1, \dots, 4$.

Note that M , is not a pervasive stable matching, as for I_3 the optimal stable matching is $\{(p_1, a_2), (p_2, a_1)\}$.

If p_1 collects more samples to decide that $a_1 \succ a_2$, then M is PSM, as it is the optimal stable for both I_1 and I_2 .

3 LEARNING PROBLEM

We consider a learning problem where agents are initially unaware of their preferences and must learn them through repeated interactions. We focus on the centralized setting, where an algorithm selects a matching at each time step and observes stochastic rewards for the matched pairs².

More formally, the algorithm interacts with an instance $\nu = \{\nu_{a,i}\}_{(a,i) \in \mathcal{P} \times \mathcal{A}} \in \mathcal{I}$, where each $\nu_{a,i}$ corresponds to the reward distribution of agent $a \in \mathcal{A}$ when matched with agent $i \in OS(a)$, and has mean $\mu_{a,i} = \mathbb{E}_{X \sim \nu_{a,i}}[X]$.

Assumption 1 (1-sub-Gaussian). For every agent $a \in \mathcal{A}$ and $i \in OS(a)$, the reward distribution $\nu_{a,i}$ is 1-sub-Gaussian.

At each time step $t \geq 1$, the algorithm selects a matching $m_t \in \mathcal{M}$ out of the set of all matchings $\mathcal{M}$, and observes stochastic rewards $X_{a,m_t(a)} \sim \nu_{a,m_t(a)}$ for each matched pair $(a, m_t(a))$ for all $a \in \mathcal{A}$. Specifically, the feedback at time t is represented by the vector $X_t = (X_{a,m_t(a)})_{a \in \mathcal{A}} \in \mathbb{R}^{K \times N}$. The filtration $\mathcal{F}_t := \sigma(m_1, X_1, \dots, m_t, X_t)$ contains all the information available at step $t + 1$, where $\sigma(\cdot)$ denotes the σ -algebra generated by its arguments.

The algorithm is defined by:

1. A **Sampling Rule**: A rule determining which matching to sample at each step t , formally a sequence $(m_t)_{t \in \mathbb{N}}$ where each m_t is $\mathcal{F}_{t-1}$ -measurable.
2. A **Stopping Rule**: A stopping time τ with respect to the filtration $(\mathcal{F}_t)_{t \in \mathbb{N}}$, which determines when the exploration phase ends.

²Equivalently, we can assume that agents receive the rewards and truthfully report them to the algorithm.

3. A **Recommendation Rule**: An $\mathcal{F}_\tau$ -measurable rule that selects a final matching m_τ to be returned upon termination.

We consider the fixed-confidence setting, where the objective is to design an algorithm that terminates and outputs the correct optimal stable matching with high probability.

Definition 3.1 (Probably Correct Optimal Stable-matching). Given a fixed confidence parameter $\delta \in (0, 1)$, an algorithm is said to be δ -PCOS for every $\nu \in \mathcal{E}_{\mathbb{M}}$ if it terminates after $\tau < \infty$ steps and returns the optimal stable matching m_ν^* with high probability:

$$\mathbb{P}_\nu(m_\tau = m_\nu^*) \geq 1 - \delta.$$

Remark 3.2. Beyond the probability of correctness, we are also interested in providing guarantees for the stopping time τ , which represents the total number of matchings sampled before termination. We refer to the expected number of steps $\mathbb{E}[\tau]$ as the sample complexity of the algorithm.

Learning with Partial Preferences. Our previous discussion in Section 2.1 indicates that an algorithm only needs to learn a valid partial preference profile P_A that admits a pervasive stable matching. This insight motivates the design of our algorithms in the subsequent sections.

Definition 3.3 (Valid Preferences). We say that a partial preference profile P_A is *valid* for the full ranking profile F_A of an instance ν if:

1. P_A is compatible with F_A , i.e., $P_A \mapsto F_A$, and
2. P_A admits a pervasive stable matching.

We denote the set of valid preference profiles by $\mathcal{P}_{\text{valid}}(F_A^\nu)$, which is never empty as it always contains the true profile F_A . While the work of Rastegari et al. [2013] proves that identifying the set of valid preference profiles is NP-hard, their follow-up research provides a polynomial-time algorithm to return the pervasive stable matching whenever it exists, given a partial ranking [Rastegari et al., 2014].

To analyze the sample complexity of our algorithms, we introduce an admissible gap Δ_F associated with an instance and the set of valid partial preference profiles $\mathcal{P}_{\text{valid}}(F_A^\nu)$.

We define the minimum gap induced by a partial ranking P_a as $\Delta_{\min}(P_a) = \min_{(i,j) \in P_a} |\mu_{a,i} - \mu_{a,j}|$. Thus $\Delta_{\min}(P_a)$ indicates how hard it is for agent a to learn their preference between pair $(i, j) \in P_a$ in the worst case. For a preference profile $P_A = \{P_a\}_{a \in \mathcal{A}}$, we extend this to $\Delta_{\min}(P_A) = \min_{a \in \mathcal{A}} \Delta_{\min}(P_a)$. Finally, the *admissible gap of an instance* is defined as the maximum such minimum gaps over all valid profiles, i.e.,

$$\Delta_F = \max_{P_A \in \mathcal{P}_{\text{valid}}(F_A^\nu)} \Delta_{\min}(P_A). \quad (1)$$

Relation to the minimum gap: Previous work typically characterizes the performance of an algorithm in terms of the minimum reward gap $\Delta_{\min} = \min_{a \in \mathcal{A}} \Delta_{\min}(F_a)$ [Liu et al., 2020]. In contrast, our admissible gap Δ_F , defined with respect to the minimum valid partial profile, satisfies $\Delta_F \geq \Delta_{\min}$ by definition.

4 UNIFORM STRATEGIES

Our first contribution using partial preferences is the analysis of a uniform strategy that samples matchings covering all possible pairs $\mathcal{P} \times \mathcal{A}$ uniformly for a fixed number of h exploration rounds. We analyze this algorithm for two cases, where the number of exploration rounds h uses $\Delta_{\min}$ or Δ_F , to ensure the identification of the correct optimal stable matching with high probability.

After uniformly sampling pairs in $\mathcal{P} \times \mathcal{A}$ for h rounds, Algorithm 1 commits to a matching by executing GS on the estimated preferences, $m_\tau = \text{GS}(\{\hat{F}_a\}_{a \in \mathcal{A}})$. Here the preference ranking for an agent $a \in \mathcal{A}$ is determined by sorting the empirical mean rewards, i.e., $\hat{F}_a = \arg \text{sort}_{i \in OS(a)} \hat{\mu}_{a,i}(t)$, where $\hat{\mu}_{a,i}(t) = \frac{1}{N_{a,i}(t)} \sum_{s=1}^t \mathbb{I}(m_s(a) = i) X_{a,i}(s)$ and $N_{a,i}(t)$ denotes the number of times the pair (a, i) has been sampled up to time t .

Sampling from Minimal Matching Covers: To sample all pairs in a set $X \subseteq \mathcal{P} \times \mathcal{A}$, we sample matchings from a minimal matching cover $\mathfrak{M}(X)$ — the smallest set of matchings such that $X \subseteq \bigcup_{m \in \mathfrak{M}(X)} m$. For a bipartite graph, finding this cover is equivalent to a minimum edge coloring problem, where the edges of each color correspond to a matching in the cover [Gabow and Kariv, 1978]. According to König’s Theorem, the minimal number of matchings in a cover is equal to the maximum degree $\deg(G(X))$ of the bipartite graph $G(X)$ with edges X . Thus, in our setting, a full exploration of all pairs $\mathcal{P} \times \mathcal{A}$ requires $\max\{K, N\}$ matchings in each exploration round h . Sampling from a matching cover is essential for our later elimination algorithms, where only a specific subset of pairs needs to be sampled in each step. It relaxes the requirement for round-robin strategies and extends the analysis beyond the usual assumption of $N < K$.

We can now bound the sample complexity of the uniform strategy where the number of exploration rounds is defined as a function of $\Delta_{\min}$. This algorithm samples pairs until all agents *complete* preference rankings can be determined.

Theorem 4.1. *Algorithm 1 with parameter $h = \lceil \frac{8 \ln(2KN/\delta)}{\Delta_{\min}^2} \rceil$ is a δ -PCOS algorithm, and the number of sampled matchings is bounded by:*

$$O\left(\max\{K, N\} \frac{\ln(KN/\delta)}{\Delta_{\min}^2}\right).$$

Next, we introduce a variant where h depends on Δ_F . In this

Algorithm 1 Uniform Exploration

Require: $h, \mathcal{P}, \mathcal{A}, n = \max\{|\mathcal{P}|, |\mathcal{A}|\}$

- 1: Compute a minimal matching cover
 $\{\mathbf{m}_i\}_{i=1}^n = \mathfrak{M}(\{(p, a) \in \mathcal{P} \times \mathcal{A}\})$
- 2: **for** t in $\{1, \dots, h \cdot n\}$ **do**
- 3: Sample m_{i+1} where $i = (t - 1) \bmod n$.
- 4: Update $\hat{\mu}_{a, \mathbf{m}_{i+1}(a)}(t) \forall a \in \mathcal{A}$
- 5: **end for**
- 6: $\hat{F}_a = \arg \text{sort}_{i \in OS(a)} \hat{\mu}_{a,i} \forall a \in \mathcal{A}$
- 7: **return** $m_\tau = \text{GS}(\{\hat{F}_a\}_{a \in \mathcal{A}})$

version, the algorithm samples each pair often enough to correctly identify a valid preference profile $P_A \in \mathcal{P}_{\text{valid}}(F_A^\nu)$, with $\Delta_F = \max_{P_A \in \mathcal{P}_{\text{valid}}(F_A^\nu)} \Delta_{\min}(P_A)$.

Theorem 4.2. *Algorithm 1 with parameter $h = \lceil \frac{8 \ln(2KN/\delta)}{\Delta_F^2} \rceil$ is a δ -PCOS algorithm, and the number of sampled matchings is bounded by:*

$$O\left(\max\{K, N\} \frac{\ln(KN/\delta)}{\Delta_F^2}\right).$$

Remark 4.3. Note that by construction $\Delta_{\min} \leq \Delta_F$, which can lead to an arbitrarily large increase in sample complexity, as illustrated in Example 2 and Figure 1. While knowing either gap, $\Delta_{\min}$ or Δ_F , might not be possible in practice, knowing Δ_F seems to be an even more unrealistic assumption, as computing the set of valid preference profiles is computationally hard [Rakhlin and Sridharan, 2013].

Example 2. *Consider an instance with two agents on each side, with the following true preference rankings:*

p_1	p_2	a_1	a_2
$a_1 \succ a_2$	$a_1 \succ a_2$	$p_1 \succ p_2$	$p_1 \succ p_2$

For this instance, the partial preference profile $P_A = \{P_{p_1} = \{a_1 \succ a_2\}, P_{p_2} = \emptyset, P_{a_1} = \{p_1 \succ p_2\}, P_{a_2} = \emptyset\}$ is valid, i.e., the algorithm does not need to estimate the preferences of p_2 and a_2 . So the difficulty of the learning task does not depend on the gaps associated with these agents.

To illustrate, consider the following mean rewards:

Agent	Mean rewards μ	
p_1	$\mu_{p_1, a_1} = 1$	$\mu_{p_1, a_2} = 1 - \Delta_F$
a_1	$\mu_{a_1, p_1} = 1$	$\mu_{a_1, p_2} = 1 - \Delta_F$
p_2	$\mu_{p_2, a_1} = 1$	$\mu_{p_2, a_2} = 1 - \Delta_{\min}$
a_2	$\mu_{a_2, p_1} = 1$	$\mu_{a_2, p_2} = 1 - \Delta_{\min}$

Figure 1 fixes $\Delta_F = 0.5$ and varies $\Delta_{\min}$, comparing the sample complexity of the algorithms with exploration time h defined in Theorem 1 and Theorem 2, and shows that the resulting differences in the sample complexities can become arbitrarily large as $\Delta_{\min}$ goes to zero.

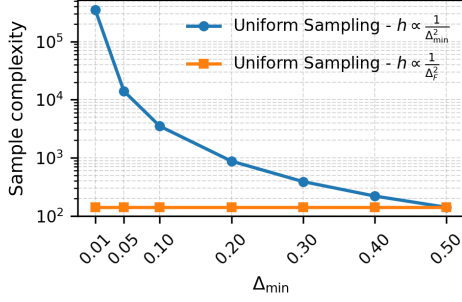


Figure 1: Sample complexity bounds from Theorem 4.1 and 4.2 for Example 2’s instance with fixed $\Delta_{\min} < \Delta_F = 0.5$.

5 ELIMINATION ALGORITHM

We now turn our attention to elimination algorithms: these have improved sample complexity and do not require prior knowledge of gaps Δ .

5.1 SIMPLE ELIMINATION ALGORITHM

The key idea of Algorithm 2 is to (i) update partial preferences once pairwise orderings are inferred with high probability, (ii) eliminate pairs from sampling, and (iii) stop when a PSM is found. High-probability correctness of the partial preferences implies identification of the optimal matching m^* with high probability.

Sampling: The elimination algorithm operates in rounds t and keeps track of active (non-eliminated) agents $S_a(t) \subseteq OS(a)$ – those for which the relative ordering in a ’s preferences is not determined with high probability yet. We initialise $S_a(0) = OS(a)$ for all agents $a \in \mathcal{A}$. Then in each round t , we sample the matchings from a *minimal matching cover* $\mathcal{M}(S_t)$, where S_t is the set of all non-eliminated pairs:

$$S_t = \{(p, a) \in \mathcal{P} \times \mathcal{A} \mid a \in S_p(t) \vee p \in S_a(t)\}$$

So at each round t , we sample n_t matchings, with n_t defined as the maximum degree of the underlying graph of the non-eliminated pairs $G(S_t)$, i.e., $n_t = \deg(G(S_t))$.

Partial Preferences: To eliminate pairs and form the partial preferences of the agents, we employ confidence intervals. Specifically, at each round t , every agent $a \in \mathcal{A}$ maintains a confidence interval for the mean reward associated with each agent $i \in OS(a)$. The corresponding confidence radius is defined as:

$$c_{a,i}(t) = \sqrt{\frac{2 \ln(8KNt^2/\delta)}{t}}. \quad (2)$$

Algorithm 2 E-PSM: Elimination PSM

Require: $\delta, \mathcal{P}, \mathcal{A}$

- 1: $S_a \leftarrow OS(a) \quad \forall a \in \mathcal{A}$ $\triangleright$ Active pairs
 - 2: $P_a \leftarrow \emptyset \quad \forall a \in \mathcal{A}$ $\triangleright$ Partial ranking
 - 3: $m_\tau = POSM(\{P_a\}_{a \in \mathcal{A}})$
 - 4: **while** $m_\tau = \text{NONE} \wedge \bigcup_{a \in \mathcal{A}} S_a \neq \emptyset$ **do**
 - 5: $S_t \leftarrow \{(p, a) \in \mathcal{P} \times \mathcal{A} \mid a \in S_p(t) \vee p \in S_a(t)\}$
 - 6: **for** $m \in \mathcal{M}(S_t)$ **do**
 - 7: Sample update $\hat{\mu}_{a,m(a)} \quad \forall a \in \mathcal{A}$
 - 8: **end for**
 - 9: Update $c_{a,i}(t) \quad \forall a \in \mathcal{A}, i \in OS(a)$ $\triangleright$ Eq. 2
 - 10: Update $P_a \quad \forall a \in \mathcal{A}$ $\triangleright$ Eq. 5
 - 11: Eliminate pairs in S_a $\triangleright$ Eq. 6
 - 12: $m_\tau = POSM(\{P_a\}_{a \in \mathcal{A}})$
 - 13: **end while**
 - 14: **return** m_τ
-

Using this radius, we define the Upper Confidence Bound (UCB) and Lower Confidence Bound (LCB) as:

$$UCB_{a,i}(t) = \hat{\mu}_{a,i}(t) + c_{a,i}(t), \quad (3)$$

$$LCB_{a,i}(t) = \hat{\mu}_{a,i}(t) - c_{a,i}(t). \quad (4)$$

We construct the partial ranking P_a by adding a directed edge from i to j for any distinct pair $i, j \in OS(a)$ whenever the confidence intervals are disjoint. Formally, at round t , we update the edge set of P_a according to:

$$P_a \leftarrow P_a \cup \{(i, j)\} \quad \text{if} \quad LCB_{a,i}(t) > UCB_{a,j}(t). \quad (5)$$

Elimination Rule: We remove an agent i from a set of active agents S_a if i ’s confidence interval is strictly disjoint from that of every other active agent $j \in S_a \setminus \{i\}$. Formally, we update $S_a \leftarrow S_a \setminus \{i\}$ if $\forall j \in S_a \setminus \{i\}$:

$$LCB_{a,i}(t) > UCB_{a,j}(t) \vee UCB_{a,i}(t) < LCB_{a,j}(t). \quad (6)$$

This implies that the relative order between i and every other active agent j has been established with high probability, rendering further sampling of i unnecessary for agent a .

Stopping Condition: We stop sampling either when all available pairs have been eliminated or when the agents’ partial preference profile admits a PSM. By construction, the partial preferences are consistent with the true underlying preferences with high probability; thus, the matching is the optimal stable matching with high probability. To compute a PSM, we use the POSM algorithm [Rastegari et al., 2014] (line 3 of Algorithm 2), which returns a PSM for the given partial preference profile if one exists, and NONE otherwise, in polynomial time.

To analyse the sample complexity of Algorithm 2, we define the following reward gaps: For an agent $a \in \mathcal{A}$ and $i \in OS(a)$. Let $\Delta_{a,i} = \min_{j \in OS(a) \setminus \{i\}} |\mu_{a,i} - \mu_{a,j}|$

denote the minimum reward gap separating i from all other candidates, and define $\Delta_{a,i}^{\min} = \min\{\Delta_{a,i}, \Delta_{i,a}\}$ and $\Delta_{a,i}^F = \max(\Delta_F, \Delta_{a,i}^{\min})$.

Theorem 5.1. *Algorithm 2 is a δ -PCOS algorithm, and the number of sampled matchings is bounded by:*

$$O\left(\max_{a \in \mathfrak{A}} \sum_{i \in OS(a)} \frac{32 \log(8KN/(\delta \Delta_{a,i}^F))}{(\Delta_{a,i}^F)^2}\right).$$

5.2 REGRET ANALYSIS

As previously discussed, almost all results bounding regret w.r.t. the optimal stable matching are based on the ETC framework. More specifically, in our setting we can employ an meta-algorithm that runs the E-PSM algorithm (Algorithm 2) during the exploration phase, and subsequently commits to the identified matching for all remaining rounds. We provide a bound for the following notion of regret:

$$R(T) = \sum_{t=1}^T \mathbb{I}\{m_t \neq m^*\}. \quad (7)$$

Basu [2025] uses a similar binary regret for a matching to be stable. In addition, our metric upper-bounds the player-optimal stable regret defined by Liu et al. [2020], a metric that reflects market stability only when it is sublinear for every agent in the market.

Theorem 5.2. *The E-PSM meta-algorithm with horizon T and $\delta = 1/T$ has an expected regret $\mathbb{E}[R(T)]$ bounded by:*

$$O\left(\min\left\{T, \max_{a \in \mathfrak{A}} \sum_{i \in OS(a)} \frac{32 \log(8KNT/\Delta_{a,i}^F)}{(\Delta_{a,i}^F)^2}\right\}\right).$$

Note that this minimum over the horizon cannot be avoided by any algorithm employing the Explore-Then-Commit framework. To our knowledge, this is the first regret bound with respect to the optimal stable matching under two-sided uncertainty that avoids a dependence on $\Delta_{\min}$.

5.3 EXTENDED ELIMINATION RULE

We now present an extended version of our previous algorithm, which further eliminates pairs upon the identification of a *super stable matching* (see Algorithm 3 in the Appendix). Specifically, if at some time step the partial preference profile $\{P_a\}_{a \in \mathfrak{A}}$ admits a super stable matching m_{ss} , we use the following elimination rule:

$$\begin{aligned} S_p &\leftarrow S_p \setminus \{a\} & \text{if } m_{ss}(p) \succ_{P_p} a \quad \forall p \in \mathcal{P}. \\ S_a &\leftarrow S_a \setminus \{p\} & \text{if } p \succ_{P_a} m_{ss}(a) \quad \forall a \in \mathcal{A}. \end{aligned}$$

To show that this elimination still allows the identification of a valid profile, and thus a pervasive stable matching, we first show that one specific profile dependant on the optimal stable matching is always valid.

Proposition 5.3. *Let F^* be a full ranking profile for agents $\mathfrak{A} = \mathcal{P} \cup \mathcal{A}$ with $\mathcal{P}$ -optimal stable matching m^* . Then the following partial profile P is a valid profile for F^* :*

- $\forall p \in \mathcal{P}, P_p = \{(a, a') \in \mathcal{A} \times \mathcal{A} \mid a \succ a' \succeq_{F_p^*} m^*(p)\} \cup \{(a, a') \in \mathcal{A} \times \mathcal{A} \mid a \succeq_{F_p^*} m^*(p) \succ_{F_p^*} a'\}$
- $\forall a \in \mathcal{A}, P_a = \{(p, p') \in \mathcal{A} \times \mathcal{A} \mid p' \prec p \preceq_{F_a^*} m^*(a)\} \cup \{(p, p') \in \mathcal{A} \times \mathcal{A} \mid p \preceq_{F_a^*} m^*(a) \prec_{F_a^*} p'\}$

Now any partial profile compatible with profile P as specified in Proposition 5.3 is again a valid profile. In particular, for a partial profile $P' = \{P'_a\}_{a \in \mathfrak{A}}$ that is compatible with full ranking profile F , $P^* = P \cup P' = \{P_a \cup P'_a\}_{a \in \mathfrak{A}}$ is a valid profile. Assume that P' admits a super stable matching m that is not the PSM. Then P^* does not contain:

- $a \succ_{P_p^*} a'$ for $a, a' \in \mathcal{A}$ and $p \in \mathcal{P}$ with $m(p) \succ_{F_p} a, a'$, unless $a \succ_{P_p} a'$:
For optimal stable matching m^* in F , we have $m^*(p) \succ_{F_p} m(p) \succ_{F_p} a, a'$. Thus $(a, a') \notin P_p$.
- $p \prec_{P_a^*} p'$ for $p, p' \in \mathcal{P}$ and $a \in \mathcal{A}$ with $m(a) \prec_{F_a} p, p'$, unless $p \prec_{P_a} p'$:
We have $m^*(a) \prec_{F_a} m(a) \prec_{F_a} p, p'$ and thus $(p, p') \notin P_a$.

As these relations are not known in P^* , and P^* is valid, we can eliminate agents as described above.

Note that just because Proposition 5.3 guarantees that a valid profile (specifically P^*) can still be identified after eliminating agents from active pairs, this does not mean that the valid profiles that are still learnable attain the best-case gap $\Delta_{a,i}^F = \max(\Delta^F, \Delta_{a,i}^{\min})$. We can therefore not transfer the sample complexity bound from Theorem 5.1. However, a super stable matching might be identifiable very fast, leading to early elimination of pairs and hence decreased sample complexity. We explore the sample complexity in our experiments to identify whether this trade-off between early elimination and restriction to valid profiles is beneficial.

6 LOWER BOUND

We now establish a non-explicit lower bound on sample complexity by adapting the analysis of Jourdan et al. [2021].

For an instance $\nu \in \mathcal{I}$, we denote with m_ν^* the optimal stable matching. To provide a lower bounds on the sample complexity, we define the set of alternative environments $\mathcal{I}_{\text{alt}}(\nu)$ as the collection of all environments in $\mathcal{I}$ where the optimal stable matching is different from that in ν :

$$\mathcal{I}_{\text{alt}}(\nu) = \{\nu' \in \mathcal{I} : m_{\nu'}^* \neq m_\nu^*\} \quad (8)$$

Now we can define the transformed simplex $\mathcal{S}_{|\mathcal{M}|} = \{\sum_{m \in \mathcal{M}} I_m w_m : w_m \in \Delta_{\mathcal{M}}\} \subset \mathbb{R}^{\mathcal{P} \times \mathcal{A}}$, where $I_m \subset \{0, 1\}^{\mathcal{P} \times \mathcal{A}}$ represents the binary indicator vector of active

pairs in matching m , and $w_m \in \Delta_{\mathcal{M}}$ denotes the proportion of rounds in which that specific matching is sampled.

Theorem 6.1. *For any δ -PACOS algorithm and any matching instance ν ,*

$$\mathbb{E}_{\nu}[\tau] \geq \ln \frac{1}{2.4\delta} c_{\star}^{-1}(\nu) \quad (9)$$

where $c_{\star}(\nu)$ is the solution of the optimization problem:

$$c_{\star}(\nu) = \sup_{w \in \mathcal{S}_{|\mathcal{M}|}} \inf_{\nu' \in \text{Alt}(\nu)} \sum_{a \in \mathcal{A}} \sum_{i \in \text{OS}(a)} w_{a,i} K L(\nu_{a,i}, \nu'_{a,i}).$$

Although the bound is not explicit, it characterizes the intrinsic difficulty of the problem. In the MAB literature, this type of lower bound typically guides the design of optimal algorithms [Garivier and Kaufmann, 2016, Degenne et al., 2019]. However, because solving the associated optimization problem is computationally hard, standard Track-and-Stop methods [Garivier and Kaufmann, 2016] cannot be directly adapted. Developing efficient approximations, such as adapting the methods of Jourdan et al. [2021], is an interesting direction that remains open.

To provide a more concrete analysis, we establish an explicit lower bound for instances with global preferences in Appendix G, similar to the approach of Kaufmann et al. [2016], using arguments based on minimum valid preference profile.

7 SIMULATIONS

In this section, we present simulations to empirically evaluate the proposed algorithms on random instances for varying market sizes n , fixing $n = N = K$ agents on each side. The code to reproduce the presented results is provided online³.

Settings: We perform two sets of experiments, covering both pure exploration and regret minimization, which we present in the following subsections.

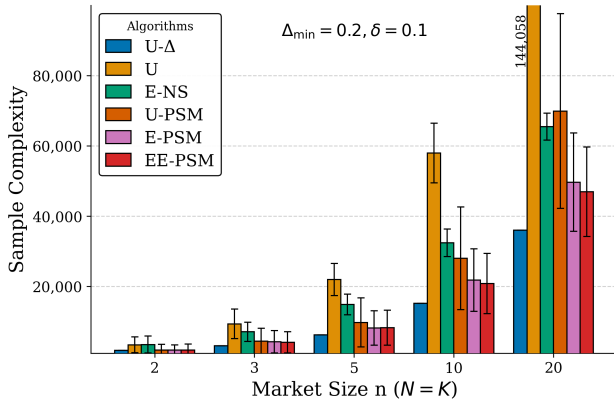


Figure 2: Sample complexity as a function of market size n .

³<https://github.com/a-athanasopoulos/Probably-Correct-Optimal-Stable-Matching>

Instances: To generate instances, we first construct the agents' preferences using random perturbations. We then assign the rewards for each agent $a \in \mathcal{A}$ by sampling the reward gaps $\Delta_{a,i,i+1}$ between consecutive agents from a uniform distribution over $[\Delta_{\min}, 1]$, where we set $\Delta_{\min} = 0.2$.

Algorithms: To isolate the individual effects of the elimination strategy and the PSM-based stopping rule, we compare:

1. **U- Δ :** The uniform sampling strategy of Section 4 with known $\Delta_{\min}$.
2. **U:** Uniform sampling of all pairs, stopping only when all confidence intervals separate (without $\Delta_{\min}$ knowledge).
3. **U-PSM:** Uniform sampling of all pairs, terminating when a pervasive stable matching is identified.
4. **E-NS:** The elimination algorithm without stopping rule, terminating only once all pairs are eliminated.
5. **E-PSM:** The elimination algorithm of Section 5, terminating when a pervasive stable matching is identified.
6. **EE-PSM (Extended PSM):** The extended elimination algorithm that further eliminates pairs after identifying a super stable matching as discussed in Section 5.3.

7.1 PURE EXPLORATION

In Figure 2, we illustrate the sample complexity of the different approaches, averaged over 100 random instances, along with their standard deviations. For these experiments, we set the confidence parameter to $\delta = 0.1$. Notably, all algorithms successfully identified the optimal stable matching in every run; this empirical correctness aligns with standard observations in the MAB literature [Jamieson et al., 2014].

Regarding sample complexity, we observe that the proposed algorithms, **E-PSM** and **EE-PSM**, significantly outperform all other baselines. Interestingly, we observe only a slight performance difference between these two specific algorithms on random instances. More specifically, upon further examining their behavior, we found that although the extended algorithm eliminated a large number of pairs, the size of the minimum matching cover did not decrease significantly, leading to near-identical sample complexities.

Finally, the results demonstrate clear, independent benefits from both the elimination strategy and the early stopping rule. For example, comparing **U-PSM** to **E-PSM** isolates the positive effect of pair elimination, while comparing **E-NS** to **E-PSM** highlights the drastic impact of the stopping rule. Furthermore, **E-NS** outperforms **U-PSM** in $n = 20$, demonstrating that the effect of the elimination rule is relatively larger on average as the market grows. To conclude, the results show that **U-PSM** exhibits larger variability across the random instances, suggesting a high sensitivity to specific reward structures. In contrast, the effect of the elimination rule in **E-NS** is much more stable.

Appendix H presents a similar experiment with different

numbers of agents on each side, further demonstrating the effectiveness of our approach.

7.2 REGRET MINIMIZATION

We further provide an ablation study for regret minimization. We measure regret using the cumulative market instability defined in Section 5.2, Eq. 7, which counts the number of rounds in which the selected matching differs from the optimal stable matching [Das and Kamenica, 2005, Basu, 2025].

To clearly demonstrate the performance of our method against existing benchmarks, we focus on the more extensively studied one-sided uncertainty setting. This allows us to compare with prior regret-minimization algorithms, while our method naturally applies to the two-sided uncertainty setting considered in this paper.

Additional benchmark: We specifically compare our approach with AETGSE, the algorithm of Kong and Li [2024], which combines adaptive elimination with the logic of GS algorithm. In the centralized setting, agents explore their currently valid partners in a coordinated round-robin manner, eliminating partners that are certified to be suboptimal. Once an agent identifies its most preferred valid partner, it stop exploring and temporarily commits to it. If this partner becomes unavailable (due to a higher-priority agent), the agent restarts the exploration process.

Figure 3 presents the results over 50 independent experiments for each algorithm, using random market instances with the number of agents on each side fixed to $K = N = 20$. The results show that our elimination-based approach with a stopping rule based on pervasive stable matching (E-PSM) achieves lower regret in this one-sided uncertainty setting. Moreover, the uniform-sampling approach (U-PSM) interpolates between AETGS-E and E-PSM, suggesting that both the elimination procedure and our stopping rule contribute to improved learning performance.

Beyond the experimental comparison, our theoretical result yields a tighter upper bound than the analysis of AETGS-E, since the relevant gap depends on the minimum valid partial profile (see Section 3). Nevertheless, we highlight that AETGS-E may perform better on specific instances where agents do not need to restart exploration, due to its adaptive sampling approach. However, in more general instances, where exploration restarts, our algorithm achieves better performance, as demonstrated by our simulations.

8 CONCLUSION

In this work, we investigated a sequential learning problem in matching markets where both sides are initially unaware of their true preferences. Focusing on the pure exploration

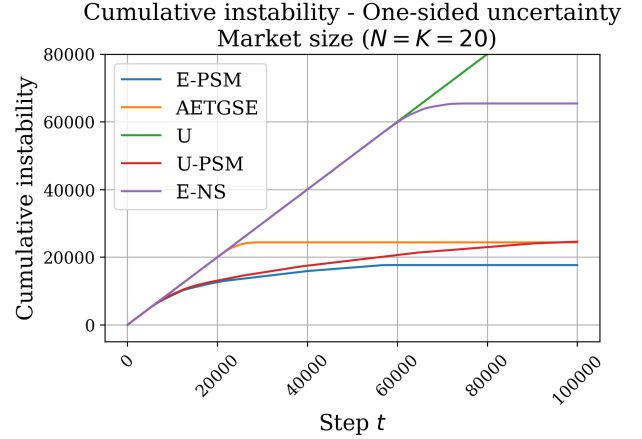


Figure 3: Cumulative regret

setting, our objective was to identify the optimal stable matching with high probability. Our primary contribution lies in leveraging partial knowledge of agents’ preferences to stop exploration once a pervasive stable matching is found. We first demonstrated the performance gains of this approach on uniform exploration strategies, and subsequently introduced practical elimination algorithms along with a refined sample complexity analysis. Further, we extended our insights to the regret minimization setting and established theoretical lower bounds on the sample complexity.

For future work, it is worth proving theoretical sample-complexity guarantees for the extended algorithm from Section 5.3, and investigating whether we can further eliminate pairs after identifying a super stable matching, specifically by removing pairs that are not part of an exposed rotation [Gusfield and Irving, 1989]. In addition, adaptive algorithms, as in [Athanasopoulos et al., 2025, Kong and Li, 2024], that do not uniformly sample active pairs are also a promising approach. Finally, designing optimal algorithms that match the lower bound established in our analysis remains an open problem.

Acknowledgements

This project was partially supported by the Research Council of Norway through its Centre of Excellence: Integreat - The Norwegian Centre for knowledge-driven machine learning, project number 332645.

References

Andreas Athanasopoulos, Anne-Marie George, and Christos Dimitrakakis. Probably correct optimal stable matching for two-sided market under uncertainty. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems, AAMAS ’25*,

- page 152–160, Richland, SC, 2025. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9798400714269.
- Soumya Basu. Competing bandits in matching markets via super stability. In *Forty-second International Conference on Machine Learning*, 2025. URL <https://openreview.net/forum?id=mNxjoMiFFf>.
- Soumya Basu, Karthik Abinav Sankararaman, and Abishek Sankararaman. Beyond $\log^2(t)$ regret for decentralized bandits in matching markets. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 705–715, Virtual, 18–24 Jul 2021. PMLR. URL <https://proceedings.mlr.press/v139/basu21a.html>.
- Ioannis Caragiannis, Aris Filos-Ratsikas, Panagiotis Kanellopoulos, and Rohit Vaish. Stable fractional matchings. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 21–39, 2019.
- Sarah H. Cen and Devavrat Shah. Regret, stability & fairness in matching markets with bandit learners. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 8938–8968, Virtual, 28–30 Mar 2022. PMLR. URL <https://proceedings.mlr.press/v151/cen22a.html>.
- Sanmay Das and Emir Kamenica. Two-sided bandits and the dating market. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence, IJCAI’05*, page 947–952, San Francisco, CA, USA, 2005. Morgan Kaufmann Publishers Inc.
- Rémy Degenne, Wouter M. Koolen, and Pierre Ménard. Non-asymptotic pure exploration by solving games. In *Neural Information Processing Systems*, 2019. URL <https://api.semanticscholar.org/CorpusID:195584182>.
- Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Pac bounds for multi-armed bandit and markov decision processes. In *Proceedings of the 15th Annual Conference on Computational Learning Theory, COLT ’02*, page 255–270, Berlin, Heidelberg, 2002. Springer-Verlag. ISBN 354043836X.
- Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *J. Mach. Learn. Res.*, 7:1079–1105, dec 2006. ISSN 1532-4435.
- Harold N. Gabow and Oded Kariv. Algorithms for edge coloring bipartite graphs. In *Proceedings of the Tenth Annual ACM Symposium on Theory of Computing, STOC ’78*, page 184–192, New York, NY, USA, 1978. Association for Computing Machinery. ISBN 9781450374378. doi: 10.1145/800133.804346. URL <https://doi.org/10.1145/800133.804346>.
- D. Gale and L. S. Shapley. College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15, 1962. ISSN 00029890, 19300972. URL <http://www.jstor.org/stable/2312726>.
- Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027. PMLR, 2016.
- Dan Gusfield and Robert W Irving. *The stable marriage problem: structure and algorithms*. MIT press, 1989.
- Hadi Hosseini, Sanjukta Roy, and Duohan Zhang. Putting gale & shapley to work: Guaranteeing stability through learning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 69043–69068. Curran Associates, Inc., 2024.
- Robert W. Irving. Stable marriage and indifference. *Discrete Applied Mathematics*, 48(3):261–272, 1994. ISSN 0166-218X. doi: [https://doi.org/10.1016/0166-218X\(92\)00179-P](https://doi.org/10.1016/0166-218X(92)00179-P). URL <https://www.sciencedirect.com/science/article/pii/0166218X9200179P>.
- Meena Jagadeesan, Alexander Wei, Yixin Wang, Michael I. Jordan, and Jacob Steinhardt. Learning equilibria in matching markets from bandit feedback. *Advances in Neural Information Processing Systems*, 34:3323–3335, 2021a.
- Meena Jagadeesan, Alexander Wei, Yixin Wang, Michael I. Jordan, and Jacob Steinhardt. Learning equilibria in matching markets from bandit feedback. *CoRR*, abs/2108.08843, 2021b. URL <https://arxiv.org/abs/2108.08843>.
- Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil’ ucb : An optimal exploration algorithm for multi-armed bandits. In Maria Florina Balcan, Vitaly Feldman, and Csaba Szepesvári, editors, *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 423–439, Barcelona, Spain, 13–15 Jun 2014. PMLR. URL <https://proceedings.mlr.press/v35/jamieson14.html>.

- Marc Jourdan, Mojmír Mutný, Johannes Kirschner, and Andreas Krause. Efficient pure exploration for combinatorial bandits with semi-bandit feedback. In *Algorithmic Learning Theory*, pages 805–849. PMLR, 2021.
- Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42, 2016.
- Fang Kong and Shuai Li. Player-optimal stable regret for bandit learning in matching markets. In *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1512–1522. SIAM, 2023.
- Fang Kong and Shuai Li. Improved bandits in many-to-one matching markets with incentive compatibility. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13256–13264, 2024.
- Fang Kong, Junming Yin, and Shuai Li. Thompson sampling for bandit learning in matching markets. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 3164–3170, Vienna, Austria, 7 2022. International Joint Conferences on Artificial Intelligence Organization. doi: 10.24963/ijcai.2022/439. URL <https://doi.org/10.24963/ijcai.2022/439>. Main Track.
- Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, Cambridge, UK, 2020.
- Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.
- Lydia T. Liu, Horia Mania, and Michael Jordan. Competing bandits in matching markets. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1618–1628. PMLR, 26–28 Aug 2020. URL <https://proceedings.mlr.press/v108/liu20c.html>.
- 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.
- David F Manlove. *Algorithmics of Matching Under Preferences*. WORLD SCIENTIFIC, Singapore, 2013. doi: 10.1142/8591.
- Shie Mannor and John N Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5(Jun):623–648, 2004.
- Tejas Pagare and Avishek Ghosh. Explore-then-commit algorithms for decentralized two-sided matching markets. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pages 2092–2097. IEEE, 2024.
- Sasha Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. *Advances in Neural Information Processing Systems*, 26, 2013.
- Baharak Rastegari, Anne Condon, Nicole Immorlica, and Kevin Leyton-Brown. Two-sided matching with partial information. In *Proceedings of the fourteenth ACM conference on Electronic Commerce*, pages 733–750, 2013.
- Baharak Rastegari, Anne Condon, Nicole Immorlica, Robert Irving, and Kevin Leyton-Brown. Reasoning about optimal stable matchings under partial information. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 431–448, 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.
- Alvin E. Roth and Marilda A. Oliveira Sotomayor. *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*. Econometric Society Monographs. Cambridge University Press, 1990.
- Peash Ranjan Saha, Salimur Choudhury, and Kai Salomaa. Altruistic bandit learning for one-to-many matching markets. In *Proceedings of the 2024 International Conference on Information Technology for Social Good*, pages 197–203, 2024.
- Abishek Sankararaman, Soumya Basu, and Karthik Abinav Sankararaman. Dominate or delete: Decentralized competing bandits in serial dictatorship. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1252–1260, Virtual, 13–15 Apr 2021. PMLR. URL <https://proceedings.mlr.press/v130/sankararaman21a.html>.
- Boris Spieker. The set of super-stable marriages forms a distributive lattice. *Discrete applied mathematics*, 58(1): 79–84, 1995.
- William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.
- Zilong Wang, Liya Guo, Junming Yin, and Shuai Li. Bandit learning in many-to-one matching markets. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 2088–2097, 2022.
- Yirui Zhang and Zhixuan Fang. Decentralized two-sided bandit learning in matching market. In Negar Kiyavash and Joris M. Mooij, editors, *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pages 4173–4191. PMLR, 2024.

Probably Correct Optimal Stable Matching under Two-Sided Uncertainty (Supplementary Material)

Andreas Athanasopoulos¹

Anne-Marie George²

Christos Dimitrakakis¹

¹University of Neuchâtel, Neuchâtel, Switzerland

²University of Oslo, Oslo, Norway

APPENDIX OUTLINE

In this appendix, we provide the detailed proofs and discussion omitted from the main text.

- **Appendix A:** Provides an extended related work section along with a discussion on different learning objectives.
- **Appendix B:** Contains the analysis of Theorems 4.1 and 4.2 from Section 4.
- **Appendix C:** Contains the proof of Theorem 5.1 from Section 5.1.
- **Appendix D:** Contains the proof of Theorem 5.2 from Section 5.2.
- **Appendix E:** Contains the proof of Proposition 5.3 from Section 5.3.
- **Appendix F:** Contains the proof of Theorem 6.1 from Section 6.
- **Appendix G:** Contains the analysis of the lower bound for specific instance and the proof of Lemma G.1.

A EXTENDED RELATED WORK

This section provides additional discussion on related work, expanding upon the overview presented in the main text.

MAB and pure exploration: The stochastic MAB problem, introduced by Thompson [1933], has been extensively studied in the past few decades, becoming a fundamental tool for analyzing online decision-making problems under uncertainty. A learner selects one action (or arm) in each round and receives a numeric reward drawn from an unknown distribution that reflects the utility of the action. In the classic version of regret minimization, the learner aims to maximize cumulative rewards by balancing exploration and exploitation: searching for optimal actions versus pulling known high-reward arms. Another popular setting is the pure exploration problem, where the learner’s objective is to identify a good action regardless of the cost incurred in the learning process. This problem is usually studied in the fixed-confidence setting, where the learner aims to identify an optimal action with high probability using as few samples as possible [Even-Dar et al., 2002]. Initial research in the field explored uniform exploration and elimination-based strategies Mannor and Tsitsiklis [2004], Even-Dar et al. [2006], focusing on their correctness and sample complexity. More recent studies Kaufmann et al. [2016], Garivier and Kaufmann [2016] establish a lower bound on sample complexity, motivating the Track-and-Stop algorithm which achieves asymptotic optimality. In this paper, we study identifying the optimal stable matching in the fixed-confidence setting. For a comprehensive overview of MAB, we refer to the book by Lattimore and Szepesvári [2020].

Stable matching problem: Two-sided matching markets model situations in which two distinct sets of agents aim to match with each other based on individual preferences [Gusfield and Irving, 1989, Roth and Sotomayor, 1990, Manlove, 2013]. Classic examples include students applying to universities, workers matching with firms, and organ donors paired with patients [Roth, 1984].

In their fundamental work, Gale and Shapley [1962] introduced the concept of stable matching, a notion of equilibrium in which no two agents have an incentive to deviate. Gale and Shapley [1962] proved that the set of stable matchings $\mathcal{S}$

is always non-empty by introducing the Gale–Shapley (GS) algorithm, which is based on sequential proposals from one side of the market to the other. The algorithm returns the (unique) *optimal stable matching* m^* for the proposing side, in the sense that every proposer receives the best partner they can obtain in any stable matching. Conversely, the resulting matching is pessimal for the side receiving the proposals, providing each agent on that side with their least-preferred partner among all stable matchings.

A core component of our approach is utilizing the agents’ *partial preference* rankings to identify an optimal stable matching, thereby circumventing the need to learn their full preference orderings. This approach is motivated by the concept of *pervasive stable matching*, introduced by Rastegari et al. [2014] to reason about optimal stable matchings under partial information in *non-learning* scenarios. Specifically, a partial preference profile of the agents admits a pervasive stable matching if that matching remains optimally stable under every possible full refinement of the given preferences. A closely related notion is that of a *super stable matching*, initially introduced to address stability in instances with indifferent preferences [Irving, 1994].

Leaning in matching markets: The problem of two-sided markets with unknown preferences has recently attracted considerable attention in various settings. Das and Kamenica were the first to introduce the marriage problem with unknown preferences on both sides of the market, framing it as a MAB problem. In their work Das and Kamenica [2005], they empirically studied several algorithms in specific preference settings. A follow-up by Liu et al. [2020] introduced the notions of *Player-Optimal* and *Player-Pessimal* stable regret. Focusing on a setting where only one side of the market has unknown preferences, they analyzed an Explore-Then-Commit algorithm, proving sublinear bounds for both regret notions. They also examined an Upper Confidence Bound approach, establishing sublinear bounds for Player-Pessimal stable regret; however, they proved an impossibility result for bounding Player-Optimal regret.

Later studies closed this gap by achieving sublinear player-optimal stable regret [Sankararaman et al., 2021, Basu et al., 2021, Kong and Li, 2023], though they primarily focused on one-sided uncertainty and specific market structures. More recent studies have begun to consider settings where both sides of the market have uncertain preferences [Pagare and Ghosh, 2024, Zhang and Fang, 2024]. However, these works employ Explore-Then-Commit type algorithms. In contrast, this work adopts a pure exploration framework that is more naturally suited to the analysis of such algorithms and enables refined results on regret with respect to the *optimal* stable matching.

The pure exploration objective of identifying an *optimal stable matching* with high probability was introduced by Athanassopoulos et al. [2025], while the work of Hosseini et al. [2024] focuses on identifying *any* stable matching. Both of these studies propose elimination-based algorithms restricted to one-sided uncertainty. We extend this literature by tackling a two-sided learning environment, providing refined arguments for the analysis of elimination algorithms.

Conceptually, the closest work to ours is Basu [2025], where the authors leverage partial preferences and the concept of super stability to bound regret in sequential learning scenarios. In contrast, our approach specifically relies on pervasive stable matchings, instead of the super stable matching. This distinction is crucial, as our objective is to identify the *optimal* stable matching, rather than *any* stable matching.

Prior work also considers decentralized settings [Liu et al., 2021, Kong et al., 2022, Sankararaman et al., 2021, Basu et al., 2021, Pagare and Ghosh, 2024, Zhang and Fang, 2024], models with transferable-utilities [Cen and Shah, 2022, Jagadeesan et al., 2021a, Basu, 2025], and the setting with many-to-one matchings [Saha et al., 2024, Wang et al., 2022, Kong and Li, 2024].

Discussion on alternative objectives: Related work typically characterizes algorithmic performance based on the cumulative rewards collected by individual agents. In particular, most existing approaches follow the notion of agent-optimal (or pessimal) stable regret, which compares accumulated rewards against a baseline that always selects the optimal stable matching, following the definitions in the seminal work of Liu et al. [2020]. This cumulative metric reflects stability only if the regret is sublinear for every agent in the market. Furthermore, this objective can be satisfied by a composition of unstable matchings over the time horizon—similar to the concept of fractional stability over probabilistic matchings that may individually be unstable [Caragiannis et al., 2019]. Another notion of regret, more compatible with our objective, is binary stable regret, which evaluates the stability of the matching at each discrete time step [Das and Kamenica, 2005, Basu, 2025]. Current studies in this area typically consider a matching to be stable without specifically requiring the identification of the optimal stable matching. Additionally, algorithms addressing these different objectives often rely on the Explore-then-Commit (ETC) framework, bounding the probability of error through sufficient exploration [Liu et al., 2020, Pagare and Ghosh, 2024, Basu, 2025, Kong and Li, 2023]. Notable exceptions to these ETC-based approaches include analyses of UCB-style algorithms [Liu et al., 2020, Jagadeesan et al., 2021b] and Thompson sampling [Kong et al., 2022],

which do not guarantee sublinear regret with respect to the optimal stable matching. Ultimately, different objectives prioritize different criteria; we emphasize that our framework is specifically designed for applications where the primary goal is to learn the optimal stable matching itself, rather than merely achieving stability in a cumulative or general sense.

B ANALYSIS OF SECTION 4

In this section, we provide the proofs for the sample complexity of our uniform exploration strategies.

B.1 PROOF OF THEOREM 4.1

Proof. We begin the proof by characterizing the sample complexity of the algorithm. We recall that the sample complexity is defined as the total number of matchings performed by the algorithm until termination.

Sample Complexity

The sample complexity follows directly from the design of Algorithm 1. In each exploration round h , the algorithm ensures that every available pair $(p, a) \in \mathcal{P} \times \mathcal{A}$ is sampled exactly once, by sampling a sequence of $n = \max\{K, N\}$ matchings. This is achieved by sampling the matching from a minimal matching cover $\mathfrak{M}(\mathcal{P} \times \mathcal{A})$. By König's Theorem, the size of this cover is equal to the maximum degree of the underlying complete bipartite graph, which is $\max\{K, N\}$.

Given that the total number of exploration rounds is h , the total number of matchings sampled is:

$$\tau = h \cdot \max\{K, N\} = \left\lceil \frac{8 \ln(2KN/\delta)}{\Delta_{\min}^2} \right\rceil \cdot \max\{K, N\}. \quad (10)$$

Thus, the sample complexity is $O\left(\max\{K, N\} \frac{\ln(KN/\delta)}{\Delta_{\min}^2}\right)$.

Correctness

We now prove that the algorithm returns the optimal stable matching with probability at least $1-\delta$, i.e., $\Pr(m_\tau = m^*) \geq 1-\delta$. If all agents' preference orderings are estimated correctly, then running GS on the estimated profile returns the true optimal stable matching; therefore,

$$\Pr(m_\tau = m^*) \geq \Pr(\hat{F}_a = F_a \ \forall a \in \mathfrak{A}) = 1 - \Pr(\exists a \in \mathfrak{A} : \hat{F}_a \neq F_a).$$

Thus, it suffices to show $\Pr(\exists a \in \mathfrak{A} : \hat{F}_a \neq F_a) \leq \delta$.

Now fix an agent $a \in \mathfrak{A}$. A sufficient condition for every empirical ranking $\hat{F}_a$ to match F_a , is that for every agent $i \in OS(a)$, $|\hat{\mu}_{a,i} - \mu_{a,i}| < \frac{\Delta_{\min}}{2}$ holds. Thus for every pair (i, j) where $\mu_{a,i} > \mu_{a,j}$ we have that:

$$\hat{\mu}_{a,i} > \mu_{a,i} - \frac{\Delta_{\min}}{2} \geq \mu_{a,j} + \frac{\Delta_{\min}}{2} > \hat{\mu}_{a,j}, \quad (11)$$

ensuring that $\hat{\mu}_{a,i} > \hat{\mu}_{a,j}$ using the definition of $\Delta_{\min}$.

By the construction of Algorithm 1, each pair $(a, i) \in \mathfrak{A} \times OS(a)$ is sampled exactly h times. Since the rewards are 1-sub-Gaussian, the empirical mean $\hat{\mu}_{a,i}$ is $(1/h)$ -sub-Gaussian. Applying the standard sub-Gaussian tail bound:

$$\Pr(\hat{F}_a \neq F_a) = \Pr\left(|\hat{\mu}_{a,i} - \mu_{a,i}| \geq \frac{\Delta_{\min}}{2}\right) \leq 2 \exp\left(-\frac{h(\Delta_{\min}/2)^2}{2}\right) = 2 \exp\left(-\frac{h\Delta_{\min}^2}{8}\right). \quad (12)$$

By applying a union bound over all KN pairs in the bipartite market:

$$\mathbb{P}\left(\exists a \in \mathfrak{A} : \hat{F}_a \neq F_a\right) \leq \sum_{a \in \mathfrak{A}} \mathbb{P}\left(\hat{F}_a \neq F_a\right) \leq 2KN \exp\left(-\frac{h\Delta_{\min}^2}{8}\right). \quad (13)$$

Now setting $h = \left\lceil \frac{8 \ln(2KN/\delta)}{\Delta_{\min}^2} \right\rceil$, results that $\mathbb{P}\left(\exists a \in \mathfrak{A} : \hat{F}_a \neq F_a\right) \leq \delta$ and thus:

$$\Pr(m_\tau = m^*) < 1 - \delta \quad (14)$$

That concludes the proof. $\square$

B.2 PROOF OF THEOREM 4.2

Proof. Now we continue with a similar analysis for the second version of the algorithm, using arguments regarding the preference profiles $\mathcal{P}_{\text{valid}}(F_{\mathfrak{A}}^\nu)$.

Sample Complexity

The sample complexity follows again directly from the definition of the algorithm, where in each exploration round h we sample $\max\{K, N\}$ number of matchings. Setting $h = \lceil \frac{8 \ln(2KN/\delta)}{\Delta_F^2} \rceil$ results in a sample complexity of $O\left(\max\{K, N\} \frac{\ln(KN/\delta)}{\Delta_F^2}\right)$.

Correctness

We now prove that the algorithm returns the optimal stable matching with probability at least $1 - \delta$. Instead of learning the full preferences F_A for the agents, the algorithm samples enough to learn a valid preference profile $P_A \in \mathcal{P}_{\text{valid}}(F_A)$. By definition, a valid preference profile admits a *pervasive stable matching*, and thus, running the $GS(\hat{F}_A)$ algorithm with any realization of the preferences P_A , i.e., $\hat{F}_A \in \mathcal{F}(P_A) = \{F_A \mid P_A \mapsto F_A\}$ results in the true optimal stable matching m^* . More precisely, we have that:

$$\Pr(m_\tau = m^*) \geq \Pr(\exists P_A \in \mathcal{P}_{\text{valid}} : \hat{F}_A \in \mathcal{F}(P_A)) \geq \Pr(\hat{F}_A \in \mathcal{F}(P^*)). \quad (15)$$

where $P_A^* = \arg \max_{P \in \mathcal{P}_{\text{valid}}(F_A)} \Delta_{\min}(P)$ be the valid preference profile that attains the admissible gap $\Delta_{\min}(P_A^*) = \Delta_F$.

A sufficient condition for the empirical profile $\hat{F}_A$ to refine P_A^* is that for every agent $a \in \mathfrak{A}$ and every $i \in OS(a)$, the condition $|\hat{\mu}_{a,i} - \mu_{a,i}| < \frac{\Delta_F}{2}$ holds. Thus, for every agent $a \in \mathfrak{A}$ and every pair $(i, j) \in P_a^*$ where $\mu_{a,i} \geq \mu_{a,j}$, we have that:

$$\hat{\mu}_{a,i} > \mu_{a,i} - \frac{\Delta_F}{2} \geq \mu_{a,j} + \frac{\Delta_F}{2} > \hat{\mu}_{a,j},$$

ensuring that the relative ordering in P_A^* is correctly estimated.

By the construction of Algorithm 1, each pair $(a, i) \in \mathfrak{A} \times OS(a)$ is sampled exactly h times. Since the rewards are 1-sub-Gaussian, the empirical mean $\hat{\mu}_{a,i}$ is $(1/h)$ -sub-Gaussian. Applying the standard sub-Gaussian tail bound:

$$\Pr\left(|\hat{\mu}_{a,i} - \mu_{a,i}| \geq \frac{\Delta_F}{2}\right) \leq 2 \exp\left(-\frac{h(\Delta_F/2)^2}{2}\right) = 2 \exp\left(-\frac{h\Delta_F^2}{8}\right)$$

By applying a union bound over all KN pairs in the bipartite market:

$$\mathbb{P}\left(\exists a \in \mathfrak{A}, i \in OS(a) : |\hat{\mu}_{a,i} - \mu_{a,i}| \geq \frac{\Delta_F}{2}\right) \leq \sum_{a \in \mathfrak{A}} \sum_{i \in OS(a)} \mathbb{P}\left(|\hat{\mu}_{a,i} - \mu_{a,i}| \geq \frac{\Delta_F}{2}\right) \leq 2KN \exp\left(-\frac{h\Delta_F^2}{8}\right)$$

Now setting $h = \lceil \frac{8 \ln(2KN/\delta)}{\Delta_F^2} \rceil$ ensures that the probability of failing to correctly estimate the valid preference profile is bounded by δ , and thus:

$$\Pr(m_\tau = m^*) \geq 1 - \delta$$

□

C ANALYSIS OF SECTION 5

C.1 PROOF OF THEOREM 5.1

Proof.

Correctness argument

We first present a formal argument that the elimination algorithm identifies the optimal stable matching with probability at least $1 - \delta$.

To begin, the algorithm is guaranteed to terminate because the confidence radius $c_{a,i}(t)$ shrinks toward zero as the number of samples t increases. Since the algorithm continues to sample every non-eliminated pair in each round, the confidence intervals will eventually shrink enough to resolve the relative ordering of every pair. Consequently, the algorithm will eventually eliminate all pairs satisfying the stopping condition.

Now, let $\mathcal{E}$ be the event that the true means $\mu_{a,i}$ lie within the confidence intervals for all estimates $\hat{\mu}_{a,i}(t)$ at all times t . Formally, $\mathcal{E}$ is the intersection of all individual events $\mathcal{E}_{t,a,i}$:

$$\mathcal{E} = \bigcap_{t=1}^{\infty} \bigcap_{a \in \mathfrak{A}} \bigcap_{i \in OS(a)} \mathcal{E}_{t,a,i}, \quad (16)$$

where $\mathcal{E}_{t,a,i}$ denotes the event that the true mean $\mu_{a,i}$ lies within the confidence interval at time t :

$$\mathcal{E}_{t,a,i} = \{|\hat{\mu}_{a,i}(t) - \mu_{a,i}| \leq c_{a,i}(t)\}. \quad (17)$$

In Lemma C.1 we prove that the event $\mathcal{E}$ holds with probability at least $1 - \delta$, i.e. $\Pr(\mathcal{E}) \geq 1 - \delta$.

Under event $\mathcal{E}$, the algorithm only adds correct preference relations to P_a because $LCB_{a,i}(t) > UCB_{a,j}(t)$ implies $\mu_{a,i} > \mu_{a,j}$. Since the algorithm terminates with a valid preference profile $\hat{P}_A \in \mathcal{P}_{\text{valid}}(F_A)$, it returns the correct optimal stable matching m^* .

To conclude, under the good event $\mathcal{E}$, which holds with probability at least $1 - \delta$, the algorithm returns the optimal stable matching.

C.2 SAMPLE COMPLEXITY

We now continue the analysis on sample complexity in terms of matching samples. Recall that in each round t , we sample every non-eliminated pair S_t by sampling n_t matchings from a minimum matching cover $\mathfrak{M}(S_t)$.

Let $\tau_{a,i}$ denote the number of rounds required to eliminate agent $i \in OS(a)$ from the active set S_a of agent $a \in \mathfrak{A}$. To eliminate agent i , we must successfully distinguish it from every other agent $j \in OS(a) \setminus \{i\}$. Therefore, the elimination time is dominated by the time required to separate i from its closest competitor, which corresponds to the minimum reward gap $\Delta_{a,i} = \min_{j \in OS(a) \setminus \{i\}} |\mu_{a,i} - \mu_{a,j}|$. By applying Lemma C.2 we have that:

$$\tau_{a,i} \leq \frac{32 \log(8KN/(\delta\Delta_{a,i}))}{\Delta_{a,i}^2}. \quad (18)$$

Since eliminating a pair $(p, a) \in \mathcal{P} \times \mathcal{A}$ requires removing p from S_a and a from S_p , the maximal number of rounds is $t_{p,a} = \max\{\tau_{p,a}, \tau_{a,p}\} \leq \frac{32 \log(8KN/(\delta\Delta_{p,a}^{\min}))}{(\Delta_{p,a}^{\min})^2}$, where $\Delta_{p,a}^{\min} = \min\{\Delta_{p,a}, \Delta_{a,p}\}$.

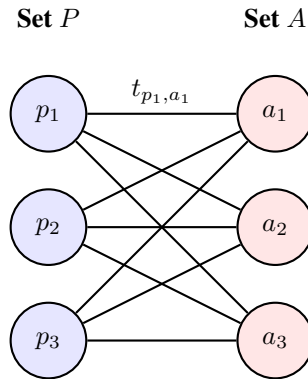


Figure 4: Illustration of the bipartite graph G .

A. Sample complexity Without Stopping Rule

Next, we give the sample complexity ignoring the stopping rule where the algorithm terminates if the partial preferences admit a pervasive stable matching.

Given the sampling requirements $t_{p,a}$ for each pair of agents $(p, a) \in \mathcal{P} \times \mathcal{A}$ (where $t_{p,a} = t_{a,p}$), we construct a bipartite multigraph $G_M = (V = \mathcal{P} \cup \mathcal{A}, E = \mathcal{P} \times \mathcal{A})$ illustrated in Figure 4, where $t_{p,a}$ denotes the edge multiplicity for the pair (p, a) . By Kőnig's Theorem, there exists a decomposition of these edges into $\deg(G_M)$ matchings, where $\deg(G_M)$ is the maximum degree of the multigraph. Consequently, the number of matchings is at most:

$$O\left(\max_{a \in \mathfrak{A}} \sum_{i \in OS(a)} t_{a,i}\right) \quad (19)$$

since $\sum_{i \in OS(a)} t_{a,i}$ denotes the degree of agent a in the multigraph.

As a note, we use this variant of the algorithm to demonstrate the isolated effect of the elimination rule in the experiments section.

B. Sample complexity With Stopping Rule

We now perform the analysis considering the stopping rule. The algorithm terminates once it correctly estimates a partial preference profile $\hat{P}_A$ that admits a pervasive stable matching, i.e., $\hat{P}_A \in \mathcal{P}_{\text{valid}}(F_A)$. By definition, there exists at least one valid preference profile $P_A^F \in \mathcal{P}_{\text{valid}}(F_A)$ that admits the maximum possible minimum gap Δ_F .

Since the algorithm uniformly samples every non-eliminated pair in each round, it is guaranteed to correctly estimate all preference relations in this optimal target profile P_A^F once all reward gaps of size at least Δ_F are resolved. By applying Lemma C.2, this requires at most τ^F rounds, where

$$\tau^F \leq \frac{32 \log(8KN/(\delta\Delta_F))}{\Delta_F^2}. \quad (20)$$

Consequently, any pair (a, i) is sampled only until it is individually eliminated from the active sets, or until the algorithm successfully identifies P_A^F and terminates globally. Thus, the number of times pair (a, i) is sampled is strictly bounded by $t_{a,i}^F = \min(t_{a,i}, \tau^F)$. Using the individual elimination bound for $t_{a,i}$ (Eq. 18) and the global termination bound τ^F , we have:

$$t_{a,i}^F \leq \frac{32 \log(8KN/(\delta\Delta_{a,i}^F))}{(\Delta_{a,i}^F)^2}, \quad (21)$$

where $\Delta_{a,i}^F = \max(\Delta_F, \Delta_{a,i}^{\min})$.

Following the logic of the previous analysis, we consider the bipartite multigraph G_M with edge multiplicities $t_{a,i}^F$. By Kőnig's Theorem, there exists a decomposition of these edges into $\deg(G_M)$ matchings, where $\deg(G_M)$ is the maximum degree. Therefore, the total number of matching samples is at most:

$$O\left(\max_{a \in \mathfrak{A}} \sum_{i \in OS(a)} t_{a,i}^F\right) \quad (22)$$

To conclude, we prove that with probability at least $1 - \delta$, the algorithm terminates with the correct optimal stable matching, and that with probability at least $1 - \delta$, the sample complexity is at most:

$$O\left(\max_{a \in \mathfrak{A}} \sum_{i \in OS(a)} \frac{32 \log(8KN/(\delta\Delta_{a,i}^F))}{(\Delta_{a,i}^F)^2}\right) \quad (23)$$

where $\Delta_{a,i}^F = \max(\Delta_F, \Delta_{a,i}^{\min})$. □

C.3 USEFUL LEMMAS

Lemma C.1. Let $\mathcal{E}$ be the event that for all time steps $t \geq 1$, all agents $a \in \mathfrak{A}$, and all agents $i \in OS(a)$, the empirical mean estimates lie within the confidence intervals defined by $c_{a,i}(t) = \sqrt{\frac{2 \ln(8KNt^2/\delta)}{t}}$. Then, the probability of this event holds satisfies:

$$\Pr(\mathcal{E}) \geq 1 - \delta. \quad (24)$$

Proof. We bound the probability of the bad event $\mathcal{E}^c$, defined as the event where there exists a time step t , an agent a , and a candidate i such that the empirical mean falls outside the confidence interval: $|\hat{\mu}_{a,i}(t) - \mu_{a,i}| \geq c_{a,i}(t)$.

Using the union bound over all time steps $t \geq 1$, all agents $a \in \mathfrak{A}$, and all candidates $i \in OS(a)$, we have:

$$\Pr(\mathcal{E}^c) \leq \sum_{t=1}^{\infty} \sum_{a \in \mathfrak{A}} \sum_{i \in OS(a)} \Pr(\mathcal{E}_{t,a,i}^c). \quad (25)$$

Since the rewards are 1-sub-Gaussian and we use the confidence radius $c_{a,i}(t) = \sqrt{\frac{2 \ln(8KNt^2/\delta)}{t}}$, applying the concentration bound for sub-Gaussian random variables yields (Corollary 5.5. [Lattimore and Szepesvári, 2020]):

$$\Pr(\mathcal{E}_{t,a,i}^c) = \Pr(|\hat{\mu}_{a,i}(t) - \mu_{a,i}| \geq c_{a,i}(t)) \quad (26)$$

$$\leq 2 \exp\left(\frac{-tc_{a,i}(t)^2}{2}\right) \quad (27)$$

$$= \frac{\delta}{4KNt^2}. \quad (28)$$

Substituting this back into the union bound sum:

$$\Pr(\mathcal{E}^c) \leq \sum_{t=1}^{\infty} \sum_{a \in \mathfrak{A}} \sum_{i \in OS(a)} \frac{\delta}{4KNt^2} \leq \delta \sum_{t=1}^{\infty} \frac{1}{2t^2} \leq \delta. \quad (29)$$

Finally, $\Pr(\mathcal{E}) = 1 - \Pr(\mathcal{E}^c) = 1 - \delta$ which concludes the proof. $\square$

Lemma C.2. Fix an agent $a \in \mathfrak{A}$ and two distinct agents $i, j \in OS(a)$. Let $\Delta_{a,i,j} = |\mu_{a,i} - \mu_{a,j}|$ be the true reward gap between i and j . Under the event $\mathcal{E}$, the number of rounds $\tau_{a,i,j}$ required to separate agent i from agent j is at most:

$$\tau_{a,i,j} \leq \frac{32 \log(8KN/(\delta \Delta_{a,i,j}))}{\Delta_{a,i,j}^2}. \quad (30)$$

Proof. Assume, without loss of generality, that $\mu_{a,i} > \mu_{a,j}$. The algorithm successfully separates agent i from agent j at round t if their confidence intervals become strictly disjoint, which occurs when $LCB_{a,i}(t) > UCB_{a,j}(t)$.

Under the good event $\mathcal{E}$, the empirical means are bounded by their confidence radii $c(t)$, meaning $\hat{\mu}_{a,i}(t) \geq \mu_{a,i} - c(t)$ and $\hat{\mu}_{a,j}(t) \leq \mu_{a,j} + c(t)$. Substituting these into the lower and upper bounds, we get:

$$LCB_{a,i}(t) = \hat{\mu}_{a,i}(t) - c(t) \geq \mu_{a,i} - 2c(t) \quad (31)$$

$$UCB_{a,j}(t) = \hat{\mu}_{a,j}(t) + c(t) \leq \mu_{a,j} + 2c(t) \quad (32)$$

Therefore, a sufficient condition to guarantee $LCB_{a,i}(t) > UCB_{a,j}(t)$ is:

$$\mu_{a,i} - 2c(t) > \mu_{a,j} + 2c(t) \implies \Delta_{a,i,j} > 4c(t). \quad (33)$$

Substituting the confidence radius $c(t) = \sqrt{\frac{2 \ln(8KNt^2/\delta)}{t}}$ into our sufficient condition yields:

$$\Delta_{a,i,j} > 4\sqrt{\frac{2 \ln(8KNt^2/\delta)}{t}} = \sqrt{\frac{32 \ln(8KNt^2/\delta)}{t}}. \quad (34)$$

Squaring both sides and solving for t , we find that the intervals are guaranteed to separate for any round t satisfying:

$$t > \frac{32}{\Delta_{a,i,j}^2} \log \left(\frac{8KNt^2}{\delta} \right). \quad (35)$$

Using standard algebraic bounds for isolating t in action-elimination algorithms, this inequality is satisfied for a number of rounds at most:

$$\tau_{a,i,j} \leq \frac{32 \log(8KN/(\delta \Delta_{a,i,j}))}{\Delta_{a,i,j}^2}. \quad (36)$$

This concludes the proof. $\square$

D PROOF OF THEOREM 5.2

Proof. We consider the meta-algorithm that runs the E-PSM algorithm during the exploration phase. If the algorithm terminates at round $\tau < T$, it commits to the identified matching m_τ for the remainder of the horizon T .

Let $\mathcal{E}$ denote the good event where the algorithm outputs the correct optimal stable matching, i.e., $\mathcal{E} = \{m_\tau = m^*\}$. By Theorem 5.1, the E-PSM algorithm is δ -PCOS. Thus, the probability of success is bounded by $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$, and the probability of failure is bounded by $\mathbb{P}(\mathcal{E}^c) \leq \delta$.

Using the law of total expectation, we can decompose the expected regret as follows:

$$\mathbb{E}[R(T)] = \mathbb{E}[R(T) \mid \mathcal{E}] \mathbb{P}(\mathcal{E}) + \mathbb{E}[R(T) \mid \mathcal{E}^c] \mathbb{P}(\mathcal{E}^c) \quad (37)$$

$$\leq \mathbb{E}[R(T) \mid \mathcal{E}] \cdot 1 + T \cdot \delta, \quad (38)$$

where we used the trivial bound $\mathbb{P}(\mathcal{E}) \leq 1$.

Conditioned on the good event $\mathcal{E}$, the algorithm correctly identifies m^* and incurs zero regret during the commitment phase. Since there is no a guarantee that the required exploration time $\tau \leq T$, we have:

$$\mathbb{E}[R(T) \mid \mathcal{E}] = \mathbb{E}[\min(\tau, T) \mid \mathcal{E}]. \quad (39)$$

Setting our confidence parameter to $\delta = 1/T$ yields to:

$$\mathbb{E}[R(T)] \leq \mathbb{E}[\min(\tau, T) \mid \mathcal{E}] + 1. \quad (40)$$

Finally, substituting the bound sample complexity derived in Theorem 5.1, that hold under the good event, we have that:

$$\mathbb{E}[R(T)] \leq \mathcal{O} \left(\min \left\{ T, \max_{a \in \mathcal{A}} \sum_{i \in OS(a)} \frac{32 \log(8KNT/\Delta_{a,i}^F)}{(\Delta_{a,i}^F)^2} \right\} \right). \quad (41)$$

which concludes the proof. $\square$

E EXTENDED ELIMINATION ALGORITHM

Now we provide a proof of our Proposition 5.3 that we restate bellow:

Proposition E.1 (Restated). *Let F^* be a full ranking profile for agents $\mathcal{A} = \mathcal{P} \cup \mathcal{A}$ with $\mathcal{P}$ -optimal stable matching $\mathbf{m}^*$. Then the following partial profile P is a valid profile for F^* :*

- $\forall p \in \mathcal{P}, \quad P_p = \{(a, a') \in \mathcal{A} \times \mathcal{A} \mid a \succ a' \succeq_{F_p^*} \mathbf{m}^*(p)\}$
 $\cup \{(a, a') \in \mathcal{A} \times \mathcal{A} \mid a \succeq_{F_p^*} \mathbf{m}^*(p) \succ_{F_p^*} a'\}$
- $\forall a \in \mathcal{A}, \quad P_a = \{(p, p') \in \mathcal{A} \times \mathcal{A} \mid p' \prec p \preceq_{F_a^*} \mathbf{m}^*(a)\}$
 $\cup \{(p, p') \in \mathcal{A} \times \mathcal{A} \mid p \preceq_{F_a^*} \mathbf{m}^*(a) \prec_{F_a^*} p'\}$

Algorithm 3 EE-PSM: Extended Elimination PSM

Require: $\delta, \mathcal{P}, \mathcal{A}$

```
1:  $S_a \leftarrow OS(a) \quad \forall a \in \mathfrak{A}$  ▷ Active pairs
2:  $P_a \leftarrow \emptyset \quad \forall a \in \mathfrak{A}$  ▷ Partial rank
3:  $m_\tau = POSM(\{P_a\}_{a \in \mathfrak{A}})$ 
4: while  $m_\tau = \text{NONE} \wedge \bigcup_{a \in \mathfrak{A}} S_a \neq \emptyset$  do
5:    $S_t \leftarrow \{(p, a) \in \mathcal{P} \times \mathcal{A} \mid a \in S_p(t) \vee p \in S_a(t)\}$ 
6:   for  $m \in \mathfrak{M}(S_t)$  do
7:     Sample update  $\hat{\mu}_{a, m(a)} \quad \forall a \in \mathfrak{A}$ 
8:   end for
9:   Update  $c_{a, i}(t) \quad \forall a \in \mathfrak{A}, i \in OS(a)$  ▷ Eq. 2
10:  Update  $P_a \quad \forall a \in \mathfrak{A}$  ▷ Eq. 5
11:  Eliminate pairs in  $S_a$  ▷ Eq. 6
12:  if  $\mathfrak{m} = SSM(\{P_a\}_{a \in \mathfrak{A}})$  exists then
13:    for all  $(p, a) \notin \mathfrak{m}$  with  $p \in S_a$  or  $a \in S_p$  do
14:      if  $\mathfrak{m}(p) \succ_{P_p} a$  then  $S_p \leftarrow S_p \setminus \{a\}$ 
15:      if  $p \succ_{P_a} \mathfrak{m}(a)$  then  $S_a \leftarrow S_a \setminus \{p\}$ 
16:    end for
17:  end if
18:   $m_\tau = POSM(\{P_a\}_{a \in \mathfrak{A}})$ 
19: end while
20: return  $m_\tau$ 
```

Proof. Consider some full ranking profile F compatible with P . We first show that $\mathfrak{m}^*$ is a stable matching in F , and then show that there exists no rotation in F that improves $\mathcal{P}$ partners, i.e., $\mathfrak{m}^*$ is optimal in F .

For contradiction, assume there exists a pair (p, a) blocking $\mathfrak{m}^*$ in F , i.e., $a \succ_{F_p} \mathfrak{m}^*(p)$ and $p \succ_{F_a} \mathfrak{m}^*(a)$. In P all relations of $\mathfrak{m}^*(i)$ to all other agents on the other side of i are known in P_i , and coincide with the relations in F^* . Because F is compatible with P , we have $a \succ_{F_p^*} \mathfrak{m}^*(p)$ and $p \succ_{F_a^*} \mathfrak{m}^*(a)$. But then (p, a) blocks $\mathfrak{m}^*$ in F^* — a contradiction to $\mathfrak{m}^*$ being stable in F^* .

Now assume, for contradiction, that there exists a rotation $(p_1, a_1), (p_2, a_2), \dots, (p_\ell, a_\ell) \in \mathfrak{m}^*$ whose elimination, i.e., matching any p_i with a_{i+1} instead of $a_i = \mathfrak{m}^*(p_i)$ while maintaining all other matches in $\mathfrak{m}^*$,

- improves all p_i 's preferences, i.e.:
 $a_{(i+1) \bmod N} \succ_{F_{p_i}} a_i = \mathfrak{m}^*(p_i)$ for all $i = 1, \dots, \ell$
- deteriorates all a_i 's preferences, i.e.:
 $p_{(i-1) \bmod N} \prec_{F_{a_i}} p_i = \mathfrak{m}^*(a_i)$ for all $i = 1, \dots, \ell$.

Again, all these relations are present in P and because of this also hold in F^* . Thus, $(p_1, a_1), (p_2, a_2), \dots, (p_\ell, a_\ell) \in \mathfrak{m}^*$ is a rotation in F^* that improves $\mathcal{P}$ -preferences in F^* — a contradiction to the optimality of $\mathfrak{m}^*$.

Because F was chosen arbitrarily, $\mathfrak{m}^*$ is the optimal stable matching in any full ranking profile compatible with P . This proves that P is a valid profile for F^* . $\square$

F LOWER BOUND

We consider a matching problem consisting of two disjoint sets of agents, $\mathcal{P}$ and $\mathcal{A}$, with $K = |\mathcal{P}|$ and $N = |\mathcal{A}|$ number of agents in each set. Let $\mathfrak{A} = \mathcal{P} \cup \mathcal{A}$ denote the set of all agents. For any agent $a \in \mathfrak{A}$, we define $OS(a)$ as the set of agents on the opposite side of the market i.e., $OS(a) = \mathcal{A}$ if $a \in \mathcal{P}$, and $OS(a) = \mathcal{P}$ if $a \in \mathcal{A}$.

Let $\mathcal{I}$ be a class of matching environments, where we define a matching instance $\nu = \{\nu_{a, i}\}_{a \in \mathfrak{A}, i \in OS(a)} \in \mathcal{I}$ as a collection of probability distributions, where each $\nu_{a, i}$ is associated with the agent-partner pair (a, i) . For every pair, the observed reward $X_{a, i}$ is a random variable drawn from $\nu_{a, i}$ with a true underlying mean reward $\mu_{a, i} = \mathbb{E}_{X \sim \nu_{a, i}}[X]$, representing the utility agent a derives from being matched with i .

For an instance $\nu \in \mathcal{I}$, we denote with m_ν^* the optimal stable matching. To provide a lower bounds on the sample complexity, we define the set of alternative environments $\mathcal{I}_{\text{alt}}(\nu)$ as the collection of all environments in $\mathcal{I}$ where the optimal stable matching is different from that in ν :

$$\mathcal{I}_{\text{alt}}(\nu) = \{\nu' \in \mathcal{I} : m_{\nu'}^* \neq m_\nu^*\} \quad (42)$$

We denote with $P_\nu(\mathcal{E})$ the probability of an event $\mathcal{E} \in \mathcal{F}_t$ under the environment ν .

F.1 PROOF OF LOWER BOUND

Theorem F.1 (Restated). *Fix $\delta \in (0, 1/2)$. For any δ -PACOS algorithm and any matching instance ν ,*

$$\mathbb{E}_\nu[\tau] \geq \ln \frac{1}{2.4\delta} c_\star^{-1}(\nu) \quad (43)$$

where $c_\star(\nu)$ is the solution of the following optimization problem:

$$c_\star(\nu) = \sup_{w \in \mathcal{S}_{|\mathcal{M}|}} \inf_{\nu' \in \mathcal{I}_{\text{alt}}(\nu)} \sum_{a \in \mathfrak{A}} \sum_{i \in OS(a)} w_{a,i} KL(\nu_{a,i}, \nu'_{a,i}) \quad (44)$$

Proof. Let $\delta \in (0, 1/2)$ and consider a δ -PACOS algorithm, where the probability of the event that the algorithm outputs the optimal stable matching is at least $1 - \delta$, i.e., the probability of the event $\mathcal{E}_\nu = \{m_\tau = m_\nu^*\}$ satisfies $P_\nu(\mathcal{E}_\nu) \geq 1 - \delta$. Now consider alternative matching instance $\nu' \in \mathcal{I}_{\text{alt}}(\nu)$, where by definition we have that $P_{\nu'}(\mathcal{E}_\nu) \leq \delta$.

Applying the Lemma F.2 to the event $\mathcal{E}_\nu$ we have that $\forall \nu' \in \mathcal{I}_{\text{alt}}(\nu)$:

$$\sum_{a \in \mathfrak{A}} \sum_{i \in OS(a)} \mathbb{E}_\nu[N_{a,i}(\tau)] KL(\nu_{a,i}, \nu'_{a,i}) \geq d(P_\nu(\mathcal{E}_\nu), P_{\nu'}(\mathcal{E}_\nu)) \geq d(1 - \delta, \delta) \quad (45)$$

where $d(x, y) := x \log(x/y) + (1 - x) \log((1 - x)/(1 - y))$ is the binary relative entropy.

Now we can further bound:

$$\sum_{a \in \mathfrak{A}} \sum_{i \in OS(a)} \mathbb{E}_\nu[N_{a,i}(\tau)] KL(\nu_{a,i}, \nu'_{a,i}) \geq \ln \frac{1}{2.4\delta} \quad (46)$$

using the lower bound on the binary relative entropy $d(1 - \delta, \delta) \geq \ln \frac{1}{2.4\delta}$ for $\delta \in (0, 1/2)$ Kaufmann et al. [2016].

Taking the infimum over the possible alternative model $\nu' \in \mathcal{I}_{\text{alt}}(\nu)$ we have that:

$$\ln \frac{1}{2.4\delta} \leq \inf_{\nu' \in \mathcal{I}_{\text{alt}}(\nu)} \sum_{a \in \mathfrak{A}} \sum_{i \in OS(a)} \mathbb{E}_\nu[N_{a,i}(\tau)] KL(\nu_{a,i}, \nu'_{a,i}) \quad (47)$$

$$\leq \mathbb{E}_\nu[\tau] \inf_{\nu' \in \mathcal{I}_{\text{alt}}(\nu)} \sum_{a \in \mathfrak{A}} \sum_{i \in OS(a)} \frac{\mathbb{E}_\nu[N_{a,i}(\tau)]}{\mathbb{E}_\nu[\tau]} KL(\nu_{a,i}, \nu'_{a,i}) \quad (48)$$

Now by the definition of the transformed simplex we have that $\frac{\mathbb{E}_\nu[N_{a,i}(\tau)]}{\mathbb{E}_\nu[\tau]} \in \mathcal{S}_{|\mathcal{M}|}$ which leads to:

$$\ln \frac{1}{2.4\delta} \leq \mathbb{E}_\nu[\tau] \sup_{w \in \mathcal{S}_{|\mathcal{M}|}} \inf_{\nu' \in \mathcal{I}_{\text{alt}}(\nu)} \sum_{a \in \mathfrak{A}} \sum_{i \in OS(a)} w_{a,i} KL(\nu_{a,i}, \nu'_{a,i}) \quad (49)$$

which concludes the proof $\square$

F.2 CHANGE OF DISTRIBUTION LEMMA

At each time step $t \geq 1$ we sample a matching $m_t \in \mathcal{M}$, and the platform observers the rewards of each agent that is matched $X_t = (X_{a,m_t(a)} I\{a \in m_t(a)\})_{\forall a \in \mathfrak{A}} \subset \mathbb{R}^{K \times N}$ similar to the semi-bandit feedback, where $X_{a,m_t(a)}$ is distributed according to $\nu_{a,m_t(a)}$.

Given a history $\mathcal{F}_t = \sigma(m_1, X_{1,m_1}, \dots, m_t, X_{t,m_t})$ up to time t , we can define the log-likelihood of the observation for the two different matching instances ν and ν' :

$$\mathcal{L}_t(m_1, X_{1,m_1}, \dots, m_t, X_{t,m_t}) = \sum_{a \in \mathfrak{A}} \sum_{i \in OS(a)} \sum_{s=1}^t I\{(a, i) \in m_t\} \ln \frac{p_{a,i}(X_{s,a,i})}{p'_{a,i}(X_{s,a,i})} \quad (50)$$

where $p_{i,j}$ denotes the density of the distribution $\nu_{i,j}$.

Now, we present a transportation lemma for two-sided matching problems, adapted from Lemma 1 originally proposed in Kaufmann et al. [2016].

Lemma F.2 (Change of Distribution for Matching Markets). *Let ν and ν' be two matching market instances where for every agent $a \in \mathfrak{A}$ and every potential partner $i \in OS(a)$, the reward distributions $\nu_{a,i}$ and $\nu'_{a,i}$ are mutually absolutely continuous. For any almost-surely finite stopping time τ with respect to $(\mathcal{F}_t)_{t \in \mathbb{N}}$, and any event $\mathcal{E} \in \mathcal{F}_\sigma$:*

$$\sum_{a \in \mathfrak{A}} \sum_{i \in OS(a)} \mathbb{E}_\nu[N_{a,i}(\tau)] KL(\nu_{a,i}, \nu'_{a,i}) \geq d(P_\nu(\mathcal{E}), P_{\nu'}(\mathcal{E}))$$

where $d(x, y) := x \log(x/y) + (1-x) \log((1-x)/(1-y))$ is the binary relative entropy.

Proof. For each $(a, i) \in \mathfrak{A} \times OS(a)$, let $\{X_{s,a,i}\}_{s \geq 1}$ be the observation sequence associated with that pair, and let $N_{a,i}(\tau)$ be the number of times it is sampled up to time τ ; thus the realized observations by time τ are $\{X_{s,a,i}\}_{s=1}^{N_{a,i}(\tau)}$.

Then we can define the the log-likelihood $\mathcal{L}_\tau$ at time τ as

$$\mathcal{L}_\tau(m_1, X_{1,m_1}, \dots, m_\tau, X_{\tau,m_\tau}) = \sum_{a \in \mathfrak{A}} \sum_{i \in OS(a)} \sum_{s=1}^{N_{a,i}(\tau)} \ln \frac{p_{a,i}(X_{s,a,i})}{p'_{a,i}(X_{s,a,i})}. \quad (51)$$

With a direct application of Wald's Identity on $\mathbb{E}_\nu[L_\tau]$ we have that:

$$\mathbb{E}_\nu[L_\tau] = \sum_{a \in \mathfrak{A}} \sum_{i \in OS(a)} \mathbb{E}_\nu[N_{a,i}(\tau)] \mathbb{E}_\nu\left[\ln \frac{p_{a,i}(X_{a,i})}{p'_{a,i}(X_{a,i})}\right] = \sum_{a \in \mathfrak{A}} \sum_{i \in OS(a)} \mathbb{E}_\nu[N_{a,i}(\tau)] KL(\nu_{a,i}, \nu'_{a,i}) \quad (52)$$

where in the second equality we use the fact that $KL(\nu, \nu') = \mathbb{E}_\nu[\ln \frac{p(X)}{p'(X)}]$.

Applying the Lemma 19 from Kaufmann et al. [2016] which state that $\mathbb{E}_\nu[L_\tau] \geq d(P_\nu(\mathcal{E}), P_{\nu'}(\mathcal{E}))$ results in:

$$\sum_{a \in \mathfrak{A}} \sum_{i \in OS(a)} \mathbb{E}_\nu[N_{a,i}(\tau)] KL(\nu_{a,i}, \nu'_{a,i}) \geq d(P_\nu(\mathcal{E}), P_{\nu'}(\mathcal{E}))$$

which concludes the proof. $\square$

G LOWER BOUNDS ON SPECIFIC INSTANCES

We can now continue by providing a lower bound based on specific instances.

Similar to the work of Kaufmann et al. [2016] we will perform an analysis changing the distribution of one pair of agents, under the identifiable class of environments:

$$\mathcal{I}_{\text{class}} = \{\nu = (\nu_{p,a})_{\forall (p,a) \in \mathcal{P} \times \mathcal{A}} : \nu_{p,a} \in \mathcal{P}_m \text{ and } m^*(\nu) \neq \emptyset\} \quad (53)$$

such that the probability measure $\mathcal{P}_m$ satisfies the following continuity assumption:

Assumption 2. For any two distribution $p, q \in \mathcal{P}_m$, such that $p \neq q$, and for all $\alpha > 0$:

1. There exist a $q_1 \in \mathcal{P}_m$ such that: $KL(p, q) < KL(p, q_1) < KL(p, q) + \alpha$ and $\mathbb{E}_{X \sim q_1}[X] > \mathbb{E}_{X \sim q}[X]$
2. There exist a $q_2 \in \mathcal{P}_m$ such that: $KL(p, q) < KL(p, q_2) < KL(p, q) + \alpha$ and $\mathbb{E}_{X \sim q_2}[X] < \mathbb{E}_{X \sim q}[X]$

G.1 GLOBAL PREFERENCES

We study instances with global preferences similar to Example 7 in Liu et al. [2020].

Lemma G.1 (Global preferences). *Consider a matching instance $\nu \in \mathcal{I}_{\text{class}}$ with equal number of agents on both sides and that satisfies Assumption 2. Suppose the agents exhibit global preferences such that for all $p_i \in \mathcal{P}$, $a_1 \succ a_2 \succ \dots \succ a_n$, and for all $a_i \in \mathcal{A}$, $p_1 \succ p_2 \succ \dots \succ p_n$. Then any δ -PACOS Algorithm, satisfies that:*

$$\mathbb{E}_\nu[\tau] \geq \log \left(\frac{1}{2.4\delta} \right) \max(\max_{p_i \in \mathcal{P}} A_i, \max_{a_i \in \mathcal{A}} B_i) \quad (54)$$

where

$$A_i = \max \left(\frac{1}{KL(\nu_{p_i, a_i}, \nu_{p_i, a_{i+1}})}, \frac{1}{KL(\nu_{a_i, p_i}, \nu_{a_i, p_{i+1}})} \right) + \sum_{j < i} \frac{1}{KL(\nu_{a_j, p_i}, \nu_{a_j, p_j})} + \sum_{j > i} \frac{1}{KL(\nu_{p_i, a_j}, \nu_{p_i, a_i})}$$

and

$$B_i = \max \left(\frac{1}{KL(\nu_{p_i, a_i}, \nu_{p_i, a_{i+1}})}, \frac{1}{KL(\nu_{a_i, p_i}, \nu_{a_i, p_{i+1}})} \right) + \sum_{j < i} \frac{1}{KL(\nu_{p_j, a_i}, \nu_{p_j, a_j})} + \sum_{j > i} \frac{1}{KL(\nu_{a_i, p_j}, \nu_{a_i, p_i})}$$

Proof. We consider an instance ν with equal number of agents on both sides and global preference profile F where for all $p_i \in \mathcal{P}$, $a_1 \succ a_2 \succ \dots \succ a_n$, and for all $a_i \in \mathcal{A}$, $p_1 \succ p_2 \succ \dots \succ p_n$, such that the optimal stable matching $\mathbf{m}$ is given by $\mathbf{m}(p_i) = a_i$ for all i .

Figure 5 illustrates the case of three agents on each side.

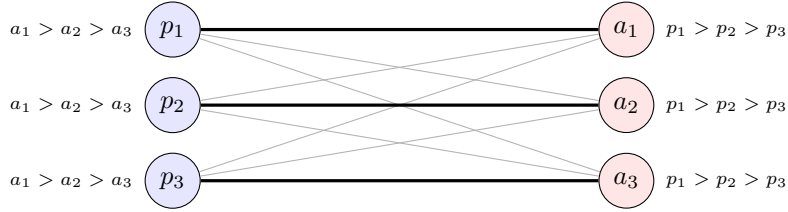


Figure 5: Instance with global preferences for $N = 3$.

We first present a valid preference profile P for this instance that is minimal, in the sense that P is a subset of every other valid partial profile P' for F , i.e., $\forall P' \in \mathcal{P}_{\text{valid}}(F)$, it holds that $P_a \subseteq P'_a$ for all $a \in \mathcal{A}$. This profile P is defined as follows: Each agent $p_i \in \mathcal{P}$ has partial preferences $P_{p_i} = \{a_i \succ a_j \mid j > i\}$, i.e., p_i prefers their optimal matching partner a_i over all higher-indexed agents in $\mathcal{A}$. Similarly, for each agent $a_i \in \mathcal{A}$, $P_{a_i} = \{p_i \succ p_j \mid j > i\}$.

Validity of P

By definition, P is compliant with F . To prove that $P \in \mathcal{P}_{\text{valid}}(F)$, we first show that $\mathbf{m}$ is a super stable matching in P , and then show that $\mathbf{m}$ is optimal in every completion of P , i.e., any full ranking profile compatible with P .

To verify that $\mathbf{m}(p_i) = a_i$ is super stable under the partial preference profile P , consider any pair (p_i, a_j) with $j \neq i$. If $i > j$, then a_j prefers their match p_j to p_i in P , since $p_j = \mathbf{m}(a_j) \succ_{P_{a_j}} p_i$. Thus, (p_i, a_j) is not a blocking pair in any completion of P . If $j > i$, then p_i prefers their match a_i to a_j in P , since $a_i = \mathbf{m}(p_i) \succ_{P_{p_i}} a_j$, and again (p_i, a_j) is not blocking $\mathbf{m}$ in any completion of P . Hence, no pair blocks $\mathbf{m}$ in any completion of P , and $\mathbf{m}$ is super stable.

We now show that $\mathbf{m}$ is also the optimal stable matching in any completion of P , i.e., a pervasive stable matching. For contradiction, assume that this is not the case, and that there exists a full ranking profile F' compatible with P that admits an optimal stable matching $\mathbf{m}' \neq \mathbf{m}$. Because $\mathbf{m}$ is super stable, it is also a stable matching in F' . Now consider some edge (p_i, a_j) in $\mathbf{m}' \setminus \mathbf{m}$. Because $\mathbf{m}'$ is assumed to be optimal in F' , we have that p_i is better off under $\mathbf{m}'$ than $\mathbf{m}$ and vice versa for a_j , i.e., $a_j = \mathbf{m}'(p_i) \succ_{F'_{p_i}} \mathbf{m}(p_i) = a_i$ and $p_j = \mathbf{m}(a_j) \prec_{F'_{a_j}} \mathbf{m}'(a_j) = p_i$. If $i > j$ this is not compliant with preferences $p_j \succ_{P_{a_j}} p_i$ in P . Similarly, if $j > i$ this is not compliant with preferences $a_i \succ_{P_{p_j}} a_j$ in P . Thus, there cannot exist any edge (p_i, a_j) in $\mathbf{m}' \setminus \mathbf{m}$. This is a contradiction to our choice of F' and $\mathbf{m}'$. Hence $\mathbf{m}$ is optimal in any full ranking profile compatible with P .

Minimality of P

We now prove that P is minimal in the sense that it is a subset of every other valid partial profile P' for F , i.e., $\forall P' \in \mathcal{P}_{\text{valid}}(F)$, it holds that $P_a \subseteq P'_a$ for all $a \in \mathcal{A}$.

To see this, consider that for any valid profile $P' \in \mathcal{P}_{\text{valid}}(F)$, in order for an edge (p_i, a_j) not to block the stable matching m in any completion of P' , we require that either $a_i \succ_{P'_i} a_j$ or $p_j \succ_{P'_{a_j}} p_i$ (not necessarily both). According to F , the former holds only when $i < j$, and the latter holds only when $j < i$. Thus:

- For all pairs (p_i, a_j) with $i < j$, a valid profile must specify $a_i \succ_{p_i} a_j$.
- For all pairs (p_i, a_j) with $i > j$, a valid profile must specify $p_j \succ_{a_j} p_i$.

These are exactly the relations specified in P . Therefore, P is both valid and minimal with respect to ensuring stability and optimality of m across all completions.

Lower bound on the expected number of pair samples $\mathbb{E}_\nu[N_{p_i, a_j}(\tau)]$

Following Kaufmann et al. [2016], we construct alternative environments to provide lower bounds on the sample counts N_{p_i, a_i} and N_{a_i, p_i} for any agent pair (p_i, a_i) .

We construct our alternative environments based on the preference profile $P = \{P_a\}_{a \in \mathcal{A}}$ defined above. Specifically, since we proved that P is valid and minimal, any completion (or alternative environment) compatible with P will yield the exact same optimal stable matching. Therefore, to create an alternative environment ν' where the optimal stable matching is different, we must construct an environment that is not consistent with P . We achieve this by minimally modifying the reward distribution of only a single pair.

We specifically consider alternative instances ν' where the distribution of one agent $p_i \in \mathcal{P}$ is modified in only one agent $a_j \in \mathcal{A}$:

$$\nu' = (\nu_{p_1}, \dots, \nu'_{p_i}, \dots, \nu_{p_K}) \quad (55)$$

$$\nu_{a_1}, \dots, \nu_{a_i}, \dots, \nu_{a_N}) \quad (56)$$

where

$$\nu'_{p_i} = (\nu_{p_i, a_1}, \dots, \nu'_{p_i, a_j}, \dots, \nu_{p_i, a_N}) \quad (57)$$

with (using Assumption 2):

- if $j = i$ then: $\mu'_{p_i, a_i} < \mu_{p_i, a_{i+1}}$ and $KL(\nu_{p_i, a_i}, \nu_{p_i, a_{i+1}}) < KL(\nu_{p_i, a_i}, \nu'_{p_i, a_i}) < KL(\nu_{p_i, a_i}, \nu_{p_i, a_{i+1}}) + \alpha$
- if $j > i$ then: $\mu'_{p_i, a_j} > \mu_{p_i, a_i}$ and $KL(\nu_{p_i, a_j}, \nu_{p_i, a_i}) < KL(\nu_{p_i, a_j}, \nu'_{p_i, a_j}) < KL(\nu_{p_i, a_j}, \nu_{p_i, a_i}) + \alpha$

Note that in every of those alternative instances we have a different optimal stable matching, i.e., $m_\nu^* \neq m_{\nu'}^*$. Thus, introducing the event $\mathcal{E} = \{m_\tau = m_\nu^*\}$, any δ -PAC algorithm stopping at time τ with matching m_τ satisfies $\Pr_\nu(\mathcal{E}) \geq 1 - \delta$ and $\Pr_{\nu'}(\mathcal{E}) \leq \delta$.

Applying Lemma F.2 on the stopping time τ for the two matching instances ν and ν' we have that:

$$\sum_{a \in \mathcal{A}} \sum_{i \in OS(a)} \mathbb{E}_\nu[N_{a, i}(\tau)] KL(\nu_{a, i}, \nu'_{a, i}) \geq d(P_\nu(\mathcal{E}), P_{\nu'}(\mathcal{E})) \geq d(1 - \delta, \delta)$$

where $N_{a, i}(\tau)$ is the number of samples of pair $(a, i) \in \mathcal{P} \times \mathcal{A}$ until stopping time τ .

Using the lower bound on the binary relative entropy $d(1 - \delta, \delta) \geq \ln \frac{1}{2.4\delta}$ for $\delta \in (0, 1/2)$ Kaufmann et al. [2016], and the fact that the two different instances ν and ν' differ only in the distribution of one pair, we obtain:

$$\mathbb{E}_\nu[N_{p_i, a_j}(\tau)] \geq \log\left(\frac{1}{2.4\delta}\right) \frac{1}{KL(\nu_{p_i, a_j}, \nu'_{p_i, a_j})}$$

Now from the definition of the alternative instances, for $i = j$:

$$\mathbb{E}_\nu[N_{p_i, a_i}(\tau)] \geq \log\left(\frac{1}{2.4\delta}\right) \frac{1}{KL(\nu_{p_i, a_i}, \nu_{p_i, a_{i+1}}) + \alpha}$$

and for $j > i$:

$$\mathbb{E}_\nu[N_{p_i, a_j}(\tau)] \geq \log\left(\frac{1}{2.4\delta}\right) \frac{1}{KL(\nu_{p_i, a_j}, \nu_{p_i, a_i}) + \alpha}$$

Furthermore, for $j < i$, the definition of P and the construction of the alternative environments imply that p_i does not require samples from a_j . Consequently:

$$\mathbb{E}_\nu[N_{p_i, a_j}(\tau)] = 0 \quad (58)$$

Now following the same reasoning for the agents $a_i \in \mathcal{A}$, by symmetry of the partial preferences P we have that for $i = j$:

$$\mathbb{E}_\nu[N_{a_i, p_i}(\tau)] \geq \log\left(\frac{1}{2.4\delta}\right) \frac{1}{KL(\nu_{a_i, p_i}, \nu_{a_i, p_{i+1}}) + \alpha}$$

and for $j > i$:

$$\mathbb{E}_\nu[N_{a_i, p_j}(\tau)] \geq \log\left(\frac{1}{2.4\delta}\right) \frac{1}{KL(\nu_{a_i, p_j}, \nu_{a_i, p_i}) + \alpha}$$

and for $j < i$:

$$\mathbb{E}_\nu[N_{a_i, p_j}(\tau)] = 0 \quad (59)$$

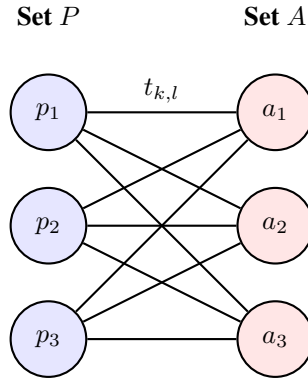


Figure 6: Illustration of the bipartite graph G .

Lower bound on the number of matchings $\mathbb{E}_\nu[\tau]$

Finally, we lower bound the number of matchings τ by analyzing the associated bipartite multigraph G with edge multiplicities $t_{i,j}$, as illustrated in Figure 6.

To define the edge multiplicities $t_{k,l}$, we observe that each pair (p_k, a_l) must be sampled at least $\max(\mathbb{E}_\nu[N_{p_k, a_l}(\tau)], \mathbb{E}_\nu[N_{a_k, p_l}(\tau)])$. Letting parameter $\alpha \rightarrow 0$, we have that for $k = l$:

$$t_{k,k} = \max(\mathbb{E}_\nu[N_{p_k, a_k}(\tau)], \mathbb{E}_\nu[N_{a_k, p_k}(\tau)]) \geq \log\left(\frac{1}{2.4\delta}\right) \max\left(\frac{1}{KL(\nu_{p_k, a_k}, \nu_{p_k, a_{k+1}})}, \frac{1}{KL(\nu_{a_k, p_k}, \nu_{a_k, p_{k+1}})}\right)$$

and for $k < l$, Eq. 59 implies that $\mathbb{E}_\nu[N_{a_l, p_k}] = 0$ and thus:

$$t_{k,l} = \mathbb{E}_\nu[N_{p_k, a_l}(\tau)] \geq \log\left(\frac{1}{2.4\delta}\right) \frac{1}{KL(\nu_{p_k, a_l}, \nu_{p_k, a_k})}$$

and for $l < k$, Eq. 58 implies that $\mathbb{E}_\nu[N_{p_k, a_l}] = 0$ and thus:

$$t_{k,l} = \mathbb{E}_\nu[N_{a_l, p_k}(\tau)] \geq \frac{1}{KL(\nu_{a_l, p_k}, \nu_{a_l, p_l})}$$

Now according to König's Theorem, the number of matchings required to cover the edges of the bipartite multigraph G , is equal to the maximum degree:

$$\mathbb{E}_\nu[\tau] = \text{def}(G) = \max\left(\max_{p_i \in \mathcal{P}} \deg(p_i), \max_{a_i \in \mathcal{A}} \deg(a_i)\right)$$

where the degree for agents $p_i \in \mathcal{P}$ is:

$$\begin{aligned} \deg(p_i) &= \sum_{j=1}^n t_{i,j} = t_{i,i} + \sum_{j<i} t_{i,j} + \sum_{j>i} t_{i,j} \\ &\geq \log\left(\frac{1}{2.4\delta}\right) \left[\max\left(\frac{1}{KL(\nu_{p_i,a_i}, \nu_{p_i,a_{i+1}})}, \frac{1}{KL(\nu_{a_i,p_i}, \nu_{a_i,p_{i+1}})}\right) + \sum_{j<i} \frac{1}{KL(\nu_{a_j,p_i}, \nu_{a_j,p_j})} + \sum_{j>i} \frac{1}{KL(\nu_{p_i,a_j}, \nu_{p_i,a_i})} \right] \end{aligned}$$

and the degree for agents for $a_i \in \mathcal{A}$:

$$\begin{aligned} \deg(a_i) &= \sum_{j=1}^n t_{j,i} = t_{i,i} + \sum_{j<i} t_{j,i} + \sum_{j>i} t_{j,i} \\ &\geq \log\left(\frac{1}{2.4\delta}\right) \left[\max\left(\frac{1}{KL(\nu_{p_i,a_i}, \nu_{p_i,a_{i+1}})}, \frac{1}{KL(\nu_{a_i,p_i}, \nu_{a_i,p_{i+1}})}\right) + \sum_{j<i} \frac{1}{KL(\nu_{p_j,a_i}, \nu_{p_j,a_j})} + \sum_{j>i} \frac{1}{KL(\nu_{a_i,p_j}, \nu_{a_i,p_i})} \right] \end{aligned}$$

To conclude we have that:

$$\mathbb{E}_\nu[\tau] \geq \log\left(\frac{1}{2.4\delta}\right) \max(\max_{p_i \in \mathcal{P}} A_i, \max_{a_i \in \mathcal{A}} B_i) \quad (60)$$

where,

$$A_i = \max\left(\frac{1}{KL(\nu_{p_i,a_i}, \nu_{p_i,a_{i+1}})}, \frac{1}{KL(\nu_{a_i,p_i}, \nu_{a_i,p_{i+1}})}\right) + \sum_{j<i} \frac{1}{KL(\nu_{a_j,p_i}, \nu_{a_j,p_j})} + \sum_{j>i} \frac{1}{KL(\nu_{p_i,a_j}, \nu_{p_i,a_i})}$$

and

$$B_i = \max\left(\frac{1}{KL(\nu_{p_i,a_i}, \nu_{p_i,a_{i+1}})}, \frac{1}{KL(\nu_{a_i,p_i}, \nu_{a_i,p_{i+1}})}\right) + \sum_{j<i} \frac{1}{KL(\nu_{p_j,a_i}, \nu_{p_j,a_j})} + \sum_{j>i} \frac{1}{KL(\nu_{a_i,p_j}, \nu_{a_i,p_i})}$$

□

H ADDITIONAL EXPERIMENTS

We present additional results for the pure exploration setting in markets with different numbers of agents on the two sides. Specifically, we fix the number of arms to $K = 10$ and vary the number of agents N from 2 to 50. For each value of N , we run 20 independent instances generated as in the main experiment. As shown in Figure 7, the observations from Section 7.1 continue to hold across different market sizes.

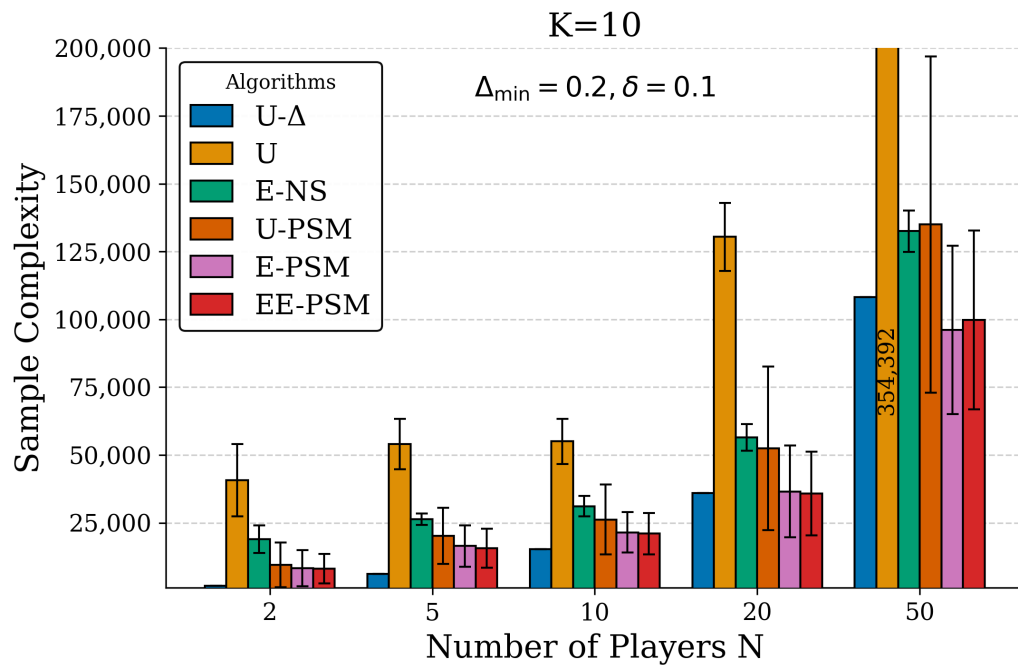


Figure 7: Sample complexity as a function of the number of players N , with $K = 10$ fixed.