Federated Multi-Armed Bandits

Chengshuai Shi, Cong Shen
Department of Electrical and Computer Engineering, University of Virginia
Charlottesville, VA 22904
{cs7ync, cong}@virginia.edu

Abstract

Federated multi-armed bandits (FMAB) is a new bandit paradigm that parallels the federated learning (FL) framework in supervised learning. It is inspired by practical applications in cognitive radio and recommender systems, and enjoys features that are analogous to FL. This paper proposes a general framework of FMAB and then studies two specific federated bandit models. We first study the approximate model where the heterogeneous local models are random realizations of the global model from an unknown distribution. This model introduces a new uncertainty of client sampling, as the global model may not be reliably learned even if the finite local models are perfectly known. Furthermore, this uncertainty cannot be quantified a priori without knowledge of the suboptimality gap. We solve the approximate model by proposing Federated Double UCB (Fed2-UCB), which constructs a novel "double UCB" principle accounting for uncertainties from both arm and client sampling. We show that gradually admitting new clients is critical in achieving an order-optimal regret while explicitly considering the communication cost. The exact model, where the global bandit model is the exact average of heterogeneous local models, is then studied as a special case. We show that, somewhat surprisingly, the order-optimal regret can be achieved independent of the number of clients with a careful choice of the update periodicity. Experiments using both synthetic and real-world datasets corroborate the theoretical analysis and provide interesting insight into the proposed algorithms.

1 Introduction

Federated learning (FL) (McMahan et al. 2017) is a new distributed machine learning (ML) paradigm that addresses new challenges in modern machine learning (ML). In particular, FL handles distributed ML with the following characteristics:

- **Non-IID local datasets.** FL caters to the growing trend that massive amount of the real-world data are generated directly at the edge devices. The local datasets are likely drawn from non-independent and identically distributed (non-IID) distributions, and do not represent the global distribution.
- **Massively distributed.** The number of participating clients can be significant, e.g., on the order of millions (Bonawitz et al. 2019).
- **Communication efficiency.** The communication cost scales with the number of clients, which becomes one of the primary bottlenecks of the FL system (McMahan et al. 2017). It is critical to minimize the communication cost while maintaining the learning accuracy.
- **Privacy.** FL protects the local data privacy by only sharing model updates instead of the raw data.

While the state of the art FL largely focuses on the supervised learning setting, we propose to extend the core principles of FL to the multi-armed bandits (MAB) problem. This is motivated by real-world applications, such as:

- **Cognitive radio.** A base station wants to select one channel from a given set of channels that is most likely to be "empty" in its coverage area. It is well-known that different geographic locations have different channel availabilities, and the (ground-truth) global channel availability is the average over the entire coverage area (see Section 2.1 for a detailed discussion). The base station, however, is fixed at one location and cannot learn the global channel availability by itself. A common solution is to utilize randomly placed devices (e.g., mobile phones) in the coverage area to sample the channels and then aggregate at the base station. Each device is at a different location and thus samples a non-IID local channel availability. In addition, the bandit problem is approximate because there are only finite devices while the global model is integrated over the entire coverage area.
- **Recommender system.** The central server wants to recommend the most popular item to new customers to maximize the expected reward. The server does not initially have the global item popularity but can learn via interacting with customers, leading to a bandit problem (Li et al. 2010). In reality, however, the server may not learn the popularity model *directly* from user behavior data due to privacy concerns or regulation requirement (e.g., private user data in some regions may not be shared outside). Instead, the user data are stored strictly on the client device or local server (and never leave) for better privacy preservation. The local view of item popularity is often biased and not representative of the overall distribution, while the global server can only access some aggregate information instead of individual data, resulting in a federated learning problem in the bandit setting.

In both applications, which are illustrated in Figs. 1 and 2, we see that the general FL characteristics still apply, and

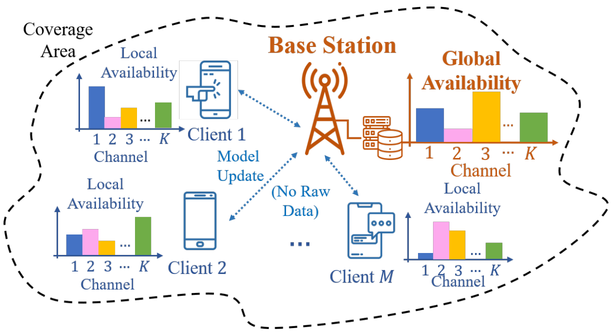

Figure 1: Motivating example–cognitive radio.

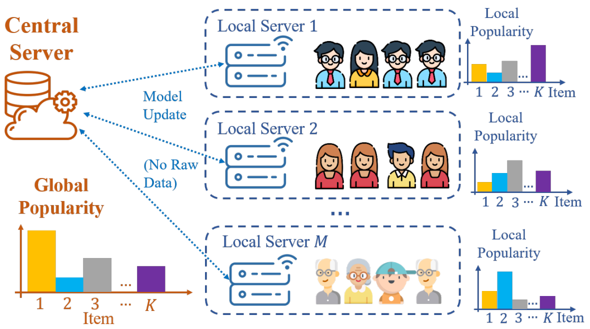

Figure 2: Motivating example–recommender system.

yet the underlying problem is a bandit one. This leads to a natural marriage of FL and MAB, and we are motivated to solve a global stochastic MAB problem from (possibly a large number of) local bandit models that are non-IID, in a communication-efficient and privacy-preserving manner.

In this work, a novel framework of Federated MAB (FMAB) is developed, which represents the first *systematic* attempt to bridge FL and MAB to the best of our knowledge. The FMAB framework is general and can incorporate a variety of bandit problems that share the FL principles. We demonstrate the merit of this framework by first studying an approximate FMAB model, where the global bandit model exists as a ground truth while local bandit models are random realizations of it. In addition to the usual reward uncertainty from arm sampling, this setting introduces a new uncertainty associated with client sampling. In particular, the approximate model does not assume any suboptimality gap knowledge, which prohibits determining the requirement for client sampling *a priori*. Mixing client sampling with arm sampling without the knowledge of suboptimality gap significantly complicates the problem, and we address these challenges by proposing a novel Federated Double UCB (Fed2-UCB) algorithm that gradually samples new clients while performing arm sampling, and thus simultaneously explores and balances both types of uncertainty. Theoretical analysis shows that Fed2-UCB achieves an $O(\log(T))$ regret (which explicitly considers communication cost) that approaches the lower bound of the standard stochastic MAB model with an additional term of communication loss. As a special case, the exact FMAB model is then studied, where the global model is the exact average of all local models. The Fed1-UCB algorithm degenerates from Fed2-UCB and achieves an order-optimal regret upper bound which, somewhat surprisingly, is independent of the number of clients with a proper choice of the update periodicity. Numerical simulations on synthetic and real-world datasets demonstrate the effectiveness and efficiency of the proposed algorithms and offer some interesting insights.

2 Problem Formulation

In the standard stochastic MAB setting, a single player directly plays K arms, with rewards X_k of arm $k \in [K]$ sampled independently from a σ-subgaussian distribution with mean μ_k. At time t, the player chooses an arm $\pi(t)$ and the goal is to receive the highest expected cumulative reward in T rounds, which is characterized by minimizing the (pseudo-)regret:

$$R(T) = \mathbb{E}\left[\sum\nolimits_{t=1}^{T} X_{k_*}(t) - \sum\nolimits_{t=1}^{T} X_{\pi(t)}(t)\right], \tag{1}$$

where k_* is the optimal arm with mean reward $\mu_* \doteq \mu_{k_*} = \max_{k\in[K]} \mu_k$, and the expectation is taken over the randomness of both policy and environment. As shown by Lai and Robbins (1985), there exists a lower bound for the regret as:

$$\liminf_{T\to\infty} \frac{R(T)}{\log(T)} \geq \sum\nolimits_{k\neq k_*} \frac{\mu_* - \mu_k}{\mathrm{kl}(\mu_k, \mu_*)}, \tag{2}$$

where $\mathrm{kl}(\mu_k, \mu_*)$ denotes the KL-divergence between the two corresponding distributions.

In this section, we present a framework of FMAB as illustrated in Fig. 3. The key aspects of the FL principles mentioned in Section 1 are also elaborated, which become more clear when algorithm designs are presented.

Clients. Multiple clients interact with the same set of K arms (referred as "local arms") in the FMAB framework. We denote M_t as the number of participating clients at time t, who are labeled from 1 to M_t to facilitate discussions (they are not used in the algorithms). A client can only interact with her own local MAB model, and there is no direct communication between clients. Arm k generates independent *observations* $X_{k,m}$ for client m following a σ-subgaussian distribution with mean $\mu_{k,m}$. Note that $X_{k,m}$ is only an observation but not a reward. For different clients $n \neq m$, their models are non-IID; hence $\mu_{k,n} \neq \mu_{k,m}$ in general.

Server. There exists a central server with a global stochastic MAB model, which has the same set of K arms (re-

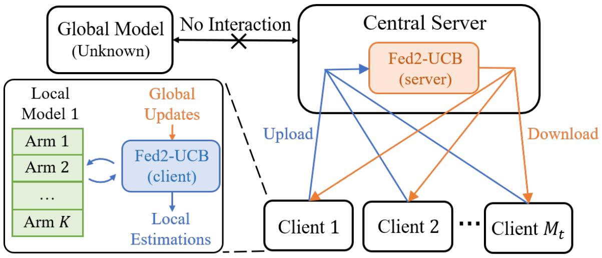

Figure 3: The FMAB framework.

ferred as "global arms") of σ-subgaussian reward distributions with mean reward μ_k for arm k. The true rewards for this system are generated on this global model, thus the learning objective is on the global arms. However, the server cannot directly observe rewards on the global model; she can only interact with clients who feed back information of their local observations. We consider the general non-IID situation where the local models are not necessarily the same as the global model, and also make the common assumption that clients and the server are fully synchronized (McMahan et al. 2017; Bonawitz et al. 2019).

Communication cost. Although clients cannot communicate with each other, after certain time, they can transmit local "model updates" based on their local observations to the server, which aggregates these updates to have a more accurate estimation of the global model. The new estimation is then sent back to the clients to replace the previous estimation for future actions. However, just like in FL, the communication resource is a major bottleneck and the algorithm has to be conscious about its usage. We incorporate this constraint in FMAB by imposing a loss C every time a client communicates to the server, which will be accounted for in the performance measure defined below.

2.1 The Approximate Model

Although the non-IID property of local models is an important feature of FMAB, there must exist some relationship between local and global models so that observations on local bandit models help the server learn the global model. Here, we propose the approximate FMAB model, where the global model is a fixed (but hidden) ground truth (i.e., exogenously generated regardless of the participating clients), and the local models are IID random realizations of it.

Specifically, the global arm k has a fixed mean reward of μ_k. For client m, the mean reward $\mu_{k,m}$ of her local arm k is a sample from an unknown distribution ϕ_k, which is a σ_c-subgaussian distribution with mean μ_k. For a different client $n \neq m$, $\mu_{k,n}$ is sampled IID from ϕ_k. Since local models are stochastic realizations of the global model, a *finite* collection of the former may not necessarily represent the latter. In other words, if there are M involving clients, although $\forall m \in [M], \mathbb{E}[\mu_{k,m}] = \mu_k$, the averaged local model $\hat{\mu}_k^M \doteq \frac{1}{M}\sum_{m=1}^{M} \mu_{k,m}$ may not be consistent with the global model. Specifically, $\hat{\mu}_k^M$ is not necessarily equal (or even close) to μ_k, which introduces significant difficulties. Intuitively, the server needs to sample sufficiently many clients to have a statistically accurate estimation of the global model, but as we show later, the required number of clients cannot be obtained *a priori* without the suboptimality gap knowledge. The need of client sampling also coincides with the property of massively distributed clients in FL.

Motivation Example. The approximate model captures the key characteristics of a practical cognitive radio system, as illustrated in Fig. 1. Assume a total of K candidate channels, indexed by $\{1, ..., K\}$. Each channel's availability is location-dependent, with $p_k(x)$ denoting the probability that channel k is available at location x. The goal of the base station is to choose one channel out of K candidates to serve all potential cellular users (e.g., control channel) in the given coverage area $\mathcal{D}$ with area D. Assuming users are uniformly randomly distributed over $\mathcal{D}$, the global channel availability is measured throughout the entire coverage area as

$$p_k = \mathbb{E}_{x\sim u(\mathcal{D})}\left[p_k(x)\right] = \oiint_{\mathcal{D}} \frac{1}{D} p_k(x) dx. \tag{3}$$

It is well known in wireless research that a base station cannot directly sample p_k by itself, because it is fixed at one location[1].In addition, Eqn. (3) requires a *continuous* sampling throughout the coverage area, which is not possible in practice. Realistically, the base station can only direct cellular user m at *discrete* location x_m to estimate $p_k(x_m)$, and then aggregate observations from finite number of users as $\hat{p}_k = \frac{1}{M}\sum_{m=1}^{M} p_k(x_m)$ to approximate p_k. Clearly, even if $p_k(x_m)$ are perfect, $\hat{p}_k$ may not necessarily represent p_k well.

Regret definition. Without loss of generality, we assume there is only one optimal global arm k_* with $\mu_* \doteq \mu_{k_*} = \max_{k\in[K]} \mu_k$, and $\Delta = \mu_* - \max_{k\neq k_*}\{\mu_k\}$ denotes the suboptimality gap of the global model (both unknown to the algorithm). We further denote $\gamma_1, \cdots, \gamma_{T_c}$ as the time slots when the clients communicate with the central server for both upload and download. The notion of (pseudo-)regret in Eqn. (1) for the single-player model can be generalized to all the clients with additional communication loss, as follows:

$$R(T) = \mathbb{E}\Bigg[\underbrace{\sum_{t=1}^{T} M_t X_{k_*}(t) - \sum_{t=1}^{T}\sum_{m=1}^{M_t} X_{\pi_m(t)}(t)}_{\text{exploration and exploitation}} + \underbrace{\sum_{\tau=1}^{T_c} C M_{\gamma_\tau}}_{\text{communication}} \Bigg], \tag{4}$$

where $\pi_m(t)$ is the arm chosen by client m at time t. In this work, we aim at designing algorithms with $O(\log(T))$ regret as in the single-player setting.

Several comments are in place for Eqn. (4). First, the reward oracle is defined with respect to the single global optimal arm but not the distinct local optimal arms. This choice is analogous to that the reward oracle of the pseudo-regret in the single-player MAB model is defined with respect to *one* optimal arm throughout the horizon but not the arms with the highest reward at every time slot (Bubeck and Cesa-Bianchi 2012). Second, the cumulative reward of the system is defined on the global model, because clients only receive *observations* from playing the local bandit game, and the *reward* is generated at the system-level global model. Taking the cognitive radio system as an example, the choice by each client only produces her observation of the channel availability, but the reward is generated by the base station when this channel is used for the entire coverage area. Lastly, regret definition in Eqn. (4) discourages the algorithm to involve too many clients. Ideally, only sufficiently many clients should be admitted to accurately reconstruct the global model, and any more clients would result in more communication loss without improving the model learning.

[1] The best it can do by itself is to estimate $p_k(x_{\text{BS}})$ where x_{BS} is the location of the base station.

3 Fed2-UCB for Approximate FMAB

3.1 Challenges and Main Ideas

The first and foremost challenge in the approximate model comes from that the local models are only stochastic realizations of the global model. Even with the perfect information of all local arms, the optimal global arm may not be produced faithfully. We refer to this new problem as the *uncertainty from client sampling*. How to simultaneously handle the two types of uncertainty (client sampling and arm sampling) is at the center of solving the approximate model.

A second issue comes from the conflict between non-IID local models and the global model. In particular, the globally optimal arm may be sub-optimal for a client's local model, and hence it cannot be correctly inferred by the client individually. Communication between clients and the server is key to address this conflict, but the challenge is how to control the communication loss and balance the overall regret.

In this section, we first characterize the uncertainty from client sampling by analyzing the probability that the averaged local model does not faithfully represent the global model, and illustrate that without knowledge of the suboptimality gap Δ, the algorithm cannot determine *a priori* the number of required clients. Then, Federated Double UCB (Fed2-UCB) is proposed, in which a novel "double UCB" principle carefully balances and trades off the two sources of uncertainty while controlling the communication cost.

3.2 Client Sampling

In the approximate model, the key to determine whether the local knowledge is sufficient lies in whether the optimal global arm can be inferred correctly. When there are M involving clients, the best approximate of the global model is the averaged local model, i.e., $\hat{\mu}_k^M$. Although the utilities of local arms may be different from the global model, if the true optimal global arm is still optimal in this averaged local model, i.e., $\hat{\mu}_{k_*}^M > \max_{k\neq k_*} \hat{\mu}_k^M$, a sub-linear regret can be achieved with local knowledge. Otherwise, arm k_* is not optimal with respect to $\hat{\mu}_k^M$, and no matter how many explorations are performed locally (even with perfect local knowledge), the global optimal arm cannot be found using the sampled M local models and thus a linear regret occurs.

The following theorem characterizes the accuracy of representing the global model by the averaged local model from a fixed number of clients.

Theorem 1. *With M involved clients, denote $P_z = \mathbb{P}\left(\hat{\mu}_{k_*}^M \leq \max_{k\in[K]} \hat{\mu}_k^M\right)$, the following result holds:*

$$P_z = O\left(\sum\nolimits_{k\neq k_*} \exp\left\{-\sigma_c^{-2} M(\mu_* - \mu_k)^2\right\}\right) = O\left(K \exp\left\{-\sigma_c^{-2} M \Delta^2\right\}\right).$$

Theorem 1 indicates that the probability that the averaged local model does not represent the global model, i.e., $\hat{\mu}_{k_*}^M \leq \max_{k\in[K]} \hat{\mu}_k^M$, decreases exponentially with respect to the number of involved clients M. Thus, it is fundamental to involve a sufficiently large number of clients in order to reconstruct the global model correctly. More specifically, to guarantee that $P_z = O(1/T)$, by which the overall regret can scale sub-linearly, it is sufficient to sample M clients with

$$M = \Omega\left(\sigma_c^2 \Delta^{-2} \log(KT)\right). \tag{5}$$

Algorithm 1 Fed2-UCB: client m

```
Initialize $p \leftarrow 1$; $[K_1] \leftarrow [K]$
while $K_p > 1$ do
    Pull each active arm $k \in [K_p]$ for $f(p)$ times
    Calculate the local sample means $\bar{\mu}_{k,m}(p), \forall k \in [K_p]$
    Send local updates $\bar{\mu}_{k,m}(p), \forall k \in [K_p]$ to the server
    Receive global update set $E_p$ from the server
    $[K_{p+1}] \leftarrow [K_p] \backslash E_p$; $p \leftarrow p + 1$
end while
$F$ ←the only element in $[K_p]$; Stay on arm $F$ until $T$
```

Algorithm 2 Fed2-UCB: central server

```
Initialize $p \leftarrow 1$; $[K_1] \leftarrow [K]$
while $K_p > 1$ do
    Admit $g(p)$ new clients                          ▷ Client sampling
    Receive local updates $\bar{\mu}_{k,m}(p), \forall k \in [K_p], \forall m \in [M(p)]$
    Calculate $\forall k \in [K_p], \bar{\mu}_k(p) \leftarrow \sum_{m=1}^{M(p)} \bar{\mu}_{k,m}(p)/M(p)$
    $E_p \leftarrow \left\{k \in [K_p] | \bar{\mu}_k(p) + B_{p,2} \leq \max_{l\in[K_p]} \bar{\mu}_l(p) - B_{p,2}\right\}$
    Send global update set $E_p$ to all involved clients
    $[K_{p+1}] \leftarrow [K_p] \backslash E_p$; $p \leftarrow p + 1$
end while
```

If Eqn. (5) is satisfied throughout the bandit game, the optimal arm can be successfully found. However, clients do not have access to the knowledge of Δ. Thus, the requirement in Eqn. (5) cannot be guaranteed in advance.

On the other hand, involving too many clients may be detrimental to the regret, as can be seen in Eqn. (4). Specifically, in order to have an $O(\log(T))$ regret, M should satisfy:

$$M = O\left(\log(T)\right). \tag{6}$$

Comparing Eqns. (5) and (6) suggests that M has to be $\Theta(\log(T))$ to achieve a correct representation of the global model while maintaining an $O(\log(T))$ regret.

3.3 The Fed2-UCB Algorithm

With the unknown requirement in Eqn. (5), it is unwise to only admit a small number of clients in the whole game. On the other hand, Eqn. (6) prohibits involving too many clients to achieve an $O(\log(T))$ regret. There are also practical system considerations that prevent having too many clients, which has been discussed in the context of FL (Bonawitz et al. 2019). We propose the Fed2-UCB algorithm where the central server gradually admits new clients into the game after each communication round while keeping local clients gathering observations. The method of gradually increasing the clients ensures that the server samples a set of small but sufficiently representative clients based on the underlying statistical structure of the bandit game. The proposed "double UCB" principle simultaneously addresses the uncertainty from both client sampling and arm sampling.

The Fed2-UCB algorithm is performed in phases simultaneously and synchronously at clients and the central server. Clients collect observations and update local estimations for the arms that have not been declared as sub-optimal, i.e., the active arms, while the server admits new clients and aggregates the local estimations as global estimations to eliminate sub-optimal active arms. We denote the set of active arms in the p-th phase by $[K_p]$ with cardinality K_p. The detailed algorithm for the clients and the central server are given in Algorithms 1 and 2, respectively.

At phase p, $g(p)$ new clients are first added into the game by the server. These clients can be viewed as interacting with newly sampled local MAB models. Each client, regardless of newly added or not, performs a sequential arm sampling among the currently active arms for $K_p f(p)$ times on their own local models, which means each active arm is pulled $f(p)$ times by each client. Thus, arm $k \in [K_p]$ is played a total of $M(p)f(p)$ times in phase p, where $M(p) = \sum_{q=1}^{p} g(q)$ is the overall number of clients at phase p. Parameters $g(p)$ and $f(p)$ are flexible and we discuss the impact of these choices on the regret in the next section. It is worth noting that the rate of admitting new clients is determined not only by $g(p)$ but also by $f(p)$, which characterizes the frequency of client sampling. With new observations from arm sampling, each client m updates her local estimations, i.e., sample mean $\bar{\mu}_{k,m}(p), k \in [K_p]$, then sends them to the central server as a local parameter update. Note that uploading sample means instead of raw samples benefits the preservation of privacy, and additional methods for better privacy protection are presented in the supplementary material.

After receiving local parameter updates from the clients, the central server first updates the global estimation as the average of them for each active arm, i.e., $\bar{\mu}_k(p) = \frac{1}{M(p)} \sum_{m=1}^{M(p)} \bar{\mu}_{k,m}(p), k \in [K_p]$. While recognizing two coexisting uncertainties, a "double" confidence bound $B_{p,2}$ is adopted to characterize them simultaneously as:

$$B_{p,2} = \underbrace{\sqrt{6\sigma^2 \eta_p \log(T)}}_{\text{arm sampling}} + \underbrace{\sqrt{6\sigma_c^2 \log(T)/M(p)}}_{\text{client sampling}},$$

where $\eta_p = \frac{1}{M(p)^2} \sum_{q=1}^{p} \frac{g(q)}{F(p)-F(q-1)}$ and $F(p) = \sum_{q=1}^{p} f(q)$ with $F(0) = 0$. The first terms in $B_{p,2}$ characterizes the uncertainty from arm sampling, which illustrates the gap between the averaged sampled local model and the exact averaged local model. The second term represents the uncertainty from client sampling, which captures the gap between the exact averaged local model and the (hidden) global model. Note that these two types of uncertainty are not independent of each other, since more admitted clients can perform more pulls on arms, thus reducing both simultaneously.

With the global estimations and the confidence bound, the elimination set E_p is determined by the server, which contains arms that are sub-optimal with a high probability:

$$E_p = \left\{k \in [K_p] \middle| \bar{\mu}_k(p) + B_{p,2} \leq \max_{l \in [K_p]} \bar{\mu}_l(p) - B_{p,2}\right\}.$$

The set $[E_p]$ is then sent back to the clients, who then remove these arms from their sets of active arms. This iteration keeps going until there is only one active arm left.

3.4 Regret Analysis

The regret of the Fed2-UCB algorithm is the combination of the exploration loss and communication loss, and relies on the design of $g(p)$ and $f(p)$.

Theorem 2. *For $k \neq k_*$, we denote $\Delta_k = \mu_* - \mu_k$ and p_k as the smallest integer p such that*

$$96\left(\sigma\sqrt{\eta_p} + \sigma_c/\sqrt{M(p)}\right)^2 \log(T) \leq \Delta_k^2, \tag{7}$$

and $p_{\max} = \max_{k \neq k_*}\{p_k\}$. If $\max_{t \leq T}\{M_t\} \leq \beta T$, where β is a constant, the regret for the Fed2-UCB algorithm satisfies

$$R_2(T) \leq \sum\nolimits_{k \neq k_*} \sum\nolimits_{q=1}^{p_k} \Delta_k M(q) f(q) + C \sum\nolimits_{q=1}^{p_{\max}} M(q) + 4\beta(1+C)K.$$

Eqn. (7) describes the requirement for phase p_k under two types of uncertainty, by which the sub-optimal arm k is guaranteed to be eliminated with a high probability. For it to hold, eventually we need at least $O(\log(T))$ clients in the game, which coincides with Eqn. (5).

Theorem 2 provides a general description, using unspecified $g(p)$ and $f(p)$. A better characterization can be had with more specific choices.

Corollary 1. *With $f(p) = \kappa$ where κ is a constant, and $g(p) = 2^p$, the asymptotic regret of Fed2-UCB is*

$$R_2(T) = O\left(\sum_{k \neq k_*} \frac{\kappa(\sigma/\sqrt{\kappa} + \sigma_c)^2 \log(T)}{\Delta_k} + C\frac{(\sigma/\sqrt{\kappa} + \sigma_c)^2 \log(T)}{\Delta^2}\right).$$

Corollary 1 shows that with carefully designed $f(p) = \kappa$ and $g(p) = 2^p$, Fed2-UCB can achieve a regret of $O(\log(T))$. The exploration loss approaches the single-player MAB lower bound[2] in Eqn. (2) (Lai and Robbins 1985), which shows the effectiveness of exploration in Fed2-UCB. Since at least $O(\log(T))$ clients need to be involved as indicated by Eqn. (5), an $O(\log(T))$ communication loss achieved in Corollary 1 is inevitable, which demonstrates the communication efficiency. The overall regret in Corollary 1 proves that Fed2-UCB can effectively deal with two types of uncertainty while balancing the communication loss.

The choice of $g(p) = 2^p$ and $f(p) = \kappa$ leads to an exponentially decreasing $B_{p,2}$, which can be viewed as maintaining an exponentially decreasing estimation $\hat{\Delta}$ of Δ and eliminating arms with a larger gap (Auer and Ortner 2010); thus, it naturally solves the difficulty associated with the unknown Δ. The regret behavior of several other choices of $f(p)$ and $g(p)$ are given in the supplementary material.

4 Special Case: Exact Model and Fed1-UCB

While the approximate model introduces two types of uncertainty simultaneously, here we study a special case of the *exact model*, where the uncertainty from client sampling does not exist. Correspondingly, the Fed1-UCB algorithm, which degenerates from Fed2-UCB, is designed and analyzed.

4.1 The Exact Model

In the exact model, the number of clients is fixed, i.e., $M_t = M, \forall t$, and the global model is the *exact* average of all the local models, which means the global arm k has a mean reward of $\mu_k = \frac{1}{M}\sum_{m=1}^{M} \mu_{k,m}$. Thus, the global model can

[2] In the case with Bernoulli rewards, it can be observed that the two terms are of the same order by invoking the Pinsker's inequality $kl(\mu_i, \mu_j) \geq 2(\mu_j - \mu_i)^2$.

be perfectly reconstructed with information of local models and there only exists the uncertainty from arm sampling. The regret expression can be simplified to $R(T) = \mathbb{E}\left[\sum_{t=1}^{T} MX_{k_*}(t) - \sum_{t=1}^{T}\sum_{m=1}^{M} X_{\pi_m(t)}(t) + CMT_c\right]$. This model focuses on optimizing the performance for a fixed group of clients that do not change throughout the T time steps. In other words, the global model is not exogenously generated but adapts to the involved clients. Taken the recommender system as an example, the overall popularity of one item is the average of its popularity over the potential clients.

4.2 The Fed1-UCB Algorithm

Without the uncertainty from client sampling, there is no need of admitting new clients. The same exploration and communication procedure of Fed2-UCB is performed in Fed1-UCB without client admitting. The confidence bound used in arm eliminations is also degenerated from $B_{p,2}$ to $B_{p,1} = \sqrt{6\sigma^2 \log(T)/(MF(p))}$, which only characterizes the uncertainty from arm sampling. A complete description of Fed1-UCB is given in the supplementary material.

4.3 Theoretical Analysis

The regret for the Fed1-UCB algorithm only relies on $f(p)$ and is characterized by the following theorem.

Theorem 3. *For $k \neq k_*$, we denote $\Delta_k = \mu_* - \mu_k$, $F(p) = \sum_{q=1}^{p} f(q)$, p_k as the smallest integer p such that*

$$MF(p) \geq 96\sigma^2 \log(T)/\Delta_k^2, \tag{8}$$

and $p_{\max} = \max_{k\neq k_*}\{p_k\}$. The regret of Fed1-UCB satisfies

$$R_1(T) \leq M \sum\nolimits_{k\neq k_*} \Delta_k F(p_k) + CMp_{\max} + 2(1+C)MK.$$

Somewhat surprisingly, Eqn. (8) shows that although involving more clients leads to a faster convergence (i.e., smaller p_k), in general the overall necessary arm pulls performed by the clients, i.e., $MF(p_k)$, are independent of M. In other words, we can trade off the convergence time with number of clients without additional exploration loss.

Corollary 2. *With $f(p) = \lceil \kappa \log(T) \rceil$ where κ is a constant, the asymptotic regret of the Fed1-UCB algorithm is*

$$R_1(T) = O\left(\sum_{k\neq k_*} \frac{\sigma^2 \log(T)}{\Delta_k}\right).$$

Corollary 2 states that the exploration loss of Fed1-UCB approaches the single-player MAB lower bound in Eqn. (2) (Lai and Robbins 1985). It is also worth noting that with $f(p) = \lceil \kappa \log(T) \rceil$, the communication loss of Fed-1UCB is a non-dominating constant, which demonstrates its communication efficiency. Furthermore, the regret is independent of M asymptotically. The regret behavior with other choices of $f(p)$ are discussed in the supplementary material.

5 Experiments

Numerical experiments have been carried out under both applications of cognitive radio and recommender system. Their results are reported in this section to demonstrate the effectiveness and efficiency of Fed2-UCB and Fed1-UCB. For the cognitive radio example, due to the lack of suitable real-world datasets, synthetic datasets are used for simulations (Avner and Mannor 2014; Bande and Veeravalli 2019). For the recommender system, real-world evaluations are performed. The performance of a (hypothetical) single-player improved UCB algorithm (Auer and Ortner 2010) directly performed at the server is used as the baseline (labeled as "baseline"). The communication cost is set to be $C = 1$.

5.1 Synthetic Dataset for Cognitive Radio

A bandit game with $K = 10$ arms is used to mimic 10 candidate channels, and Gaussian distributions with $\sigma = 0.5$ are used to generate local observations of the channel availability. The means of global arms are in the interval $[0.7, 0.8]$ with $\Delta = 0.02$. We first start with the relatively simple exact model, where $M = 5$ clients are involved while arm 1 is not the optimal arm of any of their local models. As shown in Fig. 4, with $f(p) = \lceil 10 \log(T) \rceil$, if there is no communication loss, Fed1-UCB (labeled as "expl") achieves almost the same performance as the baseline, which proves its effectiveness. When considering the communication loss, the centralized version of Fed1-UCB (labeled as "cent"), where clients send their raw data in every time slot, has a very large regret due to significant yet unnecessary communications. However, with $f(p) = \lceil 10 \log(T) \rceil$, Fed1-UCB only incurs a small communication loss, which proves its efficiency. It is also worth noting that Fed1-UCB converges faster than the baseline, which is the result of higher arm sampling rate due to multiple clients simultaneously pulling arms. In other words, the fast convergence over time is due to the increased client dimension. When increasing the number of clients to $M = 10$, the overall regret remains approximately the same as $M = 5$, but with even faster convergence, which corroborates Theorem 3 and Corollary 2.

For the approximate model, the same set of global arms is used while the local models are generated by Gaussian distributions with $\sigma_c = 0.02$. Fig. 5 shows that Fed2-UCB with $f(p) = 100$ and $g(p) = 2^p$ successfully finds the optimal global arm and, without communication loss, has a performance (labelled as "expl") slightly worse than the baseline. Furthermore, the additional communication loss is very limited. Compared with the performance of Fed1-UCB in Fig. 4, we see that Fed2-UCB achieves almost the same performance for the more challenging approximate model, and the convergence of Fed2-UCB is even faster since the impact of increasing the number of clients is already significant at the very beginning. Under a reduced time horizon $T = 10^4$, Fig. 6 provides a finer look at the shape of regret curves of Fed2-UCB and illustrates the need of balancing two types of uncertainty[3]. With a short update period $f(p) = 10$, new clients are admitted rapidly, which sharply decreases the uncertainty from client sampling, but insufficient local exploration leads to a large uncertainty from arm sampling, which causes a large regret despite the fast the convergence. On the other extreme, although local exploration is guaranteed to be sufficient with $f(p) = 100$, it admits new clients slowly,

[3] The baseline is not included in Fig. 6 since it cannot converge in such a short time period.

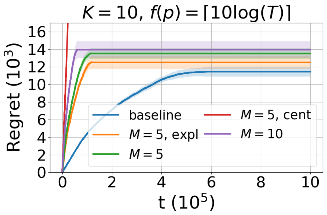

Figure 4: Fed1-UCB performance.

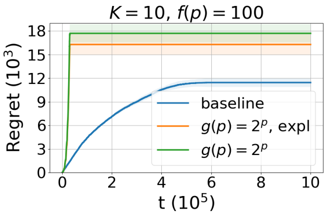

Figure 5: Fed2-UCB performance.

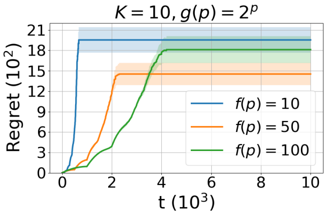

Figure 6: Fed2-UCB performance with different $f(p)$.

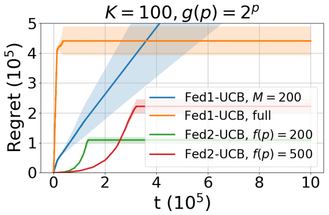

Figure 7: Performance with MovieLens dataset.

which delays the convergence and causes unnecessary local explorations. $f(p) = 50$ strikes a better balance between two types of uncertainty and thus a better performance.

5.2 Real-world Dataset for Recommender System

The MovieLens dataset (Cantador, Brusilovsky, and Kuflik 2011) is used for the real-world evaluation as an implementation of recommender system, which has been widely adopted in MAB studies (Oh and Iyengar 2019; Mahadik et al. 2020). It links the movies of MovieLens dataset with IMDb and Rotten Tomatoes movie review systems, and contains 2113 clients and 10197 movies. All the users are assumed to be available while the movies are randomly divided into 100 groups and the observations for clients are defined as their ratings of each group of movies. The sub-optimality gap of the pre-processed data is $\Delta \approx 0.0053$. The number of arms and potential clients are much larger than the synthetic dataset. First, as shown in Fig. 7, if a small fraction of clients ($M = 200$) are used for Fed1-UCB, which can be viewed as only involving a small amount of clients at the beginning of Fed2-UCB, the regret curve trends upward, meaning the global optimal arm is not found due to insufficient client sampling. Oppositely, when all clients are involved, Fed1-UCB converges to the optimal arm but with a large regret, which shows the harm of oversampling. Using Fed2-UCB and $f(p) = 200$ with $g(p) = 2^p$, a much better performance is achieved, since only the necessary amount of clients are sampled to capture the global model faithfully without unnecessary loss. With $f(p) = 500$, new clients are admitted more slowly but it still outperforms Fed1-UCB.

6 Related Works

Federated learning was introduced by McMahan et al. (2017); Konečnỳ et al. (2016a,b) and has been an active research topic with studies spanning from communication efficiency (Konečnỳ et al. 2016b; Sattler et al. 2019), security and privacy (Geyer, Klein, and Nabi 2017; Bagdasaryan et al. 2020), fairness (Li et al. 2020), to system designs (Smith et al. 2017; Bonawitz et al. 2019), with successful applications in recommender system (Ammad-Ud-Din et al. 2019), and medical treatment (Li et al. 2019).

Multi-armed bandits (Auer, Cesa-Bianchi, and Fischer 2002; Lattimore and Szepesvári 2020) is a rich research area with various applications, such as cognitive radio (Gai, Krishnamachari, and Jain 2010), recommender system (Li et al. 2010; Wu et al. 2017), and clinical trials (Shen et al. 2020; Lee et al. 2020, 2021). The research of distributed multi-player MAB (MP-MAB) are related to the proposed FMAB but with fundamental differences. The MP-MAB research either considers the “cooperative” setting, where players interact with a *common* MAB game and communicate with each other to accelerate learning (Landgren, Srivastava, and Leonard 2016, 2018; Martínez-Rubio, Kanade, and Rebeschini 2019; Wang et al. 2020), or the “competitive” setting, where there is no explicit communications and solving arm collisions is the fundamental difficulty (Liu and Zhao 2010; Avner and Mannor 2014; Rosenski, Shamir, and Szlak 2016; Boursier and Perchet 2019; Shi et al. 2020; Bubeck et al. 2020; Shi and Shen 2020). The FMAB framework differs from MP-MAB research in that clients can only access observation signals locally, which may not be sufficient to infer the true rewards and the optimal global arm. This difference results in a more fundamental role of communications.

The concept of *federated bandits* has been touched upon by a few recent works but with very different focuses than our work. IID local models are studied by Li, Song, and Fragouli (2020); Dubey and Pentland (2020) with a focus on privacy protection. Agarwal, Langford, and Wei (2020) study regression-based contextual bandits as an example of the federated residual learning framework, which does not generalize to our setting. Zhu et al. (2021) consider a similar problem as the exact model and focuses on sharing information through gossiping among clients with privacy protection. Shi, Shen, and Yang (2021) consider federated bandits with personalization, where the clients play mixed bandit games that incorporate both global and local models.

7 Conclusion

In this work, we have developed a general FMAB framework that bridges MAB and FL. In the proposed approximate model, a new source of uncertainty from client sampling was introduced. We proposed the Fed2-UCB algorithm that involves clients in an increasing manner and explores two types of uncertainty simultaneously while balancing the communication loss, which achieves an $O(\log(T))$ regret. A special case of the exact model was studied with the Fed1-UCB algorithm, which also achieves an order-optimal regret while providing an opportunity to tradeoff the convergence time and number of clients. Experiments with synthetic and real-world datasets proved the effectiveness of the proposed algorithms and corroborated the theoretical analysis.

Acknowledgements

The work was partially supported by the US National Science Foundation (NSF) under Grant CNS-2002902, ECCS-2033671 and ECCS-2029978.

Ethics Statement

Federated multi-armed bandits represents a novel framework that extends the federated learning principles to the bandit problems. Intellectually, this work broadens the scope of multi-player MAB by proposing global-local (non-IID) bandit model interactions through clients in a communication-efficient and privacy-preserving way. This new FMAB framework is general and may spark other innovations in this field. In addition, two FMAB models are considered that capture different global-local model relationships, and their corresponding algorithms and theoretical analyses may be useful for other similar problems. Practically, this work has the potential to benefit applications with sensitive data collected locally at the clients while a global bandit model is desired to be learned based on (but cannot directly access) the local data, including but not limited to the examples of cognitive radio and recommender systems given in the paper. In both examples, our research will enable the central server to avoid collecting raw data from the end users, hence mitigating risks to privacy and cost of communication. This work also provides a tangible way to leverage massively distributed clients for bandit learning. The authors do not foresee significant disadvantage for the involved parties with the proposed FMAB paradigm. As for the failure of the system, it is less likely to be an issue since the operations are fully distributed among large number of clients, but Fed2-UCB and Fed1-UCB may end up finding sub-optimal arms when it does happen. It is, however, a significant risk if the synchronization is broken, which poses an interesting research problem for future investigation. The FMAB framework also requires reliable communication infrastructure and sufficient computation power at the clients, which can be difficult in some situations (e.g., remote/rural areas or low-cost devices).

References

Agarwal, A.; Langford, J.; and Wei, C.-Y. 2020. Federated residual learning. *arXiv preprint* arXiv:2003.12880.

Ammad-Ud-Din, M.; Ivannikova, E.; Khan, S. A.; Oyomno, W.; Fu, Q.; Tan, K. E.; and Flanagan, A. 2019. Federated collaborative filtering for privacy-preserving personalized recommendation system. *arXiv preprint* arXiv:1901.09888.

Auer, P.; Cesa-Bianchi, N.; and Fischer, P. 2002. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.* 47(2-3): 235–256.

Auer, P.; and Ortner, R. 2010. UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica* 61(1-2): 55–65.

Avner, O.; and Mannor, S. 2014. Concurrent bandits and cognitive radio networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 66–81. Springer.

Bagdasaryan, E.; Veit, A.; Hua, Y.; Estrin, D.; and Shmatikov, V. 2020. How to backdoor federated learning. In *International Conference on Artificial Intelligence and Statistics*, 2938–2948. PMLR.

Bande, M.; and Veeravalli, V. V. 2019. Multi-user multi-armed bandits for uncoordinated spectrum access. In *International Conference on Computing, Networking and Communications (ICNC)*, 653–657. IEEE.

Bonawitz, K.; Eichner, H.; Grieskamp, W.; Huba, D.; Ingerman, A.; Ivanov, V.; Kiddon, C.; Konecny, J.; Mazzocchi, S.; McMahan, H. B.; Overveldt, T. V.; Petrou, D.; Ramage, D.; and Roselander, J. 2019. Towards federated learning at scale: System design. In *Proceedings of the 2nd SysML Conference*, 1–15.

Boursier, E.; and Perchet, V. 2019. SIC-MMAB: synchronisation involves communication in multiplayer multi-Armed bandits. In *Advances in Neural Information Processing Systems*, 12071–12080.

Bubeck, S.; and Cesa-Bianchi, N. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning* 5(1): 1–122.

Bubeck, S.; Li, Y.; Peres, Y.; and Sellke, M. 2020. Non-stochastic multi-player multi-armed bandits: Optimal rate with collision information, sublinear without. In *Conference on Learning Theory*, 961–987. PMLR.

Cantador, I.; Brusilovsky, P.; and Kuflik, T. 2011. 2nd Workshop on Information Heterogeneity and Fusion in Recommender Systems (HetRec 2011). In *Proceedings of the 5th ACM Conference on Recommender Systems*, RecSys 2011. New York, NY, USA: ACM.

Dubey, A.; and Pentland, A. 2020. Differentially-private federated linear bandits. *Advances in Neural Information Processing Systems* 33.

Gai, Y.; Krishnamachari, B.; and Jain, R. 2010. Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation. In *IEEE Symposium on New Frontiers in Dynamic Spectrum (DySPAN)*, 1–9. IEEE.

Geyer, R. C.; Klein, T.; and Nabi, M. 2017. Differentially private federated learning: A client level perspective. *arXiv preprint* arXiv:1712.07557.

Konečný, J.; McMahan, H. B.; Ramage, D.; and Richtárik, P. 2016a. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint* arXiv:1610.02527.

Konečný, J.; McMahan, H. B.; Yu, F. X.; Richtárik, P.; Suresh, A. T.; and Bacon, D. 2016b. Federated learning: Strategies for improving communication efficiency. In *Advances in Neural Information Processing Systems – Workshop on Private Multi-Party Machine Learning*.

Lai, T. L.; and Robbins, H. 1985. Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.* 6(1): 4–22.

Landgren, P.; Srivastava, V.; and Leonard, N. E. 2016. On distributed cooperative decision-making in multiarmed bandits. In *European Control Conference (ECC)*, 243–248. IEEE.

Landgren, P.; Srivastava, V.; and Leonard, N. E. 2018. Social imitation in cooperative multiarmed bandits: partition-based algorithms with strictly local information. In *IEEE Conference on Decision and Control (CDC)*, 5239–5244. IEEE.

Lattimore, T.; and Szepesvári, C. 2020. *Bandit algorithms*. Cambridge University Press.

Lee, H.-S.; Shen, C.; Jordon, J.; and van der Schaar, M. 2020. Contextual constrained learning for dose-finding clinical trials. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2645–2654.

Lee, H.-S.; Shen, C.; Zame, W.; Lee, J.; and van der Schaar, M. 2021. SDF-Bayes: Cautious Optimism in Safe Dose-Finding Clinical Trials with Drug Combinations and Heterogeneous Patient Groups. In *Proceedings of the 24rd International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Li, L.; Chu, W.; Langford, J.; and Schapire, R. E. 2010. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, 661–670.

Li, T.; Sanjabi, M.; Beirami, A.; and Smith, V. 2020. Fair resource allocation in federated learning. In *International Conference on Learning Representations*.

Li, T.; Song, L.; and Fragouli, C. 2020. Federated recommendation system via differential privacy. *IEEE International Symposium on Information Theory (ISIT)* .

Li, W.; Milletarì, F.; Xu, D.; Rieke, N.; Hancox, J.; Zhu, W.; Baust, M.; Cheng, Y.; Ourselin, S.; Cardoso, M. J.; et al. 2019. Privacy-preserving federated brain tumour segmentation. In *International Workshop on Machine Learning in Medical Imaging*, 133–141. Springer.

Liu, K.; and Zhao, Q. 2010. Distributed learning in multi-armed bandit with multiple players. *IEEE Transactions on Signal Processing* 58(11): 5667–5681.

Mahadik, K.; Wu, Q.; Li, S.; and Sabne, A. 2020. Fast distributed bandits for online recommendation systems. In *Proceedings of the 34th ACM International Conference on Supercomputing*, 1–13.

Martínez-Rubio, D.; Kanade, V.; and Rebeschini, P. 2019. Decentralized cooperative stochastic bandits. In *Advances in Neural Information Processing Systems*, 4529–4540.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 1273–1282. Fort Lauderdale, FL, USA.

Oh, M.-h.; and Iyengar, G. 2019. Thompson sampling for multinomial logit contextual bandits. In *Advances in Neural Information Processing Systems*, 3151–3161.

Rosenski, J.; Shamir, O.; and Szlak, L. 2016. Multi-player bandits – a musical chairs approach. In *International Conference on Machine Learning*, 155–163.

Sattler, F.; Wiedemann, S.; Müller, K.-R.; and Samek, W. 2019. Robust and communication-efficient federated learning from non-iid data. *IEEE Transactions on Neural Networks and Learning Systems* .

Shen, C.; Wang, Z.; Villar, S.; and van der Schaar, M. 2020. Learning for dose allocation in adaptive clinical trials with safety constraints. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 7310–7320.

Shi, C.; and Shen, C. 2020. On no-sensing adversarial multi-player multi-armed bandits with collision communications. *arXiv preprint* arXiv:2011.01090.

Shi, C.; Shen, C.; and Yang, J. 2021. Federated Multi-armed Bandits with Personalization. In *Proceedings of the 24rd International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Shi, C.; Xiong, W.; Shen, C.; and Yang, J. 2020. Decentralized multi-player multi-armed bandits with no collision information. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*. Palermo, Sicily, Italy.

Smith, V.; Chiang, C.-K.; Sanjabi, M.; and Talwalkar, A. S. 2017. Federated multi-task learning. In *Advances in Neural Information Processing Systems*, 4424–4434.

Wang, Y.; Hu, J.; Chen, X.; and Wang, L. 2020. Distributed bandit learning: Near-optimal regret with efficient communication. In *International Conference on Learning Representations*.

Wu, Q.; Wang, H.; Hong, L.; and Shi, Y. 2017. Returning is believing: Optimizing long-term user engagement in recommender systems. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 1927–1936.

Zhu, Z.; Zhu, J.; Liu, J.; and Liu, Y. 2021. Federated bandit: A gossiping approach. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 5(1): 1–29.