Federated Multi-armed Bandits with Personalization: Supplementary Materials

A Related Works

We provide a more comprehensive review of the related literature.

Differences to FL. Federated learning was introduced by McMahan et al. (2017); Konečnỳ et al. (2016a,b) to perform model training with data only locally available at a large number of clients. This paradigm has many attractive features, as summarized in Section 1. FL has been an active research topic over the past few years, with studies spanning from communication efficiency (Konečnỳ et al., 2016b; Reisizadeh et al., 2020; Sattler et al., 2019), security and privacy (Geyer et al., 2017; McMahan et al., 2018; Bagdasaryan et al., 2018), fairness (Li et al., 2020b), to system designs (Bonawitz et al., 2019; Wang et al., 2019b), with successful applications in recommender systems (Ammad-Ud-Din et al., 2019), and medical treatment (Li et al., 2019), among others.

Recently, FL with personalization (Kulkarni et al., 2020) draws a lot of attention, where the goal is to train an individualized model for each client, based on the client's own dataset and the datasets of other clients. One representative way of achieving personalization is to learn a mixed local and global model (Hanzely and Richtárik, 2020; Deng et al., 2020; Mansour et al., 2020), which is also adopted in PF-MAB. There are also other approaches, e.g., transfer learning (Wang et al., 2019a), multi-task learning (Smith et al., 2017), and meta-learning (Jiang et al., 2019; Fallah et al., 2020). Nevertheless, existing studies on FL are almost exclusively on supervised learning and there is limited literature considering bandit (Shi and Shen, 2021; Zhu et al., 2020).

Differences to Multi-player MAB. Multi-armed bandits is a rich research area (Lai and Robbins, 1985; Auer et al., 2002; Bubeck and Cesa-Bianchi, 2012; Lattimore and Szepesvári, 2020), with many successful applications such as cognitive radio (Gai et al., 2010; Avner and Mannor, 2016), recommender systems (Li et al., 2010; Wu et al., 2017; Oh and Iyengar, 2019; Mahadik et al., 2020), and clinical trials (Shen et al., 2020; Lee et al., 2020). The state-of-the-art decentralized multi-player MAB (MP-MAB) are related to the proposed PF-MAB but there are several fundamental differences, as elaborated below. One line of MP-MAB research considers the "cooperative" setting (Landgren et al., 2016, 2018; Martínez-Rubio et al., 2019; Wang et al., 2020), where players interact with a *common* MAB game and communicate with each other to accelerate learning under potential constraints, e.g., communication cost, privacy concern. However, with *non-IID* local models in PF-MAB, communications play a more fundamental role since global knowledge cannot be obtained by clients individually. The other line of research considers the "competitive" setting (Liu and Zhao, 2010; Avner and Mannor, 2014; Rosenski et al., 2016; Boursier and Perchet, 2019; Shi et al., 2020), where simultaneous pulls on the same arm by different players lead to a collision with zero reward for all involved players, and no explicit communications are allowed. Players in this setting focus on finding the best *allocation* of the common set of arms in a fully distributed manner, and solving arm collisions is the fundamental difficulty. Also, since explicit communications are not allowed, communication costs are not considered. However, in the PF-MAB framework, the clients are interested in finding the optimal arms on their own mixed models. Most importantly, although user-dependent local models are studied in both MP-MAB settings (Shahrampour et al., 2017; Bistritz and Leshem, 2018; Boursier et al., 2020), this is the first time that a flexible and mixed learning objective that balances generalization and personalization is studied in the field of MAB, to the best of our knowledge.

Recent Advances. There are a few very recent works that touch upon the concept of federated bandits but none of them systemically takes personalization into account. Li et al. (2020) assumes strictly IID local models and focuses on addressing privacy protection by combining differential privacy with statistics sharing. Agarwal et al. (2020) studies regression-based contextual bandits as a specific example of the federated residual learning framework, which does not generalize to the setting of our paper. The recent studies in Zhu et al. (2020); Shi and Shen (2021) are more related to this work, where federated MAB without personalization (i.e., global-only) is studied. Shi and Shen (2021) focuses on dealing with the stochastic relationship between local and global models and a similar client-server communication protocol is adopted. Zhu et al. (2020) discards the client-server structure and applies a gossiping information-sharing strategy, where privacy protection is also explicitly considered.

B Details of Choosing Exploration Lengths and Algorithm Enhancement

As stated in Section 4, the key challenge to solve PF-MAB is how to gain *sufficient but not excessive* local and global information simultaneously based on the required degree of personalization. Sections 4 and 6 provide two choices and here the details behind these choices are elaborated.

From client m's perspective on a locally active arm $k \neq k'_{*,m}$, in order to maintain the convergence rate of $1/(MF(p))$ (as specified in Section 4) while reducing the loss, an optimization problem over $N_{k,m}(p)$ and $N^g_{k,n}(p), \forall n \neq m$ can be formulated as:

$$\begin{aligned} &\text{minimize } \; N_{k,m}(p)\Delta'_{k,m} + \sum\nolimits_{n\neq m, k'_{*,n}\neq k} N^g_{k,n}(p)\Delta'_{k,n} \\ &\text{subject to } \; \frac{[\alpha+(1-\alpha)/M]^2}{N_{k,m}(p)} + \sum_{n\neq m}\frac{[(1-\alpha)/M]^2}{N^g_{k,n}(p)} \leq \frac{1}{MF(p)} \end{aligned}$$

where $N_{k,m}(p)$ is the number of pulls on arm k at client m up to phase p, and $N^g_{k,n}(p)$ is the guaranteed number of global pulls on arm k at a different client n up to phase p. The optimization objective is the loss associated with client m's local and global information estimation for arm k, while the constraint is a sufficient condition for $B_p = \sqrt{4\log(T)/(MF(p))}$ and Lemma 1 to hold. Note that the convergence rate constraint can have many forms, and the choice here is to match the discussion in the main paper.

Using the Cauchy-Schwarz inequality, the exploration length described in Section 6 can be obtained as:

$$\begin{aligned} n^l_{k,m}(p) &\propto \frac{\alpha M f(p)}{(\Delta'_{k,m})^{1/2}}, \forall k \in A_m(p), k \neq k'_{*,m}; \\ n^g_{k,m}(p) &\propto \frac{(1-\alpha) f(p)}{(\Delta'_{k,m})^{1/2}}, \forall k \in A(p), k \neq k'_{*,m}, \end{aligned}$$

and $N^l_{k,m}(p) = \sum_{q=1}^p n^l_{k,m}(q)$, $N^g_{k,m}(p) = \sum_{q=1}^p n^g_{k,m}(q)$ and $N_{k,m}(p) = N^l_{k,m}(p) + N^g_{k,m}(p)$. This result is the key to choosing exploration lengths as it builds up the relationship between local and global explorations.

The issue however is that the knowledge of $\Delta'_{k,m}$ is unavailable. An easy way to tackle this problem is to assume all the sub-optimal gaps are the same, which results in the chosen length in PF-UCB in Section 4. The alternative way proposed in Section 6 is to use $\bar{\Delta}'_{k,m}(p) = \max_{l\in[K]} \bar{\mu}'_{l,m}(p-1) - \bar{\mu}'_{k,m}(p-1) + 2B_{p-1}$ in place of $\Delta'_{k,m}(p)$. This approach leverages information collected in the game. However, $\bar{\Delta}'_{k,m}(p)$ needs to be communicated to the server and then broadcast to maintain synchronization among clients, which may increase the risk of privacy leaking.

C Proof for the Lower Bound Analysis in Theorem 1

Proof. First, the following lemma recalls the classic result from the single-player MAB (Lai and Robbins, 1985), which directly leads to the lower bound in Eqn. (2).

Lemma 6. *For any consistent policy Π, for any arm k such that $\mu_k < \mu_{k_*}$, it holds that*

$$\liminf_{T\to\infty} \frac{T_k}{\log(T)} \geq \frac{1}{\mathrm{kl}\left(X_k, X_{k_*}\right)},$$

where T_k is the expected number of pulls performed on arm k during T.

Then, from client m's perspective of her suboptimal arm $k \neq k_{*,m}$ on the mixed model, the mixed reward in Eqn. (4) can be decomposed as

$$X'_{k,m} = \left(\alpha + \frac{1-\alpha}{M}\right) X_{k,m} + \frac{1-\alpha}{M}\sum_{n\neq m} X_{k,n}.$$

The difficulty is that $X'_{k,m}$ involves the rewards from all M clients, which are M sources of randomness. Next we attempt to isolate these sources of randomness.

First, if we assume client m has perfect knowledge of $\{\mu_{k,n}\}_{n\neq m}$, a new random variable $Y_{k,m}$ is constructed as

$$Y_{k,m} = \left(\alpha + \frac{1-\alpha}{M}\right) X_{k,m} + \frac{1-\alpha}{M}\sum_{n\neq m}\mu_{k,n} = \left(\alpha + \frac{1-\alpha}{M}\right) X_{k,m} + \mu'_{k,m} - \left(\alpha + \frac{1-\alpha}{M}\right)\mu_{k,m}.$$

Under this construction, $Y_{k,m}$ shares the same mean with $X'_{k,m}$ while the randomness only comes from $X_{k,m}$. Then, $Y_{k,m}$ forms a new hypothetical bandit game degenerated from client m's mixed model, where the mean rewards and the optimal arm remain the same. With Lemma 6, if client m individually interacts with this new game, her pulls on arm k can be bounded as

$$\liminf_{T\to\infty}\frac{T_{k,m}}{\log(T)} \geq \frac{1}{\mathrm{kl}\left(Y_{k,m}, Y_{k'_{*,m},m}\right)}.$$

On the other hand, from a different client n's perspective, whose arm k is also sub-optimal, she also needs information of client m's arm k. However, client n's mixed reward is constructed as

$$X'_{k,n} = \left(\alpha + \frac{1-\alpha}{M}\right) X_{k,n} + \frac{1-\alpha}{M}X_{k,m} + \frac{1-\alpha}{M}\sum_{l\neq m,n} X_{k,l},$$

which is different from $X'_{k,m}$. Following a similar idea of isolating randomness, if we assume client n has perfect knowledge of $l \neq m, \mu_{k,l}$, including $\mu_{k,n}$, a new random variable $Z^m_{k,n}$ can be constructed as

$$Z^m_{k,n} = \left(\alpha + \frac{1-\alpha}{M}\right)\mu_{k,n} + \frac{1-\alpha}{M}X_{k,m} + \frac{1-\alpha}{M}\sum_{l\neq m,n}\mu_{k,l} = \frac{1-\alpha}{M}X_{k,m} + \mu'_{k,n} - \frac{1-\alpha}{M}\mu_{k,m}.$$

Under this construction, $Z^m_{k,n}$ shares the same mean as $X_{k,n}$ while the randomness only comes from $X_{k,m}$. Then $Z^m_{k,n}$ forms another new hypothetical bandit game degenerated from client n's mixed model, where the optimal arm remains the same and client m has to provide information to help client n distinguish arm k. Similarly, with Lemma 6, if client m individually interacts with this new game, her pulls on arm k can be bounded as

$$\liminf_{T\to\infty}\frac{T_{k,m}}{\log(T)} \geq \frac{1}{\mathrm{kl}\left(Z^m_{k,n}, Z^m_{k'_{*,n},n}\right)}.$$

Since $Z^m_{k,n}$ can be constructed for any client, it must hold that

$$\liminf_{T\to\infty}\frac{T_{k,m}}{\log(T)} \geq \max_{n:n\neq m, k'_{*,n}\neq k}\left\{\frac{1}{\mathrm{kl}\left(Z^m_{k,n}, Z^m_{k'_{*,n},n}\right)}\right\} = \frac{1}{\min_{n:n\neq m,k'_{*,n}\neq k}\left\{\mathrm{kl}\left(Z^m_{k,n}, Z^m_{k'_{*,n},n}\right)\right\}}.$$

Combining the above results, we can have

$$\liminf_{T\to\infty}\frac{T_{k,m}}{\log(T)} \geq \max\left\{\frac{1}{\mathrm{kl}\left(Y_{k,m}, Y_{k'_{*,m},m}\right)}, \frac{1}{\min_{n:n\neq m,k'_{*,n}\neq k}\left\{\mathrm{kl}\left(Z^m_{k,n}, Z^m_{k'_{*,n},n}\right)\right\}}\right\}.$$

Since the regret can be decomposed as

$$R(T) = \sum_{m=1}^{M}\sum_{k:k\neq k'_{*,m}} T_{k,m}\Delta'_{k,m},$$

Theorem 1 can be established. $\square$

Note that the randomness isolation utilized in the proof reduces the hardness of the problem, which results in a relaxed lower bound. Although it can recover the single-player stochastic MAB lower bound with $\alpha = 1$, when α moves away from 1, the lower bound becomes less tight.

$f(p)$	$p_{k,m}$, $k \neq k'_{*,m}$	$R(T)$
λ	$O\left(\frac{\log(T)}{M\lambda(\Delta'_{k,m})^2}\right)$	$O\left(\sum_{m=1}^{M}\sum_{k\neq k'_{*,m}}\left[\frac{\alpha}{\Delta'_{k,m}}+\frac{\frac{1-\alpha}{M}\Delta'_{k,m}}{\Delta_k'^2}\right]\log(T)+\frac{C\log(T)}{\lambda(\Delta'_{\min})^2}\right)$
$\lambda\log(T)$	$O\left(\frac{1}{M\lambda(\Delta'_{k,m})^2}\right)$	$O\left(\sum_{m=1}^{M}\sum_{k\neq k'_{*,m}}\left[\frac{\alpha}{\Delta'_{k,m}}+\frac{\frac{1-\alpha}{M}\Delta'_{k,m}}{\Delta_k'^2}\right]\log(T)+\frac{C}{\lambda(\Delta'_{\min})^2}\right)$
2^p	$O\left(\log\left(\frac{\log(T)}{M(\Delta'_{k,m})^2}\right)\right)$	$O\left(\sum_{m=1}^{M}\sum_{k\neq k'_{*,m}}\left[\frac{\alpha}{\Delta'_{k,m}}+\frac{\frac{1-\alpha}{M}\Delta'_{k,m}}{\Delta_k'^2}\right]\log(T)+CM\log\left(\frac{\log(T)}{M(\Delta'_{\min})^2}\right)\right)$
$2^p\log(T)$	$O\left(\log\left(\frac{1}{M(\Delta'_{k,m})^2}\right)\right)$	$O\left(\sum_{m=1}^{M}\sum_{k\neq k'_{*,m}}\left[\frac{\alpha}{\Delta'_{k,m}}+\frac{\frac{1-\alpha}{M}\Delta'_{k,m}}{(\Delta'_k)^2}\right]\log(T)+CM\log\left(\frac{1}{M(\Delta'_{\min})^2}\right)\right)$

Table 1: Regret of PF-UCB algorithm with different choices of $f(p)$
λ is a constant; $\Delta'_k=\min_{n:k'_{*,n}\neq k}\{\Delta'_{k,n}\}$; $\Delta'_{\min}=\min_k\{\Delta'_k\}$.

D Discussions for Theorem 2

Table 1 summarizes the regrets under several different choices of $f(p)$, including $f(p)=2^p\log(T)$ in Corollary 2. All choices listed in Table 1 achieve a similar exploration regret and a non-dominating exploitation loss (which is omitted in the regret expression). However, they lead to varying communication losses. With $f(p)=\lambda$, the communication loss is of order $O(\log(T))$ and scales with $1/(\Delta'_{\min})^2$, which actually dominates the exploration loss. This is the result of the unnecessary communications with $f(p)=\lambda$. With $f(p)=\lambda\log(T)$, the communication loss is no longer of order $O(\log(T))$; however, it still scales with $1/(\Delta'_{\min})^2$. The dependency of communication loss on $\Delta'_{\min}$ is improved with an exponential $f(p)$, as both $f(p)=2^p$ and $f(p)=2^p\log(T)$ have communication losses that scale only with $\log(1/\Delta'_{\min})$, which greatly reduces the communication burden. Furthermore, with $f(p)=2^p\log(T)$, the communication cost is a constant that is independent of T. Thus, among all considered choices of $f(p)$, the most preferable one is $f(p)=2^p\log(T)$.

We further note that all the choices of $f(p)$ listed in Table 1 do not depend on the communication loss parameter C. This is made to simplify the problem, as otherwise the analysis will have a convoluted relationship between the exploration loss and the communication loss. Intuitively, with a larger C, it is better to increase $f(p)$ to reduce the communication frequency and lower the communication loss, e.g., adding a $1/C$ multiplicative factor to the listed choice of $f(p)$.

E Proofs for Regret Analysis

E.1 Proof of Lemma 1

Proof. To decouple the randomness of $A_m(p)$, we assume a virtual system without elimination, i.e., in this virtual system $\forall m\in[M],\forall p, A_m(p)=[K]$. At phase p, $\forall m\in[M],\forall k\in A_m(p)$, $\bar{\mu}'_{k,m}(p)$ can be decomposed as

$$\bar{\mu}'_{k,m}(p)=\left(\alpha+\frac{1-\alpha}{M}\right)\bar{\mu}_{k,m}(p)+\frac{1-\alpha}{M}\sum_{n\neq m}\bar{\mu}_{k,n}(p).$$

It can be shown that $\bar{\mu}_{k,m}(p)$ is a $\sqrt{\frac{1}{N_{k,m}(p)}}$-subgaussian random variable, since client m has explored arm k for $N_{k,m}(p)=\sum_{q=1}^{p}n_{k,m}(q)$ times in the global and local exploration sub-phases. However, $\forall n\in[M],n\neq m$, client m can only make sure that $\bar{\mu}_{k,n}(p)$ is a $\sqrt{\frac{1}{N^g_{k,n}(p)}}$-subgaussian random variable, where $N^g_{k,n}(p)=\sum_{q=1}^{p}n^g_{k,n}(q)$, since she is only assured that each other client has explored arm k in the global exploration sub-phases. Overall, we can claim that $\bar{\mu}'_{k,m}(p)$ is a $\sigma'_{k,m}(p)$-subgaussian random variable where

$$\begin{aligned}\sigma'_{k,m}(p)&=\sqrt{\left(\alpha+\frac{1-\alpha}{M}\right)^2\frac{1}{N_{k,m}(p)}+\left(\frac{1-\alpha}{M}\right)^2\sum_{n\neq m}\frac{1}{N^g_{k,n}(p)}}\\&\leq\sqrt{\left(\alpha+\frac{1-\alpha}{M}\right)^2\frac{1}{[(1-\alpha)+M\alpha]F(p)}+\left(\frac{1-\alpha}{M}\right)^2\sum_{n\neq m}\frac{1}{(1-\alpha)F(p)}}\end{aligned}$$

$$= \sqrt{\frac{1}{MF(p)}}.$$

With the concentration inequality for subgaussian random variables, we have

$$\mathbb{P}\left(|\bar{\mu}'_{k,m}(p) - \mu'_{k,m}| \geq B_p\right) \leq 2\exp\left\{-\frac{B_p^2}{2(\sigma'_{k,m}(p))^2}\right\} \leq 2\exp\left\{-\frac{\frac{4\log(T)}{MF(p)}}{2\frac{1}{MF(p)}}\right\} = \frac{2}{T^2}.$$

Thus, with the union bound, P_G can be bounded as

$$\begin{aligned}
P_G &= 1 - \mathbb{P}\left\{\exists p, \exists m \in [M], \exists k \in A_m(p), |\bar{\mu}'_{k,m}(p) - \mu'_{k,m}| \geq B_p\right\}\\
&\geq 1 - \sum_{p=1}^{T}\sum_{m=1}^{M}\sum_{k=1}^{K}\mathbb{P}\left(|\bar{\mu}'_{k,m}(p) - \mu'_{k,m}| \geq B_p\right)\\
&\geq 1 - \frac{2MK}{T}.
\end{aligned}$$

Since this argument applies to $k \in [K]$, it also applies to all arms in the local active arm set $A_m(p)$ of the real system, which concludes the proof. □

E.2 Proof of Lemma 2

Proof. Recall that $\forall k \neq k'_{*,m}$, $p'_{k,m}$ is the smallest integer such that

$$MF(p'_{k,m}) \geq \frac{64\log(T)}{(\Delta'_{k,m})^2},$$

which ensures that $\forall p \geq p'_{k,m}, B_p \leq \frac{\Delta'_{k,m}}{4}$. Thus, based on that event G happens, at phase $p'_{k,m}$, we have

$$\begin{aligned}
\bar{\mu}'_{k,m}(p'_{k,m}) + B_{p'_{k,m}} &\overset{(i)}{\leq} \mu'_{k,m} + 2B_{p'_{k,m}} \leq \mu'_{k,m} + \frac{\Delta'_{k,m}}{2}\\
&= \mu'_{*,m} - \frac{\Delta'_{k,m}}{2} \overset{(ii)}{\leq} \bar{\mu}'_{k'_{*,m},m}(p'_{k'_{*,m},m}) + B_{p'_{k,m}} - \frac{\Delta'_{k,m}}{2} \leq \bar{\mu}'_{k'_{*,m},m}(p'_{k'_{*,m},m}) - B_{p'_{k,m}},
\end{aligned}$$

where inequalities (i) and (ii) are guaranteed by event G. Thus, arm k is guaranteed to be eliminated at phase $p'_{k,m}$ by client m. □

E.3 Proof of Lemma 3

Proof. Lemma 2 indicates for a sub-optimal arm k, after phase $p'_{k,m}$, it is guaranteed to be eliminated from set $A_m(p)$. Thus, it is pulled for at most $\sum_{p=1}^{p'_{k,m}}\lceil\alpha Mf(p)\rceil$ times in the local exploration sub-phases, which leads to the local exploration loss as

$$R_l^{expr}(T) \leq \sum_{m=1}^{M}\sum_{k\neq k'_{*,m}}\Delta'_{k,m}\sum_{p=1}^{p'_{k,m}}\lceil\alpha Mf(p)\rceil.$$

However, arm k is still pulled in the global exploration sub-phases until $k \notin A(p)$, i.e., arm k is eliminated by all of the clients whose optimal arm is not it. Since arm k is guaranteed to be eliminated globally by phase $p'_k = \max_{m\in[M]}\{p'_{k,m}\}$, it is pulled for at most $\sum_{p=1}^{p'_k}\lceil(1-\alpha)f(p)\rceil$ times in the global exploration sub-phases. Thus, the global exploration loss can be bounded as:

$$R_g^{expr}(T) \leq \sum_{m=1}^{M}\sum_{k\neq k'_{*,m}}\Delta'_{k,m}\sum_{p=1}^{p'_k}\lceil(1-\alpha)f(p)\rceil.$$

□

E.4 Proof of Lemma 4

Proof. At phase p, the exploitation time for client m is at most $\max_n\{|A_n(p)| - A_m(p)\}\lceil M\alpha f(p)\rceil$, which is the difference between the longest local exploration duration and her local exploration duration. The probability that the exploited arm in the exploitation phase, i.e., arm $\bar{k}'_{*,m}$, is arm k instead of $k'_{*,m}$ can be bounded as:

$$\begin{aligned}
\mathbb{P}\left(\bar{k}'_{*,m} = k\right) &\leq P\left(\bar{\mu}'_{k'_{*,m},m}(p-1) \leq \bar{\mu}_{k,m}(p-1)\right) \\
&= P\left(\bar{\mu}'_{k'_{*,m},m}(p-1) - \bar{\mu}_{k,m}(p-1) - \Delta'_{k,m} \leq -\Delta'_{k,m}\right) \\
&\overset{(i)}{\leq} 2\exp\left\{-\frac{(\Delta'_{k,m})^2}{2(\sigma'^2_{k,m}(p-1) + \sigma'^2_{k'_{*,m},m}(p-1))}\right\} \\
&\leq 2\exp\left\{-\frac{(\Delta'_{k,m})^2 MF(p-1)}{4}\right\} \\
&= P'_{k,m}(p).
\end{aligned}$$

Thus, it can be shown that the exploration loss caused by arm k for client m is bounded as

$$\begin{aligned}
R^{expt}_{k,m}(T) &\leq \Delta'_{k,m} \sum_{p=1}^{p'_{k,m}} \left(\max_n\{|A_n(p)| - A_m(p)\}\right)\lceil M\alpha f(p)\rceil P'_{k,m}(p) \\
&\leq \Delta'_{k,m} \sum_{p=1}^{p'_{k,m}} K\lceil M\alpha f(p)\rceil \exp\left\{-\frac{(\Delta'_{k,m})^2 MF(p-1)}{4}\right\}.
\end{aligned}$$

The overall exploration loss can be obtained by summing over all of the clients and arms:

$$R^{expt}(T) = \sum_{m=1}^{M}\sum_{k=1}^{K} \Delta'_{k,m} R^{expt}_{k,m}(T) \leq \sum_{m=1}^{M}\sum_{k\neq k'_{*,m}}\sum_{p=1}^{p'_{k,m}} K\lceil M\alpha f(p)\rceil \Delta'_{k,m} \exp\left\{-\frac{(\Delta'_{k,m})^2 MF(p-1)}{4}\right\}.$$

In addition, we note that in phase $p = 1$, all the players share the same global and local active arm sets, i.e., $\forall m \in [M], A_m(p) = A(p) = [K]$, which means there would be no exploration loss. Thus, the sum of index p in the exploitation loss above can start from 2 instead of 1. This fact does not change the scaling of the overall regret, but would be useful in deriving Corollary 2 from Theorem 2. □

E.5 Proof of Lemma 5

Proof. As designed in the PF-UCB algorithm, clients do not communicate any more after they find their optimal arms. Thus, there is no more communication after phase $p'_{\max} = \max_{k\in[K]}\{p'_{k,m}\} = \max_{m\in[M]}\max_{k\neq k'_{*,m}}\{p'_{k,m}\}$. Before phase $p'_{\max}$, there are two communications in each phase for arm statistics and active sets, respectively, which leads to the communication loss upper bound as:

$$R^{comm}(T) \leq 2CMp'_{\max}.$$

□

E.6 Proof of Theorem 2

Proof. Lemmas 3, 4 and 5 are all based on the condition that event G happens, which has probability P_G as shown in Lemma 1. When event G does not happen, the regret is directly upper bounded by $MT + 2CMT$, which assumes full exploration and communication loss. Thus, Theorem 2 follows by putting everything together as:

$$R(T) = P_G\left(R^{expr}(T) + R^{expt}(T) + R^{comm}(T)\right) + (1 - P_G)(1 + 2C)MT$$

$$\leq R_l^{expr}(T) + R_g^{expr}(T) + R^{expt}(T) + R^{comm}(T) + 2M^2K(1+2C)$$

$$\leq \sum_{m=1}^{M} \sum_{k \neq k'_{*,m}} \Delta'_{k,m} \sum_{p=1}^{p'_{k,m}} \lceil \alpha M f(p) \rceil + \sum_{m=1}^{M} \sum_{k \neq k'_{*,m}} \Delta'_{k,m} \sum_{p=1}^{p'_{k}} \lceil (1-\alpha) f(p) \rceil$$

$$+ \sum_{m=1}^{M} \sum_{k \neq k'_{*,m}} \Delta'_{k,m} \sum_{p=1}^{p'_{k,m}} K \lceil M\alpha f(p) \rceil \exp\left\{ -\frac{(\Delta'_{k,m})^2 M F(p-1)}{4} \right\} + 2CMp'_{\max} + 2M^2K(1+2C).$$

□

E.7 Proof of Corollary 2

Proof. With $f(p) = 2^p \log(T)$, $p'_{k,m}$ can be bounded from Eqn. (9) as

$$p'_{k,m} = O\left(\log_2\left(\frac{64}{M(\Delta'_{k,m})^2}\right)\right).$$

Plugging this into Theorem 2, Corollary 2 follows. □

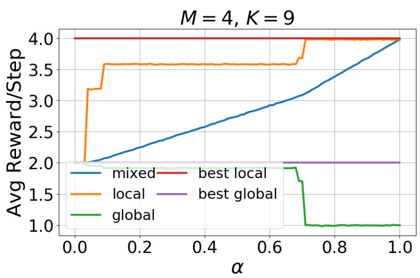

Figure 5: Synthetic Reward

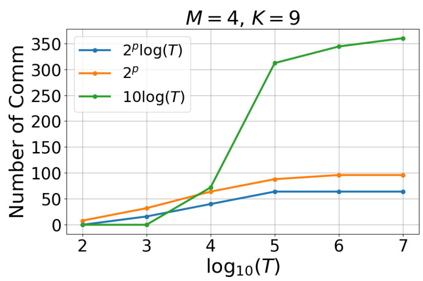

Figure 6: Number of Communications

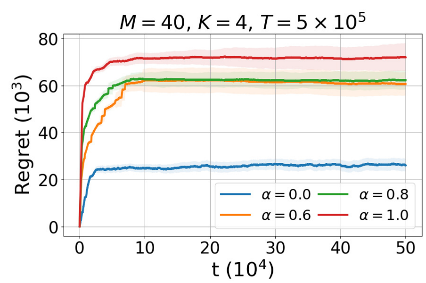

Figure 7: Large M and Small K

F Additional Experimental Results

The implementation codes of the PF-UCB and its enhancement used for simulations have been made publicly available at `https://github.com/ShenGroup/PF_MAB`, which also contains the synthetic dataset and the pre-processed real-world MovieLens dataset. The original version of the MovieLens dataset is publicly available at `https://grouplens.org/datasets/hetrec-2011/`.

Experimental details and additional experiment results are provided here. First, for the synthetic dataset used in Fig. 1, the specific arm statistics are given as follows:

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0.9 & 0.4 & 0.35 & 0.35 & 0.5 \\ 0 & 1 & 0 & 0 & 0.3 & 0.9 & 0.35 & 0.3 & 0.5 \\ 0 & 0 & 1 & 0 & 0.35 & 0.35 & 0.9 & 0.3 & 0.5 \\ 0 & 0 & 0 & 1 & 0.4 & 0.3 & 0.35 & 0.9 & 0.5 \end{bmatrix},$$

where the rows and columns correspond to the clients and arms, respectively. This dataset is specially designed so that the local optimal arm for client $m \in \{1, 2, 3, 4\}$ is arm m, while the global optimal arm is arm 9. Moreover, each of the local optimal arms perform poorly at other clients. All remaining arms share similar global utilities, but diverge locally. The averaged per-step reward with PF-UCB under this synthetic dataset is reported in Fig. 5, which shows a similar trend as in Fig. 3.

The communication times in the horizon of 10^6 for the synthetic game are provided in Table 2. Compared with the time horizon, the overall communication times are almost negligible, which shows the efficiency of communication under the choice of $f(p) = 2^p \log(T)$. The communication times under different time horizons for different choices of $f(p)$ are reported in Fig. 6 with the same synthetic game and $\alpha = 0.5$, which illustrates

α	Comm Times
0	104
0.2	64
0.5	72
0.9	80
1	56

Table 2: Synthetic Communication Times

that $f(p) = 10\log(T)$ leads to more communications for large T than the other two choices and $f(p) = 2^p \log(T)$ is the most efficient one. This observation coincides with the results in Table 1.

As in real-world FL systems, it is common to have a small K (number of arms) and a large M (number of clients). Additional experiments are performed with a small $K = 4$ and a large $M = 40$ with results reported in Fig. 7. It can be observed that PF-UCB still achieves stable performance in this scenario.

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. *Machine learning*, 47(2-3):235–256.

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

Avner, O. and Mannor, S. (2016). Multi-user lax communications: a multi-armed bandit approach. In *IEEE INFOCOM 2016*, pages 1–9. IEEE.

Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., and Shmatikov, V. (2018). How to backdoor federated learning. *arXiv preprint arXiv:1807.00459*.

Bistritz, I. and Leshem, A. (2018). Distributed multi-player bandits-a game of thrones approach. In *Advances in Neural Information Processing Systems*, pages 7222–7232.

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*, pages 1–15.

Boursier, E., Kaufmann, E., Mehrabian, A., and Perchet, V. (2020). A practical algorithm for multiplayer bandits when arm means vary among players. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, Palermo, Sicily, Italy.

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

Brânzei, S. and Peres, Y. (2019). Multiplayer bandit learning, from competition to cooperation. *arXiv preprint arXiv:1908.01135*.

Bubeck, S. and Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*.

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.

Deng, Y., Kamani, M. M., and Mahdavi, M. (2020). Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461*.

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

Fallah, A., Mokhtari, A., and Ozdaglar, A. (2020). Personalized federated learning: A meta-learning approach. *arXiv preprint arXiv:2002.07948*.

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

Hanzely, F. and Richtárik, P. (2020). Federated learning of a mixture of global and local models. *arXiv preprint arXiv:2002.05516*.

Jiang, Y., Konečný, J., Rush, K., and Kannan, S. (2019). Improving federated learning personalization via model agnostic meta learning. *arXiv preprint arXiv:1909.12488*.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. (2019). Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*.

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*.

Kulkarni, V., Kulkarni, M., and Pant, A. (2020). Survey of personalization techniques for federated learning. *arXiv preprint arXiv:2003.08673*.

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 multi-armed bandits. In *2016 European Control Conference (ECC)*, pages 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 *2018 IEEE Conference on Decision and Control (CDC)*, pages 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. *arXiv preprint arXiv:2001.02463*.

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*, pages 661–670.

Li, T., Sahu, A. K., Talwalkar, A., and Smith, V. (2020a). Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60.

Li, T., Sanjabi, M., Beirami, A., and Smith, V. (2020b). 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. In *IEEE International Symposium on Information Theory (ISIT)*, pages 2592–2597.

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*, pages 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*, pages 1–13.

Mansour, Y., Mohri, M., Ro, J., and Suresh, A. T. (2020). Three approaches for personalization with applications to federated learning. *arXiv preprint arXiv:2002.10619*.

Martínez-Rubio, D., Kanade, V., and Rebeschini, P. (2019). Decentralized cooperative stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 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)*, pages 1273–1282, Fort Lauderdale, FL, USA.

McMahan, H. B., Ramage, D., Talwar, K., and Zhang, L. (2018). Learning differentially private recurrent language models. In *International Conference on Learning Representations*.

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

Reisizadeh, A., Mokhtari, A., Hassani, H., Jadbabaie, A., and Pedarsani, R. (2020). FedPAQ: A communication-efficient federated learning method with periodic averaging and quantization. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, Palermo, Sicily, Italy.

Rosenski, J., Shamir, O., and Szlak, L. (2016). Multi-player bandits–a musical chairs approach. In *International Conference on Machine Learning*, pages 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*, 31(9):3400–3413.

Shahrampour, S., Rakhlin, A., and Jadbabaie, A. (2017). Multi-armed bandits in multi-agent networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2786–2790. IEEE.

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)*, pages 7310–7320.

Shi, C. and Shen, C. (2021). Federated multi-armed bandits. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI)*.

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*, pages 4424–4434.

Wang, K., Mathews, R., Kiddon, C., Eichner, H., Beaufays, F., and Ramage, D. (2019a). Federated evaluation of on-device personalization. *arXiv preprint arXiv:1910.10252*.

Wang, S., Tuor, T., Salonidis, T., Leung, K. K., Makaya, C., He, T., and Chan, K. (2019b). Adaptive federated learning in resource constrained edge computing systems. *IEEE Journal on Selected Areas in Communications*, 37(6):1205–1221.

Wang, Y., Hu, J., Chen, X., and Wang, L. (2020). Distributed bandit learning: Near-optimal regret with efficient communication. In *2020 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*, pages 1927–1936.

Zhu, Z., Zhu, J., Liu, J., and Liu, Y. (2020). Federated bandit: A gossiping approach. *arXiv preprint*, arXiv:2010.12763.