Doubly Adversarial Federated Bandits

Jialin Yi[1] **Milan Vojnović**[1]

Abstract

We study a new non-stochastic federated multi-armed bandit problem with multiple agents collaborating via a communication network. The losses of the arms are assigned by an oblivious adversary that specifies the loss of each arm not only for each time step but also for each agent, which we call "doubly adversarial". In this setting, different agents may choose the same arm in the same time step but observe different feedback. The goal of each agent is to find a globally best arm in hindsight that has the lowest cumulative loss averaged over all agents, which necessities the communication among agents. We provide regret lower bounds for any federated bandit algorithm under different settings, when agents have access to full-information feedback, or the bandit feedback. For the bandit feedback setting, we propose a near-optimal federated bandit algorithm called FEDEXP3. Our algorithm gives a positive answer to an open question proposed in (Cesa-Bianchi et al., 2016): FEDEXP3 can guarantee a sub-linear regret without exchanging sequences of selected arm identities or loss sequences among agents. We also provide numerical evaluations of our algorithm to validate our theoretical results and demonstrate its effectiveness on synthetic and real-world datasets.

1. Introduction

There is a rising trend of research on federated learning, which coordinates multiple *heterogeneous* agents to collectively train a learning algorithm, while keeping the raw data decentralized (Kairouz et al., 2021). We consider the federated learning variant of a multi-armed bandit problem which is one of the most fundamental sequential decision making problems. In standard multi-armed bandit problems, a learning agent needs to balance the trade-off between exploring various arms in order to learn how much rewarding they are and selecting high-rewarding arms. In federated bandit problems, multiple heterogeneous agents collaborate with each other to maximize their cumulative rewards. The challenge here is to design decentralized collaborative algorithms to find a *globally* best arm for all agents while keeping their raw data decentralized.

Finding a globally best arm with raw arm or loss sequences stored in a distributed system has ubiquitous applications in many systems built with a network of learning agents. One application is in recommender systems where different recommendation app clients (i.e. agents) in a communication network collaborate with each other to find news articles (i.e. arms) that are popular among all users within a specific region, which can be helpful to solve the *cold start* problem (Li et al., 2010; Yi et al., 2021). In this setting, the system avoids the exchange of users' browsing history (i.e. arm or loss sequences) between different clients for better privacy protection. Another motivation is in international collaborative drug discovery research, where different countries (i.e. agents) cooperate with each other to find a drug (i.e. arm) that is uniformly effective for all patients across the world (Varatharajah & Berry, 2022). To protect the privacy of the patients involved in the research, the exact treatment history of specific patients (i.e. arm or loss sequences) should not be shared during the collaboration.

The federated bandit problems are focused on identifying a globally best arm (pure exploration) or maximizing the cumulative group reward (regret minimization) in face of heterogeneous feedback from different agents for the same arm, which has gained much attention in recent years (Dubey & Pentland, 2020; Zhu et al., 2021; Huang et al., 2021; Shi et al., 2021; Réda et al., 2022). In prior work, heterogeneous feedbacks received by different agents are modeled as samples from some unknown but fixed distributions. Though this formulation of heterogeneous feedback allows elegant statistical analysis of the regret, it may not be adequate for dynamic (non-stationary) environments. For example, consider the task of finding popular news articles within a region mentioned above. The popularity of news articles on different topics can be time-varying, e.g. the news on football may become most popular during the FIFA World

[1] Department of Statistics, London School of Economics and Political Science, London, United Kingdom. Correspondence to: Jialin Yi <j.yi8@lse.ac.uk>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

Cup but may be less popular afterwards.

In contrast with the prior work, we introduce a new *non-stochastic* federated multi-armed bandit problem in which the heterogeneous feedback received by different agents are chosen by an oblivious adversary. We consider a federated bandit problem with K arms and N agents. The agents can share their information via a communication network. At each time step, each agent will choose one arm, receive the feedback and exchange their information with their neighbors in the network. The problem is *doubly adversarial*, i.e. the losses are determined by an oblivious adversary which specifies the loss of each arm not only for each time step but also for each agent. As a result, the agents which choose the same arm at the same time step may observe different losses. The goal is to find the *globally* best arm in hindsight, whose cumulative loss averaged over all agents is lowest, without exchanging raw information consisting of arm identity or loss value sequences among agents. As standard in online learning problems, we focus on regret minimization over an arbitrary time horizon.

1.1. Related work

The doubly adversarial federated bandit problem is related to several lines of research, namely that on federated bandits, multi-agent cooperative adversarial bandits, and distributed optimization. Here we briefly discuss these related works.

Federated bandits Solving bandit problems in the federated learning setting has gained attention in recent years. (Dubey & Pentland, 2020) and (Huang et al., 2021) considered the linear contextual bandit problem and extended the LinUCB algorithm (Li et al., 2010) to the federated learning setting. (Zhu et al., 2021) and (Shi et al., 2021) studied a federated multi-armed bandit problem where the losses observed by different agents are i.i.d. samples from some common unknown distribution. (Réda et al., 2022) considered the problem of identifying a globally best arm for multi-armed bandit problems in a centralized federated learning setting. All these works focus on the stochastic setting, i.e. the reward or loss of an arm is sampled from some unknown but fixed distribution. Our work considers the non-stochastic setting, i.e. losses are chosen by an oblivious adversary, which is a more appropriate assumption for non-stationary environments.

Multi-agent cooperative adversarial bandit (Cesa-Bianchi et al., 2016; Bar-On & Mansour, 2019; Yi & Vojnovic, 2022) studied the adversarial case where agents receive the same loss for the same action chosen at the same time step, whose algorithms require the agents to exchange their raw data with neighbors. (Cesa-Bianchi et al., 2020) discussed the cooperative online learning setting where the agents have access to the full-information feedback and the communication is asynchronous. In these works, the agents that choose the same action at the same time step receive the same reward or loss value and agents aggregate messages received from their neighbors. Our work relaxes this assumption by allowing agents to receive different losses even for the same action in a time step. Besides, we propose a new algorithm that uses a different aggregation of messages than in the aforementioned papers, which is based on distributed dual averaging method in (Nesterov, 2009; Xiao, 2009; Duchi et al., 2011).

Distributed optimization (Duchi et al., 2011) proposed the dual averaging algorithm for distributed convex optimization via a gossip communication mechanism. Subsequently, (Hosseini et al., 2013) extended this algorithm to the online optimization setting. (Scaman et al., 2019) found optimal distributed algorithms for distributed convex optimization and a lower bound which applies to strongly convex functions. The doubly adversarial federated bandit problem with full-information feedback is a special case of distributed online linear optimization problems. Our work complements these existing studies by providing a lower bound for the distributed online linear optimization problems. Moreover, our work proposes a near-optimal algorithm for the more challenging bandit feedback setting.

1.2. Organization of the paper and our contributions

We first formally formulate the doubly adversarial federated bandit problem and the federated bandit algorithms we study in Section 2. Then, in Section 3, we provide two regret lower bounds for any federated bandit algorithm under the full-information and bandit feedback setting, respectively. In Section 4, we present a federated bandit algorithm adapted from the celebrated Exp3 algorithm for the bandit-feedback setting, together with its regret upper bound. Finally, we show results of our numerical experiments in Section 5.

Our contributions can be summarized as follows:

(i) We introduce a new federated bandit setting, doubly adversarial federated bandits, in which no stochastic assumptions are made for the heterogeneous losses received by the agents. This adversarial setting complements the prior work focuses on the stochastic setting.

(ii) For both the full-information and bandit feedback setting, we provide regret lower bounds for any federated bandit algorithm. The regret lower bound for the full-information setting also applies to distributed online linear optimization problems, and, to the best of our knowledge, is the first lower bound result for this problem.

(iii) For the bandit feedback setting, we propose a new near-optimal federated bandit Exp3 algorithm (FEDEXP3) with a sub-linear regret upper bound. Our FEDEXP3 algorithm resolves an open question proposed in (Cesa-Bianchi et al., 2016): it is possible to achieve a sub-linear regret for each agent simultaneously without exchanging both the action or loss information and the distribution information among agents.

2. Problem setting

Consider a communication network defined by an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of N agents and $(u, v) \in \mathcal{E}$ if agent u and agent v can directly exchange messages. We assume that $\mathcal{G}$ is simple, i.e. it contains no self loops nor multiple edges. The agents in the communication network collaboratively aim to solve a non-stochastic multi-armed bandit problem. In this problem, there is a fixed set $\mathcal{A}$ of K arms and a fixed time horizon T. Each instance of the problem is parameterized by a tensor $L = (\ell_t^v(i)) \in [0,1]^{T\times N\times K}$ where $\ell_t^v(i)$ is the loss associated with agent $v \in \mathcal{V}$ if it chooses arm $i \in \mathcal{A}$ at time step t.

At each time step t, each agent v will choose its action $a_t^v = i$, observe the feedback I_t^v and incur a loss defined as the average of losses of arm i over all agents, i.e.,

$$\bar{\ell}_t(i) = \frac{1}{N}\sum_{v\in\mathcal{V}} \ell_t^v(i). \tag{1}$$

At the end of each time step, each agent $v \in \mathcal{V}$ can communicate with their neighbors $\mathcal{N}(v) = \{u \in \mathcal{V} : (u, v) \in \mathcal{E}\}$. We assume a *non-stochastic* setting, i.e. the loss tensor L is determined by an oblivious adversary. In this setting, the adversary has the knowledge of the description of the algorithm running by the agents but the losses in L do not depend on the specific arms selected by the agents.

The performance of each agent $v \in \mathcal{V}$ is measured by its *regret*, defined as the difference of the expected cumulative loss incurred and the cumulative loss of a *globally* best fixed arm in hindsight, i.e.

$$R_T^v(\pi, L) = \mathbb{E}\left[\sum_{t=1}^{T} \bar{\ell}_t(a_t^v) - \min_{i\in\mathcal{A}}\left\{\sum_{t=1}^{T}\bar{\ell}_t(i)\right\}\right] \tag{2}$$

where the expectation is taken over the action of all agents under algorithm π on instance L. We will abbreviate $R_T^v(\pi, L)$ as R_T^v when the algorithm π and instance L have no ambiguity in the context. We aim to characterize $\max_L R_T^v(\pi, L)$ for each agent $v \in \mathcal{V}$ under two feedback settings,

- full-information feedback: $I_t^v = \ell_t^v$, and
- bandit feedback: $I_t^v = \ell_t^v(a_t^v)$.

Let $\mathcal{F}_t^v$ be the sequence of agent v's actions and feedback up to time step t, i.e., $\mathcal{F}_t^v = \bigcup_{s=1}^{t}\{a_s^v, I_s^v\}$. For a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, we denote as $d(u, v)$ the number of edges of a shortest path connecting nodes u and v in $\mathcal{V}$ and $d(v, v) = 0$.

We focus on the case when π is a *federated bandit algorithm* in which each agent $v \in \mathcal{V}$ can only communicate with their neighbors within a time step.

Definition 2.1 (federated bandit algorithm)**.** A federated bandit algorithm π is a multi-agent learning algorithm such that for each round t and each agent $v \in \mathcal{V}$, the action selection distribution p_t^v only depends on $\bigcup_{u\in\mathcal{V}} \mathcal{F}^u_{t-d(u,v)-1}$.

From Definition 2.1, the communication between any two agents u and v in $\mathcal{G}$ comes with a delay equal to $d(u, v) + 1$. Here we give some examples of π in different settings:

- when $|\mathcal{V}| = 1$ and $I_t^v = \ell_t^v$, π is an online learning algorithm for learning with expert advice problems (Cesa-Bianchi et al., 1997),
- when $|\mathcal{V}| = 1$ and $I_t^v = \ell_t^v(a_t^v)$, π is a sequential algorithm for a multi-armed bandit problem (Auer et al., 2002),
- when $I_t^v \in \partial f_i(x_t^v)$ for some convex function $f(x)$, π belongs to the black-box procedure for distributed convex optimization over a simplex (Scaman et al., 2019), and
- when $\mathcal{G}$ is a star graph, π is a centralized federated bandit algorithm discussed in (Réda et al., 2022).

3. Lower bounds

In this section, we show two lower bounds on the cumulative regret of any federated bandit algorithm π in which all agents exchange their messages through the communication network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, for full-information and bandit feedback setting. Both lower bounds highlight how the cumulative regret of the federated algorithm π is related to the minimum time it takes for all agents in $\mathcal{G}$ to reach an agreement on a globally best arm.

Agents reaching an agreement about a globally best arm in hindsight is to find $i^* \in \arg\min_{i\in\mathcal{A}} \sum_{t=1}^{T} \bar{\ell}_t(i)$ by each agent v exchanging their private information about $\{\sum_{t=1}^{T} \ell_t^v(i) : i \in \mathcal{A}\}$ with their neighbors. This is known as a distributed consensus averaging problem (Boyd et al., 2004). Let $d_v = |\mathcal{N}(v)|$ and $d_{\max} = \max_{v\in\mathcal{V}} d_v$ and $d_{\min} = \min_{v\in\mathcal{V}} d_v$. The dynamics of a consensus averaging procedure is usually characterized by spectrum of the Laplacian matrix M of graph $\mathcal{G}$ defined as

$$M_{u,v} := \begin{cases} d_u & \text{if } u = v \\ -1 & \text{if } u \neq v \text{ and } (u, v) \in \mathcal{E} \\ 0 & \text{otherwise.} \end{cases}$$

Let $\lambda_1(M) \geq \cdots \geq \lambda_N(M) = 0$ be the eigenvalues of the Laplacian matrix M. The second smallest eigenvalue $\lambda_{N-1}(M)$ is the *algebraic connectivity* which approximates the sparest-cut of graph $\mathcal{G}$ (Arora et al., 2009). In the following two theorems, we show that for any federated bandit algorithm π, there always exists a problem instance and an agent whose worst-case cumulative regret is $\Omega(\lambda_{N-1}(M)^{-1/4}\sqrt{T})$.

Theorem 3.1 (Full-information feedback)**.** *For any federated bandit algorithm π, there exists a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with Laplacian matrix M and a full-information feedback instance $L \in [0,1]^{T\times N\times K}$ such that for some $v_1 \in \mathcal{V}$,*

$$R_T^{v_1}(\pi, L) = \Omega\left(\sqrt[4]{\frac{1+d_{\max}}{\lambda_{N-1}(M)}}\sqrt{T\log K}\right). \tag{3}$$

The proof, in Appendix A.1, relies on the existence of a graph in which there exist two clusters of agents, A and B, with distance $d(A,B) = \min_{u\in A, v\in B} d(u,v) = \Omega\left(\sqrt{(d_{\max}+1)/\lambda_{N-1}(M)}\right)$. Then, we consider an instance where only agents in cluster A receive non-zero losses. Based on a reduction argument, the cumulative regrets for agents in cluster B are the same as (up to a constant factor) the cumulative regret in a single-agent adversarial bandit problem with feedback of delay $d(A,B)$ (see Lemma A.4 in Appendix A.1). Hence, one can show that the cumulative regret of agents in cluster B is $\Omega\left(\sqrt{d(A,B)}\sqrt{T\log K}\right)$.

Note that the doubly adversarial federated bandit with full-information feedback is a special case of distributed online linear optimization, with the decision set being a $K-1$-dimensional simplex. Hence, Theorem 3.1 immediately implies a regret lower bound for the distributed online linear optimization problem. To the best of our knowledge, this is the first non-trivial lower bound that relates the hardness of distributed online linear optimization problem to the algebraic connectivity of the communication network.

Leveraging the lower bound for the full-information setting, we show a lower bound for the bandit feedback setting.

Theorem 3.2 (Bandit feedback)**.** *For any federated bandit algorithm π, there exists a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with Laplacian matrix M and a bandit feedback instance $L \in [0,1]^{T\times N\times K}$ such that for some $v_1 \in \mathcal{V}$,*

$$R_T^{v_1}(\pi, L) = \Omega\left(\max\left\{\sqrt{\frac{1}{1+d_{v_1}}}\sqrt{KT},\right.\right. \\ \left.\left.\sqrt[4]{\frac{1+d_{\max}}{\lambda_{N-1}(M)}}\sqrt{T\log K}\right\}\right). \tag{4}$$

The proof is provided in Appendix A.2. The lower bound contains two parts. The first part, derived from the information-theoretic argument in (Shamir, 2014), captures the effect from bandit feedback. The second part is inherited from Theorem 3.1 by the fact that the regret of an agent in bandit feedback setting cannot be smaller than its regret in full-information setting.

4. FEDEXP3: a federated regret-minimization algorithm

Inspired by the fact that the cumulative regret is related to the time need to reach consensus about a globally best arm, we introduce a new federated bandit algorithm based on the gossip communication mechanism, called FEDEXP3. The details of FEDEXP3 are described in Algorithm 1. We shall also show that FEDEXP3 has a sub-linear cumulative regret upper bound which holds for all agents simultaneously.

The FEDEXP3 algorithm is adapted from the Exp3 algorithm, in which each agent v maintains an estimator $z_t^v \in \mathbb{R}^K$ of the cumulative losses for all arms and a tentative action selection distribution $x_t^v \in [0,1]^K$. At the beginning of each time step t, each agent follows the action selection distribution x_t^v with probability $1-\gamma_t$, or performs a uniform random exploration with probability γ_t. Once the action a_t^v is sampled, the agent observes the associated loss $\ell_t^v(a_t^v)$ and then computes an importance-weighted loss estimator $g_t^v \in \mathbb{R}^K$ using the sampling probability $p_t^v(a_t^v)$. Before g_t^v is integrated into the cumulative loss estimator z_{t+1}^v, the agent communicates with its neighbors to average its cumulative loss estimator z_t^v.

The communication step is characterized by the *gossip* matrix which is a doubly stochastic matrix $W \in [0,1]^{N\times N}$ satisfying the following constraints

$$\sum_{v\in\mathcal{V}} W_{u,v} = \sum_{u\in\mathcal{V}} W_{u,v} = 1$$

and $W_{u,v} \geq 0$ where equality holds when $(u,v) \notin \mathcal{E}$. This gossip communication step facilitates the agents to reach a consensus on the estimators of the cumulative losses of all arms, and hence allows the agents to identify a globally best arm in hindsight. We present below an upper bound on the cumulative regret of each agent in FEDEXP3.

Theorem 4.1. *Assume that the network runs Algorithm 1 with*

$$\gamma_t = \sqrt[3]{\frac{\left(C_W + \frac{1}{2}\right)K^2\log K}{t}}$$

and

$$\eta_t = \frac{\log K}{T\gamma_T} = \sqrt[3]{\frac{(\log K)^2}{\left(C_W + \frac{1}{2}\right)K^2T^2}}$$

with $C_W = \min\{2\log T + \log N, \sqrt{N}\}/(1-\sigma_2(W)) + 3$.

Algorithm 1 FEDEXP3

```
Input: Learning rates $\{\eta_t > 0\}$, Exploration ratios $\{\gamma_t > 0\}$, and a gossip matrix $W \in [0,1]^{N\times N}$.
Initialize: $z_1^v(i) = 0, x_1^v(i) = 1/K$ for all $i \in \mathcal{A}$ and $v \in \mathcal{V}$.
for each time step $t = 1, 2, \ldots, T$ do
    for each agent $v \in \mathcal{V}$ do
        compute the action distribution
            $p_t^v(i) = (1-\gamma_t)x_t^v(i) + \gamma_t/K;$
        choose the action $a_t^v \sim p_t^v$;
        compute the loss estimators
            $g_t^v(i) = \ell_t^v(i)\mathbb{I}\{a_t^v = i\}/p_t^v(i);$
        update the gossip accumulative loss
            $z_{t+1}^v = \sum_{u:(u,v)\in\mathcal{E}} W_{u,v} z_t^u + g_t^v;$
        update the tentative action selection distribution
            $x_{t+1}^v = \frac{\exp\{-\eta_t z_{t+1}^v(i)\}}{\sum_{j\in A}\exp\{-\eta_t z_{t+1}^v(j)\}};$
    end for
end for
```

Then, the expected regret of each agent $v \in \mathcal{V}$ is bounded as

$$R_T^v = \tilde{O}\left(\frac{1}{\sqrt[3]{1-\sigma_2(W)}}K^{2/3}T^{2/3}\right)$$

where $\sigma_2(W)$ is the second largest singular value of W.

Proof sketch Let $\hat{\ell}_t$ and $\bar{z}_t$ be the average instant loss estimator and average cumulative loss estimator,

$$f_t = \frac{1}{N}\sum_{v\in\mathcal{V}} g_t^v \quad \text{and} \quad \bar{z}_t = \frac{1}{N}\sum_{v\in\mathcal{V}} z_t^v,$$

and let y_t be action distribution that minimizes the regularized average cumulative loss estimator

$$y_t(i) = \frac{\exp\{-\eta_{t-1}\bar{z}_t(i)\}}{\sum_{j\in A}\exp\{-\eta_{t-1}\bar{z}_t(j)\}}.$$

The cumulative regret can be bounded by the sum of three terms

$$R_t^v \le \underbrace{\mathbb{E}\left[\sum_{t=1}^{T}\left(\langle f_t, y_t^v\rangle - f_t(i^*)\right)\right]}_{\text{FTRL}} + \underbrace{K\sum_{t=1}^{T}\eta_{t-1}\mathbb{E}\|z_t^v - \bar{z}_t\|_*}_{\text{CONSENSUS}} + \underbrace{\sum_{t=1}^{T}\gamma_t}_{\text{EXPLORATION}}$$

where $i^* \in \arg\min_{i\in\mathcal{A}}\sum_{t=1}^{T}\bar{\ell}_t(i)$ is a globally best arm in hindsight.

The FTRL term is a typical regret term from the classic analysis for the Follow-The-Regularized-Leader algorithm (Lattimore & Szepesvári, 2020). The CONSENSUS term measures the cumulative approximation error generated during the consensus reaching process, which can be bounded using the convergence analysis of distributed averaging algorithm based on doubly stochastic matrices (Duchi et al., 2011; Hosseini et al., 2013). The last EXPLORATION term can be bounded by specifying the time-decaying exploration ratio γ_t. The full proof of Theorem 4.1 is provided in Appendix A.4.

The FEDEXP3 algorithm is also a valid algorithm for the multi-agent adversarial bandit problem (Cesa-Bianchi et al., 2016) which is a special case of the doubly adversarial federated bandit problem when $\ell_t^v(i) = \ell_t(i)$ for all $v \in \mathcal{V}$. According to the distributed consensus process of FEDEXP3, each agent $v \in \mathcal{V}$ only communicates cumulative loss estimator values z_t^v, instead of the actual loss values $\ell_t^v(a_t^v)$, and the selection distribution p_t^v. FEDEXP3 can guarantee a sub-linear regret without the exchange of sequences of selected arm identities or loss sequences of agents, which resolves an open question raised in (Cesa-Bianchi et al., 2016).

Choice of the gossip matrix The gossip matrix W can be constructed using the *max-degree* trick in (Duchi et al., 2011), i.e.,

$$W = I - \frac{D-A}{2(1+d_{\max})}$$

where $D = \operatorname{diag}(d_1, \ldots, d_N)$ and A is the adjacency matrix of $\mathcal{G}$. This construction of W requires that all agents have knowledge of the maximum degree $d_{\max}$, which can indeed be easily computed in a distributed system by nodes exchanging messages and updating their states using the maximum reduce operator.

Another choice of W comes from the effort to minimize the cumulative regret. The leading factor $1/\sqrt[3]{1-\sigma_2(W)}$ in the regret upper bound of FEDEXP3 can be minimized by choosing a gossip matrix W with smallest $\sigma_2(W)$. Suppose that the agents have knowledge of the topology structure of the communication graph $\mathcal{G}$, then the agents can choose the gossip matrix to minimize their regret by solving the following convex optimization problem:

$$\begin{array}{rl} \text{minimize} & \left\|W - (1/n)\mathbf{1}\mathbf{1}^T\right\|_2 \\ \text{subject to} & W \geq 0, W\mathbf{1} = \mathbf{1}, W = W^T, \\ & W_{ij} = 0, \text{ for } (i,j) \notin \mathcal{E} \end{array}$$

which has an equivalent semi-definite programming formulation as noted in (Boyd et al., 2004).

Gap between upper and lower bounds The regret upper bound of FEDEXP3 algorithm in Theorem 4.1 grows sublinearly in the number of arms K and horizon time T. There is only a small polynomial gap between the regret upper bound and the lower bound in Theorem 3.2 with respect to these two parameters. The regret upper bound depends also on the second largest singular value $\sigma_2(W)$ of W. The related term in the lower bound in Theorem 3.2 is the second smallest eigenvalue $\lambda_{N-1}(M)$ of the Laplacian matrix M. To compare these two terms, we point that when the gossip matrix is constructed using the max-degree method, as discussed in Corollary 1 in (Duchi et al., 2011),

$$\frac{1}{\sqrt[3]{1-\sigma_2(W)}} \leq \sqrt[3]{2\frac{d_{\max}+1}{\lambda_{N-1}(M)}}.$$

With respect to $\sqrt[4]{(d_{\max}+1)/\lambda_{N-1}(M)}$ in Theorem 3.2, there is only a small polynomial gap between the regret upper bound and the lower bound. We note that a similar dependence on $\sigma_2(W)$ is present in the analysis of distributed optimization algorithms (Duchi et al., 2011; Hosseini et al., 2013).

5. Numerical experiments

We present experimental results for the FEDEXP3 algorithm (W constructed by the max-degree method) using both synthetic and real-world datasets. We aim to validate our theoretical analysis and demonstrate the effectiveness of FEDEXP3 on finding a globally best arm in non-stationary environments. All the experiments are performed with 10 independent runs. The code for producing our experimental results is available online in an anonymous Github repository: https://github.com/jialinyi94/doubly-stochastic-federataed-bandit.

5.1. Synthetic datasets

We validate our theoretical analysis of the FEDEXP3 algorithm on synthetic datasets. The objective is two-fold. First, we demonstrate that the cumulative regret of FEDEXP3 grows sub-linearly with time. Second, we examine the dependence of the regret on the second largest singular value of the gossip matrix.

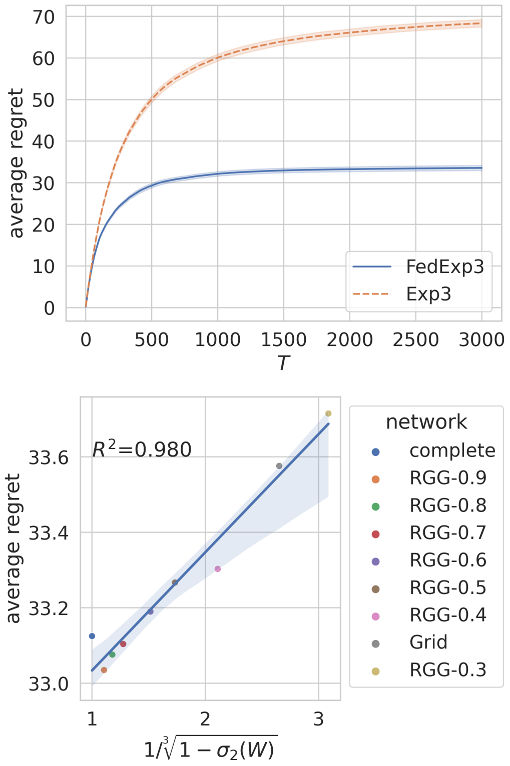

Figure 1. (Top) The average cumulative regret, i.e. $\sum_{v\in\mathcal{V}} R_T^v/N$, versus T, for FEDEXP3 and Exp3 on the grid graph. (Down) The average cumulative regret versus $(1-\sigma_2(W))^{-1/3}$ for FEDEXP3 on different networks at $T = 3000$.

A motivation for finding a globally best arm in recommender systems is to provide recommendations for those users whose feedback is sparse. In this setting, we construct a federated bandit setting in which a subset of agents will be activated at each time step and only activated agents may receive non-zero loss. Specifically, we set $T = 3,000$ with $N = 36$ and $K = 20$. At each time step t, a subset U_t of $N/2$ agents are selected from $\mathcal{V}$ with replacement. For all activated agents U_t, the loss for arm i is sampled independently from Bernoulli distribution with mean $\mu_i = (i-1)/(K-1)$. All non-activated agents receive a loss of 0 for any arm they choose at time step t.

We evaluate the performance of FEDEXP3 on different networks, i.e. for a complete graph, a $\sqrt{N}$ by $\sqrt{N}$ grid network, and random geometric graphs. The random geometric graph RGG(d) is constructed by uniform random placement of each node in $[0,1]^2$ and connecting any two nodes whose distance is less or equal to d (Penrose, 2003). Random geometric graphs are commonly used for modeling spatial networks.

In our experiments, we set $d \in \{0.3, \ldots, 0.9\}$. The results in Figure 1 confirm that the cumulative regret of the FED-

Exp3 algorithm grows sub-linearly with respect to time and suggest that the cumulative regret of FedExp3 is proportional to $(1-\sigma_2(W))^{-1/3}$. This is compatible with the regret upper bound in Theorem 4.1.

5.2. MovieLens dataset: recommending popular movie genres

We compare FedExp3 with a UCB-based federated bandit algorithm in a movie recommendation scenario using a real-world dataset. In movie recommendation settings, users' preferences over different genres of movies can change over time. In such non-stationary environments, we demonstrate that a significant performance improvement can be achieved by FedExp3 against the GossipUCB algorithm (we refer to as GUCB) proposed in (Zhu et al., 2021), which is defined for stationary settings.

We evaluate the two algorithms using the MovieLens-Latest-full dataset which contains 58,000 movies, classified into 20 genres, with 27,000,000 ratings (rating scores in $\{0.5, 1, \ldots, 5\}$) from 280,000 users. Among all the users, there are 3,364 users who rated at least one movie for every genre. We select these users as our agents, i.e. $N = 3,364$, and the 20 genres as the arms to be recommended, i.e. $K = 20$.

We create a federated bandit environment for movie recommendation based on this dataset. Let $m^v(i)$ be the number of ratings that agent v has for genre i. We set the horizon $T = \max_{v\in\mathcal{N}} m^v(i) = 12,800$. To reflect the changes in agents' preferences over genres as time evolves, we sort the ratings in an increasing order by their Unix timestamps and construct the loss tensor in the following way. Let $r_j^v(i)$ be the j-th rating of agent v on genre i, the loss of recommending an movie to agent v of genre i at time step t is defined as

$$\ell_t^v(i) = \frac{5.5 - r_j^v(i)}{5.5}$$

for $t \in \left[(j-1)\left\lfloor \frac{T}{m^v(i)} \right\rfloor, j\left\lfloor \frac{T}{m^v(i)} \right\rfloor\right)$. The performance of FedExp3 and GUCB is shown in Figure 2. The results demonstrate that FedExp3 can outperform GUCB by a significant margin for different communication networks.

6. Conclusion and future research

We studied doubly adversarial federated bandits, a new adversarial (non-stochastic) setting for federated bandits, which complement prior study on stochastic federated bandits. Firstly, we derived regret lower bounds for any federated bandit algorithm when the agents have access to full-information or bandit feedback. These regret lower bounds relate the hardness of the problem to the algebraic connectivity of the network through which the agents communicate. Then we proposed the FedExp3 algorithm which is a federated version of the Exp3 algorithm. We showed that there is only a small polynomial gap between the regret upper bound of FedExp3 and the lower bound. Numerical experiments performed by using both synthetic and real-word datasets demonstrated that FedExp3 can outperform the state-of-the-art stochastic federated bandit algorithm by a significant margin in non-stationary environments.

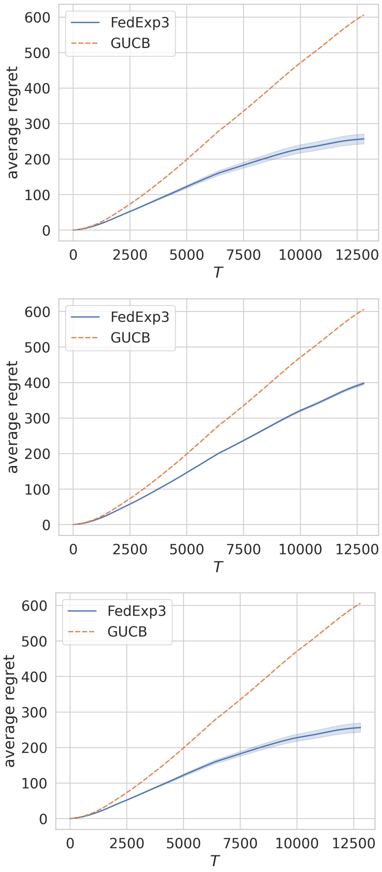

Figure 2. The average cumulative regret versus horizon time for FedExp3 and GUCB in the movie recommendation setting with the communication networks: (top) complete graph, (mid) the grid network, and (down) RGG(0.5).

We point out some interesting avenues for future research on doubly adversarial federated bandits. The first is to close the gap between the regret upper bound of FedExp3 algorithm and the lower bounds shown in this paper. The second is to extend the doubly adversarial assumption to federated linear bandit problems, where the doubly adversarial assumption could replace the stochastic assumption on the noise in the linear model.

References

Arora, S., Rao, S., and Vazirani, U. Expander flows, geometric embeddings and graph partitioning. *Journal of the*

ACM, 56(2):1–37, 2009.

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

Bar-On, Y. and Mansour, Y. Individual regret in cooperative nonstochastic multi-armed bandits. *Advances in Neural Information Processing Systems*, 32, 2019.

Boyd, S., Diaconis, P., and Xiao, L. Fastest mixing markov chain on a graph. *SIAM review*, 46(4):667–689, 2004.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge university press, 2006.

Cesa-Bianchi, N., Freund, Y., Haussler, D., Helmbold, D. P., Schapire, R. E., and Warmuth, M. K. How to use expert advice. *Journal of the ACM*, 44(3):427–485, 1997.

Cesa-Bianchi, N., Gentile, C., Mansour, Y., and Minora, A. Delay and cooperation in nonstochastic bandits. In *Conference on Learning Theory*, pp. 605–622. PMLR, 2016.

Cesa-Bianchi, N., Cesari, T., and Monteleoni, C. Cooperative online learning: Keeping your neighbors updated. In *Algorithmic Learning Theory*, pp. 234–250. PMLR, 2020.

Dubey, A. and Pentland, A. Differentially-private federated linear bandits. *Advances in Neural Information Processing Systems*, 33:6003–6014, 2020.

Duchi, J. C., Agarwal, A., and Wainwright, M. J. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic control*, 57(3):592–606, 2011.

Hagberg, A., Swart, P., and S Chult, D. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., et al. Array programming with numpy. *Nature*, 585(7825):357–362, 2020.

Hiriart-Urruty, J.-B. and Lemarechal, C. *Convex Analysis and Minimization Algorithms II: Advanced Theory and Bundle Methods: 306*. Springer, Berlin Heidelberg, softcover reprint of hardcover 1st ed. 1993 edition, December 2010.

Hosseini, S., Chapman, A., and Mesbahi, M. Online distributed optimization via dual averaging. In *52nd IEEE Conference on Decision and Control*, pp. 1484–1489. IEEE, 2013.

Huang, R., Wu, W., Yang, J., and Shen, C. Federated linear contextual bandits. *Advances in Neural Information Processing Systems*, 34:27057–27068, 2021.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

Lattimore, T. and Szepesvári, C. *Bandit Algorithms*. Cambridge University Press, 2020. doi: 10.1017/9781108571401.

Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 661–670, 2010.

Nesterov, Y. Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120(1):221–259, 2009.

Penrose, M. *Random geometric graphs*, volume 5. OUP Oxford, 2003.

Réda, C., Vakili, S., and Kaufmann, E. Near-optimal collaborative learning in bandits. *arXiv preprint arXiv:2206.00121*, 2022.

Scaman, K., Bach, F., Bubeck, S., Lee, Y., and Massoulié, L. Optimal convergence rates for convex distributed optimization in networks. *Journal of Machine Learning Research*, 20:1–31, 2019.

Shamir, O. Fundamental limits of online and distributed algorithms for statistical learning and estimation. *Advances in Neural Information Processing Systems*, 27, 2014.

Shi, C., Shen, C., and Yang, J. Federated multi-armed bandits with personalization. In *International Conference on Artificial Intelligence and Statistics*, pp. 2917–2925. PMLR, 2021.

Varatharajah, Y. and Berry, B. A contextual-bandit-based approach for informed decision-making in clinical trials. *Life*, 12(8):1277, 2022.

Xiao, L. Dual averaging method for regularized stochastic learning and online optimization. *Advances in Neural Information Processing Systems*, 22, 2009.

Yi, J. and Vojnovic, M. On regret-optimal cooperative nonstochastic multi-armed bandits. *arXiv preprint arXiv:2211.17154*, 2022.

Yi, J., Wu, F., Wu, C., Liu, R., Sun, G., and Xie, X. Efficient-fedrec: Efficient federated learning framework for privacy-preserving news recommendation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 2814–2824, 2021.

Zhu, Z., Zhu, J., Liu, J., and Liu, Y. Federated bandit: A gossiping approach. In *Abstract Proceedings of the 2021 ACM SIGMETRICS/International Conference on Measurement and Modeling of Computer Systems*, pp. 3–4, 2021.

A. Appendix

Notation For a vector x, we use $x(i)$ to denote the i-th coordinate of x. We define $\mathcal{F}_t = \bigcup_{v \in \mathcal{V}} \mathcal{F}_t^v$ where $\mathcal{F}_t^v$ is the sequence of agent v's actions and feedback up to time step t, i.e., $\mathcal{F}_t^v = \bigcup_{s=1}^t \{a_s^v, I_s^v\}$.

A.1. Proof of Theorem 3.1

We first define a new class of cluster-based distributed online learning procedure, referred to as *cluster-based federated algorithms*, in which the delay only occurs when the communication is between different clusters. The regret lower bound for federated bandit algorithms will be no less than the regret lower bound for cluster-based federated algorithms, as shown in Lemma A.2. Then we show in Lemma A.3 that there exists a special graph in which there exist two clusters of agents A and B with distance $d(A, B) = \min_{u \in A, v \in B} d(u, v) = \Omega\left(\sqrt{(d_{\max} + 1)/\lambda_{N-1}(M)}\right)$. Then, we consider an instance where only agents in cluster A receive non-zero losses. Based on a reduction argument, the cumulative regrets of agents in cluster B are the same as (up to a constant factor) the cumulative regrets in a single-agent adversarial bandit setting with feedback delay $d(A, B)$ (see Lemma A.4 in Appendix A.1). Hence, one can show that the cumulative regret of agents in cluster B is $\Omega\left(\sqrt{d(A, B)}\sqrt{T \log K}\right)$.

We denote with $d(\mathcal{U}, \mathcal{U}')$ the smallest distance between any two nodes in $\mathcal{U}, \mathcal{U}' \subset \mathcal{V}$, i.e.

$$d(\mathcal{U}, \mathcal{U}') = \min_{u \in U, u' \in \mathcal{U}'} d(u, u')$$

where $d(u, v)$ is the length of a shortest path connecting u and u'.

Definition A.1 (Cluster-based federated algorithms)**.** A cluster-based federated algorithm is a multi-agent learning algorithm defined by a partition of graph $\bigcup_r \mathcal{U}_r = \mathcal{V}$ where $\mathcal{U}_r$ is called cluster. In the cluster-based federated algorithm, at each round t, the action selection probability p_t^v of agent $v \in \mathcal{U}_r$ depends on the history information up to round $t - d(\mathcal{U}_r, \mathcal{U}_{r'}) - 1$ of all agents $u' \in \mathcal{U}_{r'}$.

Note that when all agents are in the same cluster $\mathcal{V}$, the centralized federated algorithm in (Réda et al., 2022) is an instance of a cluster-based federated algorithm.

Lemma A.2 (Monotonicity)**.** *Let Π and Π' be two sets of all cluster-based federated algorithms with two partitions $\bigcup_r \mathcal{U}_r$ and $\bigcup_s \mathcal{U}'_s$, respectively. Suppose for any cluster $\mathcal{U}'_s$ of π', there exists a cluster $\mathcal{U}_r$ of π such that $\mathcal{U}'_s \subset \mathcal{U}_r$, then*

$$\Pi' \subset \Pi \quad \textit{and} \quad \min_{\pi \in \Pi} R_T^v(\pi, L) \leq \min_{\pi' \in \Pi'} R_T^v(\pi', L)$$

for any $L \in [0, 1]^{T \times N \times K}$ and any $v \in \mathcal{V}$.

Proof. It suffices to show $\Pi' \subset \Pi$. Consider a cluster-based federated algorithm $\pi' \in \Pi'$. For any agent $v \in \mathcal{V}$, let $\mathcal{U}'_s$ be the cluster of v in Π'. By definition of cluster-based procedure, agent v's action selection distribution probability p_t^v depends on the history information up to round $t - d(\mathcal{U}'_s, \mathcal{U}'_h) - 1$ of all agents $u' \in \mathcal{U}'_h$.

By the assumption, there exists two subset $\mathcal{U}_{r_1}, \mathcal{U}_{r_2} \subset \mathcal{V}$ such that $\mathcal{U}'_s \subset \mathcal{U}_{r_1}$ and $\mathcal{U}'_h \subset \mathcal{U}_{r_2}$. Hence $d(\mathcal{U}_{r_1}, \mathcal{U}_{r_2}) \leq d(\mathcal{U}'_s, \mathcal{U}'_h)$, from which it follows $t - d(\mathcal{U}'_s, \mathcal{U}'_h) - 1 \leq t - d(\mathcal{U}_{r_1}, \mathcal{U}_{r_2}) - 1$. Hence $\pi' \in \Pi$ which completes the proof. □

Lemma A.3. *There exists a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with N nodes and a matrix $M \in \mathcal{M}_{\mathcal{G}}$, together with two subsets of nodes $I_0, I_1 \subset \mathcal{V}$ of size $|I_0| = |I_1| \geq N/4$ and such that*

$$d\left(I_0, I_1\right) \geq \tilde{\Delta},$$

where $d\left(I_0, I_1\right)$ is the shortest-path distance in $\mathcal{G}$ between the two sets and

$$\tilde{\Delta} = \frac{\sqrt{2}}{3}\sqrt{\frac{1 + d_{\max}}{\lambda_{N-1}(M)}}.$$

Proof. From Lemma 24 in (Scaman et al., 2019), there exists exists a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with N nodes and a matrix $M \in \mathcal{M}_{\mathcal{G}}$, together with two subsets of nodes $I_0, I_1 \subset \mathcal{V}$ of size $|I_0| = |I_1| \geq N/4$ and such that

$$d\left(I_0, I_1\right) \geq \frac{\sqrt{2}}{3} \sqrt{\frac{\lambda_1(M)}{\lambda_{N-1}(M)}}.$$

We show that $\lambda_1(M) \geq 1 + d_{\max}$. To see this, note that $\lambda_1(M)$ is the Rayleigh quotient $\max_{x \neq 0} \frac{x^T M x}{x^T x}$. By the definition of the Laplacian matrix,

$$x^T M x = \sum_{(v,u) \in \mathcal{E}} \left(x_v - x_u\right)^2.$$

Let v be a vertex whose degree is $d_{\max}$ and

$$x_u := \begin{cases} \sqrt{\frac{d_{\max}}{1+d_{\max}}} & \text{if } u = v \\ -\frac{1}{\sqrt{d_{\max}}\sqrt{1+d_{\max}}} & \text{if } u \neq v \text{ and } v_i \text{ is adjacent to } v_j \\ 0 & \text{otherwise} \end{cases}$$

then

$$\sum_{(v,u) \in \mathcal{E}} \left(x_v - x_u\right)^2 = d_{\max} \left(\sqrt{\frac{d_{\max}}{1 + d_{\max}}} + \frac{1}{\sqrt{d_{\max}}\sqrt{1 + d_{\max}}} \right)^2 = d_{\max} \left(\frac{1}{\sqrt{d_{\max}}\sqrt{1 + d_{\max}}} \right)^2 \left(d_{\max} + 1\right)^2 = 1 + d_{\max}$$

and

$$\sum_{u \in \mathcal{V}} x_u^2 = \frac{d_{\max}}{1 + d_{\max}} + d_{\max} \frac{1}{d_{\max}(1 + d_{\max})} = 1.$$

Hence, $\lambda_1(M) \geq 1 + d_{\max}$. □

Let I_0, I_1 be two subsets of nodes satisfying

$$d(I_0, I_1) \geq \tilde{\Delta} \quad \text{and} \quad |I_0| = |I_1| = N/4.$$

The number of rounds needed to communicate between any node in I_0 and any node I_1 is at least $\tilde{\Delta}$.

Lemma A.4. *Let $v_0 \in I_0$ and $v_1 \in \mathcal{V} \backslash I_0$. Consider a cluster-based federated algorithm with clusters I_0 and $V \backslash I_0$. Then, any distributed online learning algorithm σ for full information feedback setting has an expected regret*

$$R_T^{v_1} \geq \frac{1 - o(1)}{4} \sqrt{\frac{\left(\tilde{\Delta} + 1\right)}{2} T \log K}$$

as $T \to \infty$.

Proof. Consider an online learning with expert advice problem with the action set $\mathcal{A}$ over B rounds (Cesa-Bianchi et al., 1997). Let $\ell'_1, \ldots, \ell'_B$ be an arbitrary sequence of losses and p'_b be the action selection distribution at round b. We show that σ can be used to design an algorithm for this online learning with expert advice problem, adapted from (Cesa-Bianchi et al., 2016).

Consider the loss sequences $\{\ell_t^v\}_{t=1}^T$ for each $v \in \mathcal{V}$ with $T = (\tilde{\Delta} + 1)B$ such that

$$\ell_t^v = \begin{cases} \ell'_{\lceil t/(\tilde{\Delta}+1) \rceil}, & v \in I_0 \\ 0 & \text{otherwise.} \end{cases}$$

Let p_t^v be the action select distribution of agent $v \in \mathcal{V}$ running the algorithm σ. Define the algorithm for the online learning with expert advice problem as follows:

$$p'_b = \frac{1}{\tilde{\Delta} + 1} \sum_{s=1}^{\tilde{\Delta}+1} p^{v_1}_{(\tilde{\Delta}+1)(b-1)+s}$$

where $p_t^v = (1/k, \dots, 1/k)$ for all $t \leq 1$ and $v \in \mathcal{V}$.

Note that p'_b is defined by $p^{v_1}_{(\tilde{\Delta}+1)(b-1)+1}, \dots, p^{v_1}_{(\tilde{\Delta}+1)b}$. These are in turn defined by $\ell_1^{v_0}, \dots, \ell^{v_0}_{(\tilde{\Delta}+1)(b-1)}$ by the definition of cluster-based communication protocol. Also note that $\lceil t/(\tilde{\Delta}+1)\rceil \leq b-1$ for $t \leq (\tilde{\Delta}+1)b$, hence p'_b is determined by $\ell'_1, \dots, \ell'_{b-1}$. Therefore $p'_1, \dots, p'_B$ are generated by a legitimate algorithm for online learning with expert advice problem.

Note that the cumulative regret of agent v_1 is

$$\begin{aligned}
\sum_{t=1}^{T} \langle p_t^{v_1}, \bar{\ell}_t \rangle &= \frac{1}{N} \sum_{t=1}^{T} \left[\sum_{v \in I_0} \langle p_t^{v_1}, \ell_t^v \rangle + \sum_{v \notin I_0} \langle p_t^{v_1}, \ell_t^v \rangle \right] \\
&= \frac{1}{4} \sum_{t=1}^{T} \langle p_t^{v_1}, \ell'_{\lceil t/(\tilde{\Delta}+1) \rceil} \rangle \\
&= \frac{1}{4} \sum_{b=1}^{B} \sum_{s=1}^{\tilde{\Delta}+1} \langle p^{v_1}_{(\tilde{\Delta}+1)(b-1)+s}, \ell'_b \rangle \\
&= \frac{\tilde{\Delta}+1}{4} \sum_{b=1}^{B} \langle p'_b, \ell'_b \rangle \qquad (5)
\end{aligned}$$

where the second equality comes from the definition of ℓ_t^v and the fourth equality comes from the definition of p'_b.

Also note that

$$\begin{aligned}
\min_{i \in \mathcal{A}} \sum_{t=1}^{T} \bar{\ell}_t(i) &= \frac{1}{4} \min_{i \in \mathcal{A}} \sum_{t=1}^{T} \ell'_{\lceil t/(\tilde{\Delta}+1) \rceil}(i) \\
&= \frac{\tilde{\Delta}+1}{4} \min_{i \in \mathcal{A}} \sum_{b=1}^{B} \ell'_b(i). \qquad (6)
\end{aligned}$$

From (5) and (6), it follows that

$$\sum_{t=1}^{T} \langle p_t^{v_1}, \bar{\ell}_t \rangle - \min_{i \in \mathcal{A}} \sum_{t=1}^{T} \bar{\ell}_t(i) = \frac{\tilde{\Delta}+1}{4} \left[\sum_{b=1}^{B} \langle p'_b, \ell'_b \rangle - \min_{i \in \mathcal{A}} \sum_{b=1}^{B} \ell'_b(i) \right].$$

There exists a sequence of losses $\ell'_1, \dots, \ell'_B$ such that for any algorithm for online learning with expert advice problem, the expected regret satisfies (Cesa-Bianchi & Lugosi, 2006, Theorem 3.7),

$$\sum_{b=1}^{B} \langle p'_b, \ell'_b \rangle - \min_{i \in \mathcal{A}} \sum_{b=1}^{B} \ell'_b(i) \geq (1 - o(1)) \sqrt{\frac{B}{2} \ln K}.$$

Hence, we have

$$\sum_{t=1}^{T} \langle p_t^{v_1}, \bar{\ell}_t \rangle - \min_{i \in \mathcal{A}} \sum_{t=1}^{T} \bar{\ell}_t(i) \geq \frac{1 - o(1)}{4} \sqrt{(\tilde{\Delta}+1) \frac{T}{2} \ln K}.$$

□

A.2. Proof of Theorem 3.2

The lower bound contains two parts. The first part is derived by using information-theoretic arguments in (Shamir, 2014) and it captures the effect of bandit feedback. The second part is inherited from the full-information feedback lower bound in Theorem 3.1 by the fact that the regret of an agent in the bandit feedback setting cannot be smaller than the regret in the full-information setting.

Consider a centralized federated algorithm with all the agents in the same cluster $\mathcal{V}$, denoted as Π^C. Note that by Lemma A.2, for a federated bandit algorithm Π^G,

$$\Pi^G \subset \Pi^C \quad \text{and} \quad \min_{\pi' \in \Pi^C} R_T^v(\pi', L) \leq \min_{\pi \in \Pi^G} R_T^v(\pi, L)$$

for any $L \in [0,1]^{T\times N\times K}$ and any $v \in \mathcal{V}$.

For any $\pi' \in \Pi^C$, at each round t, every agent $v \in \mathcal{V}$ receives $O(|\mathcal{N}(v)|)$ bits since its neighboring agents can choose at most $|\mathcal{N}(v)|$ distinct actions. By Theorem 4 in (Shamir, 2014), there exists some distribution $\mathcal{D}$ over $[0,1]^K$ such that loss vectors $\bar{\ell}_t \overset{i.i.d}{\sim} \mathcal{D}$ for all $t = 1, 2, \ldots, T$ and $\min_{i\in\mathcal{A}} \mathbb{E}\left[\sum_{t=1}^T \bar{\ell}_t(a_t(v)) - \sum_{t=1}^T \bar{\ell}_t(i)\right] = \Omega\left(\min\{T, \sqrt{KT/(1+|\mathcal{N}(v)|)}\}\right)$.

Hence, it follows that

$$\begin{aligned}
\min_L R_T^v(\pi, L) &\geq \min_L R_T^v(\pi', L) \\
&\geq \mathbb{E}_{\bar{\ell}_t\sim\mathcal{D}}\left[\sum_{t=1}^T \bar{\ell}_t(a_t(v)) - \min_{i\in\mathcal{A}}\sum_{t=1}^T \bar{\ell}_t(i)\right] \\
&\geq \max_{i\in\mathcal{A}} \mathbb{E}_{\bar{\ell}_t\sim\mathcal{D}}\left[\sum_{t=1}^T \bar{\ell}_t(a_t(v)) - \sum_{t=1}^T \bar{\ell}_t(i)\right] \\
&\geq \min_{i\in\mathcal{A}} \mathbb{E}_{\bar{\ell}_t\sim\mathcal{D}}\left[\sum_{t=1}^T \bar{\ell}_t(a_t(v)) - \sum_{t=1}^T \bar{\ell}_t(i)\right] \\
&= \Omega\left(\min\{T, \sqrt{KT/(1+|\mathcal{N}(v)|)}\}\right)
\end{aligned}$$

where the third inequality comes from Jensen's inequality.

Also note that any federated bandit algorithm for bandit feedback setting is also a federated bandit algorithm for full-information setting, from which it follows

$$\begin{aligned}
\min_L R_T^v(\pi, L) &\geq \max\left\{\Omega\left(\min\{T, \sqrt{KT/(1+|\mathcal{N}(v)|)}\}\right), \Omega\left(\sqrt[4]{\frac{1+d_{\max}}{\lambda_{N-1}(M)}}\sqrt{T\log K}\right)\right\} \\
&= \Omega\left(\min\left\{T, \max\left\{\sqrt{K/(1+|\mathcal{N}(v)|)}, \sqrt[4]{\frac{1+d_{\max}}{\lambda_{N-1}(M)}}\sqrt{\log K}\right\}\sqrt{T}\right\}\right).
\end{aligned}$$

A.3. Auxiliary lemmas

Here we present some auxiliary lemmas which are used in the proof of Theorem 4.1. Recall that $\hat{\ell}_t$ and $\bar{z}_t$ are the average instant loss estimator and average cumulative loss,

$$f_t = \frac{1}{N}\sum_{v\in\mathcal{V}} g_t^v \quad \text{and} \quad \bar{z}_t = \frac{1}{N}\sum_{v\in\mathcal{V}} z_t^v$$

and y_t is action distribution to minimize the regularized average cumulative loss

$$y_t(i) = \frac{\exp\{-\eta_{t-1}\bar{z}_t(i)\}}{\sum_{j\in A}\exp\{-\eta_{t-1}\bar{z}_t(j)\}}.$$

Lemma A.5. *For each time step* $t = 1, \ldots, T$,

$$\bar{z}_{t+1} = \bar{z}_t + f_t$$

and

$$\max\{\|g_t^v\|_*, \|f_t\|_*\} \leq \frac{K}{\gamma_t}.$$

Proof.

$$\begin{aligned} \bar{z}_{t+1} &= \frac{1}{N}\sum_{v\in\mathcal{V}} z^v_{t+1} \\ &= \frac{1}{N}\sum_{v\in\mathcal{V}}\sum_{u:(u,v)\in\mathcal{E}} W_{u,v} z^u_t + \frac{1}{N}\sum_{v\in\mathcal{V}} g^v_t \\ &= \frac{1}{N}\sum_{v\in\mathcal{V}} z^v_t + \frac{1}{N}\sum_{v\in\mathcal{V}} g^v_t \\ &= \bar{z}_t + f_t \end{aligned}$$

where the second equality comes from Line 7 in Algorithm 1 and the third equality comes from the double-stochasticity of W.

Noting that $p^v_t(i) \geq \gamma/K$ for all $v \in \mathcal{V}$, $i \in \mathcal{A}$ and $t \in \{1, \ldots, T\}$, it follows that

$$\|g^v_t\|_* = \frac{\ell^v_t(a^v_t)}{p^v_t(a^v_t)} \leq \frac{K}{\gamma_t} \quad \text{and} \quad \|f_t\|_* \leq \frac{1}{N}\sum_{v\in\mathcal{V}} \|g^v_t\|_* \leq \frac{K}{\gamma_t}.$$

□

Lemma A.6. *For any $v \in \mathcal{V}$ and $t \geq 1$, it holds that*

$$\mathbb{E}\left[g^v_t \mid \mathcal{F}_{t-1}\right] = \ell^v_t \textit{ and } \mathbb{E}\left[f_t \mid \mathcal{F}_{t-1}\right] = \bar{\ell}_t$$

with

$$\mathbb{E}\left[\|f_t\|_*\right] \leq K \textit{ and } \mathbb{E}\left[\|f_t\|^2_*\right] \leq \frac{K^2}{\gamma_t}.$$

Proof. Note that p^v_t is determined by $\mathcal{F}_{t-1}$, hence

$$\mathbb{E}\left[g^v_t(i) \mid \mathcal{F}_{t-1}\right] = \frac{\ell^v_t(i)}{p^v_t(i)}\mathbb{E}\left[\mathbb{I}\left\{a^v_t = i\right\} \mid \mathcal{F}_{t-1}\right] = \frac{\ell^v_t(i)}{p^v_t(i)} p^v_t(i) = \ell^v_t(i)$$

and

$$\mathbb{E}\left[\|g^v_t\|_*\right] = \mathbb{E}\left[\frac{\ell^v_t(a^v_t)}{p^v_t(a^v_t)}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{\ell^v_t(a^v_t)}{p^v_t(a^v_t)} \mid \mathcal{F}_{t-1}\right]\right] = \mathbb{E}\left[\sum_{i\in\mathcal{A}} p^v_t(i)\frac{\ell^v_t(i)}{p^v_t(i)}\right] = \sum_{i\in\mathcal{A}} \ell^v_t(i) \leq K$$

where the last inequality comes from $\ell^v_t(i) \leq 1$. Since $f_t(i) = \frac{1}{N}\sum_{v\in\mathcal{V}} g^v_t$, it follows that

$$\mathbb{E}\left[f_t(i) \mid \mathcal{F}_{t-1}\right] = \frac{1}{N}\sum_{v\in\mathcal{V}} \ell^v_t(i) = \bar{\ell}_t(i)$$

and

$$\mathbb{E}\left[\|f_t\|_*\right] \leq \frac{1}{N}\sum_{v\in\mathcal{V}} \mathbb{E}\left[\|g^v_t\|_*\right] \leq K$$

which comes from Jensen's inequality. Notice that

$$\mathbb{E}\left[\|g^v_t\|^2_*\right] = \mathbb{E}\left[\frac{\ell^v_t(a^v_t)^2}{p^v_t(a^v_t)^2}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{\ell^v_t(a^v_t)^2}{p^v_t(a^v_t)^2} \mid \mathcal{F}_{t-1}\right]\right] = \mathbb{E}\left[\sum_{i\in\mathcal{A}} p^v_t(i)\frac{\ell^v_t(i)^2}{p^v_t(i)^2}\right] \leq \mathbb{E}\left[\sum_{i\in\mathcal{A}} \frac{1}{p^v_t(i)}\right] \leq \frac{K^2}{\gamma_t}$$

where the last inequality comes from $p^v_t(i) \geq \gamma_t/K$. Again, from Jensen's inequality, it follows

$$\mathbb{E}\left[\|f_t\|^2_*\right] \leq \frac{1}{N}\sum_{v\in\mathcal{V}} \mathbb{E}\left[\|g^v_t\|^2_*\right] \leq \frac{K^2}{\gamma_t}.$$

□

Before presenting the next lemma, we recall the definition of strongly-convex functions and Fenchel duality. A function ϕ is said to be α-strongly convex function on a convex set $\mathcal{X}$ if

$$\phi(x') \geq \phi(x) + \langle \nabla \phi(x), x' - x \rangle + \frac{1}{2}\alpha \|x' - x\|^2$$

for all $x', x \in \mathcal{X}$, for some $\alpha \geq 0$.

Let ϕ^* denote the *Fenchel conjugate* of ϕ, i.e.,

$$\phi^*(y) = \max_{x \in \mathcal{X}} \{\langle x, y \rangle - \phi(x)\}$$

with the projection,

$$\nabla \phi^*(y) = \arg\max_{x \in \mathcal{X}} \{\langle x, y \rangle - \phi(x)\}.$$

Lemma A.7. *Let ψ the normalized negative entropy function (Lattimore & Szepesvári, 2020) on $\mathcal{P}_{K-1} = \{x \in [0,1]^K : \sum_{i=1}^K x(i) = 1\}$,*

$$\psi_\eta(x) = \frac{1}{\eta} \sum_{i=1}^{k} x(i) \left(\log(x(i)) - 1\right).$$

For all $t = 1, \ldots, T$, it holds that

$$x_t^v = \operatorname*{argmin}_{x \in \mathcal{P}_{K-1}} \{\langle x, z_t^v \rangle + \psi_{\eta_t}(x)\} = \nabla \psi^*_{\eta_{t-1}}(-z_t^v)$$

with $\mathcal{X} = \mathcal{P}_{K-1}$ and

$$y_t = \operatorname*{argmin}_{x \in \mathcal{P}_{K-1}} \{\langle x, \bar{z}_t \rangle + \psi_{\eta_{t-1}}(x)\} = \nabla \psi^*_{\eta_{t-1}}(-\bar{z}_t).$$

Furthermore, it holds

$$\|x_t^v - y_t\| \leq \eta_{t-1} \|z_t^v - \bar{z}_t\|_*.$$

Proof. We prove for $y_t = \operatorname*{argmin}_{x \in \mathcal{P}_{K-1}} \{\langle x, \bar{z}_t \rangle + \psi_{\eta_{t-1}}(x)\}$ whose argument also applies to x_t^v.

Notice it suffices to consider the minimization problem

$$\begin{array}{cc} \min\limits_{x \in \mathcal{P}_{K-1}} & \eta_{t-1} \sum_{k=1}^K x(i) \bar{z}_t(i) + \sum_{i=1}^k x(i) \log(x(i)) \\ \text{subject to} & \sum_{k=1}^K x(i) = 1. \end{array}$$

It suffices to consider the Lagrangian,

$$\mathcal{L} = -\eta_{t-1} \sum_{k=1}^K x(i) \bar{z}_t(i) - \sum_{i=1}^k x(i) \log(x(i)) - \lambda \left(\sum_{k=1}^K x(i) - 1 \right).$$

Consider the first-order conditions for all $i = 1, \ldots, K$

$$\frac{\partial \mathcal{L}}{\partial x(i)} = -\eta_{t-1} \bar{z}_t(i) - \log(x(i)) - 1 - \lambda = 0$$

which gives $x(i) = \exp\{-\eta_{t-1} \bar{z}_t(i)\} / \exp\{1 + \lambda\}$ for all $i = 1, \ldots, K$. Plugging into the constraint $\sum_{k=1}^K x(i) = 1$ together with the definition of Fenchel duality (Hiriart-Urruty & Lemarechal, 2010) completes the proof for y_t.

Note that the normalized negative entropy $\psi(x)$ is 1-strongly convex,

$$\psi(x') \geq \psi(x) + \langle \nabla \psi(x), x' - x \rangle + \frac{1}{2} \|x' - x\|^2.$$

Multiplying $1/\eta_{t-1}$ both sides of the inequality yields that $\psi_{\eta_{t-1}}(x)$ is $1/\eta_{t-1}$-strongly convex. By Theorem 4.2.1 in (Hiriart-Urruty & Lemarechal, 2010), we have that $\nabla \psi^*_{\eta_{t-1}}(z)$ is η_{t-1}-Lipschitz.

It follows that

$$\|p_t^v - \bar{p}_t\| = \|\nabla \psi^*_\eta(-z_t^v) - \nabla \psi^*_\eta(-\bar{z}_t)\| \leq \eta_{t-1} \|\bar{z}_t - z_t^v\|_*.$$

□

We state an upper bound on the network disagreement on the cumulative loss estimators from (Duchi et al., 2011) and (Hosseini et al., 2013).

Lemma A.8. *For any $v \in \mathcal{V}$ and $t = 1, 2, \ldots, T$,*

$$\|\bar{z}_t - z_t^v\|_* \leq \frac{K}{\gamma_T}\left(\frac{\min\{2\log T + \log n, \sqrt{n}\}}{1 - \sigma_2(W)} + 3\right) = \frac{K}{\gamma_T} C_W$$

where $\sigma_2(W)$ is the second largest singular value of W.

Proof. From Lemma A.5, it follows that $\|g_t^v\|_* \leq K/\gamma_t$. Since $\{\gamma_t\}$ is non-increasing, let $L = K/\gamma_T$ in Eq. (29) in (Duchi et al., 2011) and Lemma 6 in (Hosseini et al., 2013) completes the proof. □

A.4. Proof of Theorem 4.1

Let $i^* = \arg\min_{i \in \mathcal{A}} \sum_{t=1}^{T} \bar{\ell}_t(i)$. Note that p_t^v is determined by $\mathcal{F}_{t-1}$ and $\mathbb{E}\left[f_t \mid \mathcal{F}_{t-1}\right] = \bar{\ell}_t$ from Lemma A.6. It follows that for each agent $v \in \mathcal{V}$

$$\begin{aligned} R_T^v &= \mathbb{E}\left[\sum_{t=1}^{T} \langle \bar{\ell}_t, p_t^v \rangle - \sum_{t=1}^{T} \bar{\ell}_t(i^*)\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{T} \langle \mathbb{E}\left[f_t \mid \mathcal{F}_{t-1}\right], p_t^v \rangle - \sum_{t=1}^{T} \mathbb{E}\left[f_t(i^*) \mid \mathcal{F}_{t-1}\right]\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{T} \langle f_t, p_t^v \rangle - \sum_{t=1}^{T} f_t(i^*)\right]. \end{aligned} \tag{7}$$

By the definition of p_t^v, it follows

$$\begin{aligned} R_T^v &= \mathbb{E}\left[\sum_{t=1}^{T} \left(\langle f_t, (1-\gamma)x_t^v + \gamma x_1^v \rangle - f_t(i^*)\right)\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{T} (1-\gamma_t)\left(\langle f_t, x_t^v \rangle - f_t(i^*)\right)\right] + \sum_{t=1}^{T} \gamma_t \mathbb{E}\left[\left(\langle f_t, x_1^v \rangle - f_t(i^*)\right)\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{T} (1-\gamma_t)\left(\langle f_t, x_t^v \rangle - f_t(i^*)\right)\right] + \sum_{t=1}^{T} \gamma_t \left(\langle \bar{\ell}_t, x_1^v \rangle - \bar{\ell}_t(i^*)\right) \\ &\leq \mathbb{E}\left[\sum_{t=1}^{T} \left(\langle f_t, x_t^v \rangle - f_t(i^*)\right)\right] + \sum_{t=1}^{T} \gamma_t \\ &= \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \left(\langle f_t, y_t^v \rangle - f_t(i^*)\right)\right]}_{\text{(I)}} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \langle f_t, x_t^v - y_t^v \rangle\right]}_{\text{(II)}} + \sum_{t=1}^{T} \gamma_t \end{aligned}$$

where the first inequality comes from the fact that $\gamma_t > 0$ and the fact that $\|\bar{\ell}_t\|_* \leq \sum_{v \in \mathcal{V}} 1/N \|\bar{\ell}_t^v\|_* \leq 1$.

From Lemma A.5, it follows $\bar{z}_t = \sum_{s=1}^{t-1} f_s$. Hence, it follows from Lemma A.7 that

$$y_t = \arg\min_{x \in \mathcal{P}_{K-1}} \left\{\sum_{s=1}^{t} \langle f_s, x \rangle + \frac{1}{\eta_{t-1}} \psi(x)\right\}.$$

From Lemma 3 in (Duchi et al., 2011) and Corollary 28.8 in (Lattimore & Szepesvári, 2020), we have

$$\text{(I)} \leq \frac{1}{2} \sum_{t=1}^{T} \eta_{t-1} \mathbb{E}\left[\|f_t\|_*^2\right] + \frac{1}{\eta_T} \log(K) \tag{8}$$

which is because $\{\eta_t\}$ is a non-increasing sequence. Note that

$$\|x_t^v - y_t\| \leq \eta_{t-1}\|z_t^v - \bar{z}_t\|_*$$

by Lemma A.7. This yields that

$$(\mathrm{II}) \leq \sum_{t=1}^{T} \eta_{t-1}\mathbb{E}\left[\|f_t\|_*\|z_t^v - \bar{z}_t\|_*\right]. \tag{9}$$

Plugging Equations (8) and (9) into (I) yields that

$$\begin{aligned} R_T^v &\leq \frac{1}{2}\sum_{t=1}^{T}\eta_{t-1}\mathbb{E}\left[\|f_t\|_*^2\right] + \frac{1}{\eta_T}\log(K) + \sum_{t=1}^{T}\eta_{t-1}\mathbb{E}\left[\|f_t\|_*\|z_t^v - \bar{z}_t\|_*\right] + \sum_{t=1}^{T}\gamma_t \\ &\leq \frac{1}{2}\sum_{t=1}^{T}\eta_{t-1}\mathbb{E}\left[\|f_t\|_*^2\right] + \frac{K}{\gamma_T}C_W\sum_{t=1}^{T}\eta_{t-1}\mathbb{E}\left[\|f_t\|_*\right] + \sum_{t=1}^{T}\gamma_t + \frac{1}{\eta_T}\log(K) \\ &\leq \frac{K^2}{2}\sum_{t=1}^{T}\frac{\eta_{t-1}}{\gamma_t} + \frac{K^2}{\gamma_T}C_W\sum_{t=1}^{T}\eta_{t-1} + \sum_{t=1}^{T}\gamma_t + \frac{1}{\eta_T}\log(K) \end{aligned}$$

where the second inequality comes from Lemma A.8 and the third inequality comes from Lemma A.6.

Let

$$\gamma_t = \sqrt[3]{\frac{\left(C_W + \frac{1}{2}\right)K^2\log K}{t}} \quad \text{and} \quad \eta_t = \frac{\log K}{T\gamma_T} = \sqrt[3]{\frac{(\log K)^2}{\left(C_W + \frac{1}{2}\right)K^2T^2}}.$$

Then, for every $v \in \mathcal{V}$, we have

$$\begin{aligned} R_T^v &\leq \frac{3}{8}\sqrt[3]{\frac{K^2\log K}{\left(C_W + \frac{1}{2}\right)^2}}T^{\frac{2}{3}} + \sqrt[3]{K^2\log K\frac{C_W^3}{\left(C_W + \frac{1}{2}\right)^2}}T^{\frac{2}{3}} \\ &+ \frac{3}{2}\sqrt[3]{\left(C_W + \frac{1}{2}\right)K^2\log K}T^{\frac{2}{3}} + \sqrt[3]{\left(C_W + \frac{1}{2}\right)K^2\log K}T^{\frac{2}{3}} \\ &\leq \frac{3}{4}\sqrt[3]{K^2\log K}T^{\frac{2}{3}} + \sqrt[3]{C_WK^2\log K}T^{\frac{2}{3}} + \frac{5\sqrt[3]{2}}{2}\sqrt[3]{C_WK^2\log K}T^{\frac{2}{3}} \\ &\leq 5\sqrt[3]{C_WK^2\log K}T^{\frac{2}{3}}. \end{aligned}$$

A.5. Numerical experiments

All the experiments are run on a desktop with AMD Ryzen 5 2600 Six-Core Processor and 16GB memory. Each experiment took less than 6 hours to finish.

The code is written in Python and uses Numpy package (Harris et al., 2020) and NetworkX package (Hagberg et al., 2008) for numerical calculation and graph operations. The Numpy package and the NetworkX package are distributed with the BSD license.