MAXIMUM PROBABILITY-DRIVEN BANDIT LEARNING FOR MATCHING MARKETS

Anonymous authors
Paper under double-blind review

ABSTRACT

Security and robustness are crucial for ensuring stable and fair transactions in two-sided markets, given the complexity of preferences and uncertain returns experienced by the participants. In contrast to traditional competing bandits in two-sided markets that focus on maximum returns, we propose a maximum probability-driven bandit learning (P-learning) model that emphasizes risk quantification. Since one side of the market lacks prior knowledge about its preferences for the other, the proposed P-learning algorithm maximizes the probability of Mean-Volatility statistics lying in a preferred and attainable interval. A scalable and stable matching rule was proposed by combining P-learning with the Gale-Shapley matching algorithm that ensures secure and efficient outcomes. A detailed exploration-exploitation procedure of the matching algorithm has been presented with the support of a centralized platform. In both the single-agent setting and the multi-agent setting, our model achieves sublinear regret of $\mathcal{O}(\sqrt{n})$, under different conditions. This paper theoretically proves that the P-learning generates stronger statistical power than classical tests based on normality. Simulation studies demonstrate the superiority of our algorithm over the existing works.

1 INTRODUCTION

Two-sided markets are crucial in modern economies, spanning sectors like online marketplaces Shi et al. (2022), sharing economy platforms Iasevoli et al. (2018); Mittendorf (2018), and financial markets Wright (2004). They involve two distinct participant groups, like Uber drivers and passengers, eBay sellers and buyers, LinkedIn employers and job seekers. Real-world decision-making introduces scarcity and competition Liu et al. (2020), adding uncertainty and complexity. Security and robustness become pivotal for matching market integrity.

Bandit learning's integration with two-sided markets, where players and arms express preferences, was pioneered Das & Kamenica (2005). Bandit algorithms have typically formulated strategies with the goal of maximizing cumulative rewards for agents, aiming to achieve the highest rewards (i.e., minimal regret) in various competitive environments. Extensions include diverse preferences Bistritz & Leshem (2018) and self-interested players Boursier & Perchet (2020). Centralized algorithms Liu et al. (2020) blend Explore-then-Commit Lattimore & Szepesvári (2020) and Upper Confidence Bound Lai et al. (1985) with Gale-Shapley Gale & Shapley (1962). Both two algorithms achieved low stable regret $\mathcal{O}(\log(n))$, which is order-optimal. The work in Liu et al. (2021) discussed a decentralized version of the problem. Work such as eliminating irrelevant choices and conflicting resolutions is further expanded Basu et al. (2021); Wang et al. (2022); Maheshwari et al. (2022). However, in competitive markets such as ride-hailing, existing strategies tend to focus solely on immediate profits, neglecting risk quantification during the matching process. Platform decisions must consider economic factors and be influenced by the actions of other decision-makers. Issues of fairness Graham & Joe (2017); Jia et al. (2017) have gained attention, as reports show that a small fraction of drivers earn most of the income, while others earn very little or even incur losses Zoepf et al. (2018).

Recently, scholars explored strategy-driven central limit theorems for the multi-armed bandit problem Chen et al. (2023a;b). Combinatorial multi-armed bandit (CMAB) with nonlinear rewards Chen et al. (2016) was studied. Robust limit theorems Lan & Zhang (2017) and dynamic allocation Cohen & Treetanthiploet (2022) under nonlinear expectations were examined. Chen et al.'s work Chen

et al. (2023a) contributed strategy-driven limit theorems. In this paper, our research is motivated by a model for the two-sided market that employs a Multi-Armed Bandit (MAB) approach within a non-linear expectation framework. We develop models to capture market dynamics, involving agents (e.g., passengers) and arms (e.g., drivers). Assuming arms have explicit preferences for agents, we introduce the **Maximum Probability-driven Learning** (P-learning) algorithm. (i) P-learning algorithm incorporates volatility measures, shifting the focus from maximizing individual rewards to optimizing the distribution, to ensure the security and robustness of transactions, thereby enabling the platform to make decisions with greater foresight. (ii) P-learning algorithm aims to maximize the probability of **Mean-Volatility** (MV) statistics falling within expected utility $[\alpha_i, \beta_i]$ ranges, thereby achieving effective risk control. Expected utility is a preferences-based interval that reflects the range of psychological expectations estimated by the platform based on user needs, leveraging previous data and operational mechanisms. The specific form of MV statistics is as follows Chen et al. (2022):

$$T_{i,m,n}^{\boldsymbol{\vartheta}_i} = \underbrace{\frac{1}{n}\sum_{j=1}^{m} X_{i,j}^{\boldsymbol{\vartheta}_i}}_{\text{mean}} + \underbrace{\frac{1}{\sqrt{n}}\sum_{j=1}^{m} \frac{X_{i,j}^{\boldsymbol{\vartheta}_i} - \hat{\mu}_j^{\boldsymbol{\vartheta}_i}}{\hat{\sigma}_j^{\boldsymbol{\vartheta}_i}}}_{\text{volatility}}. \tag{1}$$

$T_{i,m,n}^{\boldsymbol{\vartheta}_i}$ can be seen as a historical statistics that determines the decision rule for the agent p_i to construct the values of statistics under the strategy $\boldsymbol{\vartheta}_i$. The first component of the above equation denotes average rewards (the first moment), while the second component characterizes average volatility (the second-moment). Figure 1 visualizes the difference between traditional algorithms (left) and P-learning algorithms (right). Leveraging the rich market information embedded in the distribution of MV statistics, we perceive the P-learning algorithm as maximum probability-driven and suitable for real-world applications. Subsequently, combining the P-learning algorithm with Gale-Shapley matching Gale & Shapley (1962) achieves stable matching. Furthermore, we introduce the concept of stable regret within the probability framework and demonstrate its application in both single-agent and multi-agent contexts.

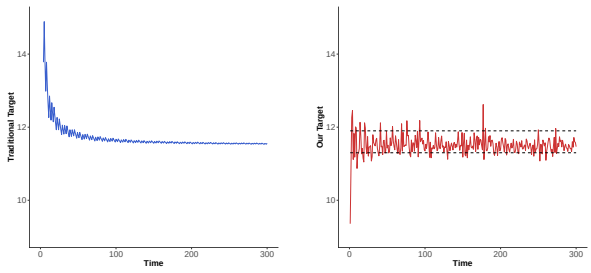


Figure 1: Illustration of comparing P-learning algorithm and traditional algorithms. The blue line depicts the mean reward of the ETC method, which steadily converges toward the target value as time progresses. The red line represents the P-learning algorithm, showcasing the fluctuation of its MV statistics within the target interval.

Our merits of the proposed secure method of matching two-sided markets in the probabilistic framework are multiple aspects, including:

- This work develops a **faster** bandit process for matching market than traditional algorithms of UCB, ϵ-greedy and so on, because the novel bandit learning considers more information about the volatility of average return.
- Different from classical bandit processes that just pursue maximum average reward, the proposed bandit learning is **adaptive** to the attainable goal of one interval by using the central limit theory of the developed Mean-Volatility statistics.
- Motivated by the **responsible** artificial intelligence in the significance of safety, this work firstly balances the controlled risk and average return in a single statistical quantity.

2 PRELIMINARIES

Agents and arms. Let $\mathcal{N} = \{p_1, p_2, ..., p_n\}$ be the set of agents, and $\mathcal{K} = \{a_1, a_2, ..., a_k\}$ be the set of arms. Each agent p_i possesses its own expected utility $[\alpha_i, \beta_i]$ and each arm a_i has predetermined and known rankings of agents. We denote the notation $p_j \succ p_{j'}$ to indicate that arm a_i exhibits a strict preference for agent p_j over agent $p_{j'}$. When the specific arm under consideration is evident from the context, we simply write $a_j \succ_i a_{j'}$, indicating that agent p_i prefers arm $\boldsymbol{a}_j$ over arm $\boldsymbol{a}'_j$.

Stable matching. Given the comprehensive preference rankings of both the arms and the agents, we employ the Gale-Shapley algorithm to achieve a stable matching. This algorithm operates by having one side of the market propose repeatedly to the other side until a stable match is found. At each time step n, the arm successfully matched to agent p_i are represented by $m_n(i)$. In instances where multiple agents express a preference for the same arm, a competitive scenario arises, and only one agent can effectively engage with the arm $m_n(i)$ at time n. Following a successful match, the agent p_i will receive a stochastic reward $X_{i,n}$, which is sampled from a Gaussian distribution.

Centralized bandit. In our study, we focus on a centralized setting in which the players are able to communicate with a central platform that computes matchings for the entire market. Agents do not need to communicate with each other. More specifically, at each time step n, $\boldsymbol{m_n} = \{m_n(1), m_n(2), \ldots, m_n(n)\}$ represents the platform's computation of a matching vector, which determines the assignment of agents to arms. It is worth emphasizing that the agents' selection of arms must rely on their historical rewards since there is no direct information exchange between the agents and the arms.

Stable regret. We define a notion of stable regret in the framework of probability, which measures the deviation between an agent's actual performance and the performance they could have achieved by making optimal choices. The probability that the MV statistics obtained by any agent falls within its expected utility is compared to that obtained by playing the agent's optimal stable match in all rounds. The n-round individual instantaneous regret for an agent, denoted as $R_i(n)$, is defined as follows:

$$R_i(n) = \sup_{\boldsymbol{\vartheta}_i \in \Theta} P\left(\alpha_i \leq T^{\boldsymbol{\vartheta}_i}_{i,\infty,\infty} \leq \beta_i\right) - P\left(\alpha_i \leq T^{\boldsymbol{\vartheta_i}}_{i,n,n} \leq \beta_i\right), \tag{2}$$

where $\boldsymbol{\vartheta}_i$ represents the set of all strategy choices made by the agent p_i over n rounds, Θ represents the collection of all possible strategy sets, we employ the parameter $\boldsymbol{\vartheta}_i$ to capture specific characteristics of interest. The MV statistic $T^{\boldsymbol{\vartheta}_i}_{i,n,n}$ represents the actual value obtained under the strategy $\boldsymbol{\vartheta}_i$ based on historical information in round n. As the time horizon n becomes sufficiently large, $T^{\boldsymbol{\vartheta}_i}_{i,\infty,\infty}$ represents the theoretical value of the MV statistics under the strategy $\boldsymbol{\vartheta}_i$. Therefore, the cumulative regret, denoted as $CR_i(n)$, is defined as follows:

$$\begin{aligned} CR_i(n) = n \sup_{\boldsymbol{\vartheta}_i \in \Theta} P\left(\alpha_i \leq T^{\boldsymbol{\vartheta}_i}_{i,\infty,\infty} \leq \beta_i\right) \\ - \sum_{m=1}^{n} P\left(\alpha_i \leq T^{\boldsymbol{\vartheta}_i}_{i,m,n} \leq \beta_i\right). \end{aligned} \tag{3}$$

Here, m represents the current round, and n denotes the horizon length. Our primary objective is to design a protocol that guides all agents to minimize their single regret.

3 MULTI-AGENT BANDITS WITH A PLATFORM

3.1 P-LEARNING ALGORITHM

In this section, we present an algorithm to formalize the decision problem. In each round, n, an agent p_i chooses an arm a_i based on a strategy that maximizes the probability of meeting their expected utility, which is known as the P-learning algorithm. This process involves a strategic variable $\boldsymbol{\vartheta}_i$ remembering historical information to decide how to integrate current rewards into the previously summarized statistics. Unlike traditional algorithms that rely on the law of large numbers, our algorithm considers the distribution of prior information, which allows us to fully utilize the statistical

characteristics and patterns embedded in the historical rewards and maximizes the probability of MV statistics within the expected utility.

When agent p_i emerges as the winner in the competition and successfully matches with arm $m_n(i)$, they are rewarded with a random variable $X_{i,n}^{\boldsymbol{\vartheta}_i}$. To be more specific, we employ the Two-Armed Bandit (TAB) Rothschild (1974) to describe how agent p_i can either match with arm a_1 (referred to as the left arm, denoted as L) and receive a reward denoted as $W_{i,n}^L$, or match with arm a_2 (referred to as the right arm, denoted as R) and receive a reward denoted as $W_{i,n}^R$. It is important to note that $W_{i,n}^L$ and $W_{i,n}^R$ are generated from distinct probability distributions. This can be expressed as follows:

$$X_{i,n}^{\boldsymbol{\vartheta}_i} = \begin{cases} W_{i,n}^L, \ m_n(i) = a_1 \\ W_{i,n}^R, \ m_n(i) = a_2. \end{cases} \tag{4}$$

To aggregate these rewards $\boldsymbol{X_i}$ and incorporate the defined strategy $\boldsymbol{\vartheta}_i = \{\vartheta_{i,1}, \cdots, \vartheta_{i,n}\}$, we propose a bandit inference learning approach that uses MV statistics $T_{i,m,n}^{\boldsymbol{\vartheta}_i}$ (see Equation 1). $X_{i,j}^{\boldsymbol{\vartheta}_i}$ represents the reward obtained under strategy $\boldsymbol{\vartheta}_i$ at time j. $\hat{\mu}_j^{\boldsymbol{\vartheta}_i}$ and $\hat{\sigma}_j^{\boldsymbol{\vartheta}_i}$ are the estimated mean and standard deviation, respectively, for the sequence of observed rewards $\boldsymbol{X_i}$ up to time n under the strategy $\boldsymbol{\vartheta}_i$. The mean estimate is defined as:

$$\hat{\mu}_n^{\boldsymbol{\vartheta}_i} = \mathbb{I}\left(m_n(i) = a_1\right) \hat{\mu}_{n,L}^{\boldsymbol{\vartheta}_i} + \mathbb{I}\left(m_n(i) = a_2\right) \hat{\mu}_{n,R}^{\boldsymbol{\vartheta}_i}, \tag{5}$$

and the standard deviation estimate is defined as:

$$\hat{\sigma}_n^{\boldsymbol{\vartheta}_i} = \sqrt{\mathbb{I}\left(m_n(i) = a_1\right) (\hat{\sigma}_{n,L}^{\boldsymbol{\vartheta}_i})^2 + \mathbb{I}\left(m_n(i) = a_2\right) (\hat{\sigma}_{n,R}^{\boldsymbol{\vartheta}_i})^2}, \tag{6}$$

where $\hat{\mu}_{n,L}^{\boldsymbol{\vartheta}_i}$ and $\hat{\mu}_{n,R}^{\boldsymbol{\vartheta}_i}$ are the estimated means for the left and right arms, respectively, and $\hat{\sigma}_{n,L}^{\boldsymbol{\vartheta}_i}$ and $\hat{\sigma}_{n,R}^{\boldsymbol{\vartheta}_i}$ are the estimated standard deviations for the left and right arms, respectively. By utilizing these statistical measures, we can effectively assess the performance of different strategies. The specific form of the strategy $\vartheta_{i,m}, 1 \leq m \leq n$ is as follows:

$$\vartheta_{i,m} = \mathbb{I}\left(T_{i,m-1,n}^{\boldsymbol{\vartheta}_i} \leq c_i - (1 - \frac{m-1}{n}) \frac{\hat{\mu}_{m-1,L}^{\boldsymbol{\vartheta}_i} + \hat{\mu}_{m-1,R}^{\boldsymbol{\vartheta}_i}}{2}\right), \tag{7}$$

where $T_{i,m-1,n}^{\boldsymbol{\vartheta}_i}$ denotes the cumulative MV statistics up to $m-1$ with a horizon length of n and the symmetry center $c_i = (\alpha_i + \beta_i)/2$, which is the center of expected utility $[\alpha_i, \beta_i]$.

The P-learning algorithm, guided by the strategy $\boldsymbol{\vartheta}_i$, balances exploration and exploitation by leveraging the relationship between the MV statistic $T_{i,m-1,n}^{\boldsymbol{\vartheta}_i}$ and the symmetrical center c_i. The decision-making process of agent p_i is influenced by the position of the MV statistics $T_{i,m-1,n}^{\boldsymbol{\vartheta}_i}$ and the symmetrical center c_i. When $T_{i,m-1,n}^{\boldsymbol{\vartheta}_i}$ is belower than $c_i - (1 - \frac{m-1}{n}) \frac{\hat{\mu}_{m-1,L}^{\boldsymbol{\vartheta}_i} + \hat{\mu}_{m-1,R}^{\boldsymbol{\vartheta}_i}}{2}$, the agent p_i tends to favor the arm with higher reward.

Since the per-step strategy $\vartheta_{i,m}$ is dependent on the horizon length n, the parameterization of n enables us to meticulously consider various constraints and limitations, such as resource constraints, time limitations, or specific rule requirements, during the process of algorithm design and implementation and provide more feasible solutions.

Next, we present the optimal results of the P-learning algorithm. Assuming a higher average reward μ_L for the left arm, according to the central limit theorem for strategies, the probability that the MV statistics satisfies the expected utility of agent p_i is given by the following formula for $\alpha_i < \beta_i \in \mathbb{R}$:

$$\begin{aligned} &\lim_{n \to \infty} P\left(\alpha_i \leq T_{i,n,n}^{\boldsymbol{\vartheta}_i} \leq \beta_i\right) \\ &= \begin{cases} \Phi(\mu_L - \alpha_i) - e^{\frac{(\mu_R - \mu_L)(\beta_i - \alpha_i)}{2}} \Phi(\mu_L - \beta_i), & q_i = 1 \\ \Phi(\beta_i - \mu_R) - e^{\frac{(\mu_R - \mu_L)(\beta_i - \alpha_i)}{2}} \Phi(\alpha_i - \mu_R), & q_i = 0, \end{cases} \end{aligned}$$

where $q_i = \mathbb{I}\left(\alpha_i + \beta_i \geq \mu_L + \mu_R\right)$ and Φ denotes the distribution function of standard normal distribution. The result is elegant and can be utilized to obtain the theoretical probability value for any given expected utility.

P-learning algorithm maximizes the probability of Mean-Volatility statistics lying in a preferred and attainable interval. Efficient convergence is observed when the expected utility falls within the range defined by the means of the two arms. However, regret escalates significantly beyond this range. In Figure 2, we investigate the convergence rates of regret across varying interval lengths and symmetric centers. The means of both arms are sampled from Gaussian distributions with means of 10 and 20, respectively. Our analysis highlights that larger interval lengths contribute to swifter convergence rates. Additionally, the algorithm demonstrates greater robustness when the expected utility is closer to the means of the arms.

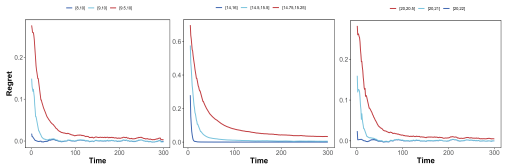


Figure 2: Illustration of single-agent regret under different expected utilities. Greater interval lengths facilitate faster convergence rates, with the algorithm exhibiting increased robustness particularly when the expected utility aligns closely with the means of the arms.

3.2 Markets with Competition

In this section, we introduce market competition by considering the setup involving multiple arms and agents. Assuming arm preferences are known, we employ the P-learning algorithm to ascertain the preference ranking of agents. Subsequently, we utilize the Gale-Shapley algorithm to merge the comprehensive preference rankings of both arms and agents to achieve stable matching. Additionally, we compute instantaneous regret values to assess the effectiveness of the performance evaluation of the algorithm.

Example 1 (2 vs 2) Let $\mathcal{N} = \{p_1, p_2\}$ and $\mathcal{K} = \{a_1, a_2\}$ with true preferences in Table 1 in the Appendix.

Table 1 demonstrates the scenarios of diverse arm preferences when two agents have consistent expected utility. The different arm preferences lead to varying matching outcomes, which affect the regrets of agent p_1 and agent p_2 to different extents.

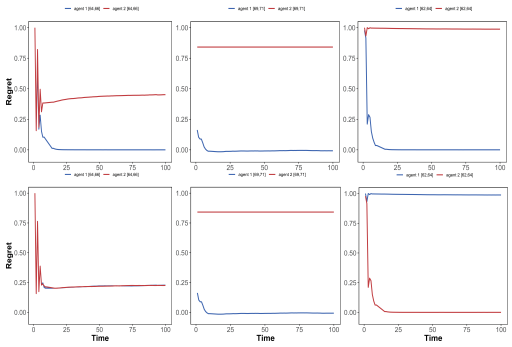


Figure 3: Illustration of the regret of two agents in the case of agents having consistent expected utility. When the arm preferences are consistent, agent p_1's regret always converges to 0, while agent p_2's regret increases to varying degrees. When the arm preferences are inconsistent, the choices made by agents mutually influence each other. The regret of p_2 is also influenced by the symmetric centers c_1 and c_2.

As depicted in Algorithm 1, we employ the P-learning algorithm to ascertain the preference ranking of agents. We specified that the means of arms were drawn from Gaussian distributions with means of 60 and 70 and a variance of 1. Following 300 replicates, we investigated the convergence rates of regret across various expected utility and arm preferences over 100 steps. In Figure 3, the

regrets of two agents are presented. In scenarios where arm preferences are consistent (first row), agent p_1 consistently converges to zero regret, while agent p_2 experiences varying degrees of regret increment. Interestingly, when c_2 and c_1' are close, within the symmetric interval (Lemma 4.4), the regret of agent p_2 decreases. Conversely, when c_2 and c_1' are further apart, the regret of agent p_2 significantly increases. In instances of inconsistent arm preferences (second row), due to the presence of the matching mechanism, agents engage in continuous strategic interactions, resulting in diverse outcomes. In conclusion, regret is influenced by the expected utilities and relative positions of the two agents.

Algorithm 1 P-learning with Competition (2 vs 2)

Input: horizon length n, preferences of arms
Output: regret $R_1(n)$, $R_2(n)$
Initial MV statistics $T_{1,0}^{\boldsymbol{\vartheta_1}} = 0$, $T_{2,0}^{\boldsymbol{\vartheta_2}} = 0$
while $m \leq n$ **do**
  Get agent p_1's and p_2's preference based on Equation (7) using strategy $\vartheta_{1,m}$ and $\vartheta_{2,m}$
  Get $m_m(1)$, $m_m(2)$ matching with Gale-Shapley algorithm
  if $m_m(i) = a_1$ **then**
    Obtain reward $X_{1,m} = W_{1,m}^L$, $X_{2,m} = W_{2,m}^R$
    Update $\hat{\mu}_m^{\boldsymbol{\vartheta_1}}$, $\hat{\sigma}_m^{\boldsymbol{\vartheta_1}}$, $\hat{\mu}_m^{\boldsymbol{\vartheta_2}}$, $\hat{\sigma}_m^{\boldsymbol{\vartheta_2}}$
    Update $T_{1,m}^{\boldsymbol{\vartheta_1}}$, $T_{2,m}^{\boldsymbol{\vartheta_2}}$ based on Equation (1)
  else
    Vice versa
  end if
  Compute $R_1(n)$, $R_2(n)$ based on Equation (2)
end while
return regret $R_1(n)$, $R_2(n)$

Example 2 (3 vs 3) Let $\mathcal{N} = \{p_1, p_2, p_3\}$ and $\mathcal{K} = \{a_1, a_2, a_3\}$ with true preferences in Table 2 in the Appendix.

Table 2 demonstrates the scenarios of diverse arm preferences when three agents have inconsistent expected utility. Our results show that the P-learning algorithm performs well in this larger-scale scenario. It is worth noting that the MAB and TAB (two-agent bandit) problems discussed in our paper are consistent in terms of algorithm and theoretical contributions. As outlined in Algorithm 2, the determination of agents' preference order involves progressively eliminating the best and worst choices. Specifically, when there are k arms, we need to group them to determine the preference order of agents. In each round, after identifying the best and worst arms, we remove these two arms from the set of candidate arms and continue to find the next best and worst arms among the remaining ones, and so forth. Furthermore, algorithm 3 demonstrates matching rules with multiple agents and arms.

We specified that the means of arms were drawn from Gaussian distributions with means of 60, 65 and 70 and a variance of 1. Following 300 replicates, we investigated the convergence rates of regret across various expected utility and arm preferences over 100 steps. As shown in Figure 4, when the preferences for each arm are consistent, the regret for agent p_1 consistently tends towards 0, while the regrets for agents p_2 and p_3 vary in degree (left and middle). However, when arm preferences

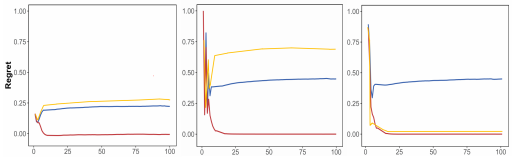


Figure 4: Illustration of the regret of three agents in the case of agents having different expected utilities. When the arm preferences are consistent, agent p_1's regret always converges to 0, while agent p_2's and p_3's regret increases to varying degrees. When the arm preferences are inconsistent, the choices made by agents mutually influence each other.

are inconsistent, the choices of agents influence each other, leading to higher regret for agent p_2 due to the inability to match preferred arms (right).

4 Achieving $\mathcal{O}(\sqrt{n})$ Regret with Two Different Settings

In this section, we provide two different settings for the cumulative regret bounds. In the single-agent setting, we achieve a problem-independent sublinear cumulative regret of $\mathcal{O}(\sqrt{n})$. In the multi-agent setting, a similar sublinear cumulative regret order of $\mathcal{O}(\sqrt{n})$ persists, where agents' selections mutually influence each other. However, the upper bound is influenced by factors such as symmetrical center deviation Δ_c, expected utility length d, and the mean and variance of the reward distributions.

4.1 Single-Agent Regret

In a single-agent setting, there is no collision of choices. Therefore, we consider a scenario where $\mathcal{N} = \{p_1\}$ and $\mathcal{K} = \{a_1, a_2\}$. The objective of agent p_1 is to learn a strategy $\boldsymbol{\vartheta}_i$ that observes the state of the environment and selects actions to maximize the probability of MV statistics in the expected utility during the interaction with the environment. Subsequently, we will provide the asymptotic distribution of $T^{\boldsymbol{\vartheta}_i}_{i,n,n}$ under this setting. The proof of Bandit distribution $\mathcal{B}(\alpha, \beta, c)$ in Lemma 4.1 and its density function are provided in the appendix.

Theorem 4.1. ***(Instantaneous Regret Bound)****. Assume that the agent ranks the arms based on MV statistics and select the top-ranked arm as their preferred match. Then, the expected individual single-step regret of agent p_i is upper bounded by*

$$R_i(n) \leq \frac{M}{\sqrt{n}}, \tag{8}$$

where M depends solely on the parameters σ_L, σ_R and $\alpha = -(\mu_L - \mu_R)/2$ of the reward distribution associated with the arms.

Theorem 4.1 demonstrates that the P-learning algorithm can achieve a problem-independent regret bound of $\mathcal{O}(1/\sqrt{n})$ in the single-agent setting. Importantly, this upper bound is independent of the agents' expected utility length d. Therefore, the symmetrical center c_i plays a crucial role, describing the growth rate of regret values for the shortest interval, determining the unattainability of slower growth rates. This indicates the wide applicability of the regret bounds we provide. With a focus on cumulative regret, subsequent theorems provide an upper bound.

Corollary 4.2. ***(Cumulative Regret Bound).*** *Under the same assumptions as Theorem 4.1, the upper bound of the expected cumulative regret for a single agent p_i is $\mathcal{O}(\sqrt{n})$.*

If the growth rate of the single-step regret is $\mathcal{O}(1/\sqrt{n})$, then the cumulative regret growth rate for the first n steps is bounded. The results indicate that the cumulative regret growth rate for the first n steps is $\mathcal{O}(\sqrt{n})$. P-learning algorithm exhibits commendable performance, showcasing convergence rates comparable to traditional bandit algorithms Lattimore & Szepesvári (2020).

4.2 Multi-Agent Regret

In a multi-agent environment, we consider a scenario involving multiple agents. We assume that all arms are equally desirable to the agents, indicating that there are no inherent preferences for any specific arm. However, the preferences for arms are consistent across all agents.

Definition 4.3. (Consistent Ranked Arms). For any a_i, if $p_1 \succ p_2$, then all arms prefer agent p_1 over agent p_2; conversely, if $p_2 \succ p_1$ then all arms prefer agent p_2 over agent p_1.

We assume that the agents maintain consistent rankings among the arms. For any given arm a_i, the preference order for agent p_i follows a consistent pattern, denoted as $p_1 \succ p_2$. This ensures that the competition remains absent for the more favored agent, resulting in outcomes similar to those in the scenario with a single agent. Conversely, for the other agent, the regret may be substantial. The lemma 4.4 describes this situation in detail.

Lemma 4.4. ***(Anti-Strategy Center).*** *Assume that all arms prefer p_1 more and let $c_1 = (\alpha_1+\beta_1)/2$, which leads to a distribution of the MV statistics consisting of the reward information obtained by agent p_2 with $c_1' = \mu_1 + \mu_2 - c_1$ as the center of symmetry, also known as the center of the anti strategy distribution.*

Given that the strategy $\vartheta_{i,n}$ is determined based on the symmetry center, a strategy-driven algorithm ensures that the MV statistics of agent p_1 are centered around c_1. However, the less favored agent p_2 may encounter strategy failures. Lemma 4.4 provides valuable insights by revealing that, in the worst-case scenario, the reward distribution for agent p_2 corresponds to the anti-strategy distribution of agent p_1. Consequently, the regret experienced by agent p_2 is significantly influenced if the values of c_2 and c_1' are considerably distant from each other, considering the expected utility interval length denoted as d.

Theorem 4.5. ***(Instantaneous Regret Bound).*** *Assume that the agents rank the arms based on their MV statistics and select the highest-ranked arm as their preferred matching choice, with agent p_1 being preferred for the arms. In this setting, the expected individual single-step regret of player p_i is upper bounded by*

$$R_i(n) \leq \sqrt{\frac{K \ln(1/\Delta_n^2)}{n}}, \tag{9}$$

where the constant K is depending only on μ_L, σ_L, μ_R and σ_R of reward distribution associated with arms, Δ_n denotes the maximum probability difference between distributions under different strategies within the range length of n.

To be more specific, we can define $\Delta_n = F_n(c_1) - \gamma$, where F_n represents the distribution under our strategy up to n, P_n represents the distribution under the anti-strategy, and $\gamma = \min\{P_n(\Delta_c - d/2), P_n(\Delta_c + d/2)\}$, with $\Delta_c = |c_1' - c_2|$ measuring the distance between the anti-strategy distribution and the ideal distribution. Here, d denotes the length of the expected utility interval. The constants K in the previous equation depend only on μ_L, σ_L, μ_R and σ_R, which is the reward distribution associated with the arms.

Theorem 4.5 provides us with an upper bound on the single-step regret, which is $\mathcal{O}(1/\sqrt{n})$, under the assumption of consistent arm preferences. This lemma reveals that the upper bound is determined by the center of symmetry offset Δ_c, the length of the interval d, as well as the mean and variance of the reward distribution.

Corollary 4.6. ***(Cumulative Regret Bound).*** *Under the same assumptions as in Theorem 4.5, the expected multi-agent cumulative regret of player p_i is upper bounded by*

$$CR_i(n) \leq \sqrt{nK \ln(1/\Delta_n^2)}. \tag{10}$$

Corollary 4.6 provides us with valuable insights into the growth rate of $R_i(n)$ as the value of n increases. It demonstrates that the growth rate of $\sqrt{n}$ surpasses that of n. Consequently, we can conclude that $\sqrt{n}$ serves as an upper bound for the growth of $R_i(n)$. By expressing the growth upper bound of $R_i(n)$ as $\mathcal{O}(\sqrt{n})$, we establish a clear understanding of the relationship between the regret and the number of steps, highlighting the sublinear nature of the regret growth.

5 SIMULATION

In this section, we present our numerical simulations involving two agents and two arms.Additionally, we provide a detailed hypothesis testing framework in the Appendix (subsection A.1) to validate the statistical significance of our results.

Baseline. We use three baselines with their respective feedback, which are Explore-Then-Commit (ETC), Upper-Confidence-Bound (UCB), and ϵ-greedy. Each agent submits its preference order to a centralized platform in each round, and the platform assigns the best match for the agent under this preference.

Results. We generated random instances to compare the performance of ETC, UCB, ϵ-greedy, and P-learning under their respective settings with standard deviation $\sigma = 1.1$, $\sigma = 1.3$, and $\sigma = 1.5$ and reported their regrets in Figure 5. We used the parameter settings mentioned in Liu et al. (2020) for ETC and UCB and set ϵ to 0.1 according to Sutton & Barto. (1998) for ϵ-greedy. We selected three

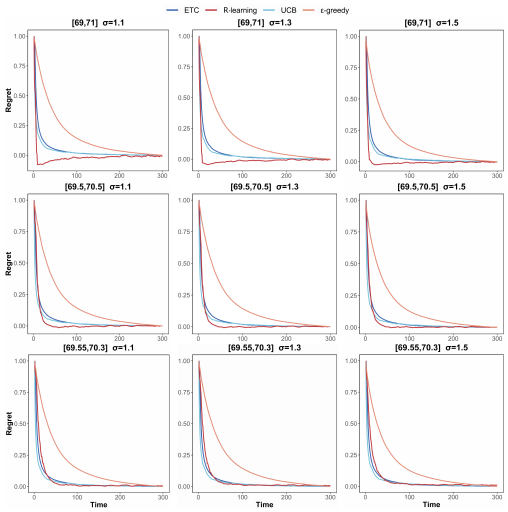


Figure 5: Illustration of comparing regret. Random instances were generated to compare ETC, UCB, ϵ-greedy, and **P-learning** across various standard deviations. The results demonstrate that **P-learning** exhibits the fastest convergence rate.

intervals with different lengths d, namely [69,71], [69.5,70.5] and [69.55,70.3], for P-learning. The means of both arms were drawn from Gaussian distributions with means of 60 and 70, respectively. To ensure consistency, we simulated all the algorithms on the same 100 sample paths and reported their normalized means.

Different variances can be used to measure the uncertainty of trading objects in two-sided markets, which affects trading volume and market price fluctuations. As shown in Figure 5, P-learning demonstrates the fastest convergence rate among the algorithms. Even when considering the interval [69.55,70.3] and $\sigma_L = \sigma_R = 1.5$, P-learning performs comparably to UCB and exhibits slightly superior performance compared to UCB. Furthermore, Table 3 in appendix reports the number of iterations required for different algorithms to achieve a convergence accuracy of 0.01 for regret errors.

6 Discussion

The P-learning algorithm effectively balances exploration and exploitation in resource-constrained market environments by maximizing the probability of meeting the expected utility (interval) for market participants. In single-agent scenarios, our model consistently achieves a sublinear regret of $\mathcal{O}(\sqrt{n})$. In multi-agent settings, agents' decisions influence each other, we maintain this regret bound. However, factors such as symmetric center offset and interval length affect the algorithm's performance.

The setup of our work is centralized, and agents don't need to communicate with each other; they interact directly with the platform. In a decentralized scenario, the potential issue is how to ensure fairness and efficiency in resource competition among agents in the context of unrestricted communication. By facilitating more efficient resource allocation and better matching of goods and services based on market participants' preferences, this algorithm may enhance overall market efficiency and consumer welfare.

REFERENCES

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

Ilai Bistritz and Amir Leshem. Distributed multi-player bandits-a game of thrones approach. *Advances in Neural Information Processing Systems*, 31, 2018.

Etienne Boursier and Vianney Perchet. Selfish robustness and equilibria in multi-player bandits. In *Conference on Learning Theory*, pp. 530–581. PMLR, 2020.

Wei Chen, Wei Hu, Fu Li, Jian Li, Yu Liu, and Pinyan Lu. Combinatorial multi-armed bandit with general reward functions. *Advances in Neural Information Processing Systems*, 29, 2016.

Zengjing Chen, Shui Feng, and Guodong Zhang. Strategy-driven limit theorems associated bandit problems, 2022.

Zengjing Chen, Larry G Epstein, and Guodong Zhang. A central limit theorem, loss aversion and multi-armed bandits. *Journal of Economic Theory*, 209:105645, 2023a.

Zengjing Chen, Xinwei Feng, Shuhui Liu, and Xiaodong Yan. Optimal distributions of rewards for a two-armed slot machine. *Neurocomputing*, 518:401–407, 2023b. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2022.11.019. URL https://www.sciencedirect.com/science/article/pii/S092523122201400X.

Samuel N Cohen and Tanut Treetanthiploet. Gittins' theorem under uncertainty. *Electronic Journal of Probability*, 27:1–48, 2022.

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

Ronald Aylmer Fisher. Statistical methods for research workers. In *Breakthroughs in statistics: Methodology and distribution*, pp. 66–70. Springer, 1970.

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

Mark Graham and Shaw Joe. *Towards a fairer gig economy*. Meatspace Press, 2017.

Gennaro Iasevoli, Laura Michelini, Cecilia Grieco, and Ludovica Principato. Mapping the sharing economy: a two-sided markets perspective. *Sinergie Italian Journal of Management*, 106:181–201, 2018.

Yongzheng Jia, Wei Xu, and Xue Liu. An optimization framework for online ride-sharing markets. In *2017 IEEE 37th international conference on distributed computing systems (ICDCS)*, pp. 826–835. IEEE, 2017.

Tze Leung Lai, Herbert Robbins, et al. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Yuting Lan and Ning Zhang. Strong limit theorems for weighted sums of negatively associated random variables in nonlinear probability. *arXiv preprint arXiv:1706.05788*, 2017.

Tor Lattimore and Csaba Szepesvári. *The Explore-Then-Commit Algorithm*, pp. 75–83. Cambridge University Press, 2020. doi: 10.1017/9781108571401.009.

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

Lydia T Liu, Feng Ruan, Horia Mania, and Michael I Jordan. Bandit learning in decentralized matching markets. *The Journal of Machine Learning Research*, 22(1):9612–9645, 2021.

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

Christoph Mittendorf. Trust and distrust in two-sided markets: An example in the sharing economy. 01 2018. doi: 10.24251/HICSS.2018.673.

Michael Rothschild. A two-armed bandit theory of market pricing. *Journal of Economic Theory*, 9 (2):185–202, 1974.

Chengchun Shi, Runzhe Wan, Ge Song, Shikai Luo, Rui Song, and Hongtu Zhu. A multi-agent reinforcement learning framework for off-policy evaluation in two-sided markets. *Annals of Applied Statistics.*, 2022.

Richard S. Sutton and Andrew G. Barto. Reinforcement learning: An introduction. *IEEE Transactions on Neural Networks*, 9(5):1054–1054, 1998. doi: 10.1109/TNN.1998.712192.

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

Julian Wright. One-sided logic in two-sided markets. *Review of Network Economics*, 3(1), 2004. doi: doi:10.2202/1446-9022.1042. URL https://doi.org/10.2202/1446-9022.1042.

Stephen M Zoepf, Stella Chen, Paa Adu, and Gonzalo Pozo. The economics of ride-hailing: Driver revenue, expenses and taxes. *CEEPR WP*, 5(2018):1–38, 2018.

A Appendix

A.1 Hypothesis Test

Considering that the data obtained in market matching is sequential, this section introduces sequential testing within a utility-driven multi-armed bandit framework to more reliably identify the optimal choice.

The platform takes into account the gap $d_1 = |\mu_L - \mu_R|$ between the optimal and suboptimal arms when formulating strategies and effectively managing risk. Our goal is to conduct a hypothesis test to determine whether this gap remains within an acceptable range, ensuring that the disparity between the best and worst arms in the matching set is not excessively large.

In other words, assuming $\mu_L > \mu_R$, we would like to conduct the hypothesis test:

$$\mathbf{H}_0 : \mu_L - \mu_R \geq d_0; \quad \mathbf{H}_1 : \mu_L - \mu_R < d_0 \tag{11}$$

Without loss of generality, we assume that the means of the optimal and suboptimal arms, μ_L and μ_R, satisfy $\mu_L + \mu_R = d > 0$, where d is a constant. This constant d represents our focus on understanding how the distance between the two arms affects the concentration of information. In other words, when the total reward of the two arms reaches a certain level, the agent will place more emphasis on the differences in arm rewards to manage risk. For example, when the total is 7, the situation with reward distributions of means 1 and 6 differs from the situation with means 4 and 3. Each corresponds to a left or right margin:

$$\begin{aligned} &\mathbf{H}_{a0} : \mu_L \geq (d + d_0)/2; \quad \mathbf{H}_{a1} : \mu_L < (d + d_0)/2 \\ &\mathbf{H}_{b0} : \mu_R \leq (d - d_0)/2; \quad \mathbf{H}_{b1} : \mu_R > (d - d_0)/2. \end{aligned}$$

The agent can naturally utilize traditional statistical tests based on the normal distribution Fisher (1970). However, the test statistic does not consider the strategy or sample performance, nor does it leverage prior information. This simple approach exhibits lower statistical power and requires more samples.

Given the challenges of performing statistical inference on sequential data using normal tests, we provide the corresponding test statistic and its asymptotic distribution based on the objective function of the utility-driven algorithm:

$$T^{\boldsymbol{\vartheta}_i}_{test,i,m,n} = \frac{1}{n}\sum_{j=1}^{m} X^{\boldsymbol{\vartheta}_i}_{i,j} + \frac{1}{\sqrt{n}}\sum_{j=1}^{m} \frac{X^{\boldsymbol{\vartheta}_i}_{i,j} - \mu^{\boldsymbol{\vartheta}_i}_{\text{test},j}}{\sigma^{\boldsymbol{\vartheta}_i}_{j}}, \tag{12}$$

where

$$\mu_{test,j}^{\vartheta_i} = \frac{d+d_0}{2} I_{\{m_n(i)=a_1\}} + \frac{d-d_0}{2} I_{\{m_n(i)=a_2\}}. \tag{13}$$

Theorem A.1. *Let $\varphi \in C(\overline{\mathbb{R}})$ be a continuous function on $\mathbb{R}$ with finite limits at $\pm\infty$, and be symmetric with centre $c \in \mathbb{R}$ and monotone on (c, ∞). The limit distributions of $\left\{T_{test,i,n,n}^{\vartheta_i}\right\}$ are Bandit distributed. That is*

$$\lim_{n\to\infty} \mathbb{E}\left[\varphi\left(T_{test,i,n,n}^{\vartheta_i}\right)\right] = \mathbb{E}\left[\varphi\left(\eta_n\right)\right]$$

where $\eta_n \sim \mathcal{B}\left(\alpha_n, \frac{d}{2}, c\right)$ and $\alpha_n = -\frac{d_1}{2} - \frac{\sqrt{n}(d_1-d_0)}{2\sigma}$.

When assuming $\mathbf{H}_0$ has a true value of $\mu_L - \mu_R = d_0$, the test statistic $T_{\text{test},i,n,n}^{\vartheta^*}$ follows the spike distribution $\mathcal{B}\left(-\frac{d_0}{2}, \frac{d}{2}, \frac{d}{2}\right)$. Consequently, we can reject the null hypothesis by the occurrence of event

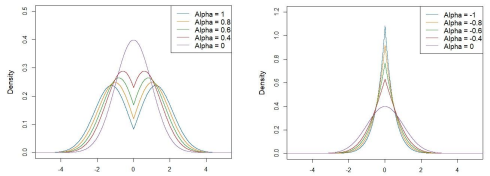
$$\left\{\left|T_{test,i,n,n}^{\vartheta_i} - \frac{d}{2}\right| > z_{\frac{\alpha}{2}}\right\}, \tag{14}$$

where $z_{\frac{\alpha}{2}}$ is the upper α th of the distribution $\mathcal{B}\left(-\frac{d_0}{2}, 0, 0\right)$. The related statistical efficiency can be calculated by

$$1-\alpha = \Phi\left(\frac{d_0}{2} + z_{\frac{\alpha}{2}}\right) - e^{-d_0 z_{\frac{\alpha}{2}}} \Phi\left(\frac{d_0}{2} - z_{\frac{\alpha}{2}}\right). \tag{15}$$

When the distance between the two arms, $\mu_L - \mu_R = d_1 > d_0$, the first parameter

$$\alpha_n = -\frac{d_1}{2} - \frac{\sqrt{n}\left(d_1 - d_0\right)}{2\sigma} < -\frac{d_1}{2}. \tag{16}$$



(a) $\alpha_n \geq 0, \beta = c = 0$ (b) $\alpha_n \leq 0, \beta = c = 0$

Figure 6: The different parameters affecting the density of the bandit distribution.

At this juncture, depicted in Figure 6, as α_n diminishes, the Bandit distribution undergoes increased steepness, resulting in heightened concentration of information and thereby enhancing statistical efficiency.

When $\mathbf{H}_1$ is true with a value of $\mu_L - \mu_R = d_1 < d_0$, the associated statistical power is calculated as follows:

$$\begin{aligned} 1-\beta_1 &= \lim_{n\to\infty} P\left(\left|T_{test,i,n,n}^{\vartheta_i} - \frac{d}{2}\right| > z_{\frac{\alpha}{2}} \,\middle|\, \mathbf{H}_1\right) \\ &= 1 - \Phi\left(-\alpha_n + z_{\alpha/2}\right) + e^{2\alpha_n z_{\alpha/2}} \Phi\left(-\alpha_n - z_{\alpha/2}\right). \end{aligned} \tag{17}$$

When α_n is positive, the statistic follows a bimodal distribution, as illustrated in Figure 6, which significantly amplifies the tail probability. This marked reduction in noise interference makes it a

more effective option for hypothesis testing compared to the conventional normal test. In the context of this utility-driven bandit learning framework, hypothesis testing is performed by integrating both prior information and the adaptive nature of the bandit process, thereby improving the test's power and convergence rate.

This paper addresses the challenge of efficiently allocating resources between two distinct agent sets (e.g., buyers and sellers or drivers and passengers) with differing preferences. The utility-driven bandit algorithm provides a real-time learning framework that adapts strategies to maximize overall utility. By introducing sequential hypothesis testing within this framework, we offer a more reliable method to identify the optimal choice in market matching. This approach improves statistical power, convergence speed, and risk management, outperforming traditional normal tests by better handling dynamic data and optimizing decision-making.

A.2 TABLES

Table 1: Illustration of the preferences of agents and arms when agents have consistent expected utility.

p_1	p_2	a_1	a_2
[64,66]	[64,66]	$p_1 \succ p_2$	$p_1 \succ p_2$
[64,66]	[64,66]	$p_1 \succ p_2$	$p_2 \succ p_1$
[69,71]	[69,71]	$p_1 \succ p_2$	$p_1 \succ p_2$
[69,71]	[69,71]	$p_1 \succ p_2$	$p_2 \succ p_1$
[62,64]	[62,64]	$p_1 \succ p_2$	$p_1 \succ p_2$
[62,64]	[62,64]	$p_1 \succ p_2$	$p_2 \succ p_1$

Table 2: Illustration of the preferences of agents and arms when arms have consistent preferences.

p_1	p_2	p_3	a_1	a_2	a_3
[69,71]	[64,66]	[59,61]	$p_1 \succ p_2 \succ p_3$	$p_1 \succ p_2 \succ p_3$	$p_1 \succ p_2 \succ p_3$
[69,71]	[68,72]	[67,73]	$p_1 \succ p_2 \succ p_3$	$p_1 \succ p_2 \succ p_3$	$p_1 \succ p_2 \succ p_3$
[69,71]	[64,66]	[64,66]	$p_1 \succ p_2 \succ p_3$	$p_3 \succ p_2 \succ p_1$	$p_1 \succ p_3 \succ p_2$

Table 3: Illustration of comparing convergence steps of ETC, UCB, and P-learning across various standard deviations.

	ETC	UCB	P-learning		
			d=0.75	d=1	d=2
$\sigma = 1.1$	142	140	120	54	9
$\sigma = 1.3$	150	147	126	57	10
$\sigma = 1.5$	156	152	130	61	10

A.3 ASSUMPTION

Assume the preferences of the arms are fixed and known.

Assume that the agent ranks the arms based on MV statistics and selects the highest-ranked as the preferred matching arm.

Assume that all arms are equally desirable to the agents in a multi-agent setting.

A.4 PROOF OF REGRET BOUND

Since the regret of the P-learning algorithm is formulated within a probabilistic framework, replacing probabilities with frequencies can introduce biases when the horizon length is n. Therefore, we first establish the proof that when the number of simulation iterations S is sufficiently large, using frequencies as a substitute for probabilities becomes reliable, ensuring the preservation of stability.

Lemma A.2. ***(Preservation of Stability.)*** *$\forall \epsilon \geq 0$, we let S denotes the number of simulation iterations and we have $\lim_{S\to+\infty} P\left(\mid \frac{1}{S}\sum_{j=1}^{S} X_j - P\left(\alpha_i \leq T_{i,n,n}^{\boldsymbol{\vartheta_i}} \leq \beta_i\right) \mid \geq \varepsilon \right) = 0$, where $X_j = \mathbb{I}\left(\alpha_i \leq T_{i,j,n}^{\vartheta_{i,j}} \leq \beta_i\right)$, measuring whether the MV statistic falls in the expected utility.*

PROOF. At time step n, we denote $T_{i,j,n}^{\vartheta_{i,j}}$, for all $j = 1, \ldots, S$, as the value of MV statistics obtained from the jth simulation for agent p_i. Since the variables $T_{i,j,n}^{\vartheta_{i,j}}$ are mutually independent, the sequence of random variables $X_1, \ldots, X_S$ is i.i.d.. Given the definition of random variable X, it can only take on the values 0 or 1. Consequently, X_j follows a Bernoulli distribution, denoted as $X_j \sim B(1, P)$, where the probability parameter P corresponds to $P\left(\alpha_i \leq T_{i,n,n}^{\boldsymbol{\vartheta}_i} \leq \beta_i\right)$. Considering the formulas for the mean and variance of the binomial distribution, as well as the finiteness of probabilities, we can easily establish the validity of the conclusion using Chebyshev's Law of Large Numbers. The following proofs are constructed on the assumption that S is sufficiently large.

A.4.1 SINGLE-AGENT REGRET

We begin by constructing the proof for the single-step regret bound and subsequently extend it to the cumulative regret.

PROOF OF THEOREM 4.1. Firstly, we acknowledge that the expectation of an indicator function is equivalent to the probability. Nevertheless, considering the necessary conditions established by Chen et al. (2022) concerning the thrice differentiability and boundedness of the function φ_h, we commence by approximating the indicator function using a trinomial model Chen et al. (2023a). This approximation allows us to subsequently leverage Chen et al. (2022) established proof framework.

Therefore, we continue to use $\{H_t(x)\}_{t\in[0,1]}$ to denote the functions defined in Chen et al. (2023a) with φ_h and $\alpha = -(\mu_L - \mu_R)/2 = \mu_R$ there. And let $T_{i,m-1,n}$ and$\{L_{m,n}(x)\}_{m=1}^{n}$ be functions defined in Chen et al. (2022) with $\{H_t(x)\}_{t\in[0,1]}$ here.

We first consider the case that $\mu_L = -\mu_R$ and let μ_L is the larger mean of the arms. For agent p_i, let $\boldsymbol{\vartheta}_i$ be the strategy defined in Equation (6) and naturally $\eta = T_{i,\infty,\infty}^{\boldsymbol{\vartheta}_i} \sim \mathcal{B}(\mu_R, 0, c)$ by direct calculation we obtain

$$
\begin{aligned}
R_i(n) &\leq \left| \sup_{\boldsymbol{\vartheta}_i \in \Theta} P\left(\alpha_i \leq T^{\boldsymbol{\vartheta}_i}_{i,\infty,\infty} \leq \beta_i\right) - P\left(\alpha_i \leq T^{\boldsymbol{\vartheta}_i}_{i,n,n} \leq \beta_i\right) \right| \\
&= \left| E_P\left[\varphi_h\left(T^{\boldsymbol{\vartheta}_i}_{i,n,n}\right)\right] - E_P\left[\varphi_h(\eta)\right] \right| \\
&= \left| E_P\left[H_1\left(T^{\boldsymbol{\vartheta}_i}_{i,n,n}\right)\right] - H_0(0) \right| \\
&\leq \left| E_P\left[H_1\left(T^{\boldsymbol{\vartheta}_i}_{i,n,n}\right)\right] - E_P\left[H_{\frac{n-1}{n}}\left(T^{\boldsymbol{\vartheta}_i}_{i,n-1,n}\right)\right] \right| \\
&+ \left| E_P\left[H_{\frac{n-1}{n}}\left(T^{\boldsymbol{\vartheta}_i}_{i,n-1,n}\right)\right] - E_P\left[H_{\frac{n-2}{n}}\left(T^{\vartheta_i}_{i,n-2,n}\right)\right] \right| + \dots \\
&+ \left| E_P\left[H_{\frac{m}{n}}\left(T^{\boldsymbol{\vartheta}_i}_{i,m,n}\right)\right] - E_P\left[H_{\frac{m-1}{n}}\left(T^{\boldsymbol{\vartheta}_i}_{i,m-1,n}\right)\right] \right| + \dots \\
&+ \left| E_P\left[H_{\frac{1}{n}}\left(T^{\boldsymbol{\vartheta}_i}_{i,1,n}\right)\right] - H_0\left(T^{\boldsymbol{\vartheta}_i}_{i,0,n}\right) \right| \\
&\leq \sum_{m=1}^{n} \left| E_P\left[H_{\frac{m}{n}}\left(T^{\boldsymbol{\vartheta}_i}_{i,m,n}\right)\right] - E_P\left[H_{\frac{m-1}{n}}\left(T^{\boldsymbol{\vartheta}_i}_{i,m-1,n}\right)\right] \right| \\
&\leq \sum_{m=1}^{n} \left| E_P\left[H_{\frac{m}{n}}\left(T^{\boldsymbol{\vartheta}_i}_{i,m,n}\right)\right] - E_P\left[L_{m,n}\left(T^{\boldsymbol{\vartheta}_i}_{i,m-1,n}\right)\right] \right| \\
&+ \sum_{m=1}^{n} \left| E_P\left[L_{m,n}\left(T^{\boldsymbol{\vartheta}_i}_{i,m-1,n}\right)\right] - E_P\left[H_{\frac{m-1}{n}}\left(T^{\boldsymbol{\vartheta}_i}_{i,m-1,n}\right)\right] \right| \\
&:= I_{1n} + I_{2n}.
\end{aligned} \tag{18}
$$

The aforementioned inequality will be decomposed into two separate components.

An application of in Chen et al. (2022) implies that $E_P\left[\Gamma\left(m,n,\boldsymbol{\vartheta}_i\right)\right] = E_P\left[L_{m,n}\left(T^{\boldsymbol{\vartheta}_i}_{i,m-1,n}\right)\right]$, we obtain that

$$
\begin{aligned}
|I_{1n}| &= \sum_{m=1}^{n} \left| E_P\left[H_{\frac{m}{n}}\left(T^{\boldsymbol{\vartheta}_i}_{i,m,n}\right)\right] - E_P\left[L_{m,n}\left(T^{\boldsymbol{\vartheta}_i}_{i,m-1,n}\right)\right] \right| \\
&= \sum_{m=1}^{n} \left| E_P\left[H_{\frac{m}{n}}\left(T^{\boldsymbol{\vartheta}_i}_{i,m,n}\right)\right] - E_P\left[\Gamma\left(m,n,\boldsymbol{\vartheta}_i\right)\right] \right| \\
&\leq \sum_{m=1}^{n} \frac{L_1}{2} \sup_{\boldsymbol{\vartheta}_i \in \Theta} E_P\left[\left| \frac{X^{\boldsymbol{\vartheta}_i}_{i,m}}{n} \right|^2 + 2 \left| \frac{X^{\boldsymbol{\vartheta}_i}_{i,m}}{n} \right| \left| \frac{X^{\boldsymbol{\vartheta}_i}_{i,m} - \xi_{i,m}}{\sigma\sqrt{n}} \right| \right. \\
&\left. + \left| \frac{X^{\boldsymbol{\vartheta}_i}_{i,m}}{n} + \frac{X^{\boldsymbol{\vartheta}_i}_{i,m} - \xi_{i,m}}{\sigma\sqrt{n}} \right|^3 \right] \\
&\leq \sum_{m=1}^{n} \frac{L_1}{2} \left(\frac{1}{n^2} + \frac{4}{\sigma n^{\frac{3}{2}}} + \frac{4}{n^3} + \frac{32}{\sigma^3 n^{\frac{3}{2}}} \right) \leq \frac{C_1}{\sqrt{n}},
\end{aligned} \tag{19}
$$

where the penultimate inequality is due to the uniform boundness of $\left\{X^{\boldsymbol{\vartheta}_i}_{i,n}\right\}$, and C_1 depends only on upper boundary of $H_t(x)$ and σ within the reward of distribution.

Further, we obtain that

$$
\begin{aligned}
&|I_{2n}| \leq \sum_{m=1}^{n} \sup_{x\in\mathbb{R}} \left| L_{m,n}(x) - H_{\frac{m-1}{n}}(x) \right| \\
&= \sum_{m=1}^{n} \sup_{x\in\mathbb{R}} \left| H_{\frac{m-1}{n}}(x) - H_{\frac{m}{n}}(x) - \frac{\bar{\mu}}{n} \right| \dot{H}_{\frac{m}{n}}(x) \left| - \frac{1}{2n} \ddot{H}_{\frac{m}{n}}(x) \right| \\
&\leq \sum_{m=1}^{n} \sup_{x\in\mathbb{R}} E_P \left[\int_{\frac{m-1}{n}}^{\frac{m}{n}} |\alpha| \left| \dot{H}_{\frac{m}{n}} \left(Y_s^{\frac{m-1}{n},x,\alpha,c} \right) - \dot{H}_{\frac{m}{n}}(x) \right| ds \right. \\
&\left. + \frac{1}{2} \int_{\frac{m-1}{n}}^{\frac{m}{n}} \left| \ddot{H} \frac{m}{n} \left(Y_s^{\frac{m-1}{n},x,\alpha,c} \right) - \ddot{H} \frac{m}{n}(x) \right| ds \right] | \\
&\leq \sum_{m=1}^{n} \sup_{x\in\mathbb{R}} \frac{L_2}{n} E_P \left[\sup_{s\in\left[\frac{m-1}{n},\frac{m}{n}\right]} \left| Y_s^{\frac{m-1}{n},x,\alpha,c} - x \right| \right] \\
&\leq \sum_{m=1}^{n} \frac{L_2}{n} E_P \left[\frac{|\alpha|}{n} + \sup_{s\in\left[\frac{m-1}{n},\frac{m}{n}\right]} \left| B_s - B_{\frac{m-1}{n}} \right| \right] \\
&\leq L_2 \left(\frac{|\alpha|}{n} + \frac{1}{\sqrt{n}} \right) \\
&:= \frac{C_2}{n} + \frac{C_3}{\sqrt{n}},
\end{aligned}
\tag{20}
$$

where C_2 and C_3 is a constant depending only on α, L and the bound of $\ddot{H}_t(x)$.

Above all, we obtain that

$$
R_i(n) \leq \frac{K1}{n} + \frac{K2}{\sqrt{n}} \leq \frac{M}{\sqrt{n}}, \tag{21}
$$

where K_1, K_2 are constants depending only on σ and $\alpha = -(\mu_L - \mu_R)/2$ within the reward distribution of the arms.

Then the results asserted for general μ_L and μ_R are established by applying the preceding special case to $\left\{ Y_{i,n}^{\vartheta_i} : n \geq 1 \right\}$, where $Y_{i,n}^{\vartheta_i} = X_{i,n}^{\vartheta_i} - (\mu_L + \mu_R)/2$.

Therefore, Theorem 4.1 shows us that single-agent P-learning achieves $\mathcal{O}(1/\sqrt{n})$ problem-independent regret. The following theorem leads to the upper limit of cumulative regret.

PROOF OF COROLLARY 4.2. This is due to if the growth rate of single step regret is $O(1/\sqrt{n})$, then the cumulative regret growth in the first n steps is upper-bounded. We can derive this conclusion by summing up the growth of single-step regrets. The result shows that the growth rate of cumulative regret in the first n steps is $O(\sqrt{n})$. In other words, as the time step n increases, the growth rate of cumulative regret will remain at the level of $O(\sqrt{n})$.

A.4.2 MULTI-AGENT REGRET

PROOF OF LEMMA 4.4. Since the strategy $\vartheta_{i,n}$ is formulated based on the symmetry center, under a strategy-driven algorithm, the distribution of the MV statistics of agent p_1 should be centered around c_1. Therefore, there exists $\gamma_1, \gamma_2 \geq 0, s.t. c_1 = \gamma_1\mu_1 + \gamma_2\mu_2$ with $\gamma_1 + \gamma_2 = 1$. Naturally, $c_1' = \gamma_2\mu_1 + \gamma_1\mu_2$. It follows from the direct calculation that

$$
\begin{aligned}
c_1 &= \gamma_1\mu_1 + \gamma_2\mu_2 \\
&= (1-\gamma_2)\mu_1 + (1-\gamma_1)\mu_2 \\
&= \mu_1 + \mu_2 - (\gamma_2\mu_1 + \gamma_1\mu_2) \\
&= \mu_1 + \mu_2 - c_1'.
\end{aligned}
\tag{22}
$$

The unpreferred agent p_2 will experience strategy failure. Lemma 4.4 tells us that in the worst-case scenario, the actual reward distribution obtained by agent p_2 corresponds to the counter-strategy distribution of agent p_1. It can be observed that if there is a significant difference between c_2 and c_1', taking into account the impact of the expected utility interval length d, this will affect the regret of p_2.

PROOF OF THEOREM 4.5. The following proof is established under the assumption of consistent ranked arms. And we let arms prioritize selecting agent p_1. Therefore, the agent p_1's regret regarding his expected utility will not be affected by the matching. On the contrary, as agent p_2's strategy may be subject to interference, we can denote $\widetilde{\vartheta}_2$ as the actual strategy implemented by agent p_2. The inequality will be decomposed into two separate components.

$$\begin{aligned} R_i(n) &\leq \left| \sup_{\boldsymbol{\vartheta}_i \in \Theta} P\left(\alpha_i \leq T_{i,\infty,\infty}^{\boldsymbol{\vartheta}_i} \leq \beta_i\right) - P\left(\alpha_i \leq T_{i,n,n}^{\tilde{\boldsymbol{\vartheta}}_i} \leq \beta_i\right) \right| \\ &\leq \left| \sup_{\boldsymbol{\vartheta}_i \in \Theta} P\left(\alpha_i \leq T_{i,\infty,\infty}^{\boldsymbol{\vartheta}_i} \leq \beta_i\right) - P\left(\alpha_i \leq T_{i,n,n}^{\boldsymbol{\vartheta}_i} \leq \beta_i\right) \right| \\ &+ \left| P\left(\alpha_i \leq T_{i,n,n}^{\boldsymbol{\vartheta}_i} \leq \beta_i\right) - P\left(\alpha_i \leq T_{i,n,n}^{\tilde{\boldsymbol{\vartheta}}_i} \leq \beta_i\right) \right| \\ &:= I_{3n} + I_{4n}. \end{aligned} \tag{23}$$

An application of Lemma 4.4 implies that $I_{3n} = I_{1n} \leq \frac{K_1}{\sqrt{n}} + \frac{K_2}{n}$. And we let F_n and P_n denote the distribution function of MV statistics by $\boldsymbol{\vartheta}_i$ and $\widetilde{\vartheta}_i$. Utilizing the Kolmogorov inequality divergence inequality, we obtain

$$\begin{aligned} I_{4n} &\leq \sup_{x \in \lfloor \alpha_i, \beta_i \rfloor} |F_n(x) - P_n(x)| \\ &\leq \frac{1}{2} \int_{-\infty}^{+\infty} \frac{|dF_n(x) - dP_n(x)|}{1 - F_n(x) + P_n(x)} \\ &\leq \frac{1}{2\sqrt{n}} \sqrt{\int_{-\infty}^{+\infty} \frac{|dF_n(x) - dP_n(x)|}{1 - F_n(x) + P_n(x)}} \\ &\leq \frac{K3}{\sqrt{n}} \sqrt{\ln\left(\frac{1}{\Delta_n^2}\right)}, \end{aligned} \tag{24}$$

where K_3 denotes problem-independent constants and $\Delta_n = F_n(c_1) - \gamma$ measures the distance between the anti distribution and the ideal distribution, and $\gamma = \min\{P_n(\Delta_c - d/2), P_n(\Delta_c + d/2)\}$ with $\Delta_c = |c_1' - c_2|$.

Above all, we obtain that

$$R_i(n) \leqslant \frac{K_1}{\sqrt{n}} + \frac{K_2}{n} + \frac{K_3}{\sqrt{n}} \sqrt{\ln(\frac{1}{\Delta_n^2})} \leq \sqrt{\frac{K \ln(1/\Delta_n^2)}{n}}, \tag{25}$$

PROOF OF COROLLARY 4.6. Proof same as in Theroem 4.1. Therefore, we can express the growth upper bound of $CR_i(n)$ as $O(\sqrt{n})$.

Although the growth rate of both single-agent and multi-agent systems is $\mathcal{O}(\sqrt{n})$, we have demonstrated that the constant factor limitation in single-agent scenarios is irrelevant to the problem, i.e., in terms of the maximum meaningful growth rate. Under the constraint of consistent agents, the growth rate of multiple agents is also $\mathcal{O}(\sqrt{n})$, and it is influenced by the mutual interaction Δ_n among the agents.

A.5 ALGORITHM

Algorithm 2 P-learning with Competition (1 vs k)

Input: current round m, the current grouping r , the set of arms $\mathcal{K}$ and the preferences of arms
Initialize the set of alternatives $\mathcal{K}' = \mathcal{K}$
while $r \leq \lfloor (k+1)/2 \rfloor$ **do**
  Get agent p_i's preference of best and worst arm a_1^*, a_2^* in set $\mathcal{K}'$ based on Equation 7
  Reset $\mathcal{K}' \leftarrow \mathcal{K}/\{a_1^*, a_2^*\}$
  $r \leftarrow r + 1$
end while
Get final preference $\vartheta_{1,m}$order with all arms for agent p_i

Algorithm 3 P-learning with Competition (k vs k)

Input: horizon length n, current round m, the set of arms $\mathcal{K}$, the set of agents $\mathcal{N}$ and the preferences of arms
Output: regret $R_i(n)$
while $m \leq n$ **do**
  Get the $v_{i,m}$ fot the agent p_i based on Algorithm 2
  Get matching vector m_m matching with Gale-Shapley algorithm
end while
return $R_i(n)$ based on Equation (2)