You Can Trade Your Experience in Distributed Multi-Agent Multi-Armed Bandits

Guoju Gao†, He Huang†*, Jie Wu‡, Sijie Huang†, and Yang Du†
†School of Computer Science and Technology, Soochow University, P. R. China
‡Department of Computer and Information Sciences, Temple University
*Correspondence to: huangh@suda.edu.cn

***Abstract*—Multi-Armed Bandit (MAB) that solves the sequential decision-making to the prior-unknown settings has been extensively studied and adopted in various applications such as online recommendation, transmission rate allocation, *etc.* Although some recent work has investigated the multi-agent MAB model, they supposed that agents share their bandit information based on social networks but neglected the incentives and arm-pulling budget for heterogeneous agents. In this paper, we propose a *transaction-based multi-agent MAB* framework, where agents can trade their bandit experience with each other to improve their total individual rewards. Agents not only face the dilemma between exploitation and exploration, but also decide to post a suitable price for their bandit experience. Meanwhile, as a buyer, the agent accepts another agent whose experience will help her the most, according to the posted price and her risk-tolerance level. The key challenge lies in that the arm-pulling and experience-trading decisions affect each other. To this end, we design the *transaction-based upper confidence bound* to estimate the prior-unknown rewards of arms, based on which the agents pull arms or trade their experience. We prove the regret bound of the proposed algorithm for each independent agent and conduct extensive experiments to verify the performance of our solution.**

***Index Terms*—Multi-agent multi-armed bandits, upper confidence bound, experience transaction, posted pricing mechanism.**

I. Introduction

Recently, the Multi-Armed Bandit (MAB) model has been extensively studied due to its wide range of applications, *e.g.*, online recommendation [1], transmission rate allocation [2], crowdsourcing user selection [3], *etc.* In the basic MAB model, the learning agent (*a.k.a.* player or decision-maker) can pull one arm in each round and obtain the *i.i.d.* (independent and identically distributed) random rewards. The objective of the agent is to maximize the total cumulative rewards within a finite time horizon. The learning agent has to face the dilemma between exploitation and exploration in the MAB model. The exploitation means that the agent prefers to pull the arm that had the best performance in the past, while the exploration indicates that the agent will also try some other arms so as to find the potentially optimal arm which will generate the highest rewards in the future. Much effort has been devoted to the basic MAB problem, and some famous algorithms such as Upper Confidence Bound (UCB) [4, 5], epoch-based UCB [6], Thompson sampling [7], *etc.*, have been proposed.

Moreover, lots of research has considered various extensions to the basic MAB model, *e.g.*, the combinatorial concept [8, 9], fairness [10], delayed arm-pulling feedback [11], contextual bandits [12], long-term returns [13], *etc.* Particularly, a variant of the basic MAB problem, called multi-agent MAB model [14, 15], has been recently put forward. Many practical application scenarios such as the multi-user multi-channel selection problem [2, 16] can be modeled as a multi-agent MAB setting. In the multi-agent MAB model, multiple agents face the same instance of a multi-armed bandit problem and study how to maximize the total aggregate rewards. In general, there are two common settings in the multi-agent MAB model: competition [17, 18] and collaboration [19, 20]. The difference is when multiple agents pull the same arm, each agent will obtain discounted or no rewards in the former setting, while she will receive independent rewards in the latter setting.

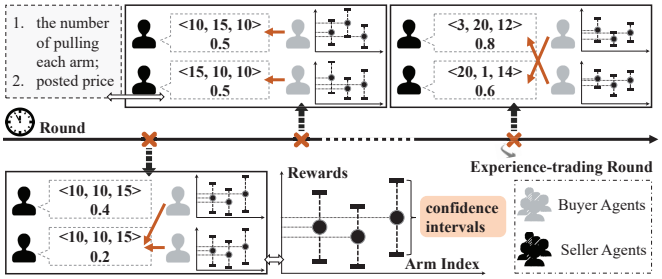

Fig. 1. The experience transaction in the TMA^{2}B framework.

Although some existing work has studied the multi-agent MAB model, that work was built on the premise that each agent will share their bandit experience via the social networks that already exist by default [15, 20–22]. On the one hand, we argue that social-network-based communication cannot totally stimulate agents to share their MAB information. One agent is considered to only communicate with her network neighbors, which might result in two agents with similar or redundant arm-pulling experiences communicating with each other. At the same time, one agent is required to share her experience with all of her network neighbors rather than selectively communicating. These will not only waste agents' communication resources but also may not improve their learning performance. On the other hand, the existing work neglects the incentive and arm-pulling budget for heterogeneous agents. They consider that pulling arms is cost-free, which is not practical. Also, the agent who contributes more valuable bandit information should be given more extra rewards so that all agents have incentives to exchange bandit messages, but this significant setting is ignored in the current work.

To this end, we propose a *transaction-based multi-agent MAB* (called TMA^{2}B) model in the collaboration setting, in which heterogeneous agents can trade their bandit experience to others or directly pull arms. The term "heterogeneous" means that each agent can enter the TMA2B model at different

979-8-3503-9973-8/23/$31.00 ©2023 IEEE

moments and has different budgets, risk-tolerance levels, *etc*. Agents with different risk-tolerance levels (*e.g.*, risk-tolerant or risk-averse) will adopt different experience-trading strategies. The TMA^{2}B model consists of arm-pulling and experience-trading rounds, in which a given parameter controls the trading frequency. In the arm-pulling rounds, each agent considers how to combine the bought experience from others into their individual arm-pulling decisions to balance exploitation and exploration. In the experience-trading rounds, each agent can simultaneously act as a seller and buyer. Since the generated rewards of each arm are *i.i.d.* over time, the arm-pulling decisions for agents are generally different. It is difficult for seller agents to post reasonable prices for their bandit experience in the past arm-pulling rounds. As a buyer, an agent can only observe each seller's number of times of pulling each arm in a period and her posted price, as shown in Fig. 1. Each buyer agent with different risk-tolerance levels and local empirical information will have different valuations on every seller's bandit experience so that she will make different trading decisions. It is challenging for buyer agents to select the most appropriate seller to complete the transaction.

For each agent, we combine her bought bandit experience in the past trading rounds with her local empirical information and thus devise a transaction-based UCB (called T-UCB) index for the prior unknown arms. Each agent will pull the arm with the highest ratio of the T-UCB-based index and the corresponding cost in the arm-pulling rounds. In the experience-trading rounds, we model the seller agents' pricing problem as an MAB problem, where each candidate price is seen as a normal arm. For each buyer, we first evaluate the expected confidence interval increase for any one arm based on one seller's published information. Then, each buyer agent will determine the seller to trade with according to her risk-tolerance level and local empirical information.

The contributions of this paper are summarized as follows:

1) We propose a novel transaction-based multi-agent MAB framework, in which each heterogeneous agent can pull arms directly or trade her bandit experience with others to improve learning performance. The arm-pulling and experience-trading decisions influence each other. To the best of our knowledge, we are the first to consider the experience transaction in the multi-agent MAB setting.
2) We re-define the total number of potential arm-pulling opportunities each agent faces by considering this agent's bought bandit experience. We then devise the transaction-based UCB (*i.e.*, T-UCB) for the prior unknown arms to solve the dilemma between exploitation and exploration. Each agent will independently pull arms according to their T-UCB-based index and the arm-pulling cost.
3) We design each agent's selling and buying strategy in the experience-trading rounds. We adopt the UCB idea to solve seller agents' posted pricing problem. We propose an efficient method to calculate the buyer agent's expected incremental rewards and determine the most appropriate seller by considering the buyer's risk-tolerance level and the seller's posted price.
4) We prove the regret bound of the proposed algorithm on each independent agent, showing that one tradeoff exists between the arm-pulling and experience-trading decisions. We conduct extensive experiments to verify the performance of TMA^{2}B, where agents in TMA^{2}B can obtain higher rewards when compared with the classical algorithms (*i.e.*, fractional KUBE and ϵ-first).

The remainder of the paper is organized as follows. We present the TMA^{2}B model and the solution in Section II and Section III, respectively. Next, we evaluate the performance of the proposed solution in Section IV. After reviewing related work in Section V, we conclude this paper in Section VI.

II. Preliminaries

A. TMA^{2}B Model

In the transaction-based multi-agent multi-armed bandit (TMA^{2}B) model, all agents face the same instance of a multi-armed stochastic bandit problem and each agent's objective is to maximize her own total rewards under a given unique budget. We follow the traditional setting that the arm-pulling process is time-slotted. Let t denote the t-th round. In this TMA^{2}B model, we consider that the agents' experience trading actions only happen in certain rounds. Thus, the slotted rounds are divided into two types: arm-pulling rounds and experience-trading rounds. We let ω denote the trading frequency, that is, the agents can trade their bandit experience in the rounds $t = k\cdot(\omega+1)$ for $k \geq 1$. In the rounds $t \neq k\cdot(\omega+1)$, every agent can individually pull one arm from the whole set of arms. We consider a collaboration/cooperation setting [19–21] in which the agents will obtain the independent rewards when they select the same arm.

We use $\mathcal{A} = \{1, 2, \cdots, A\}$ to denote the set of A arms in the TMA^{2}B setting. Since the rewards of arms are *i.i.d.* over time and are agent-independent, we let r_a ($a \in \mathcal{A}$) denote the a-th arm's expected reward. The value of r_a is not observed by any agents, so these agents need to learn the unknown parameters while maximizing their total rewards under the budget constraint. In the round t, if the arm a is pulled by one agent, the reward obtained by the agent is denoted as r_a^t. Here, we have $\mathbb{E}[r_a^t]_{t\geq 1} = r_a$ which indicates that $\{r_a^t | t \geq 1\}$ is *i.i.d.* with an unknown expectation r_a. At the same time, pulling one arm will certainly consume some of the budget resources. Let c_a denote the cost of pulling the arm $a \in \mathcal{A}$.

Moreover, we use $\mathcal{N} = \{1, 2, \cdots, N\}$ to denote the set of N independent agents. Each agent $i \in \mathcal{N}$ has a unique budget and we let B_i denote this budget. In the experience-trading rounds, each agent i can sell her bandit experience of the past ω rounds. We consider the posted pricing mechanism and use the set $\mathcal{P}_i = \{p_{i,1}, \cdots, p_{i,L}\}$ to denote L discrete price values of agent i where $p_{i,1} < \cdots < p_{i,L}$. We define the maximum posted price as $p_{max} = \max_{i\in\mathcal{N}}\{p_{i,L}\}$ and have $p_{max} < \min_{a\in\mathcal{A}} c_a$. Otherwise, the agents have no incentive to buy others' bandit experience. On the one hand, each agent can sell their bandit experience to gain extra revenue. On the other hand, one agent can also improve her confidence intervals in the prior-unknown

rewards of arms by buying others' experience, although this process will consume some of the budget.

According to the existing work [6, 7], we observe that one agent's confidence interval in any one arm will undoubtedly improve with the increase of the number of times this arm is pulled. When facing the arm with a higher observed average reward but a worse confidence interval, each agent with different characteristics (*i.e.*, risk-tolerant or risk-averse) will have different choices. We here use $\varphi_i \in [0,1]$ (for $i \in \mathcal{N}$) to denote the agent i's risk-tolerance level. The larger φ_i, the higher the risks i can take. The parameter φ_i plays a decisive role in the experience-trading process for each agent. An agent will exit the TMA^{2}B framework when she exhausts her budget.

B. Problem Formalization

In the TMA^{2}B model, each agent can accumulate her achieved revenues by two methods: 1) pulling one arm under some budget consumption in one round, and 2) trading her bandit experience of the past ω rounds. In the former case, the agent has to face the dilemma between exploitation and exploration in the traditional MAB problem [6]. Note that an agent can improve her confidence intervals in the prior-unknown reward distribution by buying others' bandit experience. In the latter case, an agent intending to sell her bandit experience needs to decide on the posted price so that she can maximize her extra rewards. For simplicity of following description, the revenue achieved in pulling arms will be called "arm-pulling rewards", while the revenue from trading the bandit experience will be called "experience-trading rewards".

We first introduce the computation of the arm-pulling rewards. For the arm-pulling rounds $t \geq 1$ and $t \neq k \cdot (\omega+1)$, we let $\pi_{i,a}^t = 1$ denote that the arm $a \in \mathcal{A}$ is pulled by the agent i, and $\pi_{i,a}^t = 0$ otherwise. Consider that the agent i will use her part of budget for pulling arms. Let B_i^1 ($\leq B_i$) denote this part, and we have:

$$\sum\nolimits_{t\geq 1 \ \& \ t\neq k\cdot(\omega+1)} \sum\nolimits_{a\in\mathcal{A}} \pi_{i,a}^t \cdot c_a \leq B_i^1.$$

In the experience-trading rounds, *i.e.*, $t = k\cdot(\omega+1)$ for $k \geq 1$, each agent can sell her bandit experience to gain extra rewards or buy others' experience to improve the learning performance. Without loss of generality, we use $s \in \mathcal{N}$ and $i \in \mathcal{N}$ to denote one seller agent and one buyer agent, respectively. Note that each agent can act as a seller and buyer in one round simultaneously. Any one agent s can sell her bandit experience of the past ω rounds. At the beginning of the trading round t, the agent s first publishes the number of times each arm was selected in the past ω rounds, and meanwhile determines her posted price $p_{s,l}^t \in \mathcal{P}_s$. In the next section, we will introduce how to post the suitable price so that each distributed agent can maximize her total experience-trading rewards.

After one seller posts her price, each buyer agent will determine whether to accept this seller's price. If so, the buyer will pay the seller $p_{s,l}^t$ and get the further information, *i.e.*, this seller's empirical rewards for each arm in the past ω rounds. We use $\gamma_{s,i}(p_{s,l}^t) = 1$ to denote the indicator that the buyer i will accept the price $p_{s,l}^t$ posted by the seller s in the trading round t, and $\gamma_{s,i}(p_{s,l}^t) = 0$ otherwise. Based on this, we can determine the total number of buyers that accept the price posted by agent s, that is, $\sum_{i\in\mathcal{N}} \gamma_{s,i}(p_{s,l}^t)$. Further, we can calculate the rewards from trading experience in this round, *i.e.*, $p_{s,l}^t \cdot \sum_{i\in\mathcal{N}} \gamma_{s,i}(p_{s,l}^t)$. Next, we compute the total experience-trading rewards in the trading rounds, denoted as $\widetilde{R}_s$, that is,

$$\widetilde{R}_s = \sum\nolimits_{t=k\cdot(\omega+1)} \Big(p_{s,l}^t \cdot \sum\nolimits_{i\in\mathcal{N}} \gamma_{s,i}(p_{s,l}^t) \Big),$$

where $k \geq 1$ and $p_{s,l}^t \in \mathcal{P}_s$. If an agent does not pull any arms in the past ω rounds, she cannot act as a seller. However, since TMA^{2}B supports the asynchronous start for agents, the agent who just enters the TMA^{2}B framework can buy others' bandit experience to complete her local initialization. The trading action terminates when the agent exits the TMA^{2}B model, *i.e.*, her budget B_i exhausts. At the same time, we also calculate a buyer agent's total cost of buying bandit experience from others (denoted as B_i^2), *i.e.*,

$$B_i^2 = \sum\nolimits_{t=k\cdot(\omega+1)} \sum\nolimits_{s\in\mathcal{N}} \gamma_{s,i}(p_{s,l}^t) \cdot p_{s,l}^t.$$

We let $R_i(B_i^1)$ denote the total arm-pulling rewards under the budget B_i^1 and have

$$R_i(B_i^1) = \sum\nolimits_{t\geq 1 \ \& \ t\neq k\cdot(\omega+1)} \sum\nolimits_{a\in\mathcal{A}} \pi_{i,a}^t \cdot r_a^t.$$

The objective of each agent is to maximize the total expected rewards under the given budget by independently pulling arms and trading her periodical bandit experience. Note that the arm-pulling and experience-trading stages are not mutually independent, because other agents' bandit experience will help the agent better pull the arms with high rewards. In the TMA^{2}B model, each agent (*e.g.*, i) has the same goal, and we formalize the optimization problem as follows:

$$\begin{aligned}
Maximize: \quad & \widetilde{R}_i + R_i(B_i^1) && (1)\\
Subject\ to: \quad & B_i^1 + B_i^2 \leq B_i && (2)\\
& \textstyle\sum_{a\in\mathcal{A}} \pi_{i,a}^t = 1 && (3)\\
& \textstyle\sum_{s\in\mathcal{N}} \gamma_{s,i}(p_{s,l}^t) \leq 1 && (4)\\
& \gamma_{s,i}(p_{s,l}^t), \pi_{i,a}^t \in \{0,1\},\ p_{s,l}^t \in \mathcal{P}_s && (5)
\end{aligned}$$

Eq. (2) means that the sum of arm-pulling cost and the cost of buying bandit experience from others cannot exceed the agent's given budget; Eq. (3) indicates that each agent can only pull one arm in each round; Eq. (4) shows that each buyer agent can select at most one seller to trade with in one round.

In the TMA^{2}B model, the arm-pulling process and the experience-trading process are mutually influenced. When one agent consumes more budget during the aspect of her independent arm-pulling, she might miss others' valuable but cheap bandit experience. In other words, she must put more effort into the tradeoff between exploitation and exploration in the traditional MAB problem. When one agent allocates more budget to the trading process, she will undoubtedly decrease the budget allocated to the arm-pulling process. The advantage is that the agent can observe others' experience,

thus increasing the probability of selecting the optimal arms in the arm-pulling process. The following section will introduce how to make optimal decisions for each independent agent. Additionally, we present the commonly-used notations of this paper in Table I.

III. Solution to TMA^{2}B

A. Arm-pulling Stage

Since all independent and rational agents face the same instance of A-armed stochastic bandit settings, without loss of generality, we let i denote any one agent. In any round t, we adopt the idea of upper confidence bound (UCB) to determine the selected arm. Unlike the traditional MAB problem, where the agent can only depend on her local empirical arm-pulling information, in the TMA^{2}B model, one agent can expend little cost to obtain others' bandit experience as her prior knowledge before round t. In any experience-trading round t, when the agents trade successfully, *i.e.*, $\gamma_{s,i}(p_{s,l}^t)=1$, we let $R_{s,a}^\star(t)$ and $n_{s,a}^\star(t)$ denote the total rewards and the total number of times of selecting the arm $a\in\mathcal{A}$ by the seller s in the past ω rounds. Now, by combining the arm-pulling empirical information and the experience-trading data, we can calculate the number of times each arm is pulled and deduce the average rewards until the round t, from the perspective of each independent agent i.

We let $\overline{r}_{i,a}(t)$ and $n_{i,a}(t)$ denote the average rewards and the total number of times of selecting the arm $a\in\mathcal{A}$. $n_{i,a}(t)$ and $\overline{r}_{i,a}(t)$ are updated as follows:

$$n_{i,a}(t)=\sum_{\tau=k\cdot(\omega+1)}\sum_{s\in\mathcal{N}}\gamma_{s,i}(p_{s,l}^\tau)\cdot n_{s,a}^\star(\tau)+\sum_{\tau=1}^{t-1}\pi_{i,a}^\tau, \tag{6}$$

$$\overline{r}_{i,a}(t)=\frac{\overbrace{\sum_{\tau=k\cdot(\omega+1)}\sum_{s\in\mathcal{N}}\gamma_{s,i}(p_{s,l}^\tau)\cdot R_{s,a}^\star(\tau)}^{\text{others' bandit experience}}+\overbrace{\sum_{\tau=1}^{t-1}\pi_{i,a}^\tau\cdot r_a^\tau}^{\text{arm-pulling reward}}}{n_{i,a}(t)} \tag{7}$$

where $k=1,2,\cdots,\lfloor\frac{t}{\omega+1}\rfloor$.

At the beginning of the round t, we can get the average empirical rewards of each arm over the last $t-1$ rounds from each agent i's perspective. At the end of this round t, the average rewards will update as $\overline{r}_{i,a}(t+1)$. In the TMA^{2}B model, we combine one agent's local arm-pulling information and her bought bandit experience from others to design the transaction-based UCB (called T-UCB) to the prior-unknown rewards of arms. We first calculate the total number of potential opportunities for i to face the arm a (including other agents' pulling opportunities) until the round t as:

$$T_{i,a}(t)=t+\omega\cdot\sum_{\tau=k\cdot(\omega+1)}\sum_{s\in\mathcal{N}}\gamma_{s,i}(p_{s,l}^\tau)-\lfloor\frac{t}{\omega+1}\rfloor, \tag{8}$$

in which $k=1,2,\cdots,\lfloor\frac{t}{\omega+1}\rfloor$.

In Eq. (8), the second term means the total number of rounds in which other agents have the opportunities to select the arm a according to the historical experience-trading results. The third term indicates that agents are not allowed to pull arms in the experience-trading rounds. We then define the T-UCB based index for each arm $a\in\mathcal{A}$ from the perspective of the agent $i\in\mathcal{N}$ until the round t, denoted as $\widehat{r}_{i,a}(t)$, *i.e.*,

$$\widehat{r}_{i,a}(t)=\overline{r}_{i,a}(t)+\sqrt{\frac{\alpha\cdot\ln(T_{i,a}(t))}{n_{i,a}(t)}}, \tag{9}$$

where α is a given parameter. Here, the T-UCB expression has combined an agent's own arm-pulling information and her bought bandit experience from others.

Each agent independently makes the arm-pulling decision in each round based on her obtained bandit information (including others' and her own). In the arm-pulling round $t\geq 1$ and $t\neq k\cdot(\omega+1)$, each agent will select the arm with the highest ratio of the T-UCB-based index and the corresponding cost. Unlike the traditional MAB settings where the agent needs to initialize by pulling every arm once, the proposed TMA^{2}B framework allows agents to trade other agents' bandit experience to complete the initialization.

TABLE I
Description of major notations.

Variable	Description
$\mathcal{N}$, $\mathcal{A}$	the sets of agents and arms, respectively.
i, a, t	the indexes for agents, arms, and rounds (slots).
c_a	the cost of pulling the arm $a\in\mathcal{A}$.
B_i	the limited budget of the agent $i\in\mathcal{N}$.
φ_i	the agent i's level of risk-tolerance.
$\mathcal{P}_i$	the set of posted price for i, $\mathcal{P}_i=\{p_{i,1},\cdots,p_{i,L}\}$.
ω	the agents' experience-trading frequency.
r_a^t	the observed rewards of the arm a in round t.
r_a	the expected reward of a, *i.e.*, $\mathbb{E}[r_a^t]=r_a$.
$\pi_{i,a}^t$	$\pi_{i,a}^t\in\{0,1\}$ is the agent i's pulling decision.
$\overline{r}_{i,a}(t)$	the average arm-pulling reward until round t.
$n_{i,a}(t)$	the total times a is pulled (observed by i).
$T_{i,a}(t)$	the total number of opportunities of i facing a.
$\widehat{r}_{i,a}(t)$	the agent i's T-UCB-based index until round t.
$R_i(B_i^1)$	i's total arm-pulling rewards under the budget B_i^1.
$\widetilde{R}_i$	the agent i's total experience-trading rewards.
$\gamma_{s,i}(p_{s,l}^t)$	the trading decision indicator between s and i.
$n_{s,a}^\star(t)$	the number of times a is pulled by s in past ω rounds.
$R_{s,a}^\star(t)$	the total rewards of s pulling a in past ω rounds.
$\nabla_{s,i}^a$	i's expected incremental confidence on a based on s.
$v_{s,i}^a$	i's expected risk-tolerant incremental reward for a.

B. Experience-trading Stage

In the TMA^{2}B model, the distributed agents can share their bandit experience in the past ω rounds. Each rational agent will ask for some payment as their rewards, so the agents in the trading rounds are divided into two kinds: buyer and seller. We adopt the posted pricing strategy in the trading process. At the beginning of the experience-trading round t, each seller $s\in\mathcal{N}$ first publishes her posted price and the number of times each arm was selected in the past ω rounds, denoted as $p_{s,l}^t\in\mathcal{P}_s$ and $n_{s,a}^\star(t)$ in which $1\leq l\leq L$ and $a\in\mathcal{A}$. After observing the displayed information of the seller s, other agents will determine whether to accept this seller's posted price according to their local information already on hand. Each agent can simultaneously be a buyer and seller, so multiple options exist for every buyer.

We first introduce the experience-trading decisions from the perspective of a buyer. Despite adopting the same T-

UCB-based arm-pulling strategy, each agent will still make different arm-pulling decisions in the past ω rounds. This is because each arm will generate the *i.i.d.* random rewards when selected by an agent in one round. In other words, each agent has different confidence intervals for each arm. When facing multiple sellers' bandit experience, one buyer will compute her expected incremental confidence intervals for each arm. At the beginning of the trading round $t = k(\omega+1)$, a buyer has the following local information: the number of times each arm is selected until t (including the experience bought in the past), *i.e.*, $n_{i,a}(t)$, and the agent i's average rewards combining the experience-trading information and her own arm-pulling rewards, *i.e.*, $\overline{r}_{i,a}(t)$.

Recall the T-UCB expression in Eq. (9). With the increase of the number of times each arm is pulled by agents, *i.e.*, $n_{i,a}(t)$, the degree of confidence that average empirical reward $\overline{r}_{i,a}(t)$ approaches the expected reward r_a will certainly increase. Until the round t, we let $\Lambda^t_{i,a}$ denote the obtained information of the agent i about the arm a, including her local empirical arm-pulling values and the bought experience. Specifically, $\Lambda^t_{i,a}$ consists of three parts: $\overline{r}_{i,a}(t)$, $T_{i,a}(t)$ and $n_{i,a}(t)$. Now, for each buyer $i \in \mathcal{N}$, she knows each seller's information (*e.g.*, $s \in \mathcal{N}$) including the posted price $p^t_{s,l} \in \mathcal{P}_s$ and the number of times each arm was selected in the past ω rounds, *i.e.*, $n^\star_{s,a}(t)$, and her local information $\Lambda^t_{i,a}$. Before the successful trade between sellers and buyers, each buyer has no knowledge about the average rewards $R^\star_{s,a}(t)/n^\star_{s,a}(t)$, so we let $\overline{r}_{i,a}(t)$ be the estimate of it at first. When given the information $n^\star_{s,a}(t)$ of the seller s in the past ω rounds, we first compute the buyer i's incremental confidence intervals for the arm $a \in \mathcal{A}$, denoted as $\nabla^a_{s,i}$, as follows:

$$\nabla^a_{s,i} = \sqrt{\frac{\alpha \ln(T_{i,a}(t))}{n_{i,a}(t)}} - \sqrt{\frac{\alpha \ln(T_{i,a}(t)+\omega)}{n_{i,a}(t)+n^\star_{s,a}(t)}}. \tag{10}$$

Recall that each agent i has different levels of risk-tolerance φ_i. This indicates that the agent cares about the expected incremental confidence intervals and is also concerned with the average empirical rewards. Thus, we propose a new evaluation metric, denoted as $v^a_{s,i}$, to capture the intrinsic property about the agent i's expected risk-tolerant incremental reward, *i.e.*,

$$v^a_{s,i} = \overline{r}_{i,a}(t)^{(1-\varphi_i)} \cdot \max\{\nabla^a_{s,i}, 0\}^{\varphi_i}. \tag{11}$$

Note that the value of $\nabla^a_{s,i}$ may be less than 0. In such a case, we replace $\nabla^a_{s,i}$ with 0 in Eq. (11). Now, the buyer $i \in \mathcal{N}$ will accept the posted price of the seller $s^\dagger$ who will increase the total incremental rewards for all arms under her risk-tolerant level most quickly. The winning seller $s^\dagger$ is determined as follows:

$$s^\dagger = \text{argmax}_{s\in\mathcal{N}}\Big(\frac{\sum_{a\in\mathcal{A}} v^a_{s,i}}{p^t_{s,l}} \cdot \mathbb{I}\Big\{\sum\nolimits_{a\in\mathcal{A}} \nabla^a_{s,i} \geq \theta\Big\}\Big), \tag{12}$$

where $\mathbb{I}\{true\} = 1$ while $\mathbb{I}\{false\} = 0$, and θ is a given threshold. The second term in Eq. (12) considers the extreme scenario where one agent has high enough degrees of confidence in all arms, such that others' experience cannot improve her confidence intervals for any arms, *i.e.*, $\sum_{a\in\mathcal{A}} \nabla^a_{s,i} < \theta$. In such a case, the agent is unwilling to buy others' bandit experience but only pulls arms by using the remaining budget. After the buyer i selects the agent $s^\dagger$, she will pay $s^\dagger$ the value of $p^t_{s^\dagger,l}$ and further obtain the seller's empirical rewards, *i.e.*, $R^\star_{s^\dagger,a}(t)$ for $a \in \mathcal{A}$. Afterwards, the buyer will update her obtained information at hand, *i.e.*, the values of $\overline{r}_{i,a}(t)$, $T_{i,a}(t)$, $n_{i,a}(t)$, and $\widehat{r}_{i,a}(t)$.

Algorithm 1 Multi-agent Self-determining Bandit Strategy

Require: $\mathcal{N}$, $\mathcal{A}$, c_a for $\forall a \in \mathcal{A}$, B_i φ_i for $\forall i \in \mathcal{N}$, ω, θ, α
Ensure: $\{\pi^t_{i,a}, \gamma^t_{s,i}, p^t_{s,l}, \forall i, s \in \mathcal{N}, \forall a \in \mathcal{A}, 1 \leq l \leq L, \forall t \geq 1\}$

```
Initialization: B_i^0 = B_i for i ∈ N;
for t = 1, 2, ··· , do
    if t = k · (ω + 1) for k ≥ 1 then
        s ∈ N acts as a seller agent (calling Alg. 2):
        p^t_{s,l} = SA(α, P_s, β_{s,l}(t), ū_{s,l}(t));
        i ∈ N acts as a buyer agent (calling Alg. 3):
        γ_{s,i}(p^t_{s,l}) = BA(α, θ, φ_i, Λ_{i,a}(t), n*_{s,a}(t), p^t_{s,l});
        % Here, Λ_{i,a}(t) = ⟨T_{i,a}(t), n_{i,a}(t), r̄_{i,a}(t)⟩.%
        Each seller s ∈ N calculates her extra revenues u^t_{s,l}
        and updates β_{s,l}(t), ū_{s,l}(t), and û_{s,l}(t);
        Each buyer i ∈ N updates several parameters: n_{i,a}(t),
        T_{i,a}(t), r̄_{i,a}(t), and B^t_i = B^{t-1}_i − Σ_{s∈N} p^t_{s,l} · γ_{s,i}(p^t_{s,l});
    else
        i ∈ N makes arm-pulling decision (calling Alg. 4):
        π^t_{i,a} = AAP(α, Λ_{i,a}(t), c_a);
        Each agent, e.g., i, updates n_{i,a}(t), T_{i,a}(t), r̄_{i,a}(t),
        and the remaining budget B^t_i = B^{t-1}_i − Σ_{a∈A} π^t_{i,a} · c_a;
    end if
    if B^t_i < min_{a∈A} c_a then
        The agent i exits the TMA²B framework;
    end if
end for
Output: {π^t_{i,a}, γ^t_{s,i}, p^t_{s,l}, ∀i, s ∈ N, ∀a ∈ A} for ∀t ≥ 1
```

Here, we use a simple example to discover the inner idea of the seller selection process from one buyer's perspective. We consider 2 arms and 2 seller agents, and let $\{\langle 20, 5\rangle, 0.5\}$ and $\{\langle 5, 20\rangle, 0.5\}$ denote these two sellers' information. $\langle 20, 5\rangle$ means the number of times of pulling each arm, while 0.5 denotes the posted price. Suppose that the buyer's numbers of times of pulling these two arms are $n_1 = 20$ and $n_2 = 100$, and the average empirical rewards are $\overline{r}_1 = 0.8$ and $\overline{r}_2 = 0.3$, respectively. Based on this, we conduct the calculation according to Eqs. (11) and (12), and find that the buyer prefers the first seller regardless of her level of risk-tolerance, *i.e.*, $\sum_{a\in\mathcal{A}} v^a_{s_1,i} > \sum_{a\in\mathcal{A}} v^a_{s_2,i}$ for $\forall \varphi_i \in (0, 1)$. Now, we set $\overline{r}_2 = 0.5$ and keep other values unchanged. We find when $0 < \varphi_i < 0.19$, we have $\sum_{a\in\mathcal{A}} v^a_{s_2,i} > \sum_{a\in\mathcal{A}} v^a_{s_1,i}$; when $0.19 < \varphi_i \leq 1$, we get the opposite conclusion. This change reflects the intrinsic nature of the proposed computation method. When the average empirical rewards for one specific arm are high, but the confidence interval is bad, the agent with a high risk-tolerance level may prefer the seller whose arm-pulling experience can improve the confidence interval for this arm. Thus, the proposed computation method can capture the

Algorithm 2 SA: One Seller Agent's Decision

Require: α, $\mathcal{P}_s$, $\beta_{s,l}(t)$, and $\overline{u}_{s,l}(t)$
Ensure: $p^t_{s,l} \in \mathcal{P}_s$
1: **if** $\lfloor \frac{t}{\omega+1} \rfloor > L$ **then**
2: Calculate $p^t_{s,l} = \text{argmax}_{p_{s,l} \in \mathcal{P}_s} \widehat{u}_{s,l}(t)$ in Eq. (13);
3: **else**
4: Randomly select one not-posted price, *e.g.*, $p^t_{s,l} \in \mathcal{P}_s$;
5: Initialize the values of $\beta_{s,l}(t)$ and $\overline{u}_{s,l}(t)$;
6: **end if**
7: **Output**: $p^t_{s,l}$

Algorithm 3 BA: One Buyer Agent's Decision

Require: $\Lambda_{i,a}(t)$, α, θ, φ_i, $n^\star_{s,a}(t)$, $p^t_{s,l}$ for $\forall s \in \mathcal{N}$, $a \in \mathcal{A}$
Ensure: $\gamma_{s,i}(p^t_{s,l})$
1: **for** $s \in \mathcal{N}$ **do**
2: Compute the incremental confidence interval for $a \in \mathcal{A}$ given $n^\star_{s,a}(t)$ and $\Lambda_{i,a}(t)$, *i.e.*, $\nabla^a_{s,i}$ in Eq. (10);
3: Calculate the expected incremental reward for a by considering the risk-tolerant level, *i.e.*, $v^a_{s,i}$ in Eq. (11);
4: **end for**
5: Determine the posted price, *i.e.*, $p^t_{s^\dagger,l}$ in Eq. (12);
6: **Output**: $\gamma_{s^\dagger,i}(p^t_{s^\dagger,l}) = 1$ and $\gamma_{s,i}(p^t_{s,l}) = 0$ for $s \neq s^\dagger$

intrinsic characteristics of heterogeneous agents.

On the other hand, we present the experience-trading decision from the perspective of a seller agent, *i.e.*, how to post one suitable price to maximize each seller's payoff over time. We model this process as a trivial MAB problem, where the posted price is regarded as one arm. In any trading round $t = k \cdot (\omega+1)$ for $k \geq 1$, each seller $s \in \mathcal{N}$ will select the posted price according to the empirical experience-trading rewards and the number of posted price being selected until t. More specifically, each seller s will try all posted price values once in the first L trading rounds, so that $p^t_{s,l} \in \mathcal{P}_s$ where $|\mathcal{P}_s| = L$ can be initialized once. In each trading round $t = k \cdot (\omega+1)$, we compute the revenue of seller s as follows:

$$u^t_{s,l} = p^t_{s,l} \cdot \sum\nolimits_{i \in \mathcal{N}} \gamma_{s,i}(p^t_{s,l}); \quad \text{for } 1 \leq l \leq L.$$

Here, we let $\overline{u}_{s,l}(t)$ and $\beta_{s,l}(t)$ denote the average revenues from s posting the l-th price in the trading round t and the total number of times of posting this price until t. When $p^t_{s,l}$ is selected, these two values are updated as follows:

$$\beta_{s,l}(t) = \beta_{s,l}(t-(\omega+1)) + 1,$$
$$\overline{u}_{s,l}(t) = \frac{\overline{u}_{s,l}(t-(\omega+1)) \cdot \beta_{s,l}(t-(\omega+1)) + u^t_{s,l}}{\beta_{s,l}(t)}.$$

We then design the UCB-based index of s posting the price $p^t_{s,l}$, denoted as $\widehat{u}_{s,l}(t)$, *i.e.*,

$$\widehat{u}_{s,l}(t) = \overline{u}_{s,l}(t) + \sqrt{\alpha \ln(\lfloor t/(\omega+1) \rfloor)/\beta_{s,l}(t)}. \quad (13)$$

Here, α is the same parameter defined in the T-UCB expression. When the price $p^t_{s,l}$ is not posted in the trading round t, these two values $\beta_{s,l}(t)$ and $\overline{u}_{s,l}(t)$ stay the same as in the last experience-trading rounds, *i.e.*, $\beta_{s,l}(t) = \beta_{s,l}(t-(\omega+1))$ and $\overline{u}_{s,l}(t) = \overline{u}_{s,l}(t-(\omega+1))$. At the beginning of each experience-trading round $t = k \cdot (\omega+1)$, each seller agent $s \in \mathcal{N}$ posts her price according to the value of $\widehat{u}_{s,l}(t-(\omega+1))$.

Algorithm 4 AAP: One Agent's Arm-Pulling Decision

Require: α, $\Lambda_{i,a}(t) = \langle T_{i,a}(t), n_{i,a}(t), \overline{r}_{i,a}(t) \rangle$, c_a for $\forall a \in \mathcal{A}$
Ensure: $\pi^t_{i,a}$
1: **for** $a \in \mathcal{A}$ **do**
2: Calculate the T-UCB based index $\widehat{r}_{i,a}(t)$ in Eq.(9);
3: **end for**
4: Select the arm with the highest ratio of the T-UCB-based index and the cost, *i.e.*, $a^\dagger = \text{argmax}_{a \in \mathcal{A}} \widehat{r}_{i,a}(t)/c_a$;
5: **Output**: $\pi^t_{i,a^\dagger} = 1$ and $\pi^t_{i,a} = 0$ for $a \neq a^\dagger$

C. Detailed Algorithms

Now, we present the solution to TMA^{2}B in detail. We first display the distributed multi-agent self-determining bandit decisions (including the arm-pulling and experience-trading decisions) in Alg. 1. In each experience-trading round $t = k(\omega+1)$ for $k \geq 1$, each agent can act as a seller and buyer simultaneously. When one agent intends to sell her bandit experience of the past ω rounds, the seller's decision procedure (*i.e.*, Alg. 2) will be called as shown in Step 4. In Alg. 2, if all price candidates have been tried, the procedure will output the posted price with the highest UCB-based index in Eq. (13). Otherwise, Alg. 2 will output any one price value in $\mathcal{P}_s$ that has not been posted before.

On the other hand, when an agent wants to buy some bandit experience from others, the buyer's decision procedure (*i.e.*, Alg. 3) will work, as shown in Step 5. More specifically, each buyer first calculates the expected incremental confidence intervals for each arm in Eq. (10), based on which it computes the expected incremental reward for each arm by considering this buyer's risk-tolerant level in Eq. (11). At last, she determines the seller according to the criterion in Eq. (12). Note that only the indicator for the selected seller will become 1. After the trading process is completed, each agent i will update her local information at hand, including $n_{i,a}(t)$, $T_{i,a}(t)$, $\overline{r}_{i,a}(t)$, B^t_i, $\beta_{i,l}(t)$, $\overline{u}_{i,l}(t)$, and $\widehat{u}_{i,l}(t)$.

In each arm-pulling round, each heterogeneous agent will independently pull one arm according to the output of Alg. 4. More precisely, each agent determines the arm with the maximum ratio of the T-UCB-based index and the corresponding cost. After all agents pull their selected arms, several parameters such as $n_{i,a}(t)$, $T_{i,a}(t)$, $\overline{r}_{i,a}(t)$, and B^t_i will update accordingly. Every agent will check their remaining budget at the end of each round. When the remaining budget is less than a threshold (*i.e.*, the minimum arm-pulling cost), the agent will leave the TMA^{2}B framework. The whole process terminates when all agents exhaust their budgets. Note that TMA^{2}B supports the asynchronous start, in which each heterogeneous agent needs to maintain a local slotted round index.

D. Theoretical Analysis

We first simplify some notations to better analyze the regret bound of each distributed agent with a given budget. We

omit i in B_i, $n_{i,a}(t)$, $\overline{r}_{i,a}(t)$, and $\widehat{r}_{i,a}(t)$. We also define some notations as follows: $a^* = \text{argmax}_{a\in\mathcal{A}} \frac{r_a}{c_a}$ (*i.e.*, a^* is the optimal arm), $\Delta_{min} = \min_{a\neq a^*}(\frac{r_{a^*}}{c_{a^*}} - \frac{r_a}{c_a})$, $c_{min} = \min_{a\in\mathcal{A}} c_a$, $c_{max} = \max_{a\in\mathcal{A}} c_a$, $x_a = c_a - c_{a^*}$, $y_a = r_{a^*} - r_a$, and $p_{min} = \min_{i\in\mathcal{N}} p_{i,1}$ (*i.e.*, the minimum posted price of all agents). Note that the values of x_a and y_a may be negative here. Moreover, we let B_1 and B_2 ($B = B_1 + B_2$) denote the budgets for pulling arms and buying others' experience, respectively. B_1 and B_2 will affect the specific arm pulled by the agent in one round and the total number of arm-pulling rounds, respectively. Let $T(B_1)$ denote the total arm-pulling rounds under the budget B_1.

We divide the regret analysis into three steps, same as [5, 6, 23]: 1) bounding the expected number of pulls of sub-optimal arms under the rounds $T(B_1)$, denoted as $\mathbb{E}[n_a(T(B_1))]$ for $a \neq a^*$; 2) linking the regret and the total expected arm-pulling rounds under the given budget (denoted as $\mathbb{E}[T(B_1)]$); 3) deriving the worst regret bound, denoted as $R(B)$. Note that the regret analysis in the TMA^{2}B model is more complicated because the experience transactions strongly affect the arm-pulling decisions in each round and decrease the expected total arm-pulling rounds. When analyzing the bound of $\mathbb{E}[n_a(t))]$ for $t \leq T(B_1)$, we let $t' = t + \delta$ (and $n'_a = n_a + \delta'$) denote the total observed number of arm-pulling rounds (and the total observed number of times of pulling a), where δ and δ' indicate the total arm-pulling rounds and the total number of times of pulling a observed from the bought experience. Recall that t and n_a are the agent's local round index and the local number of times of pulling a. According to this, we have the following lemma.

Lemma 1: The expected number of times of pulling a under the local rounds $T(B_1)$, *i.e.*, $\mathbb{E}[n_a(T(B_1))]$, is bounded as:

$$\mathbb{E}[n_a(T(B_1))] \leq \frac{4\alpha \ln(B_1/c_{min} + B_2\omega/p_{min})}{(\Delta_{min} c_{min})^2} + 1 + \frac{\pi^2}{3}.$$

Proof: We first define A_t, which indicates the arm pulled in the t-th arm-pulling round. Then, we have $\mathbb{E}[n_a(T(B_1))] = 1 + \sum_{t=A+1}^{T(B_1)} \mathbb{I}\{A_t = a\}$ where $\mathbb{I}\{true\} = 1$. Further, we have the following results:

$$\begin{aligned}\mathbb{E}[n_a(T(B_1))] &\leq \mu + \sum\nolimits_{t=A+1}^{T(B_1)} \mathbb{I}\{A_t = a, n_a(t) \geq \mu\} \\ &\leq \mu + \sum_{t'=1}^{T(B_1)} \sum_{n'_{a^*}}^{t'-1} \sum_{n'_a}^{t'-1} \Big\{ \frac{\overline{r}_{a^*}(t') + w_{t',n'_{a^*}}}{c_{a^*}} \leq \frac{\overline{r}_a(t') + w_{t',n'_a}}{c_a} \Big\}, \quad (14)\end{aligned}$$

where we let $w_{t,n} = \sqrt{\frac{\alpha \ln t}{n}}$ for simplicity and $t' = \sum_{a\in\mathcal{A}} n'_a$ indicates the total number of arm-pulling rounds, including the agent's local information and bought bandit experience from others. According to the existing work [5, 6, 23], we find that at least of three following cases must hold: $\overline{r}_{a^*}(t') \leq r_{a^*} - w_{t',n'_{a^*}}$, $\overline{r}_a(t') \geq r_a + w_{t',n'_a}$, and $\frac{r_{a^*}}{c_{a^*}} < \frac{r_a + 2w_{t',n'_a}}{c_a}$. The probability of the first two cases being true is less than $2(t')^{-4}$ based on the Chernoff-Hoeffding bound [5, 6, 24].

However, the probability of the third case being false is different from the existing work. This is because an agent's arm-pulling decisions may be affected by others' bandit experience. When we set $\mu \geq \frac{4\alpha \ln(t+\delta)}{(\Delta_{min} c_{min})^2} - \delta'$, we conclude that the third case is false, that is,

$$\begin{aligned}&\frac{r_{a^*}}{c_{a^*}} - \frac{r_a}{c_a} - \frac{2w_{t',n'_a}}{c_a} \geq \frac{r_{a^*}}{c_{a^*}} - \frac{r_a}{c_a} - \frac{2w_{t',n'_a}}{c_{min}} \\ &\geq \Delta_{min} - \frac{2}{c_{min}} \sqrt{\frac{\alpha \ln t'}{n'_a}} \geq \Delta_{min} - \frac{2}{c_{min}} \sqrt{\frac{\alpha \ln t'}{\mu + \delta'}} \geq 0.\end{aligned}$$

Due to $\delta' \geq 0$, $t \leq B_1/c_{min}$, and $\delta \leq \omega B_2/p_{min}$, we continue Eq. (14) and get

$$\mathbb{E}[n_a(T(B_1))] \leq \frac{4\alpha \ln(B_1/c_{min} + B_2\omega/p_{min})}{(\Delta_{min} c_{min})^2} + 1 + \frac{\pi^2}{3}. \quad (15)$$

Lemma 1 holds. □

Next, we analyze the relationship between the total number of arm-pulling rounds and the budget. Based on [5], we have

$$\begin{aligned}\mathbb{E}[T(B_1)] \geq \frac{B_1}{c_{a^*}} &- \frac{4\alpha}{(\Delta_{min} c_{min})^2} \sum\nolimits_{x_a>0} \frac{x_a}{c_{a^*}} \ln(\frac{B_1}{c_{min}}) \\ &- \sum\nolimits_{x_a>0} \frac{x_a}{c_{a^*}} (1 + \pi^2/3) - 1. \quad (16)\end{aligned}$$

We omit the detailed proof process here due to the space limit. At last, we analyze the worst regret bound.

Theorem 1: The regret of our algorithm in TMA^{2}B under the budget B (*i.e.*, $R(B)$) is constrained by

$$\begin{aligned}R(B) \leq &\Big(\frac{4\alpha}{(\Delta_{min} c_{min})^2} \sum\nolimits_{x_a>0} \frac{x_a}{c_{a^*}} \ln(\frac{B_1}{c_{min}}) \\ &+ \sum\nolimits_{x_a>0} \frac{x_a}{c_{a^*}} (1 + \frac{\pi^2}{3}) + 1 + \frac{B_2}{c_{a^*}} \Big) r_{a^*} \\ &+ \sum\nolimits_{y_a>0} y_a \Big(\frac{4\alpha \ln(B_1/c_{min} + B_2\omega/p_{min})}{(\Delta_{min} c_{min})^2} + 1 + \frac{\pi^2}{3} \Big).\end{aligned}$$

Proof: According to the definition of regret (*i.e.*, the total reward gap between the optimal solution and ours), we derive the worst regret bound as follows:

$$\begin{aligned}R(B) &= \frac{B r_{a^*}}{c_{a^*}} - \sum\nolimits_{t=1}^{T(B_1)} r_a^t \\ &\leq \frac{B r_{a^*}}{c_{a^*}} - T(B_1) r_{a^*} + T(B_1) r_{a^*} - \sum\nolimits_{t=1}^{T(B_1)} r_a^t \\ &\leq \Big(\frac{B_1 + B_2}{c_{a^*}} - T(B_1) \Big) r_{a^*} + \sum_{t=1}^{T(B_1)} \Big(r_{a^*} - \sum_{a\in\mathcal{A}} r_a \mathbb{I}\{A_t = a\} \Big) \\ &\leq \Big(\frac{B_1}{c_{a^*}} - T(B_1) + \frac{B_2}{c_{a^*}} \Big) r_{a^*} + \sum_{y_a>0} y_a \mathbb{E}[n_a(T(B_1))]. \quad (17)\end{aligned}$$

Now, by substituting the results of Eq. (15) and Eq. (16) into Eq. (17), we complete the proof of Theorem 1. □

This regret bound is consistent with our observations. Due to the fact $B = B_1 + B_2$, when we put a larger proportion of the budget on the arm-pulling process (*i.e.*, increase B_1 but decrease B_2), the total number of arm-pulling rounds increases. Otherwise, the total number of pulling arms will reduce. Still, the agent will select the optimal arm in each round with a higher probability. This is because trading others' experience more will undoubtedly improve the confidence intervals for each prior unknown arm. Thus, a tradeoff exists between the arm-pulling and experience-trading stages. Additionally, the given regret bound in Theorem 1 does not consider the experience-trading rewards. Actually, the regret bound of our

algorithm involving the trading revenues will be tighter.

IV. Experiment

A. Experiment Settings

We compare our solution with several famous algorithms, called Optimal, fractional KUBE [5, 6] (frac-KUBE for short), and ϵ-first [25]. We first introduce the Optimal algorithm, which means each agent knows the expected reward of each arm (*i.e.*, r_a for $a \in \mathcal{A}$) in advance. In such a case, each agent will always pull the same arm a^* in each round, that is, $a^* = \text{argmax}_{a\in\mathcal{A}} \frac{r_a}{c_a}$. Note that in each round, the reward obtained by the agent, denoted as $r^t_{a^*}$, will follow the Gaussian distribution with the expectation r_{a^*}. The frac-KUBE algorithm indicates that each agent will independently adopt the density-ordered greedy strategy based on the traditional UCB index to determine the arm in each round. After the agent initializes several parameters by pulling all arms once, she will pull the arm in each round according to the ratio of the UCB values and the cost. The ϵ-first algorithm stipulates that each agent i will use a part of the budget, *i.e.*, $\epsilon \cdot B_i$, to explore the prior unknown rewards of arms, *i.e.*, randomly pulling arms. Then, she will always pull the arm which has the best performance in the exploring stages, *i.e.*, the highest ratio of the average empirical reward and the cost. Here, we find that when $\epsilon = 0.01$ and $\epsilon = 0.1$, the ϵ-first algorithm has relatively good performance.

Next, we introduce some detailed experiment settings. We first generate the expected reward r_a for each arm $a \in \mathcal{A}$ according to the uniform distribution from the range $[3, 7]$. We also create the cost of each arm c_a from the range $[1, 2]$. To generate the value of r^t_a, we use the Gaussian distribution with the expectation r_a. Here, we generate the variance of the Gaussian distribution for $a \in \mathcal{A}$ from the range $(0, 5]$. Then, we produce the posted price for sellers, considering that all agents have the same posted price set for simplicity. Here, we let $L = 5$ and $\mathcal{P}_L = c_{min} \cdot \{0.1, 0.2, 0.3, 0.4, 0.5\}$ in which $c_{min} = \text{argmin}_{a\in\mathcal{A}} c_a$. We generate each agent's budget from the range $[10^3, 10^5]$ and let $B_i = 2 * 10^3$ in default. Moreover, we set $\alpha = 2$ in the experiments based on the existing work [5, 6]. We generate the numbers of agents and arms, *i.e.*, N and A, from the range $[20, 140]$, and let $N = 20$ as well as $A = 40$ in default. At last, we let the transaction frequency ω be selected from the range $[19, 89]$.

B. Experiment Results

Now, we present the experiment results. We first evaluate the total achieved rewards for each agent with the change in her budget. When we increase the budget from 10^3 to 10^4, we see that our algorithm outperforms the ϵ-first algorithm on average and almost catches up with the Optimal algorithm, as shown in Fig. 2. We analyze that the rewards achieved by our algorithm are about 32% higher than that of the ϵ-first algorithm. Also, along with the increase in the agent's budget, the total revenues of all algorithms increase accordingly. These experimental results are consistent with our theoretical analysis. We then evaluate the impact of the number of agents (*i.e.*, N) in the TMA^{2}B framework, as shown in Fig. 3. We observe that our

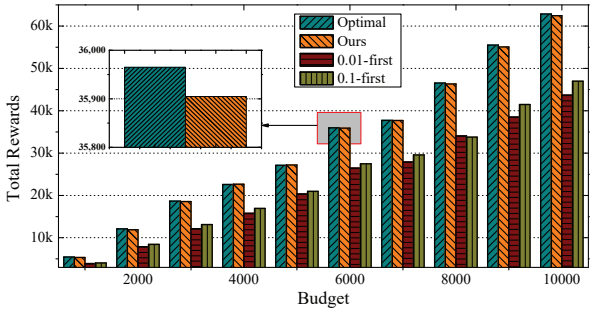

Fig. 2. Total Rewards vs. Budget.

proposed algorithms still beat the ϵ-first algorithm and almost keep up with the Optimal algorithm. These observations verify the effectiveness of the proposed solution. Moreover, we evaluate the performance of all algorithms by changing the number of arms (*i.e.*, A) and display the experiment results in Fig. 4. With the increase in the number of arms, the total rewards achieved by these several algorithms have an upward trend. This is because more arms mean more choices for each agent. The probability of pulling the arm with higher rewards will increase. The total rewards from our algorithm fall between the Optimal and frac-KUBE algorithms. The average rewards of the Optimal algorithm and our solutions are 12499 and 11889 in this setting. On the other hand, we present the reward performance when we change the trading frequency ω, as shown in Fig. 5. The total rewards obtained by our algorithm are still higher than those of frac-KUBE and ϵ-first. These experimental results are in line with expectations.

Next, for any one arm across different rounds, we compare the T-UCB values in our algorithm and the UCB values in the frac-KUBE algorithm, as shown in Fig. 6. We let the yellow line denote the expected reward (*i.e.*, 6.5) of any one arm in the TMA^{2}B framework. At the same time, we let the red and blue lines denote the T-UCB values in Eq. (9) and the trivial UCB values, respectively. We also display generated rewards for each round using the gray dots based on the expectation 6.5 in the graph. We observe that T-UCB can approach the expected reward value more quickly than UCB at the beginning of the MAB process. Moreover, the T-UCB values can estimate the prior unknown reward of this arm more accurately than UCB in the end. The T-UCB and UCB values are getting closer to the expected reward with the increase in rounds. In addition, we present the total experience-trading revenues and payments for the agents in Fig. 7. We observe that for some agents, the revenue from selling their local experience is higher than the payment used to buy others' experience in the trading rounds. We also display the changes in the trading revenue and payment for one given agent across different trading rounds in Fig. 8. In some rounds, several buyers may accept this agent's posted price, while in other rounds, no buyer will accept it. These results remain consistent with our theoretical analysis.

V. Related Work

So far, lots of research has studied Multi-Armed Bandit (MAB) problems, including the MAB applications and the variants of traditional MAB. We first review the MAB applications in various fields. [26] proposed the multi-agent

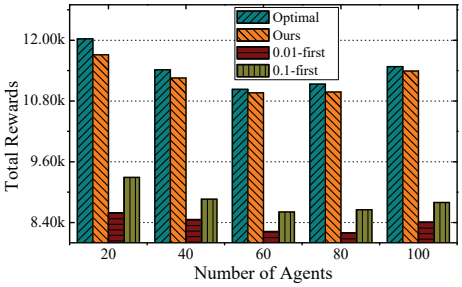

Fig. 3. Total Rewards vs. Number of Agents.

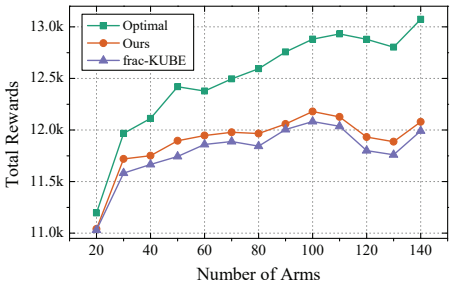

Fig. 4. Total Rewards vs. Number of Arms.

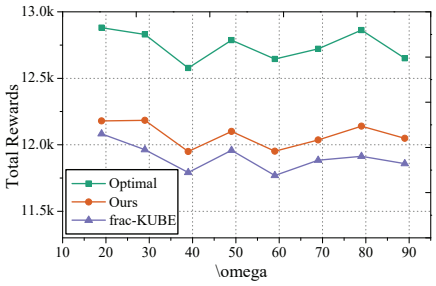

Fig. 5. Total Rewards vs. Parameter ω.

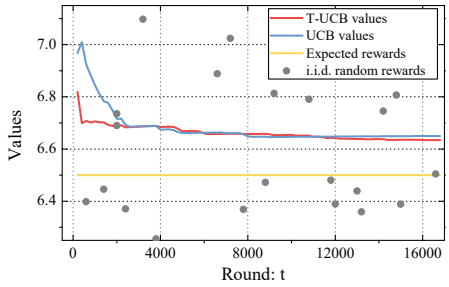

Fig. 6. T-UCB vs. UCB.

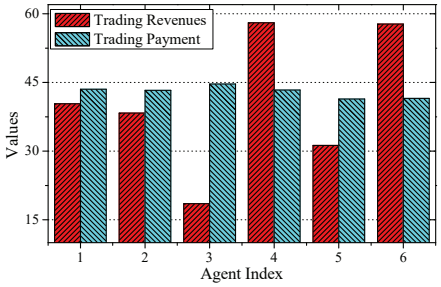

Fig. 7. Trading Revenues and Payment.

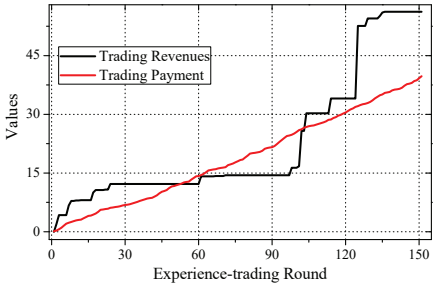

Fig. 8. One Agent's Trading Revenues and Payment.

MAB-based solution to the table orthogonal allocation in ad-hoc networks. [27] adopted the online multi-agent MAB model to solve the decentralized delay-sensitive task offloading problem in edge computing scenarios. [28] studied the cache placement problem in the mobile edge computing field and designed a multi-agent MAB solution. [29] studied the beam tracking problem in millimeter-wave systems and developed the adaptive Thompson sampling solution. Next, we review the MAB-related algorithms. [19] studied the distributed MAB model with heterogeneous agents, in which each agent's goal is to find her optimal local arm among a local subset of all arms, while [30] aimed at finding the optimal global arm. [31] investigated the optimal arm identification problem in the stochastic MAB model and proposed a reward-cost ratio based solution. The work [8, 10] studied a special MAB model with the fairness and sleeping arm constraints, where fairness means the number of pulling each arm should not be less than a given threshold, and sleeping arms indicate that some arms cannot be pulled in a round. The work [1, 32] investigated the combinatorial MAB models where multiple arms can be simultaneously pulled in each round and the agent's goal is to maximize the total rewards.

Some existing MAB work has considered multi-agent MAB scenarios. For example, the work [21, 33] studied how multiple agents cooperate with their immediate neighbors to solve a linear bandit-optimization problem. [34] considered that each agent can exchange messages through a underlying network in two cases where the communication graph is known or unknown. [22] studied the impact of cooperation and communication on the regret and communication cost in the distributed multi-agent MAB model. The work [15, 35] studied a special multi-agent MAB problem aiming to maximize all agents' total rewards and devised the distributed learning algorithms. Similarly, the work [11, 20] considered communication delays in the multi-agent MAB problem, while [14] combined the UCB idea with a message-delivering protocol to propose a decentralized algorithm. Different from these studies on the MAB models and applications, we investigate the transaction-based multi-agent MAB model in which these learning agents can trade their bandit experience with each other.

On the other hand, some literature has studied data valuation and pricing problems. Among them, [36, 37] considered data uncertainty, economic robustness, and the concept of entropy in the models, and proposed the online data valuation and pricing mechanisms; [38] supposed that each buyer has a time-sensitive valuation on the items, which follows an unknown distribution, and a seller has a limited supply; [39] investigated the VM placement and pricing problem in the load-unbalanced edge computing scenario and proposed an auction-based solution; [40] studied the online crowdsensing task pricing problem in which each worker arrives dynamically. In this paper, we adopt the easily-implemented posted pricing mechanism to evaluate the bandit experience for each distributed agent. We focus on how each agent determines the seller agent to trade with by considering her incremental confidence intervals and her level of risk tolerance.

VI. Conclusion

In this paper, we study a transaction-based multi-agent MAB (TMA^{2}B) model where not only can each agent pull an arm independently in a round, but also each agent can trade their bandit experience. The challenges lie in how to design efficient arm-pulling and experience-trading strategies. For each independent agent, we combine her individually-observed rewards and her bought bandit experience, based on which we devise a tailor-made transaction-based upper confidence bound (T-UCB) to denote the prior-unknown rewards of arms. Each agent pulls the arms or trades her bandit experience according to the T-UCB values and the corresponding cost. We analyze the regret bound of our algorithm, indicating that a tradeoff exists between the arm-pulling and experience-trading

stages. Finally, we conduct lots of experiments to verify the effectiveness of the proposed solution.

ACKNOWLEDGEMENT

This research was supported in part by the National Natural Science Foundation of China (NSFC) under Grant U20A20182, 62102275, in part by the NSF of Jiangsu in China under Grant BK20210704, and in part by the NSF of the Jiangsu Higher Education Institutions of China under Grant 21KJB520025.

REFERENCES

[1] W. Chen, Y. Wang, and Y. Yuan, "Combinatorial multi-armed bandit: General framework and applications," in *International conference on machine learning*, 2013.

[2] H. Gupta, A. Eryilmaz, and R. Srikant, "Low-complexity, low-regret link rate selection in rapidly-varying wireless channels," in *IEEE INFOCOM*, 2018.

[3] X. Gao, S. Chen, and G. Chen, "Mab-based reinforced worker selection framework for budgeted spatial crowdsensing," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 3, pp. 1303–1316, 2022.

[4] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Foundations and Trends in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.

[5] L. Tran-Thanh, A. Chapman, A. Rogers, and N. R. Jennings, "Knapsack based optimal policies for budget-limited multi-armed bandits," in *AAAI*, 2012.

[6] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, vol. 47, no. 2-3, pp. 235–256, 2002.

[7] S. Agrawal and N. Goyal, "Analysis of thompson sampling for the multi-armed bandit problem," in *Conference on learning theory*, 2012.

[8] F. Li, J. Liu, and B. Ji, "Combinatorial sleeping bandits with fairness constraints," in *IEEE INFOCOM*, 2019.

[9] G. Gao, S. Huang, H. Huang, M. Xiao, J. Wu, Y.-E. Sun, and S. Zhang, "Combination of auction theory and multi-armed bandits: Model, algorithm, and application," *IEEE Transactions on Mobile Computing*, pp. 1–14, 2023.

[10] J. Steiger, B. Li, and N. Lu, "Learning from delayed semi-bandit feedback under strong fairness guarantees," in *IEEE INFOCOM*, 2022.

[11] N. Cesa-Bianchi, C. Gentile, Y. Mansour, and A. Minora, "Delay and cooperation in nonstochastic bandits," in *Conference on Learning Theory*, 2016.

[12] T. Ouyang, X. Chen, Z. Zhou, R. Li, and X. Tang, "Adaptive user-managed service placement for mobile edge computing via contextual multi-armed bandit learning," *IEEE Transactions on Mobile Computing*, vol. 22, no. 3, pp. 1313–1326, 2023.

[13] A. Sawwan and J. Wu, "A new framework: Short-term and long-term returns in stochastic multi-armed bandit," in *IEEE INFOCOM*, 2023.

[14] A. Dubey and A. Pentland, "Cooperative multi-agent bandits with heavy tails," in *ACM ICML*, 2020.

[15] I. Bistritz and N. Bambos, "Cooperative multi-player bandit optimization," in *NeurIPS*, 2020.

[16] S. Kang and C. Joo, "Low-complexity learning for dynamic spectrum access in multi-user multi-channel networks," *IEEE Transactions on Mobile Computing*, vol. 20, no. 11, pp. 3267–3281, 2021.

[17] S. Bubeck, Y. Li, Y. Peres, and M. Sellke, "Non-stochastic multi-player multi-armed bandits: Optimal rate with collision information, sublinear without," in *Conference on Learning Theory*, 2020.

[18] K. Liu and Q. Zhao, "Distributed learning in multi-armed bandit with multiple players," *IEEE transactions on signal processing*, vol. 58, no. 11, pp. 5667–5681, 2010.

[19] L. Yang, Y.-Z. J. Chen, M. H. Hajiemaili, J. C. Luiy, and D. Towsley, "Distributed bandits with heterogeneous agents," in *IEEE INFOCOM*, 2022.

[20] U. Madhushani, A. Dubey, N. Leonard, and A. Pentland, "One more step towards reality: Cooperative bandits with imperfect communication," in *NeurIPS*, 2021.

[21] R. K. Kolla, K. Jagannathan, and A. Gopalan, "Collaborative learning of stochastic bandits over a social network," *IEEE/ACM Transactions on Networking*, vol. 26, no. 4, pp. 1782–1795, 2018.

[22] S. Buccapatnam, J. Tan, and L. Zhang, "Information sharing in distributed stochastic bandits," in *IEEE INFOCOM*, 2015.

[23] Y. Xia, T. Qin, W. Ding, H. Li, X. Zhang, N. Yu, and T.-Y. Liu, "Finite budget analysis of multi-armed bandit problems," *Neurocomputing*, vol. 258, pp. 13–29, 2017.

[24] J. P. Schmidt, A. Siegel, and A. Srinivasan, "Chernoff–hoeffding bounds for applications with limited independence," *SIAM Journal on Discrete Mathematics*, vol. 8, no. 2, pp. 223–250, 1995.

[25] L. Tran-Thanh, A. Chapman, E. M. de Cote, A. Rogers, and N. R. Jennings, "Epsilon–first policies for budget–limited multi-armed bandits," in *AAAI*, 2010.

[26] S. J. Darak and M. K. Hanawal, "Multi-player multi-armed bandits for stable allocation in heterogeneous ad-hoc networks," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 10, pp. 2350–2363, 2019.

[27] X. Wang, J. Ye, and J. C. Lui, "Decentralized task offloading in edge computing: a multi-user multi-armed bandit approach," in *IEEE INFOCOM*, 2022.

[28] Y. Han, L. Ai, R. Wang, J. Wu, D. Liu, and H. Ren, "Cache placement optimization in mobile edge computing networks with unaware environment -an extended multi-armed bandit approach," *IEEE Transactions on Wireless Communications*, vol. 20, no. 12, pp. 8119–8133, 2021.

[29] I. Aykin, B. Akgun, M. Feng, and M. Krunz, "Mamba: A multi-armed bandit framework for beam tracking in millimeter-wave systems," in *IEEE INFOCOM*, 2020.

[30] L. Yang, Y.-Z. J. Chen, S. Pasteris, M. Hajiesmaili, J. Lui, and D. Towsley, "Cooperative stochastic bandits with asynchronous agents and constrained feedback," in *NeurIPS*, 2021.

[31] Z. Qin, X. Gan, J. Liu, H. Wu, H. Jin, and L. Fu, "Exploring best arm with top reward-cost ratio in stochastic bandits," in *IEEE INFOCOM*, 2020.

[32] D. P. Zhou and C. J. Tomlin, "Budget-constrained multi-armed bandits with multiple plays," in *AAAI*, 2018.

[33] S. Amani and C. Thrampoulidis, "Decentralized multi-agent linear bandits with safety constraints," in *AAAI*, 2021.

[34] Y. Bar-On and Y. Mansour, "Individual regret in cooperative nonstochastic multi-armed bandits," in *NeurIPS*, 2019.

[35] D. Martínez-Rubio, V. Kanade, and P. Rebeschini, "Decentralized cooperative stochastic bandits," in *NeurIPS*, 2019.

[36] Z. Zheng, Y. Peng, F. Wu, S. Tang, and G. Chen, "An online pricing mechanism for mobile crowdsensing data markets," in *ACM MobiHoc*, 2017.

[37] A. Xu, Z. Zheng, F. Wu, and G. Chen, "Online data valuation and pricing for machine learning tasks in mobile health," in *IEEE INFOCOM*, 2022.

[38] S. Li, L. Zhang, and X.-Y. Li, "Online pricing with limited supply and time-sensitive valuations," in *IEEE INFOCOM*, 2022.

[39] M. Siew, K. Guo, D. Cai, L. Li, and T. Q. Quek, "Let's share vms: Optimal placement and pricing across base stations in mec systems," in *IEEE INFOCOM*, 2021.

[40] H. Jin, H. Guo, L. Su, K. Nahrstedt, and X. Wang, "Dynamic task pricing in multi-requester mobile crowd sensing with markov correlated equilibrium," in *IEEE INFOCOM*, 2019.