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); Konečný 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 (Konečný 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 MP-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:

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

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:

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

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 \( \Delta'_{k,m}(p) = \max_{k \in [K]} \bar{\mu'}_{k,m}(p-1) - \bar{\mu'}_{k,m}(p-1) + 2B_{p-1} \) in place of \( \Delta'_{k,m}(p) \). This approach leverages information collected in the game. However, \( \Delta'_{k,m}(p) \) 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 \( \Pi \), for any arm \( k \) such that \( \mu_k < \mu_{k^*} \), it holds that

\[
\liminf_{T \to \infty} \frac{T_k}{\log(T)} \geq \frac{1}{\text{kl}(X_k, X_{k^*})},
\]

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

$$\liminf_{T \to \infty} \frac{T_{k,m}}{\log(T)} \geq \frac{1}{\text{kl} (Y_{k,m}, Y_{k',m,m}).}$$

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 $m$’s mixed reward is constructed as

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

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 $Z_{k,n}^m$ can be constructed as

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

Under this construction, $Z_{k,n}^m$ shares the same mean as $X_{k,n}$ while the randomness only comes from $X_{k,m}$. Then $Z_{k,n}^m$ 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

$$\liminf_{T \to \infty} \frac{T_{k,m}}{\log(T)} \geq \frac{1}{\text{kl} (Z_{k,n}^m, Z_{k',n,m}^m).}$$

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

$$\liminf_{T \to \infty} \frac{T_{k,m}}{\log(T)} \geq \max_{n: n \neq m, k', n \neq k} \left\{ \frac{1}{\text{kl} (Z_{k,n}^m, Z_{k',n,m}^m)} \right\} = \frac{1}{\text{kl} (Z_{k,n}^m, Z_{k',n,m}^m)}.$$  

Combining the above results, we can have

$$\liminf_{T \to \infty} \frac{T_{k,m}}{\log(T)} \geq \max \left\{ \frac{1}{\text{kl} (Y_{k,m}, Y_{k',m,m})}, \frac{1}{\text{kl} (Z_{k,n}^m, Z_{k',n,m}^m)} \right\}.$$  

Since the regret can be decomposed as

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

Theorem 1 can be established.

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 $\alpha$ moves away from 1, the lower bound becomes less tight.
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 $1/(\Delta_{\text{min}}^*)^2$, which actually dominates the exploration loss.

D Discussions for Theorem 2

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 $1/(\Delta_{\text{min}}^*)^2$. The dependency of communication loss on $\Delta_{\text{min}}^*$ 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 $1/(\Delta_{\text{min}}^*)$, 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 $A_n(p)$, 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 $p$, $\forall m \in [M], \forall k \in A_m(p)$, $\tilde{\mu}_{k,m}(p)$ can be decomposed as

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

It can be shown that $\tilde{\mu}_{k,m}(p)$ 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 $\tilde{\mu}_{k,n}(p)$ is a $\sqrt{\frac{1}{N_{k,n}(p)}}$-subgaussian random variable, where $N_{k,n}(p) = \sum_{q=1}^{p} n_{k,n}^2(q)$, since she is only assured that each other client has explored arm $k$ in the global exploration sub-phases. Overall, we can claim that $\tilde{\mu}_{k,m}(p)$ is a $\sigma_{k,m}(p)$-subgaussian random variable where

$$\sigma_{k,m}(p) = \sqrt{\left(\alpha + \frac{1 - \alpha}{M}\right)^2 \frac{1}{N_{k,m}(p)} + \left(\frac{1 - \alpha}{M}\right)^2 \sum_{n \neq m} \frac{1}{N_{k,n}(p)}}$$

$$\leq \sqrt{\left(\alpha + \frac{1 - \alpha}{M}\right)^2 \frac{1}{(1 - \alpha) + M\alpha} F(p) + \left(\frac{1 - \alpha}{M}\right)^2 \sum_{n \neq m} \frac{1}{(1 - \alpha) F(p)}}$$

Table 1: Regret of PF-UCB algorithm with different choices of $f(p)$

<table>
<thead>
<tr>
<th>$f(p)$</th>
<th>$p_{k,m}$, $k \neq k_{*,m}$</th>
<th>$R(T)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda$</td>
<td>$O\left(\frac{\log(T)}{M(\Delta_{k,m}^*)^2}\right)$</td>
<td>$O\left(\sum_{m=1}^{M} \sum_{k \neq k'<em>{*,m}} \frac{1}{\Delta</em>{k,m}} + \frac{1 - \lambda}{\Delta_{k,m}^<em>} \log(T) + \frac{C \log(T)}{\lambda(\Delta_{\text{min}}^</em>)^2}\right)$</td>
</tr>
<tr>
<td>$\lambda \log(T)$</td>
<td>$O\left(\frac{1}{M(\Delta_{k,m}^*)^2}\right)$</td>
<td>$O\left(\sum_{m=1}^{M} \sum_{k \neq k'<em>{*,m}} \frac{1}{\Delta</em>{k,m}} + \frac{1 - \lambda}{\Delta_{k,m}^<em>} \log(T) + \frac{C \log(T)}{\lambda(\Delta_{\text{min}}^</em>)^2}\right)$</td>
</tr>
<tr>
<td>$2^p$</td>
<td>$O\left(\log\left(\frac{\log(T)}{M(\Delta_{k,m}^*)^2}\right)\right)$</td>
<td>$O\left(\sum_{m=1}^{M} \sum_{k \neq k'<em>{*,m}} \frac{1}{\Delta</em>{k,m}} + \frac{1 - \lambda}{\Delta_{k,m}^<em>} \log(T) + CM \log\left(\frac{\log(T)}{M(\Delta_{\text{min}}^</em>)^2}\right)\right)$</td>
</tr>
<tr>
<td>$2^p \log(T)$</td>
<td>$O\left(\log\left(\frac{1}{M(\Delta_{k,m}^*)^2}\right)\right)$</td>
<td>$O\left(\sum_{m=1}^{M} \sum_{k \neq k'<em>{*,m}} \frac{1}{\Delta</em>{k,m}} + \frac{1 - \lambda}{\Delta_{k,m}^<em>} \log(T) + CM \log\left(\frac{1}{M(\Delta_{\text{min}}^</em>)^2}\right)\right)$</td>
</tr>
</tbody>
</table>
With the concentration inequality for subgaussian random variables, we have

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

Thus, with the union bound, \( P_G \) can be bounded as

\[
P_G = 1 - P \left\{ \exists p, \exists m \in [M], \exists k \in A_m(p), |\bar{\mu}_{k,m}'(p) - \mu_{k,m}'| \geq B_p \right\}
\]

\[
\geq 1 - \sum_{p=1}^{T} \sum_{m=1}^{M} \sum_{k=1}^{K} P \left( |\bar{\mu}_{k,m}'(p) - \mu_{k,m}'| \geq B_p \right)
\]

\[
\geq 1 - \frac{2MK}{T}.
\]

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.

\[\square\]

### E.2 Proof of Lemma 2

**Proof.** Recall that \( \forall k \neq k^*_m, \bar{p}_{k,m}' \) 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

\[
\bar{\mu}_{k,m}'(p_{k,m}') + B_{p_{k,m}'}(i) \leq \bar{\mu}_{k,m}' \leq \mu_{k,m}' + \frac{\Delta_{k,m}'}{2}
\]

\[
= \mu_{k,m}' - \frac{\Delta_{k,m}'}{2} \leq \bar{\mu}_{k,m}',(i) \leq B_{p_{k,m}'}(i) + B_{p_{k,m}-} - \frac{\Delta_{k,m}'}{2} \leq \bar{\mu}_{k,m}',(i) - B_{p_{k,m}-},
\]

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 \( m \).

\[\square\]

### 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 \( \sum_{p=1}^{\bar{p}_{k,m}} [\alpha M f(p)] \) times in the local exploration sub-phases, which leads to the local exploration loss as

\[
R_{l}^{exp}(T) \leq \sum_{m=1}^{M} \sum_{k \neq k^*_m} \sum_{p=1}^{\bar{p}_{k,m}} [\alpha M f(p)].
\]

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 \( \sum_{p=1}^{\bar{p}_{k}} [(1 - \alpha)f(p)] \) times in the global exploration sub-phases. Thus, the global exploration loss can be bounded as:

\[
R_{g}^{exp}(T) \leq \sum_{m=1}^{M} \sum_{k \neq k^*_m} \sum_{p=1}^{\bar{p}_{k}} [(1 - \alpha)f(p)].
\]
E.4 Proof of Lemma 4

Proof. At phase \( p \), the exploitation time for client \( m \) is at most \( \max_n \{|A_n(p)| - A_m(p)\} [M\alpha f(p)] \), 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 \( k'_{s,m} \), is arm \( k \) instead of \( k'_{s,m} \) can be bounded as:

\[
P (k'_{s,m} = k) \leq P \left( \hat{\mu}_{k',m}(p-1) \leq \mu_{k,m}(p-1) \right) = P \left( \hat{\mu}_{k',m}(p-1) - \mu_{k,m}(p-1) - \Delta_{k,m} \leq -\Delta_{k,m}' \right) \\
\leq 2 \exp \left\{ - \frac{(\Delta_{k,m}')^2}{2(\sigma_{k,m}'(p-1) + \sigma_{k,m}(p-1))} \right\} \\
\leq 2 \exp \left\{ - \frac{(\Delta_{k,m}')^2 MF(p-1)}{4} \right\} = P_{k,m}(p).
\]

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

\[
R_{k,m}^{exp}(T) \leq \Delta_{k,m}' \sum_{p=1}^{\hat{p}_{k,m}} \left( \max_n \{|A_n(p)| - A_m(p)\} [M\alpha f(p)] \right) P_{k,m}(p) \\
\leq \Delta_{k,m}' \sum_{p=1}^{\hat{p}_{k,m}} K [M\alpha f(p)] \exp \left\{ - \frac{(\Delta_{k,m}')^2 MF(p-1)}{4} \right\}.
\]

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

\[
R^{exp}(T) = \sum_{m=1}^{M} \sum_{k=1}^{K} \Delta_{k,m}' R_{k,m}^{exp}(T) \leq \sum_{m=1}^{M} \sum_{k=1}^{K} \sum_{p=1}^{\hat{p}_{k,m}} K [M\alpha f(p)] \Delta_{k,m}' \exp \left\{ - \frac{(\Delta_{k,m}')^2 MF(p-1)}{4} \right\}.
\]

In addition, we note that in phase \( p = 1 \), all the players share the same global and local active arm sets, i.e., \( \forall m \in [M], A_m(1) = A(1) = [K] \), which means there would be no exploration loss. Thus, the sum of index \( p \) in the exploration 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 \( p'_{\text{max}} = \max_{k \in [K]} \{p'_{k,m}\} = \max_{m \in [M]} \max_{k \neq k'_{s,m}} \{p'_{k,m}\} \). Before phase \( p'_{\text{max}} \), there are two communications in each phase for arm statistics and active sets, respectively, which leads to the communication loss upper bound as:

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

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:

\[
R(T) = P_G (R^{exp}(T) + R^{exp}(T) + R^{comm}(T)) + (1 - P_G)(1 + 2C)MT
\]
\[
\leq R_{1}^{expr}(T) + R_{g}^{expr}(T) + R_{p}^{expd}(T) + R_{comm}^{m}(T) + 2M^{2}K(1 + 2C)
\]
\[
\leq \sum_{m=1}^{M} \sum_{k \neq k',m} \Delta_{k,m} \sum_{p=1}^{p_{k,m}} \alpha Mf(p) + \sum_{m=1}^{M} \sum_{k \neq k',m} \Delta_{k,m} \sum_{p=1}^{p'} [(1 - \alpha)f(p)]
\]
\[
+ M \sum_{m=1}^{M} \sum_{k \neq k',m} \Delta_{k,m} \sum_{p=1}^{p_{k,m}} K [M \alpha f(p)] \exp \left\{ -\frac{(\Delta_{k,m}^{2})^{2} MF(p - 1)}{4} \right\} + 2CMp_{\max} + 2M^{2}K(1 + 2C).
\]

E.7 Proof of Corollary 2

\textbf{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.

F Additional Experimental Results

The implementation codes of the PF-UCB and its enhancement used for simulations have been made publicly available at \url{https://github.com/ShenGroup/PF_MAB}, which also contains the synthetic dataset and the preprocessed real-world MovieLens dataset. The original version of the MovieLens dataset is publicly available at \url{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
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


Chengshuai Shi, Cong Shen, Jing Yang


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.


