FEDERATED MULTI-ARMED BANDIT VIA UNCOORDINATED EXPLORATION

Zirui Yan, Quan Xiao, Tianyi Chen, Ali Tajer

Electrical, Computer, and Systems Engineering Department
Rensselaer Polytechnic Institute, Troy, NY

ABSTRACT

A wide range of multi-agent decision-making problems can be abstracted as a federated multi-armed bandit (FMAB) problem. A key challenge of the FMAB problem is that the exploration-exploitation dichotomy inherited from the multi-armed bandit aspect is compounded with data heterogeneity in federated learning. This renders the exploration and exploitation of different agents inherently entangled. This paper focuses on overcoming the difficulty of exploration in FMAB problems, and it proposes a novel federated upper confidence bound (UCB) algorithm that requires uncoordinated exploration (UE) decisions by the agents. The major distinction of this algorithm, referred to as FedUCB-UE, with the existing FMAB algorithms is that it allows the agents to explore the non-optimal arms and make personalized arm-selection decisions without coordination. While such uncoordinated exploration makes the regret analysis non-trivial, it comes with both the theoretical and empirical benefit of diversity in explorations. Under certain mild assumptions, this paper establishes that FedUCB-UE has a $\mathcal{O}(\log T)$ regret bound. Furthermore, experiments performed on synthetic datasets show that FedUCB-UE outperforms the state-of-the-art algorithms.

Index Terms— Federated learning, multi-armed bandit, upper confidence bound

1. INTRODUCTION

Multi-armed bandit (MAB) settings provide a rich context for formulating and analyzing online learning tasks. MAB settings were first introduced in [1, 2] for designing clinical trials. Recently, they have been applied to more modern applications such as recommender systems [3], cognitive radios [4], and portfolio selection [5]. In the classic MAB setting, an agent is faced with some actions (called "arms"), and each action (arm) is associated with a reward distribution. When the agent chooses an arm, it will receive a stochastic reward (unknown a priori). The objective of such problems is to design a strategy in which the agent maximizes the reward it accumulates over time. To this end, the agent needs to pull all arms sufficiently to form reliable estimates of all rewards (exploration). These estimates guide the agent to pull the arm that likely has the largest reward (exploitation). Hence, the core challenge in MAB problems is finding a trade-off between exploration and exploitation (for more details, see a recent survey in [6]).

Due to the growing computation capability of devices and the increasing privacy concerns, it is expected that many learning tasks need to be performed collaboratively in a distributed way. In this context, federated learning (FL) emerges as a unifying framework to tackle collaborative learning over distributed devices [7]. Recently, the MAB problems have also been revisited in the FL setting, referred to as the federated MAB (FMAB) problems. In FMAB problems, several agents explore the same set of actions and communicate through a server to cooperatively form an optimal global decision [8–14]. Next, we review the existing approaches to FMAB that are most relevant to this paper.

1.1. Related Works

Decentralized multi-agent MAB. FMAB considered in this paper is intimately related to the literature of multi-agent MAB. Wang *et al.* [9] and Agarwal *et al.* [14] have considered the multi-agent MAB problem and have obtained regret-communication tradeoffs when the rewards of each agent are independent and identically distributed (i.i.d.). Vial *et al.* [13] have proposed a robust algorithm to defend the malicious attack in the i.i.d. setting. Shahrampour *et al.* [12] have proposed the d-UER algorithm using majority vote in the non-i.i.d. setting, and Zhu *et al.* [15] have proposed the Gossip-UCB algorithm using delayed information about exploration time and have provided a variant of it that preserves differential privacy. Subsequently, Zhu *et al.* [16] have extended this setting to the distributed setting in which the agents are connected by a directed graph.

FMAB. Unlike decentralized multi-agent MAB, in FMAB, the agents only communicate with the server in certain communication rounds, and the agents cannot share their information directly. Thus, the transmission of information is delayed. The study of FMAB can be traced back to the studies by Li *et al.* [8] and Dubey *et al.* [10], which have focused on the privacy perspective of FMAB in the i.i.d. setting. Mitra *et al.* [11] have proposed the Fed-SEL algorithm for the i.i.d. setting to address the setting in which the agents do not have access to all arms. Recently, Shi *et al.* [17] have considered the FMAB with personalization and have proposed the PF-UCB algorithm in the non-i.i.d. setting.

1.2. Our Contribution

In this context, we propose a novel federated upper confidence bound with uncoordinated exploration (FedUCB-UE) algorithm that runs the upper confidence bound (UCB) algorithm [18] on each agent, but uniquely allows uncoordinated exploration. Compared to the existing FMAB algorithms, the key differences and contributions are summarized as follows.

C1) The FedUCB-UE algorithm gives the different agents freedom to explore the same arm for different durations, which leads to improved exploration efficiency.

C2) Compared to the existing algorithms, we establish that FedUCB-UE achieves the optimal regret bound $\mathcal{O}(\log T)$ under certain regularity conditions, and it outperforms the regret of the state-of-the-art algorithms.

This work was supported by National Science Foundation CAREER Award 2047177, and the Rensselaer-IBM AI Research Collaboration (`http://airc.rpi.edu`), part of the IBM AI Horizons Network (`http://ibm.biz/AIHorizons`).

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

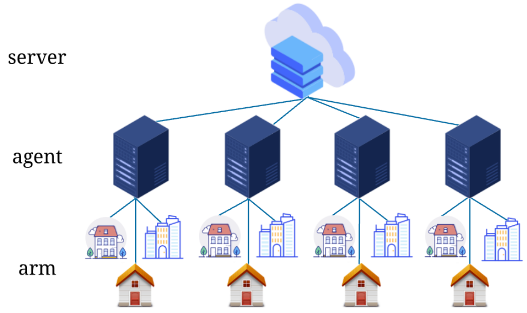


Fig. 1. Schematic diagram for FMAB setting.

C3) We evaluate our algorithm numerically and demonstrate its empirical performance gains compared to the state-of-the-art algorithms.

2. PRELIMINARIES

Consider an FMAB setting that consists of one server and N agents sharing the same set of M arms. At each round $t \in \mathbb{N}$, each agent pulls one arm. We denote the arm selected by agent $i \in [N] \triangleq \{1, 2, \cdots, N\}$ at round $t \in \mathbb{N}$ by $A_i(t) \in [M] \triangleq \{1, 2, \cdots, M\}$. We denote the reward of agent $i \in [N]$ after selecting arm $m \in [M]$ at round $t \in \mathbb{N}$ by $X_{i,m}(t)$. We assume that $X_{i,m}(t)$ is sampled from a σ-sub-Gaussian distribution with mean $\mu_{i,m}$ and the expected local rewards are distinct, i.e., $\mu_{i,m} \neq \mu_{j,m}$ if $i \neq j$. Without loss of generality, we assume that $\sigma = 1$. The expected global reward is defined as the average of the expected local rewards, i.e., $\mu_m \triangleq \frac{1}{N}\sum_{i=1}^{N} \mu_{i,m}$. Accordingly, we define $X_m(t)$ as the average of reward realization on arm m assuming that all agents pull arm m at round t, i.e., $X_m(t) \triangleq \frac{1}{N}\sum_{i=1}^{N} X_{i,m}(t)$. This serves as a measure of global reward associated with arm m.

Without loss of generality, we further assume that arm 1 is the best arm, i.e., $\mu_1 = \max_m \mu_m$ and define $\Delta_m \triangleq \mu_m - \mu_1$ as the reward gap between arm m and the best arm. Our objective is to minimize the static regret defined as

$$R(T) \triangleq NT\mu_1 - \sum_{t=1}^{T}\sum_{i=1}^{N} \mathbb{E}[X_{A_i(t)}(t)] , \tag{1}$$

which measures the difference between the cumulative reward of online decisions and the global optimal reward.

3. FEDUCB-UE ALGORITHM

In this section, we present the FedUCB-UE algorithm that allows each agent to make decisions according to its individual UCB and the global information received from the server. We first introduce the steps of FedUCB-UE in Section 3.1 and discuss the designing rationale in Section 3.2.

3.1. FedUCB-UE Algorithm

We use the local update structure of the federated average algorithm [7], in which each agent independently updates the local estimator in $E \in \mathbb{N}$ local rounds and communicates with the server at round $\{rE : r \in \mathbb{Z}_+\}$. Thus FedUCB-UE can be divided into three phases: initialization, local exploration, and communication.

Phase 1. Initialization. The algorithm starts with the training process at $t = 0$. Each agent $i \in [N]$ pulls each arm $m \in [M]$ once and gets the reward $X_{i,m}(0)$, which serves as the initial local reward estimate by $\widehat{X}_{i,m}(0) \triangleq X_{i,m}(0)$. Subsequently, each agent $i \in [N]$ sends $\widehat{X}_{i,m}(0)$ to the server and initializes the exploration time counter $n_{i,m}(0) = 1$. After receiving the local rewards from all agents, server calculates the initial global rewards for each arm m by $\widehat{X}_m(0) \triangleq \frac{1}{N}\sum_{i=1}^{N} \widehat{X}_{i,m}(0)$ and broadcasts them to all agents. Server also initializes the exploration time counter by $n_m(0) = 1$.

Phase 2. Local exploration. At times $t \in \mathbb{N}$, each agent $i \in [N]$ first identifies the set

$$S_i(t) \triangleq \left\{ m \,\middle|\, n_{i,m}(t-1)E < n_m\left(\left\lfloor \frac{t-1}{E} \right\rfloor E\right) \right\} , \tag{2}$$

where $\lfloor\cdot\rfloor$ is the floor function. If $S_i(t) \neq \emptyset$, agent i randomly selects an arm from $S_i(t)$. Otherwise, agent i chooses the arm that maximizes $\mathsf{UCB}_{i,m}$ defined as

$$\mathsf{UCB}_{i,m}(t) \triangleq B_{i,m}(t-1) + C_m(t-1) , \tag{3}$$

where $B_{i,m}(t) \triangleq \widehat{X}_m(\lfloor \frac{t}{E} \rfloor E) + \frac{1}{N}[\widehat{X}_{i,m}(t) - \widehat{X}_{i,m}(\lfloor \frac{t}{E} \rfloor E)]$ is an unbiased estimate of μ_m, and

$$C_m(t) \triangleq \min\left\{ \sqrt{\frac{8\log(t+1)}{N}}, \sqrt{\frac{8\log(t+1)}{N\left(n_m(\lfloor \frac{t}{E} \rfloor E) - 2\right)}} \right\}, \tag{4}$$

denotes the global confidence level on arm m. Once $A_i(t)$ is chosen, each agent $i \in [N]$ updates the local exploration time counter and the local reward estimator by

$$n_{i,m}(t) \triangleq \sum_{\tau=0}^{t} \mathbb{1}\{A_i(\tau) = m\} , \tag{5}$$

$$\widehat{X}_{i,m}(t) \triangleq \frac{1}{n_{i,m}(t)} \sum_{\tau=0}^{t} X_{i,m}(t)\mathbb{1}\{A_i(\tau) = m\} , \tag{6}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function.

Phase 3. Communication. At communication time $t \in \{rE : r \in \mathbb{Z}_+\}$, each agent $i \in [N]$ transmits $\widehat{X}_{i,m}(t)$ and $n_{i,m}(t)$ in (5)-(6) to the server. The sever calculates $\widehat{X}_m(t)$ and $n_m(t)$ for each arm $m \in [M]$ via

$$n_m(t) \triangleq \max_i n_{i,m}(t) , \quad \widehat{X}_m(t) \triangleq \frac{1}{N}\sum_{i=1}^{N} \widehat{X}_{i,m}(t) , \tag{7}$$

and broadcasts them to all agents.

We summarize the detailed steps of FedUCB-UE in Algorithm 1. In the next subsection, we will discuss the rationale for transmitting the exploration time counter $n_{i,m}(t)$ and $n_m(t)$.

3.2. Transmitting the Counter Value

The maximal exploration time $n_m(t)$ serves as a source of global information that guides the agents to adjust their local decisions. If $n_{i,m}(t)$ is substantially smaller than $n_m(t)$, arm m is deemed underexplored by agent i. In this case, agent i will explore arm m more frequently. We argue that accessing the global maximal pulling counter is necessary for the agents. Note that besides this counter, the agents have access only to the global reward estimates, which is insufficient for them to determine their actions. The reason is that the difference between $\widehat{X}_{i,m}(t)$ and $\widehat{X}_m(t)$ also depends on the difference between $\mu_{i,m}$ and μ_m, which is unknown to the agents.

Algorithm 1: FedUCB-UE

```
Initialization: Each agent explores each arm once and gets
  $\widehat{X}_{i,m}(0) = X_{i,m}(0)$ and communicates with server to get
  $\widehat{X}_m(0)$. Each agent sets $n_{i,m}(0) = 1$ and the server sets
  $n_m(0) = 1$
Agent $i$:
for $t = 1, 2, \cdots$ do
    Compute
      $S_i(t) = \{m \in [M] \mid n_{i,m}(t-1)E < n_m(\lfloor \frac{t-1}{E} \rfloor E)\}$
    if $S_i(t) \neq \emptyset$ then
        $A_i(t)$ is sampled randomly from $S_i(t)$
    else
        Compute $\mathsf{UCB}_{i,m}(t)$ according to (3)
        $A_i(t) = \arg\max_m \mathsf{UCB}_{i,m}(t)$
    end
    Pull arm $m = A_i(t)$ and observe $X_{i,m}(t)$
    $n_{i,m}(t) = n_{i,m}(t-1) + 1$
    Update $\widehat{X}_{i,m}(t)$ according to (6)
    if $t \mod E = 0$ then
        Send $\widehat{X}_{i,m}(t)$ and $n_{i,m}(t)$ to the server
        Receive $\widehat{X}_m(t)$ and $n_m(t)$ from the server
    end
end
Server:
for $t = 1, 2 \cdots$ do
    if $t \mod E = 0$ then
        Receive $\widehat{X}_{i,m}(t)$, $n_{i,m}(t)$ from all agents
        Update $\widehat{X}_m(t), n_m(t)$ according to (7)
        Broadcast $\widehat{X}_m(t)$, $n_m(t)$ to all agents
    end
end
```

4. THEORETICAL GUARANTEE

In this section, we first explain the reason why we choose such confidence levels in (4) based on a conjecture in Section 4.1 and then provide our regret bound and discussion in Section 4.2. The proof sketch of Theorem 1 will be given in Section 4.3.

4.1. Conjecture and its numerical verification

Besides the global maximal pulling counter $n_m(t)$, $C_m(t)$ is another source of global information. Instead of setting the agent-based confidence levels, we use the global-based one to prevent agents from underestimating the global reward. Since $B_{i,m}(t)$ is sub-Gaussian, the choice of the confidence level depends on the upper bound of the variance of $\widehat{X}_m(t)$ given by

$$\sigma_m^2(t) = \frac{1}{N^2} \sum_{i=1}^{N} \frac{1}{n_{i,m}(t)} \ . \tag{8}$$

We can bound (8) by lower bounding $\{n_{i,m}(t) : i \in [N]\}$. For simplicity, we use the shorthand $n_m(t) = n_m(\lfloor \frac{t}{E} \rfloor E)$. By noting the role of $S_i(t)$ in the adjusting rule, we have the natural lower bound $n_{i,m}(t) \geq n_m(t)/E - 1$. However, this bound is loose since the term E in the denominator in $S_i(t)$ is used to prevent the extreme cases at the initial stages that $M-1$ arms are underexplored by agent i and we need to make sure $S_i(t)$ is empty after E local rounds, i.e., agent i tries all the arms in $S_i(t)$. But the probability of this extreme case will be gradually vanishing since $\mathsf{UCB}_{i,m}(t)$ is a sub-Gaussian variable with a vanishing variance as $t \to \infty$, which means that the choices of different agents tends to achieve consensus. As a result, $n_{i,m}$ and n_m grow at almost the same rates, and $\forall H > 1$, the probability $\mathbb{P}\{n_{i,m}(t) < \frac{n_m(t)}{H} - 1\}$ diminishes as $t \to \infty$, since the coefficient of $n_{i,m}(t)$ is larger than that of $n_m(t)$. This means that the essential global consistency lower bound does not depend on E. To summarize, we have the following conjecture.

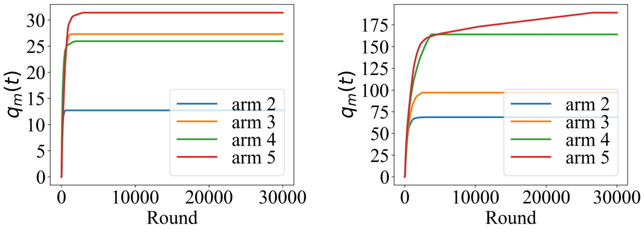
Conjecture 1 *Define the event* $D_{i,m}(t) \triangleq \{n_{i,m}(t) \geq \frac{n_m(t)}{2} - 1\}$*, and define the complementary event* $D^{\mathrm{c}}_{i,m}(t) \triangleq \{n_{i,m}(t) < \frac{n_m(t)}{2} - 1\}$*. Assume that for all non-optimal arms* $m \neq 1$*, we have*

$$q_{i,m} = \sum_{t=1}^{\infty} \mathbb{P}\{D^{\mathrm{c}}_{i,m}(t)\} < +\infty \ . \tag{9}$$

Note that by Borel-Cantelli lemma, this conjecture indicates that $n_{i,m}(t)$ is asymptotically lower bounded by $\frac{n_m(t)}{2} - 1$ almost surely as $t \to \infty$. Invoking Conjecture 1, we can choose the global confidence level as in (4).

The advantages of this confidence level are: 1) it is negatively related to the number of agents, and 2) it is time-varying, which does not require the knowledge of termination horizon T, and it permits the agents to choose arms flexibly at the initial stages.

To assess Conjecture 1, consider an FMAB setting with $N = 20$ agents and $M = 5$ arms. We set $\mu_m = (5 - m)/100$ and generate $\mu_{i,m}$ from a joint Gaussian distribution. And we run the experiment for $T = 30000$ rounds. The experiment is repeated 100 times to obtain the Monte-Carlo estimate of $\mathbb{P}\{D^{\mathrm{c}}_{i,m}(t)\}$ at each round t. We show the estimate of $\frac{1}{N}\sum_{i=1}^{N}\sum_{\tau=1}^{t}\mathbb{P}\{D^{\mathrm{c}}_{i,m}(\tau)\}$ denoted by $q_m(t)$ in Figure 2 for $E = 10$ and $E = 20$.



(a) E=10

(b) E=20

Fig. 2. Verification of Conjecture 1.

From Figure 2, for these two choices of E, $q_m(t)$ converges for all non-optimal arms m, which implies that $q_{i,m}$ is finite for all $i \in [N]$ and $m \neq 1$.

4.2. Regret Bound

In this section, we provide a regret bound for the FedUCB-UE algorithm and discuss its implications.

Theorem 1 *When Conjecture 1 holds, and* $E \geq M$*, the regret for the* FedUCB-UE *algorithm is bounded as follows.*

$$R(T) \leq \sum_{i=1}^{N} \sum_{m \neq 1} \Delta_m \left(\max\left\{ \frac{32}{N\Delta_m^2} \log T, 1 \right\} + q'_{i,m} \right) , \tag{10}$$

where $q'_{i,m} \triangleq \frac{\pi^2}{3} + 2q_{i,m}(\frac{1}{E} + 1) + 1$ *and* $q_{i,m}$ *is specified in* (9).

Theorem 1 implies that FedUCB-UE has a regret bound that scales with $\mathcal{O}(\log T)$, which is the optimal rate [17]. Even though involving more agents will not result in a lower regret $R(T)$, with more agents the algorithm will converge faster since the regret of each agent $R_i(T) \triangleq T\mu_1 - \sum_{t=1}^{T} \mathbb{E}[X_{A_i(t)}(t)]$ will decrease. Compared to Gossp_UCB [15], which has an $\mathcal{O}(N^2 \log T)$ regret bound, FedUCB-UE will significantly improve the regret bound. The PF-UCB algorithm also has $\mathcal{O}(\log T)$ regret but with a larger coefficient when the termination horizon T is given. In Section 5, through numerical evaluations, we show that FedUCB-UE outperforms these algorithms.

4.3. Proof sketch

We provide a proof sketch for Theorem 1 in this section. We start by observing that the regret can be decomposed as

$$R(T) = \sum_{i=1}^{N} \Delta_m \mathbb{E}[n_{i,m}(T)] \ . \tag{11}$$

To bound $R(T)$, we need to bound $\mathbb{E}[n_{i,m}(T)]$. If agent i chooses arm m instead of the optimal arm 1 at time t, at least one of the following four cases will happen: **case (1):** $m \in S_i(t)$; **case (2):** $B_{i,m}(t-1) - \mu_m \geq C_m(t-1)$; **case (3):** $u_1 - B_{i,m}(t-1) \geq C_1(t-1)$; **case (4):** $u_1 - u_m < 2C_m(t-1)$.

We first set t_0 to be large enough such that for all $t > t_0$ case (4) does not hold, i.e., $n_{i,m}(t_0) = \frac{32 \log T}{N\Delta_m^2} + 3$. To bound the occurrence times of case (2) and case (3), we need the following lemma.

Lemma 1 *For the* FedUCB-UE *algorithm,* $\forall i \in [N]$ *and* $m \in [M]$*, if* r_0 *is large enough such that* $n_m(r_0 E) \geq 3$*, then for all rounds* $t > r_0 E$*, we have*

$$\mathbb{P}(|B_{i,m}(t) - \mu_m| \geq C_m(t)) \leq \frac{2}{(t+1)^2}\mathbb{1}\{D_{i,m}(t)\} + 2\mathbb{1}\{D^{\mathrm{c}}_{i,m}(t)\} \ .$$

Note that $B_{i,m}(t)$ is sub-Gaussian and its variance can be bounded by $\frac{E}{N(n_m(t)-E)}\mathbb{1}\{D^c_{i,m}(t)\} + \frac{2}{N(n_m(t)-2)}\mathbb{1}\{D_{i,m}(t)\}$. The proof is complete by using the definition of sub-Gaussian variables.

The occurrence time of case (2) and case (3) after sufficiently large time t_0 can be bounded as $\frac{\pi^2}{6} - 1 + q_{i,m}$. The last step is to bound case (1) by two parts: i) the additional occurrence time of the set $S_i(t)$ is nonempty caused by case (2) and case (3) are no larger than $\frac{2q_{i,m}}{E} - 1$, ii) the adjusting rule needs the lower bound $n_{i,m}(t) \geq \frac{n_m(t)}{E}$, thus, each arm needs one additional exploration step to meet the gap.

Combining the results above, we have

$$\mathbb{E}[n_{i,m}(T)] \leq \frac{32}{N\Delta_m^2}\log T + q'_{i,m} \ , \tag{12}$$

where $q'_{i,m} = \frac{\pi^2}{3} + 2q_{i,m}(\frac{1}{E} + 1) + 1$.

5. NUMERICAL EXPERIMENTS

In this section, we conduct numerical evaluations to assess the performance of the FedUCB-UE algorithm. We compare our algorithm with the following two state-of-the-art algorithms.

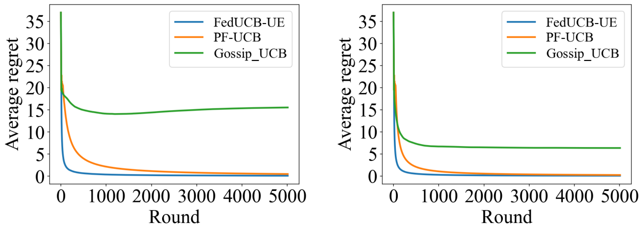
Gossip_UCB [15]: each agent communicates with only one of its neighbors at each local exploration round, i.e., $E = 1$. We set the topology of agents to be fully connected for a fair comparison.

PF-UCB [17]: since the global regret is considered, during two communication rounds, each agent of the PF-UCB algorithm explores each arm in the candidate set of the best arm for the same number of times. And we set the exploration time to be $\frac{E}{2}$ for a fair comparison.

We use a synthetic dataset to evaluate the performance of FedUCB-UE. We set $N = 20$, $M = 10$, and $E = 10$, and run the experiment with the mean rewards generated by

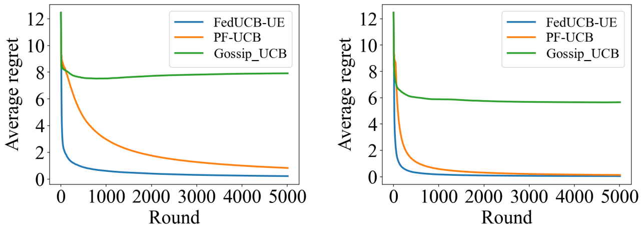
$$\textbf{Model 1: } \mu_{i,m} \sim \mathcal{N}(\tfrac{20-m}{5}, 1) \ ,$$
$$\textbf{Model 2: } \mu_{i,m} \sim \mathcal{N}(\tfrac{20-m}{20}, 1) \ .$$

Model 1 is simpler for agents to find the best arm since the non-optimal gap Δ_m is relatively large. We repeat the experiments for each model 100 times and plot the average regret $\frac{R(t)}{t}$. Besides using theoretical thresholds for all the algorithms, we also perform the grid search to find optimally-tuned thresholds for each algorithm. Figure 3 shows the average regret $\frac{R(t)}{t}$ for Model 1 with the theoretical threshold and the optimally-tuned threshold, while Figure 4 shows the results under the same metrics for Model 2.



(a) Theoretical threshold (b) Optimally-tuned threshold

Fig. 3. Average regret for different methods in Model 1.



(a) Theoretical threshold (b) Optimally-tuned threshold

Fig. 4. Average regret for different methods in Model 2.

Both Figure 3 and Figure 4 show that FedUCB-UE outperforms the other two in terms of average regret in both the cases of the theoretical or optimally-tuned threshold. Furthermore, comparing Figure 3(a) and Figure 3(b) (Figure 4(a) and Figure 4(b)) for the same algorithms implies that the theoretical threshold for FedUCB-UE almost matches the optimally-tuned threshold.

6. CONCLUSIONS

In this work, we have proposed a novel adaptive federated multi-armed bandit algorithm. Distinct from the existing algorithms, the proposed FedUCB-UE algorithm allows the agents to form local decisions and obtains the global information only intermittently. Our theoretical analysis shows that FedUCB-UE achieves the optimal $\mathcal{O}(\log T)$ regret bound asymptotically. The numerical tests demonstrate that FedUCB-UE outperforms the state-of-the-art algorithms. Future research includes extending FedUCB-UE to personalized setting and decentralized settings.

7. REFERENCES

[1] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, no. 3/4, pp. 285–294, 1933.

[2] H. Robbins, "Some aspects of the sequential design of experiments," *Bulletin of the American Mathematical Society*, vol. 58, no. 5, pp. 527–535, 1952.

[3] L. Li, W. Chu, J. Langford, and R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation," in *Proc. International Conference on World Wide Web*, Raleigh, NC, April 2010, pp. 661–670.

[4] A. Anandkumar, N. Michael, A. K. Tang, and A. Swami, "Distributed algorithms for learning and cognitive medium access with logarithmic regret," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 4, pp. 731–745, 2011.

[5] X. Huo and F. Fu, "Risk-aware multi-armed bandit problem with application to portfolio selection," *Royal Society Open Science*, vol. 4, no. 11, pp. 171377, 2017.

[6] T. Lattimore and C. Szepesvári, *Bandit Algorithms*, Cambridge University Press, Cambridge, UK, 2020.

[7] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Artificial Intelligence and Statistics*, Fort Lauderdale, FL, April 2017, vol. 54, pp. 1273–1282.

[8] T. Li, L. Song, and C. Fragouli, "Federated recommendation system via differential privacy," in *Proc. IEEE International Symposium on Information Theory*, CA, USA, June 2020, pp. 2592–2597.

[9] Yuanhao Wang, Jiachen Hu, Xiaoyu Chen, and Liwei Wang, "Distributed bandit learning: Near-optimal regret with efficient communication," in *Proc. International Conference on Learning Representations*, Addis Ababa, Ethiopia, April 2019.

[10] A. Dubey and A. Pentland, "Differentially-private federated linear bandits," in *Proc. Neural Information Processing Systems*, Vancouver, Canada, December 2020.

[11] A. Mitra, H. Hassani, and G. Pappas, "Robust federated best-arm identification in multi-armed bandits," *arXiv:2109.05700*, 2021.

[12] S. Shahrampour, A. Rakhlin, and A. Jadbabaie, "Multi-armed bandits in multi-agent networks," in *Proc. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, New Orleans, LA, March 2017, pp. 2786–2790.

[13] D. Vial, S. Shakkottai, and R. Srikant, "Robust multi-agent multi-armed bandits," in *Proc. International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, Shanghai, China, July 2021, pp. 161–170.

[14] M. Agarwal, V. Aggarwal, and K. Azizzadenesheli, "Multi-agent multi-armed bandits with limited communication," *arXiv:2102.08462*, 2021.

[15] Z. Zhu, J. Zhu, J. Liu, and Y. Liu, "Federated bandit: A gossiping approach," in *Proc. ACM International Conference on Measurement and Modeling of Computer Systems*, Beijing, China, June 2021.

[16] J. Zhu and J. Liu, "A distributed algorithm for multi-armed bandit with homogeneous rewards over directed graphs," in *Proc. IEEE American Control Conference*, New Orleans, LA, May 2021, pp. 3038–3043.

[17] C. Shi, C. Shen, and J. Yang, "Federated multi-armed bandits with personalization," in *Porc. International Conference on Artificial Intelligence and Statistics*, Virtual, April 2021, pp. 2917–2925.

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