A Simple and Provably Efficient Algorithm for Asynchronous Federated Contextual Linear Bandits

Jiafan He*
Department of Computer Science
University of California, Los Angeles
Los Angeles, CA 90095
jiafanhe19@ucla.edu

Tianhao Wang*
Department of Statistics and Data Science
Yale University
New Haven, CT 06511
tianhao.wang@yale.edu

Yifei Min*
Department of Statistics and Data Science
Yale University
New Haven, CT 06511
yifei.min@yale.edu

Quanquan Gu
Department of Computer Science
University of California, Los Angeles
Los Angeles, CA 90095
qgu@cs.ucla.edu

Abstract

We study federated contextual linear bandits, where M agents cooperate with each other to solve a global contextual linear bandit problem with the help of a central server. We consider the asynchronous setting, where all agents work independently and the communication between one agent and the server will not trigger other agents' communication. We propose a simple algorithm named `FedLinUCB` based on the principle of optimism. We prove that the regret of `FedLinUCB` is bounded by $\widetilde{O}(d\sqrt{\sum_{m=1}^{M} T_m})$ and the communication complexity is $\widetilde{O}(dM^2)$, where d is the dimension of the contextual vector and T_m is the total number of interactions with the environment by m-th agent. To the best of our knowledge, this is the first provably efficient algorithm that allows fully asynchronous communication for federated contextual linear bandits, while achieving the same regret guarantee as in the single-agent setting.

1 Introduction

Contextual linear bandit is a canonical model in sequential decision making with partial information feedback that has found vast applications in real-world domains such as recommendation systems (Li et al., 2010a,b; Gentile et al., 2014; Li et al., 2020), clinical trials (Wang, 1991; Durand et al., 2018) and economics (Jagadeesan et al., 2021; Li et al., 2022). Most existing works on contextual linear bandits focus on either the single-agent setting (Auer, 2002; Abe et al., 2003; Dani et al., 2008; Li et al., 2010a; Rusmevichientong and Tsitsiklis, 2010; Chu et al., 2011; Abbasi-Yadkori et al., 2011; Agrawal and Goyal, 2013) or multi-agent settings where communications between agents are instant and unrestricted (Cesa-Bianchi et al., 2013; Li et al., 2016; Wu et al., 2016; Li et al., 2021). Due to the increasing amount of data being distributed across a large number of local agents (e.g., clients, users, edge devices), federated learning (McMahan et al., 2017; Karimireddy et al., 2020) has become an emerging paradigm for distributed machine learning, where agents can jointly learn a global model without sharing their own localized data. This motivates the development of distributed/federated linear bandits (Wang et al., 2019; Huang et al., 2021; Li and Wang, 2022a), which enables a collection of agents to cooperate with each other to solve a global linear bandit

*Equal contribution.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

problem while enjoying performance guarantees comparable to those in the classical single-agent centralized setting.

However, most existing federated linear bandits algorithms are limited to the synchronous setting (Wang et al., 2019; Dubey and Pentland, 2020; Huang et al., 2021), where all the agents have to first upload their local data to the server upon the request of the server, and the agents will download the latest data from the server after all uploads are complete. This requires full participation of the agents and global synchronization mandated by the server, which is impractical in many real-world application scenarios. The only notable exception is Li and Wang (2022a), where an asynchronous federated linear bandit algorithm is proposed. Nevertheless, in their algorithm, the upload by one agent may trigger the download from the server to all other agents. Therefore, the communications between different agents and the server are not totally independent. Moreover, they make a stringent regularity assumption on the contexts, which basically requires the contexts to be stochastic rather than adversarial as in standard contextual linear bandits. That said, their theoretical proof is actually flawed as they ignored some unique challenges caused by asynchronous communication (see Appendix A.1 for details). Therefore, how to design a truly asynchronous contextual linear bandit algorithm remains an open problem.

In this work, we resolve the above open problem by proposing a simple algorithm for asynchronous federated contextual linear bandits over a star-shaped communication network. Our algorithm is based on the principle of optimism (Abbasi-Yadkori et al., 2011) and enjoys the following advantages: (i) Each agent can decide whether or not to participate in each round. Full participation is not required, thus it allows temporarily offline agents. This is much more flexible than existing algorithms for federated linear bandits in Wang et al. (2019); Dubey and Pentland (2020); Huang et al. (2021) where all agents are required to participate in each round; and (ii) the communication between each agent and the server is asynchronous and totally independent of other agents. There is no need of global synchronization or mandatory coordination by the server, in contrast to Li and Wang (2022a) where each agent might be asked by the server to download data. In particular, the communication between the agent and the server is triggered by a matrix determinant-based criterion that can be computed independently by each agent. Our algorithm design not only allows the agents to independently operate and synchronize with the server, but also ensures low communication complexity (i.e., total number of rounds of communication between agents and the server) and low switching cost (i.e., total number of local model updates for all agents) (Abbasi-Yadkori et al., 2011).

While being simple, our algorithm design introduces a challenge in the regret analysis. Since the order of the interaction between the agent and the environment is not fixed, standard martingale-based concentration inequality cannot be directly applied. Specifically, this challenge arises due to the mismatch between the partial data information collected by the central server and the true order of the data generated from the interaction with the environment, as is explained in detail in Section 5 and illustrated by Figure 1. We address this challenge by a novel proof technique, which first establishes the local concentration of each agent's data and then relates it to the "virtual" global concentration of all data via the determinant-based criterion. Based on this proof technique, we are able to obtain tight enough confidence bounds that lead to a nearly optimal regret. Moreover, our theoretical analysis relies only on minimal assumptions that are standard for contextual linear bandits, relaxing the strong assumptions made in Li and Wang (2022a).

Main contributions. Our contributions are highlighted as follows:

- We devise a simple algorithm named `FedLinUCB` that achieves near-optimal regret, low communication complexity and low switching cost simultaneously for asynchronous federated contextual linear bandits. In detail, we prove that our algorithms achieves a near-optimal $\widetilde{O}(d\sqrt{T})$ regret with merely $\widetilde{O}(dM^2)$ total communication complexity and $\widetilde{O}(dM^2)$ total switching cost. Here M is the number of agents, d is the dimension of the context and $T = \sum_{m=1}^{M} T_m$ is the total number of rounds with T_m being the number of rounds that agent m participates in. When degenerated to single-agent bandits, the regret of our algorithm matches the optimal regret $\widetilde{O}(d\sqrt{T})$ (Abbasi-Yadkori et al., 2011).

- We also prove an $\Omega(M/\log(T/M))$ lower bound for the communication complexity. Together with the $O(dM^2)$ upper bound of our algorithm, it suggests that there is only an $\widetilde{O}(dM)$ gap between the upper and lower bounds of the communication complexity.

- We identify the issue of ill-defined filtration caused by the unfixed order of interactions between agents and the environment, which is absent in previous synchronous or single-agent settings. We tackle this unique challenge by connecting the local concentration of each local agent's data and the global concentration of the aggregated data from all agents. We believe this proof technique is of independent interest for the analysis of other asynchronous bandit problems.

Notation. For any positive integer n, we denote the set $\{1, 2, \ldots, n\}$ by $[n]$. We use $\mathbf{I}$ to denote the $d \times d$ identity matrix. We use O to hide universal constants and $\widetilde{O}$ to further hide poly-logarithmic terms. For any vector $\mathbf{x} \in \mathbb{R}^d$ and positive semi-definite $\mathbf{\Sigma} \in \mathbb{R}^{d \times d}$, we denote $\|\mathbf{x}\|_{\mathbf{\Sigma}} = \sqrt{\mathbf{x}^\top \mathbf{\Sigma} \mathbf{x}}$.

2 Related Work

We review related work on distributed/federated bandit algorithms stratified by the type of bandits: (1) multi-armed bandits, (2) stochastic linear bandits and (3) contextual linear bandits.

Distributed/federated multi-armed bandits. There is a vast literature on distributed/federated multi-armed bandits (MABs) (Liu and Zhao, 2010; Szorenyi et al., 2013; Landgren et al., 2016; Chakraborty et al., 2017; Landgren et al., 2018; Martínez-Rubio et al., 2019; Sankararaman et al., 2019; Wang et al., 2019, 2020; Zhu et al., 2021), to mention a few. However, none of these algorithms can be directly applied to linear bandits, needless to say contextual linear bandits with infinite decision sets.

Distributed/federated stochastic linear bandits. In distributed/federated stochastic linear bandits, the decision set is fixed across all the rounds $t \in [T]$ and all the agents $m \in [M]$. Wang et al. (2019) proposed the `DELD` algorithm for distributed stochastic linear bandits on both star-shaped network and P2P network. Huang et al. (2021) proposed an arm elimination-based algorithm called `Fed-PE` for federated stochastic linear bandits on the star-shaped network. Both algorithms are in the synchronous setting and require full participation of the agents upon the server's request.

Distributed/federated contextual linear bandits. The contextual linear bandit is more general and challenging than stochastic linear bandits, because the decision sets can vary for each t and m. In this setting, Korda et al. (2016) considered a P2P network and proposed the `DCB` algorithm based on the `OFUL` algorithm in Abbasi-Yadkori et al. (2011). Wang et al. (2019) considered both star-shaped and P2P communication networks and achieved the near-optimal $\widetilde{O}(d\sqrt{T})$ regret in the synchronous setting.[2] Dubey and Pentland (2020) further introduced the differential privacy guarantee into the setting of Wang et al. (2019). Li and Wang (2022b) extended distributed contextual linear bandits to generalized linear bandits (Filippi et al., 2010; Jun et al., 2017) in the synchronous setting. Li and Wang (2022a) proposed the first asynchronous algorithm for federated contextual linear bandits with the star-shaped graph and achieved a near-optimal $\widetilde{O}(d\sqrt{T})$ regret. However, their setting is different from ours in two aspects: (1) the upload triggered by an agent will lead the server to trigger download possibly for all the agents in their setting. In contrast, the upload triggered by an agent will only lead to download to the same agent in our setting; (2) their regret guarantee relies on a stringent regularity assumption on the contexts, which basically requires the contexts to be stochastic. As a comparison, the contexts in our setting can be even adversarial, which is exactly the standard setting of contextual linear bandits (Abbasi-Yadkori et al., 2011; Li et al., 2019). This difference in the setting makes our algorithm a truly asynchronous contextual linear bandit algorithm but also makes our regret analysis more challenging.

For better comparison, we compare our work with the most related contextual linear bandit algorithms in Table 1.

3 Preliminaries

Federated contextual linear bandits. We consider the federated contextual linear bandits as follows: At each round $t \in [T]$, an arbitrary agent $m_t \in [M]$ is active for participation. This agent receives a decision set $D_t \subset \mathbb{R}^d$, picks an action $\mathbf{x}_t \in D_t$, and receives a random reward r_t. We assume that

[2] In the original paper of Wang et al. (2019), the regret bound is expressed as $\widetilde{O}(d\sqrt{MT})$. The T in their paper is equivalent to the T_m in ours, so their $d\sqrt{MT}$ should be understood as $d\sqrt{T}$ under our notation.

Setting	Algorithm	Regret	Communication	Low-switching	No extra assump. on contexts	Allow free participation
Single-agent	`OFUL` (Abbasi-Yadkori et al., 2011)	$d\sqrt{T\log T}$	N/A	✓	✓	✘
Federated (Sync.)	`DisLinUCB` (Wang et al., 2019)	$d\sqrt{T}\log^2 T$	$dM^{3/2}$	✘	✓	✘
Federated (Async.)	`Async-LinUCB` (Li and Wang, 2022a)	$d\sqrt{T}\log T$	$dM^2\log T$	✘	✘	✘
Federated (Async.)	`FedLinUCB` (Our Algorithm 1)	$d\sqrt{T}\log T$	$dM^2\log T$	✓	✓	✓

Table 1: Comparison of our result with baseline approaches for contextual linear bandits. Our result achieves near-optimal regret under low communication complexity. Here d is the dimension of the context, M is the number of agents, and $T = \sum_{m=1}^{M} T_m$ with each T_m being the number of rounds that agent m participates in.

the reward r_t satisfies $r_t = \langle \mathbf{x}_t, \boldsymbol{\theta}^* \rangle + \eta_t$ for all $t \in [T]$, where η_t is conditionally independent of $\mathbf{x}_t$ given $\mathbf{x}_{1:t-1}, m_{1:t}, r_{1:t-1}$. More specifically, we make the following assumption on η_t, $\boldsymbol{\theta}^*$ and D_t, which is a standard assumption in the contextual linear bandit literature (Abbasi-Yadkori et al., 2011; Wang et al., 2019; Dubey and Pentland, 2020).

Assumption 3.1. The noise η_t is R-sub-Gaussian conditioning on $\mathbf{x}_{1:t}$, $m_{1:t}$ and $r_{1:t-1}$, i.e.,

$$\mathbb{E}\big[e^{\lambda\eta_t} \mid \mathbf{x}_{1:t}, m_{1:t}, r_{1:t-1}\big] \le \exp(R^2\lambda^2/2), \quad \text{for any } \lambda \in \mathbb{R}.$$

We also assume that $\|\boldsymbol{\theta}^*\|_2 \le S$ and $\|\mathbf{x}\|_2 \le L$ for all action $\mathbf{x} \in \mathcal{D}_t$, for all $t \in [T]$.

Notably, we assume m_t can be arbitrary for all t, which basically says that each agent can decide whether and when to participate or not.[3] Our setting is more general than the synchronous setting in Wang et al. (2019); Dubey and Pentland (2020), which requires a round-robin participation of all agents.

Learning objective. The goal of the agents is to collaboratively minimize the cumulative regret defined as

$$\mathrm{Regret}(T) := \sum_{t=1}^{T} \Big(\max_{\mathbf{x} \in D_t} \langle \mathbf{x}, \boldsymbol{\theta}^* \rangle - \langle \mathbf{x}_t, \boldsymbol{\theta}^* \rangle \Big) = \sum_{t=1}^{T} \langle \mathbf{x}_t^* - \mathbf{x}_t, \boldsymbol{\theta}^* \rangle. \tag{3.1}$$

To achieve such a goal, we allow the agents to collaborate via communication through the central server. Below we will explain the details of the communication model.

Communication model. We consider a star-shaped communication network (Wang et al., 2019; Dubey and Pentland, 2020) consisting of a central server and M agents, where each agent can communicate with the server by uploading and downloading data. However, any pair of agents cannot directly communicate with each other. We define the communication complexity as the total number of communication rounds between agents and the server (counting both the uploads and downloads) (Wang et al., 2019; Dubey and Pentland, 2020; Li and Wang, 2022a). For simplicity, we assume that there is no latency in the communication channel.

We consider the asynchronous setting, where the communication protocol satisfies: (1) each agent can decide whether or not to participate in each round. Full participation is not required, which allows temporarily offline agents; and (2) the communication between each agent and the server is asynchronous and independent of other agents without mandatory download required by the server.

Switching cost. The notion of switching cost in online learning and bandits refers to the number of times the agent switches its policy (i.e., decision rule) (Kalai and Vempala, 2005; Abbasi-Yadkori et al., 2011; Dekel et al., 2014; Ruan et al., 2021). In the context of linear bandits, it corresponds to the number of times the agent updates its policy of selecting an action from the decision set (Abbasi-Yadkori et al., 2011). Algorithms with low switching cost are preferred in practice since each policy switching might cause additional computational overhead.

[3] Without loss of generality, we can assume that it cannot happen that more than one agent participate at the same time. Therefore, there is always a valid order of participation indexed by $t \in [T]$.

Algorithm 1 Federated linear UCB (`FedlinUCB`)

1: Initialize $\mathbf{\Sigma}_{m,1} = \mathbf{\Sigma}_1^{\text{ser}} = \lambda \mathbf{I}$, $\widehat{\boldsymbol{\theta}}_{m,1} = 0$, $\mathbf{b}_{m,0}^{\text{loc}} = 0$ and $\mathbf{\Sigma}_{m,0}^{\text{loc}} = 0$ for all $m \in [M]$
2: **for** round $t = 1, 2, \ldots, T$ **do**
3: Agent m_t is active
4: Receive D_t from the environment
5: Select $\mathbf{x}_t \leftarrow \operatorname{argmax}_{\mathbf{x} \in D_t} \langle \widehat{\boldsymbol{\theta}}_{m_t,t}, \mathbf{x} \rangle + \beta \|\mathbf{x}\|_{\mathbf{\Sigma}_{m_t,t}^{-1}}$ `/* Optimistic decision */`
6: Receive r_t from environment
7: $\mathbf{\Sigma}_{m_t,t}^{\text{loc}} \leftarrow \mathbf{\Sigma}_{m_t,t-1}^{\text{loc}} + \mathbf{x}_t \mathbf{x}_t^\top, \quad \mathbf{b}_{m_t,t}^{\text{loc}} \leftarrow \mathbf{b}_{m_t,t-1}^{\text{loc}} + r_t \mathbf{x}_t$ `/* Local update */`
8: **if** $\det(\mathbf{\Sigma}_{m_t,t} + \mathbf{\Sigma}_{m_t,t}^{\text{loc}}) > (1+\alpha) \det(\mathbf{\Sigma}_{m_t,t})$ **then**
9: Agent m_t sends $\mathbf{\Sigma}_{m_t,t}^{\text{loc}}$ and $\mathbf{b}_{m_t,t}^{\text{loc}}$ to server `/* Upload */`
10: $\mathbf{\Sigma}_t^{\text{ser}} \leftarrow \mathbf{\Sigma}_t^{\text{ser}} + \mathbf{\Sigma}_{m_t,t}^{\text{loc}}, \quad \mathbf{b}_t^{\text{ser}} \leftarrow \mathbf{b}_t^{\text{ser}} + \mathbf{b}_{m_t,t}^{\text{loc}}$ `/* Global update */`
11: $\mathbf{\Sigma}_{m_t,t}^{\text{loc}} \leftarrow 0, \quad \mathbf{b}_{m_t,t}^{\text{loc}} \leftarrow 0$
12: Server sends $\mathbf{\Sigma}_t^{\text{ser}}$ and $\mathbf{b}_t^{\text{ser}}$ back to agent m_t `/* Download */`
13: $\mathbf{\Sigma}_{m_t,t+1} \leftarrow \mathbf{\Sigma}_t^{\text{ser}}, \quad \mathbf{b}_{m_t,t+1} \leftarrow \mathbf{b}_t^{\text{ser}}$
14: $\widehat{\boldsymbol{\theta}}_{m_t,t+1} \leftarrow \mathbf{\Sigma}_{m_t,t+1}^{-1} \mathbf{b}_{m_t,t+1}$ `/* Compute estimate */`
15: **else**
16: $\mathbf{\Sigma}_{m_t,t+1} \leftarrow \mathbf{\Sigma}_{m,t}, \quad \mathbf{b}_{m_t,t+1} \leftarrow \mathbf{b}_{m,t}, \quad \widehat{\boldsymbol{\theta}}_{m_t,t+1} \leftarrow \widehat{\boldsymbol{\theta}}_{m_t,t}$
17: **end if**
18: **for** other inactive agent $m \in [M] \setminus \{m_t\}$ **do**
19: $\mathbf{\Sigma}_{m,t+1} \leftarrow \mathbf{\Sigma}_{m,t}, \quad \mathbf{b}_{m,t+1} \leftarrow \mathbf{b}_{m,t}, \quad \widehat{\boldsymbol{\theta}}_{m,t+1} \leftarrow \widehat{\boldsymbol{\theta}}_{m,t}$
20: **end for**
21: **end for**

4 The Proposed Algorithm

We propose a simple algorithm based on the principle of optimism that enables collaboration among agents through asynchronous communications with the central server. The main algorithm is displayed in Algorithm 1. For clarity, we first summarize the related notations in Table 2.

Notation	Meaning
$\widehat{\boldsymbol{\theta}}_{m,t}$	estimate of $\boldsymbol{\theta}^*$
$\mathbf{\Sigma}_{m,t}, \mathbf{b}_{m,t}$	data used to compute $\widehat{\boldsymbol{\theta}}_{m,t}$
$\mathbf{\Sigma}_{m,t}^{\text{loc}}, \mathbf{b}_{m,t}^{\text{loc}}$	local data for agent m
$\mathbf{\Sigma}_{m,t}^{\text{ser}}, \mathbf{b}_{m,t}^{\text{ser}}$	data stored at the server

Table 2: Notations used in Algorithm 1.

Specifically, in each round $t \in [T]$, agent m_t participates and interacts with the environment (Line 3). The environment specifies the decision set D_t (Line 4), and the agent m_t selects the action based on its current optimisitic estimate of the reward (Line 5). Here the bonus term $\beta \|\mathbf{x}\|_{\mathbf{\Sigma}_{m_t,t}^{-1}}$ reflects the uncertainty of the estimated reward $\langle \widehat{\boldsymbol{\theta}}_{m_t,t}, \mathbf{x} \rangle$ and encourages exploration. After receiving the true reward r_t from the environment, agent m_t then updates its local data (Line 7).

The key component of the algorithm is the matrix determinant-based criterion (Line 8), which evaluates the information accumulated in current local data. If the criterion is satisfied, it suggests that the local data would help significantly reduce the uncertainty of estimating the model $\boldsymbol{\theta}^*$. Therefore, agent m_t will share its progress by uploading the local data to the server (Line 9) so that it can benefit other agents. Then the server updates the global data accordingly (Line 10). Afterwards, agent m_t downloads the latest global data from the server (Line 12) and updates its local data and model (Line 13-14). If the criterion in Line 8 is not met, then the communication between the agent and the server will not be triggered, and the local data remains local and unshared for agent m_t (Line 16). Finally, all the other inactive agents remain unchanged (Line 19).

Note that in Algorithm 1, the communication between the agent and the server (Line 9 and 12) involves only the active agent in that round, which is completely independent of other agents. This is in sharp contrast to existing algorithms. For example, in the main algorithm in Li and Wang (2022a), upload by any agent may trigger other agents to download the latest data, while our algorithm does not mandate this. On the other hand, many existing algorithms for multi-agent settings (e.g., Wang

et al. (2019)) require all agents to interact with the environment in each round, which essentially require full participation of all the agents.

The determinant-based criterion in Line 8 has been a long-standing design trick in contextual linear bandits that can help address the issue of unknown time horizon and reduce the switching cost (Abbasi-Yadkori et al., 2011; Ruan et al., 2021). For multi-agent bandits, such a criterion has also been used to control the communications complexity (Wang et al., 2019; Dubey and Pentland, 2020; Li and Wang, 2022a). This is because the need for policy switching or communication essentially reflects the same fact: enough information has been collected and it is time to update the (local) model. Indeed, achieving low communication complexity and low switching cost are unified in our `FedlinUCB` algorithm in the sense that the communication complexity is exactly twice the switching cost. Furthermore, using lazy update makes our algorithm amenable for analysis, which will be clear later in Section 6. In addition, we leave $\alpha > 0$ as a tuning parameter as it controls the trade-off between the regret and the communication complexity.

5 Theoretical Results

We now present our main result on the theoretical guarantee of Algorithm 1.

Theorem 5.1. Under Assumption 3.1 for Algorithm 1, if we set the confidence radius $\beta = \sqrt{\lambda}S + (\sqrt{1+M\alpha} + M\sqrt{2\alpha})\Big(R\sqrt{d\log\Big(\big(1+TL^2/(\min(\alpha,1)\lambda)\big)/\delta\Big)} + \sqrt{\lambda}S\Big)$, then with probability at least $1-\delta$, the regret in the first T rounds can be upper bounded by

$$\mathrm{Regret}(T) \leq 2dSLM\log(1+TL^2/\lambda) + 2\sqrt{2(1+M\alpha)}\beta\sqrt{2dT\log(1+TL^2/\lambda)}.$$

Moreover, the communication complexity and switching cost are both bounded by $2\log 2 \cdot d(M + 1/\alpha)\log(1+TL^2/(\lambda d))$.

Remark 5.2. Theorem 5.1 suggests that if we set the parameters $\alpha = 1/M^2$ and $\lambda = 1/S^2$ in Algorithm 1, then its regret is bounded by $\widetilde{O}(Rd\sqrt{T})$ and the corresponding communication complexity and switching cost are both bounded by $\widetilde{O}(dM^2)$. This choice of parameters yields the regret bound and the communication complexity presented in Table 1.

As a complement, we also provide a lower bound for the communication complexity as stated in the following theorem. See Appendix D for the proof.

Theorem 5.3. For any algorithm **Alg** with expected communication complexity less than $O(M/\log(T/M))$, there exist a linear bandit instance with $R = L = S = 1$ such that for $T \geq Md$, the expected regret for algorithm **Alg** is at least $\Omega(d\sqrt{MT})$.

Remark 5.4. Suppose each agent runs the `OFUL` algorithm (Abbasi-Yadkori et al., 2011) separately, then each agent $m \in [M]$ admits an $\widetilde{O}(d\sqrt{T_m})$ regret, where T_m is the number of rounds that agent m participates in. Thus the total regret of M agents is upper bounded by $\sum_{m=1}^{M}\widetilde{O}(\sqrt{T_m}) = \widetilde{O}(d\sqrt{MT})$. Theorem 5.3 implies that for any algorithm **Alg**, if its communication complexity is less than $O(M/\log(T/M))$, then its regret cannot be better than naively running M independent `OFUL` algorithms. In other words, Theorem 5.3 suggests that in order to improve the performance through collaboration, an $\Omega(M)$ communication complexity is necessary.

6 Overview of the Proof

When analyzing the performance of `FedLinUCB`, we face a unique challenge caused by the asynchronous communication, as illustrated in Figure 1. Here $(\mathbf{x}_{m,t}, \eta_{m,t})$ denotes the decision and the noise for agent m in its own t-th round. Specifically, in the synchronous setting, the filtration is generated by all the data collected by all agents, i.e., $\mathcal{F}_5 = \sigma\{\mathbf{x}_{m,t}, \eta_{m,t}\}_{t=1,m=1}^{5,5}$, as marked by the green dashed rectangle. This kind of filtration is well-defined since all agents share their data with each other at the end of each round. In sharp contrast, in our asynchronous setting, the data at the server can be generated by an irregular set of data from the agents, as marked by the blue rectangles. Such data pattern can be arbitrary and depends on the data collected in all previous rounds, which prevents us from defining a fixed filtration as we can do in the synchronous setting. Since the

application of standard martingale concentration inequalities relies on the well-defined filtration, they cannot be directly applied to our asynchronous setting.

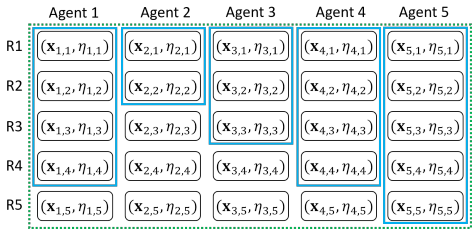


Figure 1: Illustration of ill-defined filtration.

To circumvent the above issue, we need to analyze the concentration property of the local data for each agent and then relate it to the concentration of the global data, so that we can control the sum of the bonuses and hence the regret. This requires a careful quantitative comparison of the local and global data covariance matrices, which is enabled by our design of determinant-based criterion. The details will be further explained in Section 6.2. In the next subsection, we present the key ingredients of the proof of Theorem 5.1.

Remark 6.1. Recall the notations in Table 2, where the values of those matrices and vectors might change within each round. To eliminate the possible confusion, from now on we follow the convention that all matrices and vectors in the analysis correspond to their values at the end of each round in Algorithm 1.

6.1 Analysis for communication complexity and switching cost

We first analyze the communication complexity and switching cost of Algorithm 1. For each $i \geq 0$, we define

$$\tau_i = \min\{t \in [T] \mid \det(\mathbf{\Sigma}_t^{\text{ser}}) \geq 2^i \lambda^d\}. \tag{6.1}$$

We divide the set of all rounds into epochs $\{\tau_i, \tau_i + 1, \ldots, \tau_{i+1} - 1\}$ for each $i \geq 0$. Then the communication complexity *within each epoch* can be bounded using the following lemma.

Lemma 6.2. Under the setting of Theorem 5.1, for each epoch from round τ_i to round $\tau_{i+1} - 1$, the number of communications in this epoch is upper bounded by $2(M + 1/\alpha)$.

Proof of Theorem 5.1: communication complexity and switching cost. It suffices to bound the number of epochs. By Assumption 3.1, we have $\|\mathbf{x}_t\|_2 \leq L$ for all $t \in [T]$. Since $\mathbf{\Sigma}_T^{\text{ser}}$ is positive definite, by the inequality of arithmetic and geometric means, we have

$$\begin{aligned}\det(\mathbf{\Sigma}_T^{\text{ser}}) &\leq \left(\frac{\operatorname{tr}(\mathbf{\Sigma}_T^{\text{ser}})}{d}\right)^d \leq \left(\frac{1}{d}\operatorname{tr}\left(\lambda \mathbf{I} + \sum_{t=1}^T \mathbf{x}_t \mathbf{x}_t^\top\right)\right)^d \\ &= \left(\lambda + \frac{1}{d}\sum_{t=1}^T \|\mathbf{x}_t\|_2^2\right)^d \leq \lambda^d \left(1 + \frac{TL^2}{\lambda d}\right)^d.\end{aligned}$$

Then recalling the definition of epochs based on (6.1), we have

$$\max\{i \geq 0 \mid \tau_i \neq \emptyset\} = \log_2 \frac{\det(\mathbf{\Sigma}_T^{\text{ser}})}{\lambda^d} \leq \log 2 \cdot d \log\left(1 + \frac{TL^2}{\lambda d}\right).$$

Therefore, the total number of epochs is bounded by $\log 2 \cdot d \log(1 + TL^2/(\lambda d))$. Now applying Lemma 6.2, the total communication complexity is bounded by $2 \log 2 \cdot d(M + 1/\alpha) \log(1 + TL^2/(\lambda d))$. Note that in Algorithm 1, each agent only switch its policy after communicating with the server, so the switching cost is exactly equal to half of the communication complexity. This finishes the proof for the claim on communication complexity and switching cost in Theorem 5.1. □

6.2 Analysis for regret upper bound

The regret analysis for Theorem 5.1 is much more involved, and it relies on a series of intermediate lemmas establishing the concentration.

Total information. We define the following auxiliary matrices and vectors that contain all the information up to round t:

$$\mathbf{\Sigma}_t^{\text{all}} = \lambda \mathbf{I} + \sum_{i=1}^t \mathbf{x}_i \mathbf{x}_i^\top, \qquad \mathbf{b}_t^{\text{all}} = \sum_{i=1}^t r_i \mathbf{x}_i, \qquad \mathbf{u}_t^{\text{all}} = \sum_{i=1}^t \eta_i \mathbf{x}_i, \tag{6.2}$$

where $\eta_i := r_i - \langle \mathbf{x}_i, \boldsymbol{\theta}^* \rangle$ is a R-sub-Gaussian noise by Assumption 3.1. In our setting, $\boldsymbol{\Sigma}_t^{\text{all}}, \mathbf{b}_t^{\text{all}}, \mathbf{u}_t^{\text{all}}$ are not accessible by the agents due to asynchronous communication, and we only use them to facilitate the analysis. With this notation, we can further define the following omnipotent estimate:

$$\widehat{\boldsymbol{\theta}}_t^{\text{all}} = (\boldsymbol{\Sigma}_t^{\text{all}})^{-1} \mathbf{b}_t^{\text{all}}. \tag{6.3}$$

As a direct application of the self-normalized martingale concentration inequality (Abbasi-Yadkori et al., 2011), we have the following global confidence bound due to the concentration of $\boldsymbol{\Sigma}_t^{\text{all}}$ and $\mathbf{b}_t^{\text{all}}$.

Lemma 6.3 (Global confidence bound; Theorem 2, Abbasi-Yadkori et al. 2011)**.** With probability at least $1-\delta$, for each round $t \in [T]$, the estimate $\widehat{\boldsymbol{\theta}}_t^{\text{all}}$ in (6.3) satisfies

$$\|\widehat{\boldsymbol{\theta}}_t^{\text{all}} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_t^{\text{all}}} \leq R\sqrt{d \log\big((1+TL^2/\lambda)/\delta\big)} + \sqrt{\lambda} S.$$

Per-agent information. Next, for each agent $m \in [M]$, we denote the rounds when agent m communicate with the server (i.e., upload and download data) as $\{t_{m,1}, t_{m,2}, ...\}$. For simplicity, *at the end of round* t, we denote by $N_m(t)$ the last round when agent m communicated with the server (so if agent m communicated with the server in round t, then $N_m(t) = t$). With this notation, for each round t and agent $m \in [M]$, the data that has been uploaded by agent m is then[4]

$$\boldsymbol{\Sigma}_{m,t}^{\text{up}} = \sum_{j=1, m_j=m}^{N_m(t)} \mathbf{x}_j \mathbf{x}_j^\top, \qquad \mathbf{u}_{m,t}^{\text{up}} = \sum_{j=1, m_j=m}^{N_m(t)} \mathbf{x}_j \eta_j.$$

Correspondingly, the local data that has not been uploaded to the server is

$$\boldsymbol{\Sigma}_{m,t}^{\text{loc}} = \sum_{j=N_m(t)+1, m_j=m}^{t} \mathbf{x}_j \mathbf{x}_j^\top, \qquad \mathbf{u}_{m,t}^{\text{loc}} = \sum_{j=N_m(t)+1, m_j=m}^{t} \mathbf{x}_j \eta_j.$$

Again, applying the self-normalized martingale concentration (Abbasi-Yadkori et al., 2011) together with a union bound, we can get the per-agent local concentration.

Lemma 6.4 (Local concentration)**.** Under the setting of Theorem 5.1, with probability at least $1-\delta$, for each round $t \in [T]$ and each agent $m \in [M]$, it holds that

$$\left\| \left(\alpha\lambda \mathbf{I} + \boldsymbol{\Sigma}_{m,t+1}^{\text{loc}}\right)^{-1} \mathbf{u}_{m,t}^{\text{loc}} \right\|_{\alpha\lambda\mathbf{I} + \boldsymbol{\Sigma}_{m,t}^{\text{loc}}} \leq R\sqrt{d \log\Big(\big(1+TL^2/(\alpha\lambda)\big)/\delta\Big)} + \sqrt{\lambda} S.$$

Moreover, based on our determinant-based communication criterion, we have the following lemma describing the quantitative relationship among the local data, uploaded data and global data.

Lemma 6.5 (Covariance comparison)**.** Under the setting of Theorem 5.1, it holds that

$$\lambda \mathbf{I} + \sum_{m'=1}^{M} \boldsymbol{\Sigma}_{m',t}^{\text{up}} \succeq \frac{1}{\alpha} \boldsymbol{\Sigma}_{m,t}^{\text{loc}} \tag{6.4}$$

for each agent $m \in [M]$. Moreover, for any $1 \leq t_1 < t_2 \leq T$, if agent m is the only active agent from round t_1 to $t_2 - 1$ and agent m only communicates with the server at round t_1, then for all $t_1 + 1 \leq t \leq t_2$, it further holds that

$$\boldsymbol{\Sigma}_{m,t} \succeq \frac{1}{1+M\alpha} \boldsymbol{\Sigma}_t^{\text{all}}. \tag{6.5}$$

Combining the above results, we obtain the local confidence bound, which then leads to the per-round regret in each round, as summarized in the following lemma.

Lemma 6.6 (Local confidence bound and per-round regret)**.** Under the setting of Theorem 5.1, with probability at least $1-\delta$, for each $t \in [T]$, the estimate $\widehat{\boldsymbol{\theta}}_{m,t+1}$ satisfies that $\|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_{m,t+1}\|_{\boldsymbol{\Sigma}_{m,t+1}} \leq \beta$ for all $m \in [M]$. Consequently, for each round $t \in [T]$, the regret in round t satisfies

$$\Delta_t = \max_{\mathbf{x} \in \mathcal{D}_t} \langle \boldsymbol{\theta}^*, \mathbf{x} \rangle - \langle \boldsymbol{\theta}^*, \mathbf{x}_t \rangle \leq 2\beta \sqrt{\mathbf{x}_t^\top \boldsymbol{\Sigma}_{m_t,t}^{-1} \mathbf{x}_t}.$$

[4] Strictly speaking, the uploaded data only consists of $\boldsymbol{\Sigma}_{m,t}^{\text{up}}$ and $\mathbf{b}_{m,t}^{\text{up}}$, and here we introduce $\mathbf{u}_{m,t}^{\text{up}}$ and $\mathbf{u}_{m,t}^{\text{loc}}$ solely for the purpose of analysis.

Now, we are ready to prove the regret bound in Theorem 5.1.

Proof of Theorem 5.1: regret. First, according to Lemma 6.6, the regret in first T round can be decomposed and upper bounded by

$$\text{Regret}(T) = \sum_{t=1}^{T} \langle \boldsymbol{\theta}^*, \mathbf{x}_t^* - \mathbf{x}_t \rangle \leq \sum_{t=1}^{T} 2\beta \|\mathbf{x}_t\|_{\boldsymbol{\Sigma}_{m_t,t}^{-1}}.$$

Now, we only need to control the summation of the upper confidence bonus term $2\beta\|\mathbf{x}_t\|_{\boldsymbol{\Sigma}_{m_t,t}^{-1}}$ and we focus on the agent-action sequence $\{(m_t, \mathbf{x}_t)\}_{t=1}^{T}$. Notice that if an agent m communicate with the server at time t_1 and t_2, then the order of actions between round t_1 and t_2 will not effect the covariance matrix for agent m. In addition, since agent m does not upload new data between round t_1 and t_2, the order of actions from agent m will also not affect other agents' covariance matrix. Thus, without effect the covariance matrix and the corresponding upper confidence bonus, we can always reorder the sequence of active agents such that each agent communicates with the server and stays active until the next agent kicks in to communicate with the server. Such reordering is valid according to the communication protocol as each agent has only local updates between communications with the server.

More specifically, we assume that the sequence of round that the active agent communicates with server is $0 = t_0 < t_1 < t_2 < \cdots < t_N = T + 1$[5], and from round $t_i + 1$ to $t_{i+1} - 1$ there is only one agent active, that is, $m_{t_i} = m_{t_i+1} = \cdots = m_{t_{i+1}-1}$. Then we apply Lemma 6.5 and get that the bonus term for each round $t_i < t < t_{i+1}$ can be controlled by

$$2\beta\|\mathbf{x}_t\|_{(\boldsymbol{\Sigma}_{m_t,t})^{-1}} \leq 2\beta\sqrt{1+M\alpha}\|\mathbf{x}_t\|_{(\boldsymbol{\Sigma}_t^{\text{all}})^{-1}}. \tag{6.6}$$

In addition, to control the bonus term for rounds $\{t_i\}_{i=1}^{N}$, we define $T_i = \min\{t \in [T] \mid \det(\boldsymbol{\Sigma}_t^{\text{all}}) \geq 2^i\lambda^d\}$. For each time interval from T_i to T_{i+1}, if an agent m communicate with the server more than once, e.g., agent m communicates with the server at round $T_{i,1}$ and $T_{i,2}$ such that $T_i \leq T_{i,1} < T_{i,2} < T_{i+1}$, then for the latter round $T_{i,2}$, the bonus term can be controlled by

$$\begin{aligned}
2\beta\|\mathbf{x}_{T_{i,2}}\|_{(\boldsymbol{\Sigma}_{m,T_{i,2}})^{-1}} &\leq 2\beta\sqrt{1+M\alpha}\|\mathbf{x}_{T_{i,2}}\|_{(\boldsymbol{\Sigma}_{T_{i,1}}^{\text{all}})^{-1}} \\
&\leq 2\beta\sqrt{2(1+M\alpha)}\|\mathbf{x}_{T_{i,2}}\|_{(\boldsymbol{\Sigma}_{T_{i+1}-1}^{\text{all}})^{-1}} \\
&\leq 2\beta\sqrt{2(1+M\alpha)}\|\mathbf{x}_{T_{i,2}}\|_{(\boldsymbol{\Sigma}_{T_{i,2}}^{\text{all}})^{-1}},
\end{aligned} \tag{6.7}$$

where the first inequality holds due Lemma 6.5 and agent m communicate with the server at the previous round $T_{i,1}$, the second inequality holds due to Lemma E.4 with the fact that $\det \boldsymbol{\Sigma}_{T_{i+1}-1}^{\text{all}} / \det(\boldsymbol{\Sigma}_{T_{i,1}}^{\text{all}}) \leq 2^{i+1}\lambda^d/(2^i\lambda^d) = 2$, and the last inequality holds due to the fact that $\boldsymbol{\Sigma}_{T_{i+1}-1}^{\text{all}} \succeq \boldsymbol{\Sigma}_{T_{i,2}}^{\text{all}}$. On the other hand, for each time interval from T_i to T_{i+1}, the bonus term for the first communication can always be trivially bounded by 1 for all agent m. Therefore, the summation of the regret over first communication can be upper bounded by 1. In addition, since the norm of each action $\|\mathbf{x}\|_2 \leq L$ and it implies that we have

$$\det(\boldsymbol{\Sigma}_T^{\text{all}}) \leq (\lambda + TL^2)^d, \tag{6.8}$$

which implies that the number of different intervals is at most $d\log(1 + TL^2/\lambda)$. Combining the upper bound of regret in (6.6), (6.7) and (6.8), we have

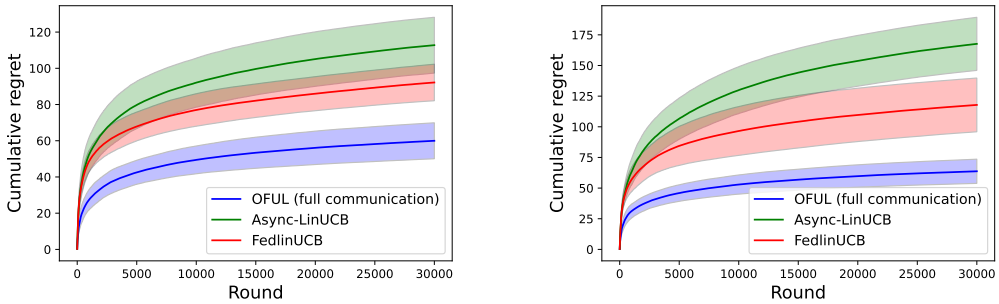
$$\begin{aligned}
\text{Regret}(T) &\leq dM\log(1+TL^2/\lambda) + \sum_{t=1}^{T} 2\sqrt{2(1+M\alpha)}\beta\|\mathbf{x}_{T_{i,2}}\|_{(\boldsymbol{\Sigma}_{T_{i,2}}^{\text{all}})^{-1}} \\
&\leq d\log(1+TL^2/\lambda) + 2\sqrt{2(1+M\alpha)}\beta\sqrt{2dT\log(1+TL^2/\lambda)},
\end{aligned}$$

where the last inequality follows from a standard elliptic potential argument (Abbasi-Yadkori et al., 2011). This completes the proof. □

[5]There is no comminucation happening at t_0 or t_N, but we include them for notational convenience.

7 Experiments

In this section, we report numerical simulation results on the comparison between our algorithm with other baselines. Specifically, we construct a contextual linear bandit instance with feature dimension $d = 25$. In each round t, the active agent m_t is uniformly sampled from all M agents. We set the total number of rounds $T = 30,000$ and test for $M = 15$ and 30. We compare our `FedLinUCB` with Async-LinUCB (Li and Wang, 2022a) and OFUL Abbasi-Yadkori et al. (2011) with full communication (i.e., the active agent communicates with the server in each round). Due to space limit, further details and more simulation results are deferred to Appendix B. The code and data for our experiments can be found on Github [6].



(a) $M = 15$. Cumulative regret versus Round

(b) $M = 30$. Cumulative regret versus Round

Figure 2: Plots of cumulative regret versus round for $M = 15$ (Fig. 2(a)) and $M = 30$ (Fig. 2(b)), comparing `FedLinUCB` (ours) with Async-LinUCB (Li and Wang, 2022a) and OFUL (Abbasi-Yadkori et al., 2011) with full communication. The results are averaged over 20 runs with the error bars chosen as the empirical one standard deviation.

It is clear from Figure 2 that `FedLinUCB` outperforms Async-LinUCB (Li and Wang, 2022a) in terms of regret. Although OFUL with full communication has the smallest regret, its communication cost ($2MT$) is much higher than ours. Furthermore, we plot the log-scaled regret in Figure 3(c) and 3(d) in Appendix B, which show that the average regret of our algorithm actually has a rate very close to the optimal rate of OFUL. Overall, the numerical simulation corroborates our theoretical results.

8 Conclusion and Future Work

In this work, we study federated contextual linear bandit problem with fully asynchronous communication. We propose a simple and provably efficient algorithm named `FedLinUCB`. We prove that `FedLinUCB` obtains a near-optimal regret of order $\widetilde{O}(d\sqrt{T})$ with $\widetilde{O}(dM^2)$ communication complexity. We also prove a lower bound on the communication complexity, which suggests that an $\Omega(M)$ communication complexity is necessary for achieving a near-optimal regret. There still exists an $O(dM)$ gap between the upper and lower bounds for the communication complexity and we leave it as a future work to close this gap. Another important direction for future work is to study federated linear bandits with a decentralized communication network without a central server (i.e., P2P networks). In addition, there are potential privacy concerns when the agents upload and download data from the server, and it remains an open problem to devise provably efficient algorithm for asynchronous federated linear bandits with privacy guarantees.

Acknowledgments and Disclosure of Funding

We thank the anonymous reviewers and area chair for their helpful comments. JH and QG are partially supported by the National Science Foundation CAREER Award 1906169 and the Sloan Research Fellowship. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

[6] https://github.com/uclaml/FedLinUCB

References

ABBASI-YADKORI, Y., PÁL, D. and SZEPESVÁRI, C. (2011). Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems* **24** 2312–2320.

ABE, N., BIERMANN, A. W. and LONG, P. M. (2003). Reinforcement learning with immediate rewards and linear hypotheses. *Algorithmica* **37** 263–293.

AGRAWAL, S. and GOYAL, N. (2013). Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*. PMLR.

AUER, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research* **3** 397–422.

CESA-BIANCHI, N., GENTILE, C. and ZAPPELLA, G. (2013). A gang of bandits. *Advances in neural information processing systems* **26**.

CHAKRABORTY, M., CHUA, K. Y. P., DAS, S. and JUBA, B. (2017). Coordinated versus decentralized exploration in multi-agent multi-armed bandits. In *IJCAI*.

CHU, W., LI, L., REYZIN, L. and SCHAPIRE, R. (2011). Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings.

DANI, V., 9, ., HAYES, T. and KAKADE, S. M. (2008). Stochastic linear optimization under bandit feedback. *21st Annual Conference on Learning Theory* 355–366.

DEKEL, O., DING, J., KOREN, T. and PERES, Y. (2014). Bandits with switching costs: T 2/3 regret. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*.

DUBEY, A. and PENTLAND, A. (2020). Differentially-private federated linear bandits. *Advances in Neural Information Processing Systems* **33** 6003–6014.

DURAND, A., ACHILLEOS, C., IACOVIDES, D., STRATI, K., MITSIS, G. D. and PINEAU, J. (2018). Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis. In *Machine learning for healthcare conference*. PMLR.

FILIPPI, S., CAPPE, O., GARIVIER, A. and SZEPESVÁRI, C. (2010). Parametric bandits: The generalized linear case. *Advances in Neural Information Processing Systems* **23**.

GENTILE, C., LI, S. and ZAPPELLA, G. (2014). Online clustering of bandits. In *International Conference on Machine Learning*. PMLR.

HUANG, R., WU, W., YANG, J. and SHEN, C. (2021). Federated linear contextual bandits. *Advances in Neural Information Processing Systems* **34**.

JAGADEESAN, M., WEI, A., WANG, Y., JORDAN, M. and STEINHARDT, J. (2021). Learning equilibria in matching markets from bandit feedback. *Advances in Neural Information Processing Systems* **34**.

JUN, K.-S., BHARGAVA, A., NOWAK, R. and WILLETT, R. (2017). Scalable generalized linear bandits: Online computation and hashing. *Advances in Neural Information Processing Systems* **30**.

KALAI, A. and VEMPALA, S. (2005). Efficient algorithms for online decision problems. *Journal of Computer and System Sciences* **71** 291–307.

KARIMIREDDY, S. P., JAGGI, M., KALE, S., MOHRI, M., REDDI, S. J., STICH, S. U. and SURESH, A. T. (2020). Mime: Mimicking centralized stochastic algorithms in federated learning. *arXiv preprint arXiv:2008.03606* .

KORDA, N., SZORENYI, B. and LI, S. (2016). Distributed clustering of linear bandits in peer to peer networks. In *International conference on machine learning*. PMLR.

LANDGREN, P., SRIVASTAVA, V. and LEONARD, N. E. (2016). On distributed cooperative decision-making in multiarmed bandits. In *2016 European Control Conference (ECC)*. 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)*. IEEE.

LATTIMORE, T. and SZEPESVÁRI, C. (2020). *Bandit algorithms*. Cambridge University Press.

LI, C. and WANG, H. (2022a). Asynchronous upper confidence bound algorithms for federated linear bandits. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

LI, C. and WANG, H. (2022b). Communication efficient federated learning for generalized linear bandits. *arXiv preprint arXiv:2202.01087* .

LI, C., WU, Q. and WANG, H. (2021). Unifying clustered and non-stationary bandits. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

LI, L., CHU, W., LANGFORD, J. and SCHAPIRE, R. E. (2010a). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*.

LI, S., KARATZOGLOU, A. and GENTILE, C. (2016). Collaborative filtering bandits. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*.

LI, T., SONG, L. and FRAGOULI, C. (2020). Federated recommendation system via differential privacy. In *2020 IEEE International Symposium on Information Theory (ISIT)*. IEEE.

LI, W., WANG, X., ZHANG, R., CUI, Y., MAO, J. and JIN, R. (2010b). Exploitation and exploration in a performance based contextual advertising system. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*.

LI, Y., WANG, C.-H., CHENG, G. and SUN, W. W. (2022). Rate-optimal contextual online matching bandit. *arXiv preprint arXiv:2205.03699* .

LI, Y., WANG, Y. and ZHOU, Y. (2019). Nearly minimax-optimal regret for linearly parameterized bandits. In *Conference on Learning Theory*. PMLR.

LIU, K. and ZHAO, Q. (2010). Distributed learning in multi-armed bandit with multiple players. *IEEE transactions on signal processing* **58** 5667–5681.

MARTÍNEZ-RUBIO, D., KANADE, V. and REBESCHINI, P. (2019). Decentralized cooperative stochastic bandits. *Advances in Neural Information Processing Systems* **32**.

MCMAHAN, B., MOORE, E., RAMAGE, D., HAMPSON, S. and Y ARCAS, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*. PMLR.

RUAN, Y., YANG, J. and ZHOU, Y. (2021). Linear bandits with limited adaptivity and learning distributional optimal design. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*.

RUSMEVICHIENTONG, P. and TSITSIKLIS, J. N. (2010). Linearly parameterized bandits. *Mathematics of Operations Research* **35** 395–411.

SANKARARAMAN, A., GANESH, A. and SHAKKOTTAI, S. (2019). Social learning in multi agent multi armed bandits. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* **3** 1–35.

SZORENYI, B., BUSA-FEKETE, R., HEGEDUS, I., ORMÁNDI, R., JELASITY, M. and KÉGL, B. (2013). Gossip-based distributed stochastic bandit algorithms. In *International Conference on Machine Learning*. PMLR.

TIE, L., CAI, K.-Y. and LIN, Y. (2011). Rearrangement inequalities for hermitian matrices. *Linear algebra and its applications* **434** 443–456.

WANG, P.-A., PROUTIERE, A., ARIU, K., JEDRA, Y. and RUSSO, A. (2020). Optimal algorithms for multiplayer multi-armed bandits. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

WANG, Y., HU, J., CHEN, X. and WANG, L. (2019). Distributed bandit learning: Near-optimal regret with efficient communication. *arXiv preprint arXiv:1904.06309* .

WANG, Y.-G. (1991). Sequential allocation in clinical trials. *Communications in Statistics-Theory and Methods* **20** 791–805.

WU, Q., WANG, H., GU, Q. and WANG, H. (2016). Contextual bandits in a collaborative environment. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*.

ZHU, Z., ZHU, J., LIU, J. and LIU, Y. (2021). Federated bandit: A gossiping approach. In *Abstract Proceedings of the 2021 ACM SIGMETRICS/International Conference on Measurement and Modeling of Computer Systems*.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] This work is seeking to develop a mathematical understanding of federated linear bandits.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We use existing theorems and cite them properly.
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Further Discussions

Here we provide further discussions on our results. We present a detailed comparison with Li and Wang (2022a) in Appendix A.1, where we first discuss the difference in the algorithmic design, and then elaborate on the concentration issue under the asynchronous setting with a simple example. In Appendix A.2, we give an alternative form of our algorithm, where we rewrite Algorithm 1 in an 'episodic' fashion. The purpose is to make it easier for readers to compare our algorithm with existing algorithms for federated linear bandits that are usually expressed in the 'episodic' form.

A.1 Comparison with Li and Wang (2022a)

Difference in algorithmic design. The Async-LinUCB algorithm proposed by Li and Wang (2022a) is not fully asynchronous since in their algorithm, if some agent uploads data to the server, the server will decide if each of the M agents needs to download the data. If the server decides that an agent needs to download the data, this agent has to first download the data from the server and then update its local policy before further interaction with the environment (i.e., taking the next action). In other words, if an agent is offline when the server requests a download, the agent cannot take any further action until it goes online and completes the required download and local model update. In contrast, under the communication protocol in our Algorithm 1, any offline agent can still take action until the trigger of the upload protocol. It is evident that their asynchronous communication protocol is very restricted.

Concentration issue. Next, we discuss the concentration issue, and we first illustrate the problem using a multi-arm bandit instance. Unlike the synchronous case, the reward estimator based on the server-end data can be biased in asynchronous federated linear bandits. To see so, let us consider the following simple example: The decision set contains two arms, A and B, and suppose for pulling arm A, the agent receives a reward equal to either 1 or -1 with equal probability. We assume that there are M agents, and each agent is active for two consecutive rounds. For each agent $m \in [M]$, if the agent has selected the arm A in the first round, then the agent will select again the arm A in the second round only if the agent receives a reward of 1 when pulling arm A in the first round. In this case, it is easy to show that with probability 0.5, an agent selects arm A one time with reward -1, and with probability 0.25, an agent selects arm A twice with total reward 2. Similarly, with probability 0.25, an agent selects arm A twice with a total reward of 0.

In the synchronous setting, all agent will upload their local data to the server at the end of each round. Thus, taking an average for all data at the server, the expected reward of arm A is 0, which equals the actual expected reward of arm A. However, in the asynchronous setting, things become more complicated. Suppose that for each agent, only selecting arm A twice will trigger the upload protocal. Then after two active rounds, an agent will upload its data to the server if and only if the agent receives reward 1 in the first round. Thus among the agents that upload the data, half of them receive a total reward of 2 and the other half receive a total reward of 0. In this case, taking an average for all data at the server, the expected reward of arm A is 0.5, which is a biased estimator compared with the actual expected reward.

Indeed, the above issue could happen in federated linear bandits with the Async-LinUCB algorithm (Li and Wang, 2022a). Specifically, let us consider a linear bandit instance with dimension $d = 2$, and we assume that arm A has context vector $\mathbf{x}_A = (3, 0)^\top$, arm B has context vector $\mathbf{x}_B = (0, 1/\sqrt{10})^\top$, the true model is $\boldsymbol{\theta}^* = \mathbf{0}$, the noise η is a Rademacher random variable, and the parameter λ is set to be 1. Therefore, the rewards for both arm A and B equal to 1 or -1 with 0.5 probability. In this case, based on the principle of optimism in the face of uncertainty, at the beginning, the optimistic estimators for the two arms A, B are 3β and $\beta/\sqrt{10}$ respectively. Thus, all agents will always choose arm A in the first round, so $\mathbf{x}_1 = \mathbf{x}_A$. After choosing arm A at the first round, the optimistic estimator for the two arms A, B in each agent's second round will be $9r/10 + 3\beta/\sqrt{10}$ and $\beta/\sqrt{10}$ respectively, where r is the reward received in the first round. Therefore, with confidence radius $\beta < 1$, each agent will select the arm A (i.e., $\mathbf{x}_2 = \mathbf{x}_A$) in the second round only if the agent receives a reward of $r = 1$ in the first round. Finally, only choosing arm A twice will increase the determinant of the covariance matrix enough to trigger the upload protocol (e.g., $\det(\lambda\mathbf{I} + \mathbf{x}_1\mathbf{x}_1^\top + \mathbf{x}_2\mathbf{x}_2^\top)/\det(\lambda\mathbf{I}) \geq 19$).

As demonstrated above, in the asynchronous setting, the reward estimator based on the server-end data can be biased, which leads to the issue that previous concentration results (e.g., Abbasi-Yadkori

Algorithm 2 Federated linear UCB (Alternative)

1: Initialize $\mathbf{\Sigma}_{m,1} = \mathbf{\Sigma}_1^{\text{ser}} = \lambda \mathbf{I}$, $\widehat{\boldsymbol{\theta}}_{m,1} = 0$, $\mathbf{b}_{m,0}^{\text{loc}} = 0$ and $\mathbf{\Sigma}_{m,0}^{\text{loc}} = 0$ for all $m \in [M]$
2: **for** $k = 1, 2, \ldots, K$ **do**
3: Participation set $P_k \subseteq [M]$ of arbitrary order
4: **for** each active agent $m \in P_k$ **do**
5: Receive $D_{m,k}$ from the environment
6: Select $\mathbf{x}_{m,k} \leftarrow \operatorname{argmax}_{\mathbf{x} \in D_{m,k}} \langle \widehat{\boldsymbol{\theta}}_{m,k}, \mathbf{x}_k \rangle + \beta \|\mathbf{x}\|_{\mathbf{\Sigma}_{m,k}^{-1}}$ `/* Optimistic decision */`
7: Receive $r_{m,k}$ from environment
8: $\mathbf{\Sigma}_{m,k}^{\text{loc}} \leftarrow \mathbf{\Sigma}_{m,k-1}^{\text{loc}} + \mathbf{x}_{m,k}\mathbf{x}_{m,k}^\top$, $\mathbf{b}_{m,k}^{\text{loc}} \leftarrow \mathbf{b}_{m,k-1}^{\text{loc}} + r_{m,k}\mathbf{x}_{m,k}$ `/*Local update*/`
9: **if** $\det(\mathbf{\Sigma}_{m,k} + \mathbf{\Sigma}_{m,k}^{\text{loc}}) > (1+\alpha)\det(\mathbf{\Sigma}_{m,k})$ **then**
10: Agent m sends $\mathbf{\Sigma}_{m,k}^{\text{loc}}$ and $\mathbf{b}_{m,k}^{\text{loc}}$ to server `/* Upload */`
11: $\mathbf{\Sigma}_k^{\text{ser}} \leftarrow \mathbf{\Sigma}_k^{\text{ser}} + \mathbf{\Sigma}_{m,k}^{\text{loc}}$, $\mathbf{b}_k^{\text{ser}} \leftarrow \mathbf{b}_k^{\text{ser}} + \mathbf{b}_{m,k}^{\text{loc}}$ `/* Global update */`
12: $\mathbf{\Sigma}_{m,k}^{\text{loc}} \leftarrow 0$, $\mathbf{b}_{m,k}^{\text{loc}} \leftarrow 0$
13: Server sends $\mathbf{\Sigma}_k^{\text{ser}}$ and $\mathbf{b}_k^{\text{ser}}$ back to agent m `/* Download */`
14: $\mathbf{\Sigma}_{m,k+1} \leftarrow \mathbf{\Sigma}_k^{\text{ser}}$, $\mathbf{b}_{m,k} \leftarrow \mathbf{b}_k^{\text{ser}}$
15: $\widehat{\boldsymbol{\theta}}_{m,k+1} \leftarrow \mathbf{\Sigma}_{m,k+1}^{-1}\mathbf{b}_{m,k+1}$ `/* Compute estimate */`
16: **else**
17: $\mathbf{\Sigma}_{m,k+1} \leftarrow \mathbf{\Sigma}_{m,k}$, $\mathbf{b}_{m,k+1} \leftarrow \mathbf{b}_{m,k}$, $\widehat{\boldsymbol{\theta}}_{m,k+1} \leftarrow \widehat{\boldsymbol{\theta}}_{m,k}$
18: **end if**
19: **end for**
20: **for** other inactive agents $m \in [M] \setminus P_k$ **do**
21: $\mathbf{\Sigma}_{m,k+1} \leftarrow \mathbf{\Sigma}_{m,k}$, $\mathbf{b}_{m,k+1} \leftarrow \mathbf{b}_{m,k}$, $\widehat{\boldsymbol{\theta}}_{m,k+1} \leftarrow \widehat{\boldsymbol{\theta}}_{m,k}$
22: **end for**
23: **end for**

et al. (2011)) cannot be directly used for the server's data. This is why we need a more dedicated analysis to control this biased error (see Lemma 6.6 for more details).

A.2 An Alternative Form of Algorithm 1

We introduce an alternative form of Algorithm 1, which is displayed in Algorithm 2. Algorithm 2 can be viewed as the episodic[7] version of Algorithm 1, and its form aligns with those of the existing algorithms for federated linear bandits (Wang et al., 2019; Dubey and Pentland, 2020; Huang et al., 2021; Li and Wang, 2022a).

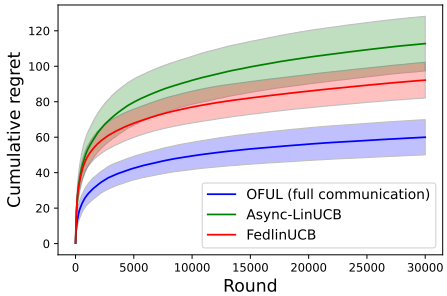
Specifically, in Algorithm 2, for each round (episode) $k \in [K]$, the set of active agents is given by P_k, where the order of agents in P_k can be arbitrary (Line 3). Then the agents in the set P_k participate according to the prefixed order (Line 4). The operations in the inner loop of Algorithm 2 (i.e., decision rule, upload/download, local/global update, and model estimates) are all identical to those in Algorithm 1. Therefore, Algorithm 2 is indeed equivalent to Algorithm 1 up to relabeling of the participation of the agents, and hence all the theoretical results for Algorithm 1 also hold for Algorithm 2.

B Experiments

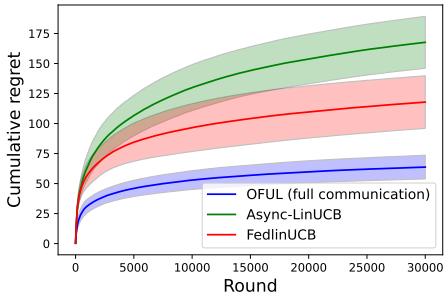
In this section, we provide the remaining details on numerical simulations.

Experiment setup. We construct two linear bandit instances with dimension $d = 25$. In the first instance, the true model parameter $\boldsymbol{\theta}^*$ is $[1/\sqrt{d}, .., 1/\sqrt{d}] \in \mathbb{R}^d$. In the second instance, the true model parameter $\boldsymbol{\theta}^*$ is generated by uniform random sampling over the space $[-1/\sqrt{d}, 1/\sqrt{d}]^d$ with normalization. For each round $t \in [T]$, the active agent m_t is uniform sampled from all M agents

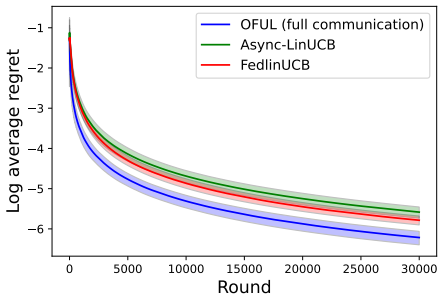
[7] Here 'episode' means a collection of every agent's interaction with the environment for one round, which is different from the usual term in online learning that refers to a sequential interaction lasting for a certain time horizon. We only use this term to differentiate Algorithm 2 from Algorithm 1.



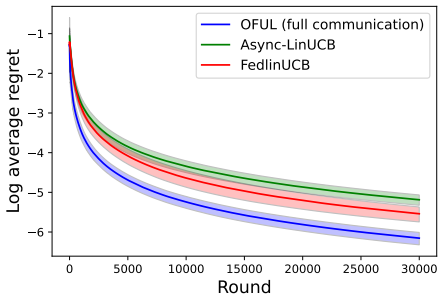
(a) $M = 15$. Cumulative regret versus Round



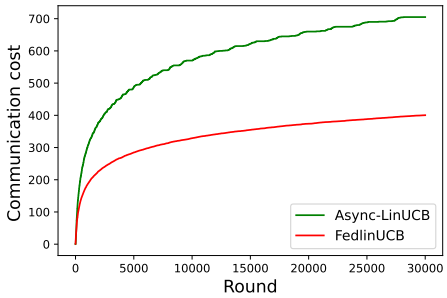
(b) $M = 30$. Cumulative regret versus Round



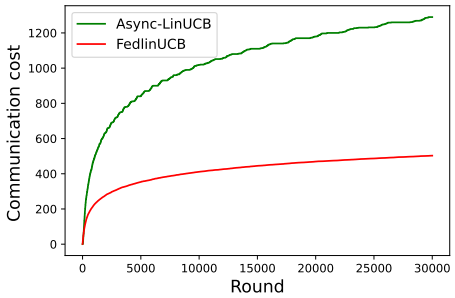
(c) $M = 15$. Log-average regret versus Round



(d) $M = 30$. Log-average regret versus Round



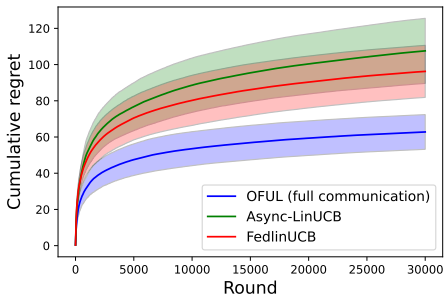
(e) $M = 15$. Communication cost versus Round



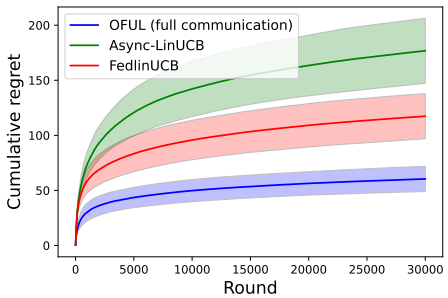
(f) $M = 30$. Communication cost versus Round

Figure 3: Comparison of FedlinUCB (ours), Async-LinUCB (Li and Wang, 2022a) and OFUL (full communication) (Abbasi-Yadkori et al., 2011) with parameter $\boldsymbol{\theta}^* = [1/\sqrt{d}, .., 1/\sqrt{d}] \in \mathbb{R}^d$. Experiments are run for $M = 15$ and $M = 30$, and results are averaged over 20 runs. Figures 3(a) and 3(b) present the cumulative regret; Figures 3(c) and 3(d) show the averaged regret (in log scale); Figures 3(e) and 3(f) compare the communication cost versus number of rounds. Note that the communication cost of OFUL with full communication is linear with the number of rounds T and is far greater than those of both FedLinUCB and Async-LinUCB. Therefore, in order to make a clearer comparison between the communication cost of FedLinUCB and Async-LinUCB, OFUL is omitted in Figure 3(e) and 3(f).

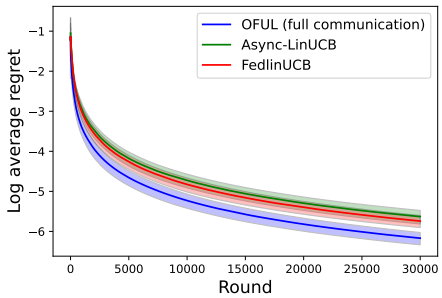
and the decision set $\mathcal{D}_t$ consists of 25 different actions uniformly randomly sampled from the space $[-1/\sqrt{d}, 1/\sqrt{d}]^d$. After the active agent m_t chooses an action $\mathbf{x}_t$, the agent m_t receives a reward given by $r_t = \langle \mathbf{x}_t, \boldsymbol{\theta}^* \rangle + \eta_t$, where η_t is a 0.3-Gaussian noise. We run simulation on the above linear bandit instance with the total number of rounds $T = 30000$ (repeating 20 times and taking the average) and the number of agents is set to be 15 or 30. We implement our `FedLinUCB` algorithm and compare its performance with Async-LinUCB (Li and Wang, 2022a) and OFUL Abbasi-Yadkori et al. (2011) with full communication (i.e., the active agent communicates with the server in each



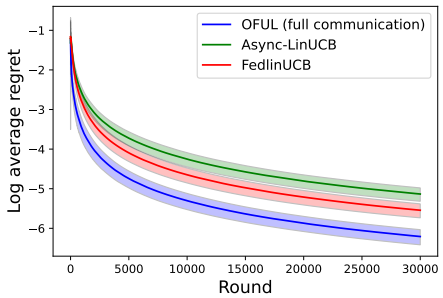
(a) $M = 15$. Cumulative regret versus Round



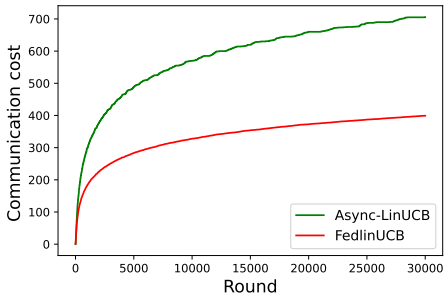
(b) $M = 30$. Cumulative regret versus Round



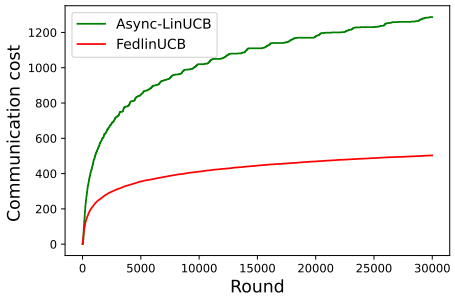
(c) $M = 15$. Log-average regret versus Round



(d) $M = 30$. Log-average regret versus Round



(e) $M = 15$. Communication cost versus Round



(f) $M = 30$. Communication cost versus Round

Figure 4: Comparison of FedlinUCB (ours), Async-LinUCB (Li and Wang, 2022a) and OFUL (full communication) (Abbasi-Yadkori et al., 2011) with random select $\boldsymbol{\theta}^*$. Experiments are run for $M = 15$ and $M = 30$, and results are averaged over 20 runs. Figures 4(a) and 4(b) present the cumulative regret; Figures 4(c) and 4(d) show the averaged regret (in log scale); Figures 4(e) and 4(f) compare the communication cost versus number of rounds. Note that the communication cost of OFUL with full communication is linear with the number of rounds T and is far greater than those of both FedLinUCB and Async-LinUCB. Therefore, in order to make a clearer comparison between the communication cost of FedLinUCB and Async-LinUCB, OFUL is omitted in Figure 4(e) and 4(f).

round). We set the parameter $\alpha = 1$ for `FedLinUCB` and $\gamma_U = \gamma_D = 5$ for Async-LinUCB to ensure that the communication costs of `FedLinUCB` and Async-LinUCB have similar magnitudes.

Results. The results are presented in Figure 3 and 4, suggesting that our algorithm significantly outperforms Async-LinUCB as our algorithm achieves smaller regret while spending less communication cost.

Specifically, Figure 3(a) and 3(b) displays the cumulative regret of our `FedLinUCB` algorithm, Async-LinUCB (Li and Wang, 2022a) and OFUL (Abbasi-Yadkori et al., 2011) with full communication.

It can be apparently seen that FedLinUCB outperforms Async-LinUCB in terms of regret. Next, Figure 3(c) and 3(d) show the average regret in log scale. These two plots show that the average regret of our algorithm has a rate very close to the optimal rate of OFUL. Finally, we plot communication cost versus number of rounds in Figure 3(e) and 3(f), which indicates that the communication cost of FedLinUCB is significantly lower than that of Async-LinUCB. Together with Figure 3(a) and 3(b), we see that our algorithm achieves lower regret with lower communication cost, compared to Async-LinUCB.

The experiment for Figure 4 adopts a different $\boldsymbol{\theta}^*$ (which is randomly selected sampled the space $[-1/\sqrt{d}, 1/\sqrt{d}]^d$ with normalization) from that of Figure 3. Figure 4 shows almost the same results as Figure 3, indicating our algorithms works in general.

Overall, the simulation corroborates our theoretical results. It also shows that our algorithm indeed outperforms Async-LinUCB (Li and Wang, 2022a).

For all the experiments, the results are averaged over 20 runs with the error bars chosen as the empirical one standard deviation. All experiments are conducted on a Macbook with 8-core CPU and 16 GB of memory.

C Missing Proofs in Section 6

Here we present the proof of the results in Section 6.

C.1 Communication complexity within each epoch

We first present the proof for the bound on the communication complexity within each epoch given in Lemma 6.2.

Proof of Lemma 6.2. For each agent $m \in [M]$, let n_m be the number of communications agent m has made during this epoch, and we denote the communication rounds as $t_1, \ldots, t_{n_m}$ for simplicity. Now we consider the data uploaded to the server, and it can be denoted by the value of covariance matrix $\boldsymbol{\Sigma}^{\text{loc}}_{m,t_j}$ before communicating with the server. For each $j = 2, \ldots, n_m$, according to the determinant-based criterion (Line 9) in Algorithm 1, we have

$$\det(\boldsymbol{\Sigma}_{m,t_j} + \boldsymbol{\Sigma}^{\text{loc}}_{m,t_j}) - \det(\boldsymbol{\Sigma}_{m,t_j}) > \alpha \cdot \det(\boldsymbol{\Sigma}_{m,t_j}),$$

which further implies that

$$\alpha \cdot \det(\boldsymbol{\Sigma}^{\text{ser}}_{T_i}) < \det(\boldsymbol{\Sigma}^{\text{ser}}_{T_i} + \boldsymbol{\Sigma}^{\text{loc}}_{m,t_j}) - \det(\boldsymbol{\Sigma}^{\text{ser}}_{T_i}), \tag{C.1}$$

where the inequality holds due to Lemma E.2 together with the fact that the communication in round t_1 updates the covariance matrix so that $\boldsymbol{\Sigma}_{m,t_j} \succeq \boldsymbol{\Sigma}^{\text{ser}}_{T_i}$. In addition, we define the sequence of all communications from T_i to $T_{i+1} - 1$ as $t'_1, \ldots, t'_L$. For each round t'_j, if the agent $m_{t'_j}$ have already communicated with the server earlier in this epoch, we have

$$\begin{aligned}
\det(\boldsymbol{\Sigma}^{\text{ser}}_{t'_j}) - \det(\boldsymbol{\Sigma}^{\text{ser}}_{t'_{j-1}}) &= \det(\boldsymbol{\Sigma}^{\text{ser}}_{t'_{j-1}} + \boldsymbol{\Sigma}^{\text{loc}}_{m_{t'_j},t'_j}) - \det(\boldsymbol{\Sigma}^{\text{ser}}_{t'_{j-1}}) \\
&\geq \det(\boldsymbol{\Sigma}^{\text{ser}}_{T_i} + \boldsymbol{\Sigma}^{\text{loc}}_{m,t_j}) - \det(\boldsymbol{\Sigma}^{\text{ser}}_{T_i}) \\
&> \alpha \cdot \det(\boldsymbol{\Sigma}^{\text{ser}}_{T_i}),
\end{aligned} \tag{C.2}$$

where the first inequality holds due to Lemma E.1 together with the fact that $\boldsymbol{\Sigma}^{\text{ser}}_{t'_{j-1}} \succeq \boldsymbol{\Sigma}^{\text{ser}}_{T_i}$, and the second inequality follows from (C.1). Now, taking the sum of (C.2) over all round t'_j, we obtain

$$\det(\boldsymbol{\Sigma}^{\text{ser}}_{T_{i+1}-1}) - \det(\boldsymbol{\Sigma}^{\text{ser}}_{T_i}) = \sum_{1 \leq j \leq L} \det(\boldsymbol{\Sigma}^{\text{ser}}_{t'_j}) - \det(\boldsymbol{\Sigma}^{\text{ser}}_{t'_{j-1}}) \geq \sum_{m=1}^{M} (n_m - 1)\alpha \cdot \det(\boldsymbol{\Sigma}^{\text{ser}}_{T_i}).$$

Since $\det(\boldsymbol{\Sigma}_{\text{ser},T_{i+1}-1}) \leq 2\det(\boldsymbol{\Sigma}_{\text{ser},T_i})$, we further have

$$\sum_{j \in M} n_j \leq M + 1/\alpha.$$

Each communication between one agent and the server includes one upload and one download, so the communication complexity within one epoch is bounded by $2(M + 1/\alpha)$. This finishes the proof. □

C.2 Proof for the covariance comparison

Next, we prove the comparison between the covariance matrices given in Lemma 6.5.

Proof of Lemma 6.5. Fix any round $t \in [T]$. Let $t_1 \leq t$ be the last round such that agent m is active at round t_1. If agent m communicated with the server at this round, then we have

$$\lambda \mathbf{I} + \sum_{m'=1}^{M} \mathbf{\Sigma}_{m',t}^{\text{up}} \succeq \mathbf{0} = \frac{1}{\alpha} \mathbf{\Sigma}_{m,t}^{\text{loc}}.$$

Otherwise, according to determinant-based criterion (Line 9) in Algorithm 1, at the end of each round t_1, we have

$$\det\left(\mathbf{\Sigma}_{m,t_1} + \mathbf{\Sigma}_{m,t_1}^{\text{loc}}\right) \leq (1+\alpha)\det(\mathbf{\Sigma}_{m,t_1}).$$

By Lemma E.4, for any non-zero vector $\mathbf{x} \in \mathbb{R}^d$, we have

$$\frac{\mathbf{x}^\top(\mathbf{\Sigma}_{m,t_1} + \mathbf{\Sigma}_{m,t_1}^{\text{loc}})\mathbf{x}}{\mathbf{x}^\top \mathbf{\Sigma}_{m,t_1}\mathbf{x}} \leq \frac{\det(\mathbf{\Sigma}_{m,t_1} + \mathbf{\Sigma}_{m,t_1}^{\text{loc}})}{\det(\mathbf{\Sigma}_{m,t_1})} \leq 1+\alpha.$$

Rearranging the above yields $\mathbf{x}^\top \mathbf{\Sigma}_{m,t_1}^{\text{loc}} \mathbf{x} \leq \alpha \mathbf{x}^\top \mathbf{\Sigma}_{m,t_1}\mathbf{x}$, which then implies that

$$\mathbf{\Sigma}_{m,t_1} \succeq \frac{1}{\alpha}\mathbf{\Sigma}_{m,t_1}^{\text{loc}}$$

Note that $\mathbf{\Sigma}_{m,t_1}$ is the downloaded covariance matrix from last communication before round t_1, so it must satisfy $\mathbf{\Sigma}_{m,t_1} \preceq \mathbf{\Sigma}_{t_1}^{\text{ser}}$. Therefore, we have

$$\lambda \mathbf{I} + \sum_{m'=1}^{M} \mathbf{\Sigma}_{m',t_1}^{\text{up}} = \mathbf{\Sigma}_{t_1}^{\text{ser}} \succeq \mathbf{\Sigma}_{m,t_1} \succeq \frac{1}{\alpha}\mathbf{\Sigma}_{m,t_1}^{\text{loc}}.$$

Now, for round t, since agent m is inactive from round t_1 to t, then we have

$$\lambda \mathbf{I} + \sum_{m'=1}^{M} \mathbf{\Sigma}_{m',t}^{\text{up}} \succeq \lambda \mathbf{I} + \sum_{m'=1}^{M} \mathbf{\Sigma}_{m',t_1}^{\text{up}} \succeq \frac{1}{\alpha}\mathbf{\Sigma}_{m,t_1}^{\text{loc}} = \frac{1}{\alpha}\mathbf{\Sigma}_{m,t}^{\text{loc}},$$

which yields the first claim in Lemma 6.5.

Next, suppose agent m is the only active agent from round t_1 to $t_2 - 1$ and agent m only communicates with the server at round t_1. Further average the above inequality over all agents $m \in [M]$, and we get

$$\lambda \mathbf{I} + \sum_{m'=1}^{M} \mathbf{\Sigma}_{m',t}^{\text{up}} \succeq \frac{1}{M\alpha} \sum_{m'=1}^{M} \mathbf{\Sigma}_{m',t}^{\text{loc}}, \tag{C.3}$$

which implies that for $t_1 + 1 \leq t \leq t_2 - 1$, we have

$$\begin{aligned}
\mathbf{\Sigma}_{m,t} &= \lambda \mathbf{I} + \sum_{m'=1}^{M} \mathbf{\Sigma}_{m',t_1}^{\text{up}} = \lambda \mathbf{I} + \sum_{m'=1}^{M} \mathbf{\Sigma}_{m',t}^{\text{up}} \\
&\succeq \frac{1}{1+M\alpha}\left(\lambda \mathbf{I} + \sum_{m'=1}^{M} \mathbf{\Sigma}_{m',t}^{\text{up}} + \sum_{m'=1}^{M} \mathbf{\Sigma}_{m',t}^{\text{loc}}\right) = \frac{1}{1+M\alpha}\mathbf{\Sigma}_{t}^{\text{all}},
\end{aligned}$$

where the second equation holds due to the fact that no agent communicate with server from round $t_1 + 1$ to $t_2 - 1$, and the inequality follows from (C.3). This yields the second claim in Lemma 6.5 and finishes the proof. □

C.3 Proof of the local concentration for agents

Recall that the global concentration and corresponding global confidence bound have been shown in Lemma 6.3. Next, we establish the concentration properties of the local data on the agents' side.

Proof of Lemma 6.4. For each agent $m \in [M]$ and any rounds $1 \le t_1 \le t_2 \le T$, consider

$$\mathbf{\Sigma}_{m,t_1,t_2} = \alpha\lambda\mathbf{I} + \sum_{i=t_1+1, m_i=m}^{t_2} \mathbf{x}_i\mathbf{x}_i^\top \quad \text{and} \quad \mathbf{u}_{m,t_1,t_2} = \sum_{i=t_1+1, m_i=m}^{t_2} \mathbf{x}_i\eta_i.$$

By Theorem 2 in Abbasi-Yadkori et al. (2011), with probability at least $1-\delta/(T^2M)$, we have

$$\|\mathbf{\Sigma}_{m,t_1,t_2}^{-1}\mathbf{u}_{m,t_1,t_2}\|_{\mathbf{\Sigma}_{m,t_1,t_2}} \le R\sqrt{d\log\Big(\big(1+TL^2/(\alpha\lambda)\big)/\delta\Big)} + \sqrt{\lambda}S.$$

Then taking an union bound over all agent $m \in [M]$ and rounds $1 \le t_1 \le t_2 \le T$ and applying to $t_1 = N_m(t)$ and $t_2 = t$ for each $t \in [T]$, we obtain the desired concentration. This finishes the proof. □

For clarity, we break Lemma 6.6 into two lemmas, Lemma C.1 for local confidence bound and Lemma C.2 for per-round regret, and prove them separately.

Lemma C.1 (Local confidence bound)**.** Under the setting of Theorem 5.1, with probability at least $1-\delta$, for each $t \in [T]$, the estimate $\widehat{\boldsymbol{\theta}}_{m,t+1}$ satisfies that $\|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_{m,t+1}\|_{\mathbf{\Sigma}_{m,t+1}} \le \beta$.

Proof of Lemma C.1. Since the estimated vector $\widehat{\boldsymbol{\theta}}_{m,t+1}$ and covariance matrix $\mathbf{\Sigma}_{m,t+1}$ will keep the same value as in the previous round if the agent m do not communicate with the server, we only need to consider for those round t where agent m communicates with the server. By the determinant-based criterion (Line 9) in Algorithm 1, if the agent m communicates with the server in round t, then at the end of this round, the covariance matrix $\mathbf{\Sigma}_{m,t+1}$ and vector $\mathbf{b}_{m,t+1}$ are given by

$$\mathbf{\Sigma}_{m,t+1} = \lambda\mathbf{I} + \sum_{m'=1}^{M} \mathbf{\Sigma}_{m',N_{m'}(t)}^{\text{up}} = \lambda\mathbf{I} + \sum_{m'=1}^{M} \mathbf{\Sigma}_{m',t}^{\text{up}}, \qquad \mathbf{b}_{m,t+1} = \sum_{m'=1}^{M} \mathbf{b}_{m',t}^{\text{up}}. \tag{C.4}$$

Therefore, the estimated vector $\widehat{\boldsymbol{\theta}}_{m,t+1}$ is

$$\begin{aligned}
\widehat{\boldsymbol{\theta}}_{m,t+1} &= \Big(\lambda\mathbf{I} + \sum_{m'=1}^{M} \mathbf{\Sigma}_{m',t}^{\text{up}}\Big)^{-1} \sum_{m'=1}^{M} \mathbf{b}_{m',t}^{\text{up}} \\
&= \Big(\lambda\mathbf{I} + \sum_{m'=1}^{M} \mathbf{\Sigma}_{m',t}^{\text{up}}\Big)^{-1} \sum_{m'=1}^{M} (\mathbf{\Sigma}_{m',t}^{\text{up}}\boldsymbol{\theta}^* + \mathbf{u}_{m',t}^{\text{up}}) \\
&= \boldsymbol{\theta}^* - \lambda\Big(\lambda\mathbf{I} + \sum_{m'=1}^{M} \mathbf{\Sigma}_{m',t}^{\text{up}}\Big)^{-1}\boldsymbol{\theta}^* + \Big(\lambda\mathbf{I} + \sum_{m'=1}^{M} \mathbf{\Sigma}_{m',k}^{\text{up}}\Big)^{-1} \sum_{m'=1}^{M} \mathbf{u}_{m',t}^{\text{up}} \\
&= \boldsymbol{\theta}^* - \lambda(\mathbf{\Sigma}_{m,t+1})^{-1}\boldsymbol{\theta}^* + \sum_{m'=1}^{M} (\mathbf{\Sigma}_{m,t+1})^{-1}\mathbf{u}_{m',t}^{\text{up}}.
\end{aligned}$$

Thus, the difference between $\widehat{\boldsymbol{\theta}}_{m,t+1}$ and the underlying truth $\boldsymbol{\theta}^*$ can be decomposed as

$$\begin{aligned}
\big\|\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_{m,t+1}\big\|_{\mathbf{\Sigma}_{m,t+1}} &\le \big\|\lambda(\mathbf{\Sigma}_{m,t+1})^{-1}\boldsymbol{\theta}^*\big\|_{\mathbf{\Sigma}_{m,t+1}} + \Big\|\sum_{m'=1}^{M} (\mathbf{\Sigma}_{m,t+1})^{-1}\mathbf{u}_{m',t}^{\text{up}}\Big\|_{\mathbf{\Sigma}_{m,t+1}} \\
&\le \sqrt{\lambda}\|\boldsymbol{\theta}^*\|_2 + \Big\|\sum_{m'=1}^{M} (\mathbf{\Sigma}_{m,t+1})^{-1}\mathbf{u}_{m',t}^{\text{up}}\Big\|_{\mathbf{\Sigma}_{m,t+1}},
\end{aligned} \tag{C.5}$$

where the first inequality holds due to that fact that $\|\mathbf{a}+\mathbf{b}\|_{\mathbf{\Sigma}} \le \|\mathbf{a}\|_{\mathbf{\Sigma}} + \|\mathbf{b}\|_{\mathbf{\Sigma}}$ and the second inequality follows from $\mathbf{\Sigma}_{m,t+1} \ge \lambda\mathbf{I}$. By the assumption that $\|\boldsymbol{\theta}^*\|_2 \le S$, the first term can be controlled by $\sqrt{\lambda}S$. For the second term in (C.5), consider the following decomposition:

$$\begin{aligned}
\sum_{m'=1}^{M} (\mathbf{\Sigma}_{m,t+1})^{-1}\mathbf{u}_{m',t}^{\text{up}} &= \sum_{m'=1}^{M} (\mathbf{\Sigma}_{m,t+1})^{-1}\big(\mathbf{u}_{m',t}^{\text{up}} + \mathbf{u}_{m',t}^{\text{loc}}\big) - \sum_{m'=1}^{M} (\mathbf{\Sigma}_{m,t+1})^{-1}\mathbf{u}_{m',t}^{\text{loc}} \\
&= \underbrace{(\mathbf{\Sigma}_{m,t+1})^{-1}\mathbf{u}_t^{\text{all}}}_{\mathcal{A}} - \sum_{m'=1}^{M} \underbrace{(\mathbf{\Sigma}_{m,t+1})^{-1}\mathbf{u}_{m',t}^{\text{loc}}}_{B_{m'}}.
\end{aligned} \tag{C.6}$$

For the term $\mathcal{A}$, it follows from (6.5) in Lemma 6.5 that

$$\begin{aligned}\big\|(\boldsymbol{\Sigma}_{m,t+1})^{-1}\mathbf{u}_t^{\text{all}}\big\|_{\boldsymbol{\Sigma}_{m,t+1}} &= \big\|(\boldsymbol{\Sigma}_{m,t+1})^{-1/2}\mathbf{u}_t^{\text{all}}\big\|_2 \\ &\le \sqrt{1+M\alpha}\cdot\big\|(\boldsymbol{\Sigma}_t^{\text{all}})^{-1/2}\mathbf{u}_t^{\text{all}}\big\|_2 \\ &\le \sqrt{1+M\alpha}\cdot\Big(R\sqrt{d\log\big((1+TL^2/\lambda)/\delta\big)}+\sqrt{\lambda}S\Big), \end{aligned} \tag{C.7}$$

where the second inequality holds due to Lemma 6.3. Next, for each term $B_{m'}$ in (C.6), by (6.4) in Lemma 6.5, we have

$$\lambda\mathbf{I}+\sum_{j=1}^{M}\boldsymbol{\Sigma}_{j,t}^{\text{up}}\succeq\frac{1}{\alpha}\boldsymbol{\Sigma}_{m',t}^{\text{loc}},$$

which further implies that

$$\lambda\mathbf{I}+\sum_{j=1}^{M}\boldsymbol{\Sigma}_{j,t}^{\text{up}}\succeq\frac{1}{2\alpha}(\alpha\lambda\mathbf{I}+\boldsymbol{\Sigma}_{m',t}^{\text{loc}}). \tag{C.8}$$

Thus, the norm of each term $B_{m'}$ can be bounded as

$$\begin{aligned}\big\|(\boldsymbol{\Sigma}_{m,t+1})^{-1}\mathbf{u}_{m',t}^{\text{loc}}\big\|_{\boldsymbol{\Sigma}_{m,t+1}} &= \big\|(\boldsymbol{\Sigma}_{m,t+1})^{-1/2}\mathbf{u}_{m',t}^{\text{loc}}\big\|_2 \\ &\le \sqrt{2\alpha}\cdot\Big\|\big(\alpha\lambda\mathbf{I}+\boldsymbol{\Sigma}_{m',t}^{\text{loc}}\big)^{-1/2}\mathbf{u}_{m',t}^{\text{loc}}\Big\|_2 \\ &\le \sqrt{2\alpha}\cdot\Big(R\sqrt{d\log\frac{\alpha\lambda+TL^2}{\alpha\lambda\delta}}+\sqrt{\lambda}S\Big), \end{aligned} \tag{C.9}$$

where the first inequality holds due to (C.8) and the second inequality follows from Lemma 6.4.

Finally, combining (C.5), (C.6), (C.7) and (C.9), we obtain

$$\big\|\boldsymbol{\theta}^*-\widehat{\boldsymbol{\theta}}_{m,t+1}\big\|_{\boldsymbol{\Sigma}_{m,t+1}}\le\sqrt{\lambda}S+\big(\sqrt{1+M\alpha}+M\sqrt{2\alpha}\big)\Big(R\sqrt{d\log\frac{\min(\alpha,1)\cdot\lambda+TL^2}{\min(\alpha,1)\cdot\lambda\delta}}+\sqrt{\lambda}S\Big).$$

Thus we finish the proof of Lemma C.1. □

Lemma C.2 (Per-round regret)**.** Under the setting of Theorem 5.1, with probability at least $1-\delta$, for each $t\in[T]$, the regret in round t satisfies

$$\Delta_t=\max_{\mathbf{x}\in\mathcal{D}_t}\langle\boldsymbol{\theta}^*,\mathbf{x}\rangle-\langle\boldsymbol{\theta}^*,\mathbf{x}_t\rangle\le 2\beta\sqrt{\mathbf{x}_t^\top\boldsymbol{\Sigma}_{m_t,t}^{-1}\mathbf{x}_t}.$$

Proof of Lemma C.2. First, by Lemma C.1, with probability at least $1-\delta$, for each round $t\in[T]$ and each action $\mathbf{x}\in\mathcal{D}_t$, we have

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_{m,t}^\top\mathbf{x}+\beta\sqrt{\mathbf{x}^\top\boldsymbol{\Sigma}_{m_t,t}^{-1}\mathbf{x}}-(\boldsymbol{\theta}^*)^\top\mathbf{x} &= (\widehat{\boldsymbol{\theta}}_{m,t}-\boldsymbol{\theta}^*)^\top\mathbf{x}+\beta\sqrt{\mathbf{x}^\top\boldsymbol{\Sigma}_{m_t,t}^{-1}\mathbf{x}} \\ &\ge -\|\widehat{\boldsymbol{\theta}}_{m,t}-\boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_{m_t,t}}\cdot\|\mathbf{x}\|_{\boldsymbol{\Sigma}_{m_t,t}^{-1}}+\beta\sqrt{\mathbf{x}^\top\boldsymbol{\Sigma}_{m_t,t}^{-1}\mathbf{x}} \\ &\ge -\beta\|\mathbf{x}\|_{\boldsymbol{\Sigma}_{m_t,t}^{-1}}+\beta\sqrt{\mathbf{x}^\top\boldsymbol{\Sigma}_{m_t,t}^{-1}\mathbf{x}} \\ &= 0, \end{aligned} \tag{C.10}$$

where the first inequality holds due to the Cauchy-Schwarz inequality and the last inequality follows from Lemma C.1. (C.10) shows that the estimator for agent m_t is always optimistic. For simplicity,

we denote the optimal action at round t as $\mathbf{x}^* = \arg\max_{\mathbf{x}\in\mathcal{D}_t} (\boldsymbol{\theta}^*)^\top \mathbf{x}$, and (C.10) further implies

$$\begin{aligned}
\Delta_t &= (\boldsymbol{\theta}^*)^\top \mathbf{x}^* - (\boldsymbol{\theta}^*)^\top \mathbf{x}_t \\
&\leq \widehat{\boldsymbol{\theta}}_{m,t}^\top \mathbf{x}^* + \beta\sqrt{(\mathbf{x}^*)^\top \boldsymbol{\Sigma}_{m_t,t}^{-1} \mathbf{x}^*} - (\boldsymbol{\theta}^*)^\top \mathbf{x}_t \\
&\leq \widehat{\boldsymbol{\theta}}_{m,t}^\top \mathbf{x}_t + \beta\sqrt{\mathbf{x}_t^\top \boldsymbol{\Sigma}_{m_t,t}^{-1} \mathbf{x}_t} - (\boldsymbol{\theta}^*)^\top \mathbf{x}_t \\
&= (\widehat{\boldsymbol{\theta}}_{m,t} - \boldsymbol{\theta}^*)^\top \mathbf{x}_t + \beta\sqrt{\mathbf{x}_t^\top \boldsymbol{\Sigma}_{m_t,t}^{-1} \mathbf{x}_t} \\
&\leq \|\widehat{\boldsymbol{\theta}}_{m,t} - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_{m_t,t}} \cdot \|\mathbf{x}_k\|_{\boldsymbol{\Sigma}_{m_t,t}^{-1}} + \beta\sqrt{\mathbf{x}_t^\top \boldsymbol{\Sigma}_{m_t,t}^{-1} \mathbf{x}_t} \\
&\leq 2\beta\sqrt{\mathbf{x}_t^\top \boldsymbol{\Sigma}_{m_t,t}^{-1} \mathbf{x}_t},
\end{aligned}$$

where the first inequality holds due to (C.10), the second inequality follows from the definition of action $\mathbf{x}_t$ in Algorithm 1, the third inequality applies the Cauchy-Schwarz inequality, and the last inequality is by Lemma C.1. Thus, we finish the proof of Lemma C.2. □

Combining Lemmas C.1 and C.2 yields Lemma 6.6

D Proof for Lower Bound

Lemma D.1 (Theorem 3 in Abbasi-Yadkori et al. 2011)**.** There exists a constant $C > 0$, such that for any normalized linear bandit instance with $R = L = S = 1$, the expectation of the regret for OFUL algorithm is upper bounded by $\mathbb{E}[\text{Regret}(T)] \leq Cd\sqrt{T}\log T$.

Lemma D.2 (Theorem 24.1 in Lattimore and Szepesvári 2020)**.** There exists a set of hard-to-learn normalized linear bandit instances with $R = L = S = 1$, such that for any algorithm **Alg** and $T \geq d$, for a uniformly random instance in the set, the regret is lower bounded by $\mathbb{E}[\text{Regret}(T)] \geq cd\sqrt{T}$ for some constant $c > 0$.

Theorem 5.3 is an extension of the lower bound result in Wang et al. (2019, Theorem 2) from multi-arm bandits to linear bandits.

Proof of Theorem 5.3. For any algorithm **Alg** for federated bandits, we construct the auxiliary **Alg1** as follows: For each agent $m \in [M]$, it performs **Alg** until there is a communication between the agent m and the server (upload or download data). After the communication, the agent m remove all previous information and perform the `OFUL` Algorithm in Abbasi-Yadkori et al. (2011). In this case, for each agent $m \in [M]$, **Alg1** do not utilize any information from other agents and it will reduce to a single agent bandit algorithm.

Now, we uniformly randomly select a hard-to-learn instance from the set given in Lemma D.2, and let each agent $m \in [M]$ be active for T/M different rounds (where we assume T/M is an integer for simplicity). Since **Alg1** reduces to a single agent bandit algorithm, Lemma D.2 implies that the expected regret for agent m with **Alg1** is lower bounded by

$$\mathbb{E}[\text{Regret}_{m,\textbf{Alg1}}] \geq cd\sqrt{T/M}. \tag{D.1}$$

Taking the sum of (D.1) over all agents $m \in [M]$, we obtain

$$\mathbb{E}[\text{Regret}_{\textbf{Alg1}}] = \sum_{m=1}^{M} \mathbb{E}[\text{Regret}_{m,\textbf{Alg1}}] \geq cd\sqrt{MT}. \tag{D.2}$$

For each agent $m \in [M]$, let δ_m denote the probability that agent m will communicate with the server. Notice that before the communication, **Alg1** has the same performance as **Alg**, while for the rounds after the communication, **Alg1** executes the `OFUL` algorithm and Lemma D.1 suggests an $O(d\sqrt{T/M}\log(T/M))$ upper bounded for the expected regret. Therefore, the expected regret for agent m with **Alg1** is upper bounded by

$$\mathbb{E}[\text{Regret}_{m,\textbf{Alg1}}] \leq \mathbb{E}[\text{Regret}_{m,\textbf{Alg}}] + \delta_m Cd\sqrt{T/M}\log(T/M). \tag{D.3}$$

Taking the sum of (D.3) over all agents $m \in [M]$, we obtain

$$\begin{aligned}\mathbb{E}[\text{Regret}_{\textbf{Alg}}(T)] &= \sum_{m=1}^{M} \mathbb{E}[\text{Regret}_{m,\textbf{Alg1}}] \\ &\leq \sum_{m=1}^{M} \mathbb{E}[\text{Regret}_{m,\textbf{Alg}}] + \bigg(\sum_{m=1}^{M} \delta_m\bigg) Cd\sqrt{T/M}\log(T/M) \\ &= \mathbb{E}[\text{Regret}_{\textbf{Alg}}] + \delta Cd\sqrt{T/M}\log(T/M), \end{aligned} \tag{D.4}$$

where $\delta = \sum_{m=1}^{M} \delta_m$ is the expected communication complexity. Combining (D.2) and (D.4), for any algorithm **Alg** with communication complexity $\delta \leq c/(2C) \cdot M/\log(T/M) = O(M/\log(T/M))$, we have

$$\mathbb{E}[\text{Regret}_{\textbf{Alg}}] \geq cd\sqrt{MT} - \delta Cd\sqrt{T/M}\log(T/M) \geq cd\sqrt{MT}/2 = \Omega(d\sqrt{MT}).$$

This finishes the proof of Theorem 5.3. □

E Auxiliary Lemmas

To make the analysis self-contained in this paper, here we include the auxiliary lemmas that have been previously used.

Lemma E.1 (Lemma 2.2 in Tie et al. 2011)**.** For any positive semi-definite matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$, it holds that $\det(\mathbf{A}+\mathbf{B}+\mathbf{C}) + \det(\mathbf{A}) \geq \det(\mathbf{A}+\mathbf{B}) + \det(\mathbf{A}+\mathbf{C})$.

Lemma E.2 (Lemma 2.3 in Tie et al. 2011)**.** For any positive semi-definite matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$, it holds that $\det(\mathbf{A}+\mathbf{B}+\mathbf{C})\det(\mathbf{A}) \leq \det(\mathbf{A}+\mathbf{B})\det(\mathbf{A}+\mathbf{C})$.

Theorem E.3 (Theorem 2 in Abbasi-Yadkori et al. 2011)**.** Suppose $\{\mathcal{F}_t\}_{t=0}^{\infty}$ is a filtration. Let $\{\eta_t\}_{t=1}^{\mathbb{R}}$ be a stochastic process in $\mathbb{R}$ such that η_t is $\mathcal{F}_t$-measurable and R-sub-Gaussian conditioning on $\mathcal{F}_{t-1}$, i.e, for any $c > 0$,

$$\mathbb{E}\left[\exp\left(c\eta_t\right)|\mathcal{F}_{t-1}\right] \leq \exp\left(\frac{c^2R^2}{2}\right).$$

Let $\{\mathbf{x}_t\}_{t=1}^{\infty}$ be a stochastic process in $\mathbb{R}^d$ such that $\mathbf{x}_t$ is $\mathcal{F}_{t-1}$-measurable and $\|\mathbf{x}_t\|_2 \leq L$. Let $y_t = \langle \mathbf{x}_t, \boldsymbol{\theta}^* \rangle + \eta_t$ for some $\boldsymbol{\theta}^* \in \mathbb{R}^d$ s.t. $\|\boldsymbol{\theta}^*\|_2 \leq S$. For any $t \geq 1$, define

$$\boldsymbol{\Sigma}_t = \lambda\mathbf{I} + \sum_{i=1}^{t} \mathbf{x}_t\mathbf{x}_t^{\top}, \quad \text{and} \quad \widehat{\boldsymbol{\theta}}_t = \boldsymbol{\Sigma}_t^{-1}\sum_{i=1}^{t}\mathbf{x}_i y_i,$$

for some $\lambda > 0$. Then for any $\delta > 0$, with probability at least $1-\delta$, for all t, we have

$$\|\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*\|_{\boldsymbol{\Sigma}_t} \leq R\sqrt{d\log\left(\frac{1+tL^2/\lambda}{\delta}\right)} + \sqrt{\lambda}S.$$

Lemma E.4 (Lemma 12 in Abbasi-Yadkori et al. (2011))**.** Suppose $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d\times d}$ are two positive definite matrices satisfying that $\mathbf{A} \succeq \mathbf{B}$, then for any $\mathbf{x} \in \mathbb{R}^d$, $\|\mathbf{x}\|_{\mathbf{A}} \leq \|\mathbf{x}\|_{\mathbf{B}} \cdot \sqrt{\det(\mathbf{A})/\det(\mathbf{B})}$.