Distributed Contextual Linear Bandits with Minimax Optimal Communication Cost

Sanae Amani [1] Tor Lattimore [2] András György [2] Lin F. Yang [1]

Abstract

We study distributed contextual linear bandits with stochastic contexts, where N agents act cooperatively to solve a linear bandit-optimization problem with d-dimensional features over the course of T rounds. For this problem, we derive the first ever information-theoretic lower bound $\Omega(dN)$ on the communication cost of any algorithm that performs optimally in a regret minimization setup. We then propose a distributed batch elimination version of the LinUCB algorithm, DisBE-LUCB, where the agents share information among each other through a central server. We prove that the communication cost of DisBE-LUCB matches our lower bound up to logarithmic factors. In particular, for scenarios with known context distribution, the communication cost of DisBE-LUCB is only $\tilde{\mathcal{O}}(dN)$ and its regret is $\tilde{\mathcal{O}}(\sqrt{dNT})$, which is of the same order as that incurred by an optimal single-agent algorithm for NT rounds. We also provide similar bounds for practical settings where the context distribution can only be estimated. Therefore, our proposed algorithm is nearly minimax optimal in terms of *both regret and communication cost*. Finally, we propose DecBE-LUCB, a fully decentralized version of DisBE-LUCB, which operates without a central server, where agents share information with their *immediate neighbors* through a carefully designed consensus procedure.

1. Introduction

In the contextual bandit problem, a learning agent repeatedly makes decisions based on contextual information, with the goal of learning a policy that maximizes their total reward over time. This model captures simple reinforcement learning tasks in which the agent must learn to make high-quality decisions in an uncertain environment, but does not need to engage in long-term planning. Contextual bandit algorithms are deployed in online personalization systems such as medical trials and product recommendation in e-commerce (Agarwal et al., 2016; Tewari and Murphy, 2017). For example, by modelling personalized recommendation of articles as a contextual bandit problem, a learning algorithm sequentially selects articles to be recommended to users based on contextual information about the users and articles, while continuously updating its article-selection strategy based on user-click feedback to maximize total user clicks (Li et al., 2010).

Distributed cooperative learning is a paradigm where multiple agents collaboratively learn a shared prediction model. More recently, researchers have explored the potential of contextual bandit algorithms in distributed systems, such as in robotics, wireless networks, the power grid and medical trials (Li et al., 2013; Avner and Mannor, 2019; Berkenkamp et al., 2016; Sui et al., 2018). For example, in sensor/wireless networks (Avner and Mannor, 2019) and channel selection in radio networks (Liu and Zhao, 2010a;b;c), a collaborative behavior is required for decision-makers/agents to select better actions as individuals.

While a distributed nature is inherent in certain systems, distributed solutions might also be preferred in broader settings, as they can lead to speed-ups of the learning process. This calls for extensions of the traditional single-agent bandit setting to networked systems. In addition to speeding up the learning process, another desirable goal of each distributed learning algorithm is *communication efficiency*. In particular, keeping the communication as rare as possible in collaborative learning is of importance. The notion of communication efficiency in distributed learning paradigms is directly related to the issue of efficient environment queries made in single-agent settings. In many practical single-agent scenarios, where the agent sequentially makes active queries about the environment, it is desirable to limit these queries to a small number of rounds of interaction, which helps to increase the parallelism of the learning process and reduce the management cost. In recent years, to address

[*]Equal contribution [1]Department of Electrical and Computer Engineering, University of California, Los Angeles. [2]DeepMind, London. Correspondence to: Sanae Amani <samani@ucla.edu>, Tor Lattimore <lattimore@google.com>, András György <agyorgy@deepmind.com>, Lin F. Yang <linyang@ee.ucla.edu>.

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

such scenarios, a surge of research activity in the area of batch online learning has shown that in many popular online learning tasks, a very small number of batches may achieve minimax optimal learning performance, and therefore it is possible to enjoy the benefits of both adaptiveity and parallelism (Ruan et al., 2021; Han et al., 2020; Gao et al., 2019). In light of the connection between communication cost in distributed settings and the number of environment queries in single-agent settings, a careful use of batch learning methods in multi-agent learning scenarios may positively affect the communication efficiency by limiting the number of necessary communication rounds. In this paper, we first prove an information-theoretic lower bound on the communication cost of distributed contextual linear bandits, and then leverage such batch learning methods to design an algorithm with a small communication cost that matches this lower bound while guaranteeing optimal regret.

Notation. Throughout this paper, we use lower-case letters for scalars, lower-case bold letters for vectors, and upper-case bold letters for matrices. The Euclidean norm of $\mathbf{x}$ is denoted by $\|\mathbf{x}\|_2$. We denote the transpose of any column vector $\mathbf{x}$ by $\mathbf{x}^\top$. For any vectors $\mathbf{x}$ and $\mathbf{y}$, we use $\langle \mathbf{x}, \mathbf{y}\rangle$ to denote their inner product. Let $\mathbf{A}$ be a positive semi-definite $d \times d$ matrix and $\boldsymbol{\nu} \in \mathbb{R}^d$. The weighted 2-norm of $\boldsymbol{\nu}$ with respect to $\mathbf{A}$ is defined by $\|\boldsymbol{\nu}\|_{\mathbf{A}} = \sqrt{\boldsymbol{\nu}^\top \mathbf{A} \boldsymbol{\nu}}$. For a positive integer n, $[n]$ denotes the set $\{1, 2, \ldots, n\}$, while for positive integers $m \leq n$, $[m : n]$ denotes the set $\{m, m+1, \ldots, n\}$. For square matrices $\mathbf{A}$ and $\mathbf{B}$, we use $\mathbf{A} \preceq \mathbf{B}$ to denote $\mathbf{B} - \mathbf{A}$ is positive semi-definite. We denote the minimum and maximum eigenvalues of $\mathbf{A}$ by $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$. We use $\mathbf{e}_i$ to denote the i-th standard basis vector. $I(X; Y)$ denotes the mutual information between two random variables X and Y. Finally, we use standard $\tilde{\mathcal{O}}$ notation for big-O notation that ignores logarithmic factors.

1.1. Problem formulation

We consider a network of N agents acting cooperatively to efficiently solve a K-armed stochastic linear bandit problem. Let T be the total number of rounds. At each round $t \in [T]$, each agent i is given a decision set $\mathcal{X}_t^i = \{\mathbf{x}_{t,a}^i : a \in [K]\} \subset \mathbb{R}^d$, drawn independently from a distribution $\mathcal{D}_t^i$. We assume that $\mathcal{D}_t^i = \mathcal{D}$ for all $(i, t) \in [N] \times [T]$. Here, $\mathbf{x}_{t,a}^i$ is a mapping from action a and the contextual information agent i receives at round t to the d-dimensional space. We call $\mathbf{x}_{t,a}^i$ the feature vector associated with action a and agent i at round t. Agent i selects action $a_{i,t} \in [K]$, and observes the reward $y_t^i = \langle \boldsymbol{\theta}, \mathbf{x}_{t,a_{i,t}}^i\rangle + \eta_t^i$, where $\boldsymbol{\theta} \in \mathbb{R}^d$ is an unknown vector and η_t^i is an independent zero-mean additive noise. The agents are also allowed to communicate with each other. Both the action selection and the communicated information of each agent may only depend on previously played actions, observed rewards, decision sets, and communication received from other agents. Throughout the paper, we rely on the following assumption.

Assumption 1. *Without loss of generality,* $\|\boldsymbol{\theta}\|_2 \leq 1$, $\|x_{t,a}^i\|_2 \leq 1$, $|y_t^i| \leq 1$ *for all* $(a, i, t) \in [K] \times [N] \times [T]$. *Also, the distribution* $\mathcal{D}$ *is known to the agents.*

The boundedness assumption is standard in the linear bandit literature (Chu et al., 2011; Dani et al., 2008; Huang et al., 2021). Moreover, our results can be readily extended to the settings where the assumption on the boundedness of y_t^i is relaxed by assuming the noise variables η_t^i are conditionally σ-subGaussiam for a constant $\sigma \geq 0$. As such, a high probability bound on η_t^i and consequently y_t^i can be established, which is desired in our analysis for establishing confidence intervals in Appendix B.1.

Our assumption on the knowledge of $\mathcal{D}$ is fairly well-motivated. A standard argument is based on having loads of unsupervised data in real-world scenarios. For example, Google, Amazon, Netflix, etc, have collected massive amounts of data about users, products, and queries, sufficiently describing the joint distributions. Given this, even if the features change (for a given user or product, etc.), their distributions can be computed/sampled from as the features are computed via a deterministic feature map. In light of this, Hanna et al. (2022a) recently studied contextual linear bandits with known context distribution. We further relax this assumption in Remark 4.3 in Section 4.2.

Goal. The performance of the network is measured via the cumulative regret of all agents in T rounds, defined as

$$R_T := \mathbb{E}[\textstyle\sum_{t=1}^T \sum_{i=1}^N \langle \boldsymbol{\theta}, \mathbf{x}_{*,t}^i\rangle - \langle \boldsymbol{\theta}, \mathbf{x}_t^i\rangle], \qquad (1)$$

where the expectation is taken over the random variables $\mathcal{X}_t^i, (i, t) \in [N] \times [T]$ with joint distribution $\bigotimes_{i,t=1}^{N,T} \mathcal{D}_t^i$, $\mathbf{x}_t^i$ and $\mathbf{x}_{*,t}^i \in \arg\max_{\mathbf{x} \in \mathcal{X}_t^i} \langle \boldsymbol{\theta}, \mathbf{x}\rangle$ are the feature vectors associated with the action chosen by agent i at round t and the best possible action, respectively.

For simplicity, in our algorithms the communication cost is measured as the number of communicated real numbers *over the course of* T *rounds*. In Section 3, we also discuss variants of our methods where the communication cost is measured as the number of communicated bits.

The goal is to design a distributed collaborative algorithm that minimizes the cumulative regret, while maintaining an efficient coordination protocol with a small communication cost. Specifically, we wish to achieve a regret close to $\tilde{\mathcal{O}}(\sqrt{dNT})$ that is incurred by an optimal *single-agent algorithm for* NT *rounds* (the total number of arm pulls) while the communication cost is $\tilde{\mathcal{O}}(dN)$ with only a mild (logarithmic) dependence on T.

A motivating example. In news article recommendation, the candidate actions correspond to K news articles. At round t, an individual user visits an online news platform that has N servers employing the same recommender systems to recommend news articles from an article pool. The

Setting	Algorithm	Regret	Communication cost	Communication cost lower bound
Contexts are fixed over time horizon and agents	DELB with server (Wang et al., 2019)	$\mathcal{O}\left(d\sqrt{NT}\log T\right)$	$\mathcal{O}\left((dN + d\log\log d)\log T\right)$	
Contexts adversarially vary over time horizon and agents	DisLinUCB with server (Wang et al., 2019)	$\mathcal{O}\left(d\sqrt{NT}\log^2 T\right)$	$\mathcal{O}\left(d^3 N^{1.5}\right)$	
	FedUCB with server (Dubey and Pentland, 2020)	$\mathcal{O}\left(d\sqrt{NT}\log^2 T\right)$	$\mathcal{O}\left(d^3 N^{1.5}\right)$	
Contexts adversarially vary over agents	Fed-PE with server (Huang et al., 2021)	$\mathcal{O}\left(\sqrt{dNT\log(KNT)}\right)$	$\mathcal{O}\left((d^2 + dK)N\log T\right)$	
Contexts stochastically vary over time horizon and agents (**this work**)	DisBE-LUCB with server	$\mathcal{O}\left(\sqrt{dNT\log d\log^2(KNT)}\right)$	$\mathcal{O}\left(dN\log\log(NT)\right)$	$\Omega(dN)$
	DecBE-LUCB without server	$\mathcal{O}\left(NS + \sqrt{dN(T+S)\log d\log^2(KNT)}\right)$	$\mathcal{O}\left(S\delta_{\max} dN\log\log(NT)\right)$	

Table 1. N: number of agents; K: number of arms; T: time horizon; d: dimension of the feature vectors; $S = \frac{\log(dN)}{\sqrt{1/|\lambda_2|}}$;$|\lambda_2|$: the second largest eigenvalue of communication matrix in absolute value; $\delta_{\max}$ is the maximum degree of the graph representing agents' network. The lower bound for the communication cost is interpreted as follows: For any algorithm with expected communication cost less than $\frac{dN}{64}$, there exists a contextual linear bandit instance with stochastic contexts, for which the algorithm's regret is $\Omega(N\sqrt{dT})$. See Theorem 3.1.

contextual information of the user, the articles and the servers at round t is modeled by $\mathcal{X}_t^i = \{\mathbf{x}_{t,a}^i : a \in [K]\}$, characterizing user's reaction to each recommended article a (e.g., click/not click) by server i, and the probability of clicking on a is modeled by $\langle \boldsymbol{\theta}, \mathbf{x}_{t,a}^i \rangle$, which corresponds to the expected reward. On the distributed side, these N servers collaborate with each other by sharing information about the feedback they receive from the users after recommending articles in an attempt to speed up learning the users' preferences. In this example, the individual users and articles can often be viewed as independent samples from the population which is characterized by distribution $\mathcal{D}$.

1.2. Contributions

We establish a lower bound on the communication cost of distributed contextual linear bandits. We propose algorithms with optimal regret and communication cost matching our lower bound (up to logarithmic factors) and growing linearly with d and N while those of previous best-known algorithms scale super linearly either in d or N. Below, we elaborate more on our contributions:

Minimax lower bound for the communication cost. As our main technical contribution, in Section 3, we prove the first information-theoretic lower bound on the communication cost (measured in bits) of any algorithm achieving an optimal regret rate for the distributed contextual linear bandit problem with stochastic contexts. In particular, we prove that for any distributed algorithm with expected communication cost less than $\frac{dN}{64}$, there exists a contextual linear bandit problem instance with stochastic contexts for which the algorithm's regret is $\Omega(N\sqrt{dT})$.

DisBE-LUCB. We propose a distributed batch elimination contextual linear bandit algorithm (DisBE-LUCB): the time steps are grouped into M pre-defined batches and at each time step, each agent first constructs confidence intervals for each action's reward, and the actions whose confidence intervals completely fall below those of other actions are eliminated. Throughout each batch, each agent uses the same policy to select actions from the surviving action sets. At the end of each batch, the agents share information through a central server and update the policy they use in the next batch. We prove that while the communication cost of DisBE-LUCB is only $\tilde{\mathcal{O}}(dN)$, it achieves a regret $\tilde{\mathcal{O}}(\sqrt{dNT})$, which is of the same order as that incurred by a near optimal *single-agent algorithm for NT rounds* . This shows that DisBE-LUCB is nearly minimax optimal in terms of *both regret and communication cost*. We highlight that while DisBE-LUCB is inspired by the single-agent batch elimination style algorithms (Ruan et al., 2021) in an attempt to save on communication as much as possible, a direct use of confidence intervals used in such algorithms would fail to guarantee optimal communication cost $\tilde{\mathcal{O}}(dN)$ and require more communication by a factor of $\mathcal{O}(d)$. We address this issue by introducing new confidence intervals in Lemma 4.4. Details are given in Section 4.

DecBE-LUCB. Finally, we propose a fully decentralized variant of DisBE-LUCB without a central server, where the agents can only communicate with their *immediate neighbors* given by a communication graph. Our algorithm, called decentralized batch elimination linear UCB (DecBE-LUCB), runs a carefully designed consensus procedure to spread information throughout the network. For this algorithm, we prove a regret bound that captures both the degree of selected actions' optimality and the inevitable delay in information-sharing due to the network structure while the communication cost still grows linearly with d and N. See Section 4.4.

We complement our theoretical results with numerical simulations under various settings in Section 5.

2. Related Work

Distributed MAB. Multi-armed bandit (MAB) in multi-agent distributed settings has received attention from several academic communities. In the context of the classical K-armed MAB, Martínez-Rubio et al. (2019); Landgren et al. (2016a;b; 2018) proposed decentralized algorithms for a network of N agents that can share information only with their immediate neighbors, while Szörényi et al. (2013) studied the MAB problem on peer-to-peer networks.

Distributed contextual linear bandits. The most closely related works on distributed linear bandits are those of Wang et al. (2019); Dubey and Pentland (2020); Huang et al. (2021); Korda et al. (2016); Hanna et al. (2022b). In particular, Wang et al. (2019) investigate communication-efficient distributed linear bandits, where the agents can communicate with a server by sending and receiving packets. They propose two algorithms, namely, DELB and DisLinUCB, for fixed and time-varying action sets, respectively. The works of Dubey and Pentland (2020); Huang et al. (2021) consider the federated linear contextual bandit model and the former focuses on federated differential privacy. In the latter, the contexts denote the specifics of the agents and are different but fixed during the entire time horizon for each agent. In the former, however, the contexts contain the information about both the environment and the agents, in the sense that contexts associated with different agents are different and vary during the time horizon. To put these in the context of an example, consider a recommender system. Both Dubey and Pentland (2020) and Huang et al. (2021) consider a multi-agent model, where each agent is associated with a different user profile. Huang et al. (2021) fix a user profile for an agent, while Dubey and Pentland (2020) consider a time-varying user profile. Therefore, Huang et al. (2021) capture the variation of contexts over agents, whereas it is captured over both agents and time horizon in Dubey and Pentland (2020). A regret and communication cost comparison between DisBE-LUCB, DecBE-LUCB and other baseline algorithms is given in Table 1.

Batch elimination in distributed bandits. An important line of work related to communication efficiency in distributed bandits studies practical single-agent scenarios using batch elimination methods, in which a very small number of batches achieve minimax optimal learning performance (Ruan et al., 2021; Han et al., 2020; Gao et al., 2019). Our proposed algorithms are inspired by the single-agent BatchLinUCB-DG proposed in Ruan et al. (2021) in an attempt to save on communication as much as possible. That said, a direct use of confidence intervals in Ruan et al. (2021) would fail to guarantee optimal communication cost $\tilde{\mathcal{O}}(dN)$ and require more communication by a factor of $\mathcal{O}(d)$. We address this issue by introducing new confidence intervals, used in our algorithms, in Lemma 4.4.

Minimax lower bound on communication cost. We are unaware of any lower bound on the communication cost scaling with both d and N for contextual linear bandits in the distributed/federated learning setting. To the best of our knowledge, our work is the first to establish such a minimax lower bound and to propose algorithms with optimal regret and communication cost matching this lower bound up to logarithmic factors. Recently, Li et al. (2022) proved a $\Omega(N)$ communication lower bound for asynchronous federated contextual linear bandits. However, their lower bound does not include the dependency on d, which is of importance in our work and emphasizes how our proposed algorithm optimally improves the communication cost of existing methods. In addition, Wang et al. (2019) previously proved a $\Omega(N)$ communication lower bound for distributed MAB.

3. Lower Bound on Communication Cost

In this section, we derive an information-theoretic lower bound on the communication cost of the distributed contextual linear bandits with stochastic contexts. In particular, we prove that for any distributed contextual linear bandit algorithm with stochastic contexts that achieves the optimal regret rate $\tilde{\mathcal{O}}(\sqrt{dNT})$, the expected amount of communication must be at least $\Omega(dN)$. This is formally stated in the following theorem.

Theorem 3.1. *Let $T \geq 4d\log(8)$. For any algorithm with expected communication cost (measured in bits) less than $\frac{dN}{64}$, there exists a contextual linear bandit instance with stochastic contexts, for which the algorithm's regret is $\Omega(N\sqrt{dT})$.*

3.1. Proof of Theorem 3.1

We start with a lower bound for a single-agent Bayesian two-armed bandit problem where the agent is given side information that contains a small amount of information about the optimal action.

Lemma 3.2. *Let $\boldsymbol{\mu}_1 = (\Delta, 0)$ and $\boldsymbol{\mu}_2 = (-\Delta, 0)$ and consider the single-agent Bayesian two-armed Gaussian bandit with mean $\boldsymbol{\mu}$ uniformly sampled from $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2\}$ and $a_* = \arg\max_{a\in\{1,2\}} \boldsymbol{\mu}_a$, which is a random variable. Suppose additionally that the agent has access to a random element M with $I(M; a_*) \leq 1/16$. Then, for any policy π,*

$$BR_T(\pi) \geq \Delta T \left(\frac{1}{2} - \sqrt{\frac{1}{2}\left(\frac{1}{16} + 4T\Delta^2\right)}\right),$$

where $BR_T(\pi) = \mathbb{E}_{\boldsymbol{\mu}\sim\mathrm{Unif}\{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2\}}[R_T(\pi, \boldsymbol{\mu})]$ and $R_T(\pi, \boldsymbol{\mu})$ is the regret suffered by policy π in the Gaussian two-armed bandit with means $\boldsymbol{\mu}$.

Remark 3.3. We assume in Lemma 3.2 that the agent has access to the message M from the beginning. The same bound continues to hold in the strictly harder problem where the agent has sequential access to a sequence of messages $M_1, \dots, M_T$ with $I(\{M_t\}_{t=1}^T; a_*) \le 1/16$.

The proof is presented in Appendix A. This lemma emphasizes the role of extra information a single agent might receive throughout the learning process on its performance, and therefore, it is key in proving Theorem 3.1. Specifically, since Lemma 3.2 makes no assumption on how the agent receives the extra information about the learning environment, we can prove Theorem 3.1 by employing this lemma and a reduction from single-agent bandit to multi-agent bandit as explained in what follows.

The construction. We consider a bandit instance where $K = 2$ and the decision sets are drawn uniformly from $\big\{(\mathbf{e}_1, \mathbf{e}_2), (\mathbf{e}_3, \mathbf{e}_4), \dots, (\mathbf{e}_{d-1}, \mathbf{e}_d)\big\}$. Let $\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^d : (\boldsymbol{\theta}_{2j-1}, \boldsymbol{\theta}_{2j}) \in \{(\Delta, 0), (-\Delta, 0)\},\ \forall j \in [\frac{d}{2}]\}$. We call $(\boldsymbol{\theta}_{2j-1}, \boldsymbol{\theta}_{2j})$ by j-th block of reward vector.

Bayesian regret. As in Lemma 3.2, we prove the minimax-style lower bound using the Bayesian regret. Let $\boldsymbol{\theta}$ be sampled uniformly from Θ and π be a fixed multi-agent policy. The multi-agent Bayesian regret is

$$BR_T = \mathbb{E}[\textstyle\sum_{t=1}^T \sum_{i=1}^N \langle \boldsymbol{\theta}, \mathbf{x}_{*,t}^i \rangle - \langle \boldsymbol{\theta}, \mathbf{x}_t^i \rangle]\,,$$

where the expectation integrates over the randomness in both $\boldsymbol{\theta}$ and the corresponding history induced by the interaction between π and the environment determined by $\boldsymbol{\theta}$. By Yao's minimax principle, there exists a $\boldsymbol{\theta} \in \Theta$ such that the expected regret is at least BR_T, so it suffices to lower bound the Bayesian regret. For the remainder of the proof $\mathbb{E}[\cdot]$ and $\mathbb{P}(\cdot)$ correspond to the expectation and probability measure on $\boldsymbol{\theta}$ and the history. For technical reasons, we assume that these probability spaces are defined to include an infinite interaction between the agents and environment. Of course, this is only used in the analysis.

Reduction from single-agent to multi-agent. Let M_{ij} be the mutual information between messages agent i receives in T rounds and $(\boldsymbol{\theta}_{2j-1}, \boldsymbol{\theta}_{2j})$. By assumption,

$$\sum_{i=1}^N \sum_{j=1}^{\frac{d}{2}} M_{ij} \le \sum_{i=1}^N \mathbb{E}[\text{Total number of bits agent } i \text{ receives}] \le \frac{dN}{64}. \tag{2}$$

Let $\mathcal{S}$ be the set of $\frac{dN}{4}$ pairs $(i, j) \in [N] \times [\frac{d}{2}]$ with smallest M_{ij}. From (2) and the definition of $\mathcal{S}$, we observe that for every pair $(i, j) \in \mathcal{S}$, we have

$$M_{ij} \le \frac{dN}{64\frac{dN}{4}} = \frac{1}{16}.$$

Let B_{ijt} be the indicator that the context is such that agent i interacts with j-th block in round t, which is

$$B_{ijt} = \mathbf{1}(\mathbf{x}_{t,1}^i = \mathbf{e}_{2j-1})\,.$$

Note that $\{B_{ijt}\}_{t=1}^\infty$ are independent and $\mathbb{E}[B_{ijt}] = 2/d$. Let $\mathcal{T}_{ij} = \{t : B_{ijt} = 1\}$ and $\mathcal{T}_{ij}^\circ$ be the first $T_\circ$ elements of $\mathcal{T}_{ij}$ with $T_\circ = T/d$. Let

$$R_{ij} = \sum_{t \in \mathcal{T}_{ij}^\circ} \langle \boldsymbol{\theta}, \mathbf{x}_{*,t}^i \rangle - \langle \boldsymbol{\theta}, \mathbf{x}_t^i \rangle$$

be the regret of agent i during the rounds in $\mathcal{T}_{ij}^\circ$ in bandit instance $\boldsymbol{\theta}$. Note that $\mathcal{T}_{ij}^\circ$ may contain rounds larger than T. Nevertheless,

$$\begin{aligned} BR_T &\ge \textstyle\sum_{i=1}^N \sum_{j=1}^{d/2} \mathbb{E}[R_{ij}\mathbf{1}(\mathcal{T}_{ij}^\circ \subset \{1, \dots, T\})] \\ &\ge \textstyle\sum_{(i,j)\in\mathcal{S}} \mathbb{E}[R_{ij}\mathbf{1}(\mathcal{T}_{ij}^\circ \subset \{1, \dots, T\})] \\ &= \textstyle\sum_{(i,j)\in\mathcal{S}} \mathbb{E}[R_{ij}] - \mathbb{E}[R_{ij}\mathbf{1}(\mathcal{T}_{ij}^\circ \not\subset \{1, \dots, T\})]\,. \end{aligned}$$

Suppose that $(i, j) \in \mathcal{S}$. Now, $\mathbb{E}[R_{ij}]$ is exactly the Bayesian regret of some policy interacting with the Bayesian two-armed bandit defined in Lemma 3.2 for $T_\circ$ rounds. Furthermore, the mutual information between the optimal action in this bandit and the messages passed to the agent is at most $M_{ij} \le 1/16$. Hence, by Lemma 3.2 and Remark 3.3,

$$\mathbb{E}[R_{ij}] \ge \Delta T_\circ \left(\frac{1}{2} - \sqrt{\frac{1}{2}\left(\frac{1}{16} + 4T_\circ\Delta^2\right)}\right)\,.$$

On the other hand,

$$\begin{aligned} \mathbb{E}[R_{ij}\mathbf{1}(\mathcal{T}_{ij}^\circ \not\subset \{1, \dots, T\})] &\le 2\Delta T_\circ \mathbb{P}(\mathcal{T}_{ij}^\circ \not\subset \{1, \dots, T\}) \\ &= 2\Delta T_\circ \mathbb{P}\left(\textstyle\sum_{t=1}^T B_{ijt} < T_\circ\right)\,. \end{aligned}$$

By Chernoff's bound, $T \ge 4d\log(8)$ and $\mathbb{E}[B_{ijt}] = 2/d$,

$$\begin{aligned} 2\mathbb{P}\left(\textstyle\sum_{t=1}^T B_{ijt} < T_\circ\right) &= 2\mathbb{P}\left(\textstyle\sum_{t=1}^T B_{ijt} < T/d\right) \\ &\le 2\exp\big(-T/(4d)\big) \le \frac{1}{4}\,. \end{aligned}$$

Therefore, with $\Delta = 0.0695\sqrt{\frac{d}{T}}$, we have

$$\begin{aligned} BR_T &\ge \frac{dNT_\circ\Delta}{4}\left(\frac{1}{4} - \sqrt{\frac{1}{2}\left(\frac{1}{16} + 4T_\circ\Delta^2\right)}\right) \\ &\ge \frac{N\sqrt{dT}}{1250} = \Omega\left(N\sqrt{dT}\right)\,, \end{aligned}$$

which concludes the proof of Theorem 3.1.

4. An Optimal Algorithm

Following the communication cost lower bound in previous section, we now present an algorithm called, *Distributed Batch Elimination Linear Upper Confidence Bound* (DisBE-LUCB), whose communication cost matches the lower bound up to logarithmic factors while achieving an optimal regret rate. DisBE-LUCB employs a central server to which, the agents send *local* updates and it then aggregates and broadcasts the updated *global* values of interest. We also discuss *Decentralized Batch Elimination Linear Upper Confidence Bound* (DecBE-LUCB), a modified version of DisBE-LUCB in the absence of a central server, where each agent can only communicate with its *immediate neighbors*.

Algorithm 1 DisBE-LUCB for agent i

```
Input: $N, d, \delta, T, M, \lambda$
Initialization: $a = \sqrt{T}(NT/d)^{\frac{1}{2(2^{M-1}-1)}}$, $T_1 = T_2 = a\sqrt{d/N}$, $T_m = \lfloor a\sqrt{T_{m-1}} \rfloor$, $\boldsymbol{\theta}_0^i = \mathbf{0}$, $\Lambda_0^i = \lambda I$, $\mathcal{T}_0 = 0$, $\mathcal{T}_m = \mathcal{T}_{m-1} + T_m$, $\lambda = 5\log(4dT/\delta)$, $\beta = 6\sqrt{\log(2KNT/\delta)} + \sqrt{\lambda}$, arbitrary policy $\pi_0^i$
for $m = 1, \dots, M$ do
    for $t = \mathcal{T}_{m-1} + 1, \dots, \min\{\mathcal{T}_m, T\}$ do
        Construct $\mathcal{X}_t^{i(m)} = \cap_{k=0}^{m-1} \mathcal{E}\left(\mathcal{X}_t^i; (\Lambda_k^i, \boldsymbol{\theta}_k^i, \beta)\right)$.
        Play arm $a_{i,t}$ associated with feature vector $\mathbf{x}_t^i \sim \pi_{m-1}^i\left(\mathcal{X}_t^{i(m)}\right)$ and observe $y_t^i$.
    end for
    Send $\mathbf{u}_m^i = \sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2} \mathbf{x}_t^i y_t^i$ to the server.
    Receive $\sum_{j=1}^N \mathbf{u}_m^j$ from the server.
    Compute/construct $\Lambda_m^i$ and $\boldsymbol{\theta}_m^i$ as in (3) and (4), respectively, $\mathcal{S}_m^i$ as in (5), and $\pi_m^i = \text{ExpPol}\left(\frac{2\lambda}{NT_m}, \mathcal{S}_m^i\right)$, where ExpPol is presented in Appendix D.
end for
```

4.1. Overview of DisBE-LUCB

Before describing how DisBE-LUCB operates for every agent $i \in [N]$, we note that all agents run DisBE-LUCB concurrently. In DisBE-LUCB, the time steps are grouped into M pre-defined batches by a grid $\mathcal{T} = \{\mathcal{T}_0, \mathcal{T}_1, \dots, \mathcal{T}_M\}$, where $0 = \mathcal{T}_0 \leq \mathcal{T}_1 \leq \dots \leq \mathcal{T}_M$, $T \leq \mathcal{T}_M$ and $T_m = \mathcal{T}_m - \mathcal{T}_{m-1}$ is the length of batch m. Our choice of grid implies that for any $m \geq 3$, we have $T_m = (a^{2^{m-1}-1} d^{\frac{1}{2}} / N^{\frac{1}{2}})^{\frac{1}{2^{m-2}}}$. Parameter a is chosen such that $T_M = T$ and $\mathcal{T}_M = \sum_{m \in [M]} T_m \geq T_M = T$, and therefore our choice of grid $\mathcal{T}$ is valid. At rounds $t \in [\mathcal{T}_{m-1} + 1 : \mathcal{T}_m]$ during batch $m \in [M]$, agent i first constructs confidence intervals for each action's reward, and the actions whose confidence intervals completely fall below those of other actions are eliminated. We denote the set of feature vectors associated with the surviving actions by $\mathcal{X}_t^{i(m)} = \cap_{k=0}^{m-1} \mathcal{E}(\mathcal{X}_t^i; (\Lambda_k^i, \boldsymbol{\theta}_k^i, \beta))$, where

$$\mathcal{E}(\mathcal{X}_t^i; (\Lambda_k^i, \boldsymbol{\theta}_k^i, \beta)) := \{\mathbf{x} \in \mathcal{X}_t^i : \langle \boldsymbol{\theta}_k^i, \mathbf{x} \rangle \\ + \beta \|\mathbf{x}\|_{(\Lambda_k^i)^{-1}} \geq \langle \boldsymbol{\theta}_k^i, \mathbf{y} \rangle - \beta \|\mathbf{y}\|_{(\Lambda_k^i)^{-1}}, \ \forall \mathbf{y} \in \mathcal{X}_t^i\}.$$

Here, $\{\Lambda_k^i\}_{k=0}^{m-1}$ and $\{\boldsymbol{\theta}_k^i\}_{k=0}^{m-1}$ are agent i's statistics used in computation of $\mathcal{X}_t^{(i)m}$ for $t \in [\mathcal{T}_{m-1} + 1 : \mathcal{T}_m]$. They are initialized to λI and $\mathbf{0}$ and will be updated at the end of each batch (will be specified how shortly). Let π_0^i be an arbitrary initial policy used in the first batch. Throughout batch $m \in [M]$, agent i uses the same policy π_{m-1}^i to select actions from the surviving actions set. At the end of batch $m \in [M]$, agent $i \in [N]$ sends $\mathbf{u}_m^i = \sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2} \mathbf{x}_t^i y_t^i$ to the server who broadcasts $\sum_{i=1}^N \mathbf{u}_m^i$ to all the agents. Then, agent i updates policy π_m^i (used in the next batch) and the following components that are key in the construction of the surviving actions set in the next batch as follows:

$$\Lambda_m^i = \lambda I + \frac{NT_m}{2} \mathbb{E}_{\mathcal{X} \sim \mathcal{D}_m^i} \mathbb{E}_{\mathbf{x} \sim \pi_{m-1}^i(\mathcal{X})}[\mathbf{x}\mathbf{x}^\top], \quad (3)$$

$$\boldsymbol{\theta}_m^i = (\Lambda_m^i)^{-1} \textstyle\sum_{j=1}^N \mathbf{u}_m^j, \quad (4)$$

where $\lambda > 0$ is a regularization constant and when conditioned on the first $(m-1)$ batches, $\mathcal{D}_m^i$ is the distribution based on which the sets of surviving feature vectors $\mathcal{X}_t^{i(m)}$ for all $t \in [\mathcal{T}_{m-1} + 1 : \mathcal{T}_m]$ are generated.

Statistics Λ_m^i and $\boldsymbol{\theta}_m^i$ are used in defining *new* confidence intervals in Lemma 4.4. We highlight that a direct use of existing standard confidence intervals in the literature such as the ones in Ruan et al. (2021) would fail to guarantee optimal communication cost $\tilde{\mathcal{O}}(dN)$ and require more communication by a factor of d [1]. Using matrix concentration inequalities, we address this issue by replacing matrix $\lambda I + \sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2} \sum_{i=1}^N \mathbf{x}_t^i \mathbf{x}_t^{i\top}$, which would have been used if Algorithm 5 in Ruan et al. (2021) had been directly extended to a multi-agent one, with $\lambda I + (NT_m/2) \mathbb{E}_{\mathcal{X} \sim \mathcal{D}_m^i} \mathbb{E}_{\mathbf{x} \sim \pi_{m-1}^i(\mathcal{X})}[\mathbf{x}\mathbf{x}^\top]$. This allows agent i to communicate only d values ($\mathbf{u}_m^i$) while achieving $\tilde{\mathcal{O}}(\sqrt{dNT})$ regret as will be shown in Theorem 4.1. As the final step of batch m, agent i implements ExpPol with inputs $\frac{2\lambda}{NT_m}, \mathcal{S}_m^i$, where

$$\mathcal{S}_m^i = \{\mathcal{X}_t^{i(m+1)}\}_{t=\mathcal{T}_{m-1}+T_m/2+1}^{\mathcal{T}_m}. \quad (5)$$

ExpPol, which is presented in Algorithm 4 in Appendix D and is inspired by Algorithm 3 in Ruan et al. (2021), computes policy π_m^i that will be used to select actions from the sets of surviving actions in the next batch. This choice of policy coupled with the definition of Λ_m^i in (3) guarantees that at all rounds $t \in [\mathcal{T}_1 + 1 : T]$, the length of the longest confidence interval in the surviving sets, which is an upper bound on the instantaneous regret of agent i at round t, can be bounded by $\mathcal{O}(\sqrt{d/NT})$. This allows us to achieve the optimal $\mathcal{O}(\sqrt{dNT})$ regret, while other exploration policies, such as the G-optimal design results in a $\mathcal{O}(d\sqrt{NT})$ regret.

[1] $d^2 + d$ values per agent, i.e., $\mathbf{u}_m^i$ and $\sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2} \mathbf{x}_t^i \mathbf{x}_t^{i\top}$.

4.2. Theoretical Results for DisBE-LUCB

We present our theoretical results for DisBE-LUCB, showing that it is nearly minimax optimal in terms of *both regret and communication cost*. The proof is given in Appendix B.

Theorem 4.1. *Fix* $M = 1 + \log(\log(NT/d)/2 + 1)$ *in Algorithm 1. Suppose Assumption 1 holds. If* $T \geq \Omega(d^{22}\log^2(NT/\delta)\log^2 d\log^2(d\lambda^{-1}))$*, then with probability at least* $1 - 2\delta$*, it holds that* $R_T \leq \mathcal{O}\left(\sqrt{dNT\log d\log^2(KNT/\delta\lambda)}\log\log\left(NT/d\right)\right)$*, and* $\textit{Communication Cost} \leq \mathcal{O}(dN\log\log(NT/d))$*, where the communication cost is measured by the number of real numbers communicated by the agents.*

We remark that simple tricks may significantly reduce the exponent constant in constraint $T \geq d^{\mathcal{O}(1)}$. For example, first running a simpler version of DisBE-LUCB, in which the exploration policy is the G-optimal design $\pi^{\mathrm{G}}(\mathcal{X}_t^{i(m)})$, for $\sqrt{T/dN}$ rounds and then switching to DisBE-LUCB would reduce the exponent to 10.

Remark 4.2. For the sake of Algorithm 1's presentation, we find it instructive to consider the communication cost as the number of real numbers communicated in the network. However, it is more realistic if we translate it into the total number of communicated bits. It would also allow us to make a fair comparison with the lower bound in Theorem 3.1 as it is stated in terms of number of communicated bits. Therefore, if we slightly modify Algorithm 1 such that instead of communicating vectors $\mathbf{u}_m^i$ in Line 8, agent i first rounds each entry of $\mathbf{u}_m^i$ with precision ϵ_0 and then sends the rounded vector to the server, then $\mathcal{O}(\log(1/\epsilon_0))$ number of bits is sufficient to communicate each entry of the rounded vectors $\mathbf{u}_m^i$. Our analysis in Appendix B.3 shows that compared to bounds in Theorem 4.1, by selecting $\epsilon_0 = \mathcal{O}(1/(N\sqrt{dT}))$, the communication cost of this slightly modified version of DisBE-LUCB, which is measured in bits, is $\mathcal{O}\left(dN\log\log\left(NT/d\right)\log(dNT)\right)$ and its regret is same as DisBE-LUCB's.

Remark 4.3. As mentioned in Section 4.1, a direct use of confidence intervals in Ruan et al. (2021) would fail to guarantee optimal communication cost $\tilde{\mathcal{O}}(dN)$ and require more communication by a factor of d. Thus, we use new confidence intervals (see Lemma 4.4) so that DisBE-LUCB would enjoy an optimal communication rate. The assumption on the knowledge of $\mathcal{D}$ is required in the computation of Λ_m^i in (3) used in these new confidence intervals. However, in practice, distribution $\mathcal{D}$ is not fully known and can only be estimated; therefore, Λ_m^i cannot be computed without any error. We relax this assumption and consider more realistic settings where each agent i can estimate matrix Λ_m^i in batch m up to an ϵ_m error, i.e., $(1-\epsilon_m)\Lambda_m^i \preceq \tilde{\Lambda}_m^i \preceq (1+\epsilon_m)\Lambda_m^i$, where $\tilde{\Lambda}_m^i$ is an estimation of Λ_m^i and $\epsilon_m \in (0,1)$[2]. In Appendix B.4, we show that for sufficiently small values of $\epsilon_m \leq 1/\sqrt{NT_m}$, a multiplicative factor $(1 - \max_{m\in[M]}\epsilon_m)^{-1}$ appears in the regret bound while the communication cost remains unchanged.

4.3. Proof Sketch of Theorem 4.1

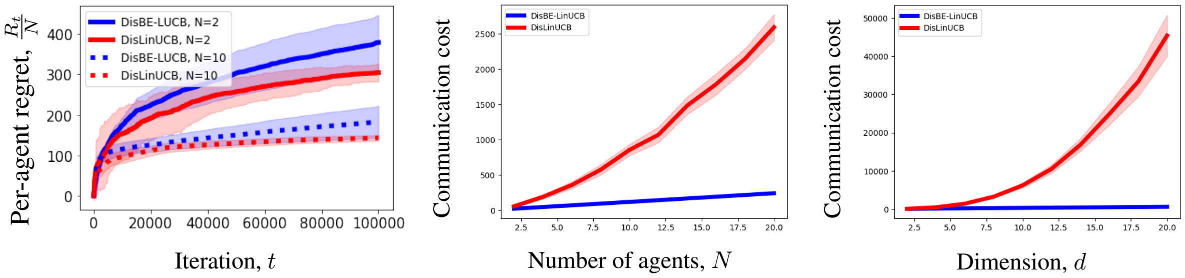
We first introduce the following lemma that constructs confidence intervals for the expected rewards.

Lemma 4.4 (Confidence intervals for DisBE-LUCB)**.** *Suppose Assumption 1 holds. For* $\delta \in (0,1)$*, let* $\beta = 6\sqrt{\log(2KNT/\delta)}+\sqrt{\lambda}$*. Then for all* $\mathbf{x}\in\mathcal{X}_t^i, i\in[N], t\in[T], m\in[M]$*, with probability at least* $1-\delta$*, it holds that* $\left|\langle\mathbf{x},\boldsymbol{\theta}_m^i-\boldsymbol{\theta}\rangle\right| \leq \beta\|\mathbf{x}\|_{(\Lambda_m^i)^{-1}}$.

We prove this lemma by first employing appropriate matrix concentration inequalities to lower bound Λ_m^i by matrix $\frac{1}{2}\sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2}\sum_{i=1}^N \mathbf{x}_t^i\mathbf{x}_t^{i\top}$. Carefully replacing Λ_m^i with its lower bound and using Azuma's inequality, we establish confidence intervals stated in the lemma. This lemma is key in ensuring an optimal communication rate $\tilde{\mathcal{O}}(dN)$, as a direct use of confidence intervals in Ruan et al. (2021) fails to guarantee optimal communication cost and requires $\tilde{\mathcal{O}}(d^2N)$ communication. See Appendix B.1 for proof.

Thanks to our choice of T_1 and T_2, and the fact that expected value of the rewards are bounded in $[-1,1]$, the regret of first two batches is bounded by $\mathcal{O}(\sqrt{dNT})$. For each batch $m \geq 3$, the confidence intervals imply that for all $t \in [\mathcal{T}_{m-1}+1 : \mathcal{T}_m]$, $\mathbf{x}_{t,*}^i \in \mathcal{X}_t^{i(m)}$ with high probability, and allow us to bound the instantaneous regret $r_t^i = \mathbb{E}[\langle\boldsymbol{\theta},\mathbf{x}_{*,t}^i\rangle - \langle\boldsymbol{\theta},\mathbf{x}_t^i\rangle]$ by $4\beta\mathbb{E}_{\mathcal{X}\sim\mathcal{D}_{m-1}^i}[\max_{\mathbf{x}\in\mathcal{X}}\sqrt{\mathbf{x}^\top(\Lambda_{m-1}^i)^{-1}\mathbf{x}}]$. Note that learning of $\boldsymbol{\theta}_m^i$ and π_m^i are done through disjoint sets of samples, i.e., $\mathcal{A} = [\mathcal{T}_{m-1}+1 : \mathcal{T}_{m-1}+T_m/2]$ and $\mathcal{B} = [\mathcal{T}_{m-1}+T_m/2+1 : \mathcal{T}_m]$, respectively. This is because $\mathcal{D}_m^i$ depends on $\boldsymbol{\theta}_m^i$, which is learned from $\mathcal{A}$, and we have to make $\mathcal{B}$ disjoint from $\mathcal{A}$ so as to ensure that elements in $\mathcal{S}_m^i$ are independently sampled from $\mathcal{D}_m^i$. Therefore, Theorem 5 in Ruan et al. (2021) guarantees that $\mathbb{E}_{\mathcal{X}\sim\mathcal{D}_{m-1}^i}[\max_{\mathbf{x}\in\mathcal{X}}\sqrt{\mathbf{x}^\top(\Lambda_{m-1}^i)^{-1}\mathbf{x}}] \leq \tilde{\mathcal{O}}(\sqrt{d/(NT_{m-1})})$. Finally, these combined with our choice of grid $\mathcal{T} = \{\mathcal{T}_0,\mathcal{T}_1,\ldots,\mathcal{T}_M\}$ and $M = 1+\log(\log(NT/d)/2+1)$ lead us to a regret bound $\tilde{\mathcal{O}}(\sqrt{dNT})$. Moreover, communications happen only at the end of each batch, whose number is M, and agents only share d-dimensional vectors $\mathbf{u}_m^i$. Therefore, communication cost is $dNM = \mathcal{O}(dN\log\log(NT/d))$.

[2] This is a weaker condition compared to the component-wise condition $(1-\epsilon_m)\Lambda_m^i \leq \tilde{\Lambda}_m^i \leq (1+\epsilon_m)\Lambda_m^i$.



(a) $d = 4$

(b) $d = 4$

(c) $N = 10$

Figure 1. The shaded regions show standard deviation around the mean. Standard deviation for communication cost of DisBE-LUCB is zero, because communication cost = dNM and parameters determining M are known upfront (see Theorem 4.1).

4.4. Fully Decentralized Batch Elimination LUCB

In a scenario where there is no server and the agents are allowed to communicate *only* with their immediate neighbors, they can be represented by nodes of a graph. Applying a carefully designed consensus procedure that guarantees sufficient information mixing among the entire network, in Appendix C, we propose a fully decentralized version of DisBE-LUCB, called DecBE-LUCB. Communication cost of DecBE-LUCB is greater than DisBE-LUCB's by an extra multiplication factor $S = \log(dN)\delta_{\max}/\sqrt{1/|\lambda_2|}$, where $\delta_{\max}$ is the maximum degree of the network's graph and $|\lambda_2|$ is the second largest eigenvalue of the communication matrix in absolute value characterizing the graph's connectivity level. This is because at the last S rounds of each batch m, agents communicate each entry of their estimations of vector $\sum_{j=1}^{N} \mathbf{u}_m^j$ with their neighbors, whose number is at most $\delta_{\max}$, to ensure enough information mixing. Moreover, this results in DecBE-LUCB having no control over the regret of the mixing rounds, and therefore an additional term $\log(dN)NM/\sqrt{1/|\lambda_2|}$, which we call the *delay effect*, in the regret bound. Note that the more connected the graph is, the smaller $|\lambda_2|$ is. This aligns with the fact that the more connected the graph is, the less number of mixing rounds S is required. For example, fixing $N = 20$, for chain, ring, star, random Erdős–Renyi graph with parameter $p = 0.5$, and complete graphs, the values of $|\lambda_2|$ are 0.9918, 0.9674, 0.97, 0.67 (average over 100 instances), and 0, respectively. As expected, for less connected graphs (Chain, Ring, Star), $|\lambda_2|$ is close to 1 and for the fully connected graph $|\lambda_2|= 0$ and for a random graph $|\lambda_2|$ is not too small nor too large. The theoretical guarantees of DecBE-LUCB are summarized in table 1 and a detailed discussion is given in Appendix C.

5. Experiments

In this section, we present numerical simulations to confirm our theoretical findings. We evaluate the performance of DisBE-LUCB on synthetic data and compare it to that of DisLinUCB proposed by Wang et al. (2019) that study the most similar setting to ours. The results shown in Figure 1 depict averages over 20 realizations, for which we have chosen $K = 20$, $\delta = 0.01$ and $T = 100000$. For each realization, the parameter $\boldsymbol{\theta}$ is drawn from $\mathcal{N}(0, I_d)$ and then normalized to unit norm and noise variables are zero-mean Gaussian random variables with variance 0.01. The decision set distribution $\mathcal{D}$ is chosen to be uniform over $\{\tilde{\mathcal{X}}_1, \tilde{\mathcal{X}}_2, \dots, \tilde{\mathcal{X}}_{100}\}$, where each $\tilde{\mathcal{X}}_i$ is a set of K vectors drawn from $\mathcal{N}(0, I_d)$ and then normalized to unit norm. While implementing DisBE-LUCB, in order to compute $\mathbb{E}_{\mathcal{X}\sim\mathcal{D}_m^i}\mathbb{E}_{\mathbf{x}\sim\pi_{m-1}^i(\mathcal{X})}[\mathbf{x}\mathbf{x}^\top]$ for agent i at batch m, we followed these steps: 1) for each $j \in [100]$, we built $\tilde{\mathcal{X}}_j^{i(m)} = \cap_{k=0}^{m-1}\mathcal{E}(\tilde{\mathcal{X}}_j; (\Lambda_k^i, \boldsymbol{\theta}_k^i, \beta))$; 2) we took average over all 100 matrices $\frac{1}{100}\sum_{j\in[100]}\mathbb{E}_{\mathbf{x}\sim\pi_{m-1}^i(\tilde{\mathcal{X}}_j^{i(m)})}[\mathbf{x}\mathbf{x}^\top]$ as $\mathcal{D}$ is a uniform distribution over $\{\tilde{\mathcal{X}}_1, \tilde{\mathcal{X}}_2, \dots, \tilde{\mathcal{X}}_{100}\}$. In Figure 1a, fixing $d = 4$, we compare the per-agent regret R_t/N of DisBE-LUCB and DisLinUCB for $t \in [T]$ and for different values of $N = 2$ and $N = 10$, where $R_t = \sum_{s=1}^{t}\sum_{i=1}^{N}\langle\boldsymbol{\theta}, \mathbf{x}_{*,s}^i\rangle - \langle\boldsymbol{\theta}, \mathbf{x}_s^i\rangle$. Figure 1b compares the communication cost of DisBE-LUCB and DisLinUCB over T rounds when both algorithms are implemented for fixed $d = 4$, and N varying from 2 to 20. Finally, Figure 1c compares the communication cost of DisBE-LUCB and DisLinUCB over T rounds when both algorithms are implemented for fixed $N = 10$, and d varying from 2 to 20. From these three comparisons, we conclude that DisBE-LUCB achieves a regret comparable with DisLinUCB, at a significantly smaller communication rate. The curves in Figures 1b and 1c verify the linear dependency of DisBE-LUCB's communication cost on N and d while communication cost of DisLinUCB grows super-linearly with N and d (see Table 1 for theoretical comparisons). Moreover, Figure 1a emphasizes the value of collaboration in speeding up the learning process. As the number of agents increases, each agent learns the environment faster as an individual.

6. Conclusion

We proved an information-theoretic lower bound on the communication cost of any algorithm achieving an optimal regret rate for the distributed contextual linear bandit problem with stochastic contexts. We then proposed DisBE-LUCB with optimal regret $\tilde{\mathcal{O}}(\sqrt{dNT})$ and communication cost $\tilde{\mathcal{O}}(dN)$ which (nearly) matches our lower bound and improves upon the previous best-known algorithms whose communication cost scale super linearly either in d or N. Finally, we proposed DecBE-LUCB, a fully decentralized variant of DisBE-LUCB, without a central server where the agents can only communicate with their immediate neighbors given by a communication graph. We showed that the structure of the network affects the regret performance via a small additive term that depends on the spectral gap of the underlying graph, while the communication cost still grows linearly with d and N. As shown in Table 1, the best communication cost achieved for settings with *adversarially* varying contexts over time horizon and agents is $\mathcal{O}(d^3N^{1.5})$. There is no formal theory proving such bounds are optimal for the adversarial context case. While our work provides optimal theoretical guarantees for stochastically varying contexts, it is not clear how to generalize these *optimal* results to settings with adversarially varying contexts. Therefore, an important future direction is to design optimal algorithms and prove communication cost lower bounds for scenarios with adversarial contexts.

Acknowledgements

This work was supported in part by DARPA under agreement HR00112190130, NSF grant #2221871, and an Amazon Research Grant. Sanae Amani was partially supported by Amazon science hub fellowship.

References

Agarwal, A., Bird, S., Cozowicz, M., Hoang, L., Langford, J., Lee, S., Li, J., Melamed, D., Oshri, G., Ribas, O., et al. (2016). Making contextual decisions with low technical debt. *arXiv preprint arXiv:1606.03966*.

Arioli, M. and Scott, J. (2014). Chebyshev acceleration of iterative refinement. *Numerical Algorithms*, 66(3):591–608.

Avner, O. and Mannor, S. (2019). Multi-user communication networks: A coordinated multi-armed bandit approach. *IEEE/ACM Transactions on Networking*, 27(6):2192–2207.

Berkenkamp, F., Krause, A., and Schoellig, A. P. (2016). Bayesian optimization with safety constraints: safe and automatic parameter tuning in robotics. *arXiv preprint arXiv:1602.04450*.

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*, pages 208–214. JMLR Workshop and Conference Proceedings.

Dani, V., Hayes, T. P., and Kakade, S. M. (2008). Stochastic linear optimization under bandit feedback.

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

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

Gao, Z., Han, Y., Ren, Z., and Zhou, Z. (2019). Batched multi-armed bandits problem. *arXiv preprint arXiv:1904.01763*.

Han, Y., Zhou, Z., Zhou, Z., Blanchet, J., Glynn, P. W., and Ye, Y. (2020). Sequential batch learning in finite-action linear contextual bandits. *arXiv preprint arXiv:2004.06321*.

Hanna, O. A., Yang, L. F., and Fragouli, C. (2022a). Contexts can be cheap: Solving stochastic contextual bandits with linear bandit algorithms. *arXiv preprint arXiv:2211.05632*.

Hanna, O. A., Yang, L. F., and Fragouli, C. (2022b). Learning in distributed contextual linear bandits without sharing the context. *arXiv preprint arXiv:2206.04180*.

Huang, R., Wu, W., Yang, J., and Shen, C. (2021). Federated linear contextual bandits. In *Thirty-Fifth Conference on Neural Information Processing Systems*.

Korda, N., Szorenyi, B., and Li, S. (2016). Distributed clustering of linear bandits in peer to peer networks. In *International conference on machine learning*, pages 1301–1309. PMLR.

Landgren, P., Srivastava, V., and Leonard, N. E. (2016a). Distributed cooperative decision-making in multiarmed bandits: Frequentist and bayesian algorithms. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 167–172. IEEE.

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

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

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

Li, C., Wang, H., Wang, M., and Wang, H. (2022). Communication efficient distributed learning for kernelized contextual bandits. *arXiv preprint arXiv:2206.04835*.

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

Li, S., Hao, F., Li, M., and Kim, H.-C. (2013). Medicine rating prediction and recommendation in mobile social networks. In *International conference on grid and pervasive computing*, pages 216–223. Springer.

Liu, K. and Zhao, Q. (2010a). Decentralized multi-armed bandit with multiple distributed players. In *2010 Information Theory and Applications Workshop (ITA)*, pages 1–10. IEEE.

Liu, K. and Zhao, Q. (2010b). Distributed learning in cognitive radio networks: Multi-armed bandit with distributed multiple players. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3010–3013. IEEE.

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

Lynch, N. A. (1996). *Distributed algorithms*. Elsevier.

Martínez-Rubio, D., Kanade, V., and Rebeschini, P. (2019). Decentralized cooperative stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 4531–4542.

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*, pages 74–87.

Seaman, K., Bach, F., Bubeck, S., Lee, Y. T., and Massoulié, L. (2017). Optimal algorithms for smooth and strongly convex distributed optimization in networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3027–3036. JMLR. org.

Sui, Y., Zhuang, V., Burdick, J. W., and Yue, Y. (2018). Stagewise safe bayesian optimization with gaussian processes. *arXiv preprint arXiv:1806.07555*.

Szörényi, B., Busa-Fekete, R., Hegedűs, I., Ormándi, R., Jelasity, M., and Kégl, B. (2013). Gossip-based distributed stochastic bandit algorithms. In *Journal of Machine Learning Research Workshop and Conference Proceedings*, volume 2, pages 1056–1064. International Machine Learning Societ.

Tewari, A. and Murphy, S. A. (2017). From ads to interventions: Contextual bandits in mobile health. In *Mobile Health*, pages 495–517. Springer.

Tropp, J. A. (2015). An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571*.

Wang, Y., Hu, J., Chen, X., and Wang, L. (2019). Distributed bandit learning: How much communication is needed to achieve (near) optimal regret. *arXiv preprint arXiv:1904.06309*.

Xiao, L. and Boyd, S. (2004). Fast linear iterations for distributed averaging. *Systems & Control Letters*, 53(1):65–78.

Young, D. M. (2014). *Iterative solution of large linear systems*. Elsevier.

A. Proof of Lemma 3.2

Let $\boldsymbol{\mu} \sim \mathrm{Unif}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$, where $\boldsymbol{\mu}_1 = [\Delta, 0]^\top, \boldsymbol{\mu}_2 = [-\Delta, 0]^\top$, $\mathbf{z} = \{z_t\}_{t=1}^T$ be the set of arm 1's reward, $H = \{a_t, y_t\}_{t=1}^T$ be the history over the course of T rounds, where a_t is the arm pulled and y_t is the observed reward at round t, $a_* = \arg\max_{a \in \{1,2\}} \boldsymbol{\mu}_a$, and $\hat{a} \sim \mathrm{Unif}(\{a_1, a_2, \ldots, a_T\})$. We have

$$\begin{aligned} BR_T(\pi) = \mathbb{E}[R_T(\pi, \boldsymbol{\mu})] &= \mathbb{E}[\sum_{t=1}^T \mathbb{1}(\hat{a} \neq a_*)\Delta] \\ &= \Delta T \mathbb{P}(\hat{a} \neq a_*). \end{aligned} \tag{$\bigstar$}$$

Now, we lower bound $\mathbb{P}(\hat{a} \neq a_*)$ as follows

$$\begin{aligned} \mathbb{P}(\hat{a} \neq a_*) &= \sum_{a \in \{1,2\}} \mathbb{P}(a_* = a)\mathbb{P}(\hat{a} \neq a | a_* = a) \\ &= \sum_{a \in \{1,2\}} \mathbb{P}(a_* = a)\left[\mathbb{P}(\hat{a} \neq a) + \mathbb{P}(\hat{a} = a) - \mathbb{P}(\hat{a} = a | a_* = a)\right] \\ &\geq \sum_{a \in \{1,2\}} \mathbb{P}(a_* = a)\left[\mathbb{P}(\hat{a} \neq a) - \sqrt{\frac{1}{2}\mathrm{D_{KL}}(\mathbb{P}_{\hat{a}|a_*=a}, \mathbb{P}_{\hat{a}})}\right] && \text{(Pinsker's inequality)} \\ &= \frac{1}{2} - \sum_{a \in \{1,2\}} \mathbb{P}(a_* = a)\sqrt{\frac{1}{2}\mathrm{D_{KL}}(\mathbb{P}_{\hat{a}|a_*=a}, \mathbb{P}_{\hat{a}})} \\ &\geq \frac{1}{2} - \sqrt{\frac{1}{2}\sum_{a \in \{1,2\}} \mathbb{P}(a_* = a)\mathrm{D_{KL}}(\mathbb{P}_{\hat{a}|a_*=a}, \mathbb{P}_{\hat{a}})} && \text{(Jensen's inequality)} \\ &= \frac{1}{2} - \sqrt{\frac{1}{2}I(\hat{a}; a_*)} \\ &\geq \frac{1}{2} - \sqrt{\frac{1}{2}I(M, H; a_*)} && \text{(Data processing)} \\ &\geq \frac{1}{2} - \sqrt{\frac{1}{2}\left(I(M; a_*) + I(H; a_*)\right)} \\ &\geq \frac{1}{2} - \sqrt{\frac{1}{2}\left(\frac{1}{16} + I(H; a_*)\right)}. && (\bigstar\bigstar) \end{aligned}$$

In our next step towards lower bounding $\mathbb{P}(\hat{a} \neq a_*)$, we upper bound $I(H; a_*)$, as follows

$$\begin{aligned} I(H; a_*) &\leq I(\mathbf{z}; a_*) && \text{(Data processing)} \\ &= \sum_{a \in \{1,2\}} \frac{1}{2}\mathrm{D_{KL}}(\mathbb{P}(\mathbf{z}|a_* = a), \mathbb{P}(\mathbf{z})) \\ &\leq \sum_{b \in \{1,2\}} \sum_{a \in \{1,2\}} \frac{1}{2}\mathrm{D_{KL}}\left(\mathbb{P}(\mathbf{z}|a_* = a), \mathbb{P}(\mathbf{z}|a_* = b)\right) \\ &= \frac{1}{2}\mathrm{D_{KL}}\left(\mathbb{P}(\mathbf{z}|a_* = 1), \mathbb{P}(\mathbf{z}|a_* = 2)\right) + \frac{1}{2}\mathrm{D_{KL}}\left(\mathbb{P}(\mathbf{z}|a_* = 2), \mathbb{P}(\mathbf{z}|a_* = 1)\right) \\ &= \frac{1}{2}\left[T(2\Delta)^2 + T(2\Delta)^2\right] \\ &= 4T\Delta^2. && (\bigstar\bigstar\bigstar) \end{aligned}$$

Combining $\bigstar$, $\bigstar\bigstar$, and $\bigstar\bigstar\bigstar$, we have

$$BR_T(\pi) \geq \Delta T\left(\frac{1}{2} - \sqrt{\frac{1}{2}\left(\frac{1}{16} + 4T\Delta^2\right)}\right),$$

which concludes the lemma.

B. Proof of Theorem 4.1

In this section, we give a complete outline of the proof of Theorem 4.1 which starts with the proof of Lemma 4.4.

B.1. Proof of Lemma 4.4

For each batch $m \in [M]$, let $\mathbf{b}_m = \sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2} \sum_{i=1}^{N} \mathbf{x}_t^i y_t^i$ and $\mathbf{V}_m = \sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2} \sum_{i=1}^{N} \mathbf{x}_t^i \mathbf{x}_t^{i\top}$. We have

$$\begin{aligned}
\Lambda_m^i &= \lambda I + \frac{NT_m}{2} \mathbb{E}_{\mathcal{X}\sim\mathcal{D}_m^i} \mathbb{E}_{\mathbf{x}\sim\pi_{m-1}^i(\mathcal{X})}[\mathbf{x}\mathbf{x}^\top] \\
&= \lambda I + \frac{NT_m}{4} \Big(2\mathbb{E}_{\mathcal{X}\sim\mathcal{D}_m^i} \mathbb{E}_{\mathbf{x}\sim\pi_{m-1}^i(\mathcal{X})}[\mathbf{x}\mathbf{x}^\top] + 6\gamma I \Big) - 1.5NT_m\gamma I. \qquad (6)
\end{aligned}$$

By choosing $\gamma = \frac{3\log(\frac{4dT}{\delta})}{NT_m}$ and $\lambda = 5\log\left(\frac{4dT}{\delta}\right)$, combining (6) and Lemma E.3, for all $m \in [M]$, with probability at least $1 - \delta/2$, we have

$$\begin{aligned}
\Lambda_m^i &\succeq \left(\lambda - 5\log\left(\frac{4dT}{\delta}\right)\right) I + \frac{1}{2} \sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2} \sum_{i=1}^{N} \mathbf{x}_t^i \mathbf{x}_t^{i\top} \\
&= \frac{1}{2}\mathbf{V}_m. \qquad (7)
\end{aligned}$$

Moreover, for a fixed $\mathbf{x} \in \mathcal{X}_t^i$ and $(i,t) \in [N]\times[T]$, let $z_{t,m}^{j,i} = \mathbf{x}^\top \left(\Lambda_m^i\right)^{-1} \Big(\mathbf{x}_t^j y_t^j - \mathbb{E}_{\mathcal{X}\sim\mathcal{D}_m^i} \mathbb{E}_{\mathbf{x}\sim\pi_{m-1}^i(\mathcal{X})}[\mathbf{x}\mathbf{x}^\top]\boldsymbol{\theta}\Big)$. Thus, we have

$$\begin{aligned}
\left|\left\langle \mathbf{x}, \boldsymbol{\theta}_m^i - \boldsymbol{\theta} \right\rangle\right| &= \left|\left\langle \mathbf{x}, \left(\Lambda_m^i\right)^{-1} \mathbf{b}_m - \boldsymbol{\theta} \right\rangle\right| \\
&= \left|\left\langle \mathbf{x}, \left(\Lambda_m^i\right)^{-1} \mathbf{b}_m \right\rangle - \left\langle \mathbf{x}, \left(\Lambda_m^i\right)^{-1} \Lambda_m^i \boldsymbol{\theta} \right\rangle\right| \\
&\le \left|\left\langle \mathbf{x}, \left(\Lambda_m^i\right)^{-1} \mathbf{b}_m \right\rangle - \left\langle \mathbf{x}, \left(\Lambda_m^i\right)^{-1} \left(\Lambda_m^i - \lambda I\right) \boldsymbol{\theta} \right\rangle\right| + \left|\lambda \langle \mathbf{x}, \left(\Lambda_m^i\right)^{-1} \boldsymbol{\theta} \rangle\right| \\
&\le \left|\mathbf{x}^\top \left(\Lambda_m^i\right)^{-1} \left(\mathbf{b}_m - \frac{NT_m}{2} \mathbb{E}_{\mathcal{X}\sim\mathcal{D}_m^i} \mathbb{E}_{\mathbf{x}\sim\pi_{m-1}^i(\mathcal{X})}[\mathbf{x}\mathbf{x}^\top]\boldsymbol{\theta}\right)\right| + \sqrt{\lambda}\|\mathbf{x}\|_{\left(\Lambda_m^i\right)^{-1}} \\
&\qquad\qquad \text{(Cauchy Schwarz inequality and Assumption 1)} \\
&= \left|\sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2} \sum_{j=1}^{N} z_{t,m}^{j,i}\right| + \sqrt{\lambda}\|\mathbf{x}\|_{\left(\Lambda_m^i\right)^{-1}}.
\end{aligned}$$

Note that

$$\mathbb{E}\left[z_{t,m}^{j,i}\right] = \mathbb{E}\left[\mathbf{x}^\top \left(\Lambda_m^i\right)^{-1} \left(\mathbf{x}_t^j (\mathbf{x}_t^{j\top}\boldsymbol{\theta} + \eta_t^j) - \mathbb{E}_{\mathcal{X}\sim\mathcal{D}_m^i} \mathbb{E}_{\mathbf{x}\sim\pi_{m-1}^i(\mathcal{X})}[\mathbf{x}\mathbf{x}^\top]\boldsymbol{\theta}\right)\right] = 0,$$

(Noise η_t^j is zero-mean and independent of $\mathbf{x}_t^j$)

By Azuma's inequality, for a fixed $\mathbf{x} \in \mathcal{X}_t^i$ and $(i,t) \in [N]\times[T]$, we have

$$\mathbb{P}\left(\left|\sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2} \sum_{j=1}^{N} z_{t,m}^{j,i}\right| \ge \alpha\|\mathbf{x}\|_{\left(\Lambda_m^i\right)^{-1}}\right) \le 2\exp\left(\frac{-\alpha^2\|\mathbf{x}\|^2_{\left(\Lambda_m^i\right)^{-1}}}{2c_m^i}\right), \qquad (8)$$

where

$$
\begin{aligned}
c_m^i &= \sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2} \sum_{j=1}^{N} \left| \mathbf{x}^\top \left(\Lambda_m^i\right)^{-1} \left(\mathbf{x}_t^j y_t^j - \mathbb{E}_{\mathcal{X}\sim\mathcal{D}_m^i} \mathbb{E}_{\mathbf{x}\sim\pi_{m-1}^i(\mathcal{X})}[\mathbf{x}\mathbf{x}^\top]\boldsymbol{\theta} \right) \right|^2 \\
&\le 2 \sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2} \sum_{j=1}^{N} \left| \mathbf{x}^\top \left(\Lambda_m^i\right)^{-1} \mathbf{x}_t^j y_t^j \right|^2 + NT_m \left| \mathbf{x}^\top \left(\Lambda_m^i\right)^{-1} \mathbb{E}_{\mathcal{X}\sim\mathcal{D}_m^i} \mathbb{E}_{\mathbf{x}\sim\pi_{m-1}^i(\mathcal{X})}[\mathbf{x}\mathbf{x}^\top]\boldsymbol{\theta} \right|^2 \\
&\le 2 \sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2} \sum_{j=1}^{N} \left| \mathbf{x}^\top \left(\Lambda_m^i\right)^{-1} \mathbf{x}_t^j \right|^2 + \frac{4}{NT_m} \left| \mathbf{x}^\top \left(\Lambda_m^i\right)^{-1} \left(\Lambda_m^i - \lambda I\right) \boldsymbol{\theta} \right|^2 && \text{(Assumption 1)} \\
&= 2 \sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2} \sum_{j=1}^{N} \mathbf{x}^\top \left(\Lambda_m^i\right)^{-1} \mathbf{x}_t^j {\mathbf{x}_t^j}^\top \left(\Lambda_m^i\right)^{-1} \mathbf{x} + \frac{4}{NT_m} \left| \mathbf{x}^\top \boldsymbol{\theta} - \lambda \mathbf{x}^\top \left(\Lambda_m^i\right)^{-1} \boldsymbol{\theta} \right|^2 \\
&\le 2\mathbf{x}^\top \left(\Lambda_m^i\right)^{-1} \mathbf{V}_m \left(\Lambda_m^i\right)^{-1} \mathbf{x} + \left(\frac{4\|\boldsymbol{\theta}\|_{\Lambda_m^i}^2}{NT_m} + \frac{4\lambda}{NT_m} \right) \|\mathbf{x}\|_{\left(\Lambda_m^i\right)^{-1}}^2 \\
& && \text{(Cauchy Schwarz inequality and Assumption 1)} \\
&\le 4\mathbf{x}^\top \left(\Lambda_m^i\right)^{-1} \Lambda_m^i \left(\Lambda_m^i\right)^{-1} \mathbf{x} + \left(\frac{4\|\boldsymbol{\theta}\|_{\Lambda_m^i}^2}{NT_m} + \frac{4\lambda}{NT_m} \right) \|\mathbf{x}\|_{\left(\Lambda_m^i\right)^{-1}}^2 && \text{(Conditioned on the event in Eqn. (7))} \\
&\le \left(6 + \frac{8\lambda}{NT_m} \right) \|\mathbf{x}\|_{\left(\Lambda_m^i\right)^{-1}}^2, && (9)
\end{aligned}
$$

where the last inequity follows from the fact that

$$
\|\boldsymbol{\theta}\|_{\Lambda_m^i}^2 \le \|\boldsymbol{\theta}\|_2^2 \, \lambda_{\max}\left(\Lambda_m^i\right) \le \lambda + \frac{NT_m}{2}. \qquad \text{(Assumption 1)}
$$

Combining (8) and (9), and by a union bound, we have

$$
\mathbb{P}\left(\left| \left\langle \mathbf{x}, \boldsymbol{\theta}_m^i - \boldsymbol{\theta} \right\rangle \right| \le \left(6\sqrt{\log\left(\frac{2KNT}{\delta}\right)} + \sqrt{\lambda} \right) \|\mathbf{x}\|_{\left(\Lambda_m^i\right)^{-1}}, \ \forall \mathbf{x} \in \mathcal{X}_t^i, i \in [N], t \in [T], m \in [M] \right) \ge 1 - \delta. \tag{10}
$$

B.2. Completing the proof of Theorem 4.1

Next, we state the following lemma, which we borrow from Theorem 5 in Ruan et al. (2021) and is used in the proof analysis of Theorem 4.1.

Lemma B.1 (Ruan et al. (2021))**.** *Let $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_L \sim \mathcal{D}$ be i.i.d drawn from a distribution $\mathcal{D}$ and input of Algorithm 4 and let π be the output policy of Algorithm 4. For any $\lambda \in (0,1)$, we have*

$$
\mathbb{P}\left[\mathbb{V}_{\mathcal{D}}^{\lambda}(\pi) \le \mathcal{O}\left(\sqrt{d \log d \log(\lambda^{-1})} \right) \right] \ge 1 - \exp\left(\mathcal{O}(d^3 \log d \log(d\lambda^{-1})) - Ld^{-2c}2^{-16} \right),
$$

where we define the λ-deviation of policy π over $\mathcal{D}$ by

$$
\mathbb{V}_{\mathcal{D}}^{\lambda}(\pi) := \mathbb{E}_{\mathcal{X}\sim\mathcal{D}}\left[\max_{\mathbf{x}\in\mathcal{X}} \sqrt{ \mathbf{x}^\top \left(\lambda I + \mathbb{E}_{\mathcal{X}\sim\mathcal{D}} \mathbb{E}_{\mathbf{y}\sim\pi(\mathcal{X})}[\mathbf{y}\mathbf{y}^\top] \right)^{-1} \mathbf{x} } \right]. \tag{11}
$$

Corollary B.2. *As a direct corollary of Lemma B.1, if $T \ge \Omega\left(d^{22} \log^2(\frac{NT}{\delta}) \log^2 d \log^2(dNT\lambda^{-1}) \right)$, then for all $m \ge 2$ and $i \in [N]$, with probability at least $1-\delta$, it holds that*

$$
\mathbb{V}_{\mathcal{D}_m^i}^{(\frac{2\lambda}{NT_m})}(\pi_{m-1}^i) \le \mathcal{O}(\sqrt{d \log d \log(NT\lambda^{-1})}). \tag{12}
$$

Now, we focus on the regret of the i-th agent at m-th batch for any $m \geq 3$. Let $\mathcal{D}_m^i$ be the distribution based on which the surviving sets $\mathcal{X}_t^{i(m)}$ for all $t \in [\mathcal{T}_{m-1}+1 : \mathcal{T}_m]$ are generated when conditioned on the first $m-1$ batches. For any $t \in [\mathcal{T}_{m-1}+1 : \mathcal{T}_m]$, conditioned on the event that the confidence intervals in Lemma 4.4 hold, we have

$$\begin{aligned}
r_t^i &= \mathbb{E}\left[\langle \boldsymbol{\theta}, \mathbf{x}_{*,t}^i \rangle - \langle \boldsymbol{\theta}, \mathbf{x}_t^i \rangle\right] \\
&\leq \mathbb{E}\left[\langle \boldsymbol{\theta}_{m-1}^i, \mathbf{x}_{*,t}^i \rangle - \langle \boldsymbol{\theta}_{m-1}^i, \mathbf{x}_t^i \rangle + \beta \left\|\mathbf{x}_{*,t}^i\right\|_{\left(\Lambda_{m-1}^i\right)^{-1}} + \beta \left\|\mathbf{x}_t^i\right\|_{\left(\Lambda_{m-1}^i\right)^{-1}}\right] && \text{(Lemma 4.4)} \\
&\leq 2\beta \mathbb{E}\left[\left\|\mathbf{x}_{*,t}^i\right\|_{\left(\Lambda_{m-1}^i\right)^{-1}} + \left\|\mathbf{x}_t^i\right\|_{\left(\Lambda_{m-1}^i\right)^{-1}}\right] && (\mathbf{x}_{*,t}^i \in \mathcal{X}_t^{i(m)}) \\
&\leq 4\beta \mathbb{E}\left[\max_{\mathbf{x} \in \mathcal{X}_t^{i(m)}} \|\mathbf{x}\|_{\left(\Lambda_{m-1}^i\right)^{-1}}\right] \\
&\leq 4\beta \mathbb{E}_{\mathcal{X} \sim \mathcal{D}_m^i}\left[\max_{\mathbf{x} \in \mathcal{X}} \|\mathbf{x}\|_{\left(\Lambda_{m-1}^i\right)^{-1}}\right] \\
&\leq 4\beta \mathbb{E}_{\mathcal{X} \sim \mathcal{D}_{m-1}^i}\left[\max_{\mathbf{x} \in \mathcal{X}} \|\mathbf{x}\|_{\left(\Lambda_{m-1}^i\right)^{-1}}\right] \\
&\leq \frac{8\beta}{\sqrt{NT_{m-1}}} \mathbb{E}_{\mathcal{X} \sim \mathcal{D}_{m-1}^i}\left[\max_{\mathbf{x} \in \mathcal{X}} \sqrt{\mathbf{x}^\top \left(\frac{2\lambda}{NT_{m-1}} I + \mathbb{E}_{\mathcal{X} \sim \mathcal{D}_{m-1}^i} \mathbb{E}_{\mathbf{y} \sim \pi_{m-2}^i(\mathcal{X})}[\mathbf{y}\mathbf{y}^\top]\right)^{-1} \mathbf{x}}\right] \\
&= \frac{8\beta}{\sqrt{NT_{m-1}}} \mathbb{V}_{\mathcal{D}_{m-1}^i}^{\left(\frac{2\lambda}{NT_{m-1}}\right)}(\pi_{m-2}^i), && (13)
\end{aligned}$$

where the third inequality follows from our established confidence intervals in Lemma 4.4 guaranteeing that $\mathbf{x}_{*,t}^i \in \mathcal{X}_t^{i(m)}$ for all $(i, t, m) \in [N] \times [\mathcal{T}_{m-1}+1 : \mathcal{T}_m] \times [M]$ with probability at least $1-\delta$. Now, continuing form (13), we bound the cumulative regret of batches $m \geq 3$, as follows:

$$\begin{aligned}
\sum_{t=\mathcal{T}_2+1}^{T} \sum_{i=1}^{N} r_t^i &\leq \sum_{m=3}^{M} \frac{8\beta N T_m}{\sqrt{NT_{m-1}}} \mathbb{V}_{\mathcal{D}_{m-1}^i}^{\left(\frac{2\lambda}{NT_{m-1}}\right)}(\pi_{m-2}^i) \\
&\leq 8\beta \sqrt{dN \log d \log(NT\lambda^{-1})} \sum_{m=2}^{M} \frac{T_m}{\sqrt{T_{m-1}}} && \text{(Conditioned on the event in Eqn. (12))} \\
&= 8\beta M a \sqrt{dN \log d \log(NT\lambda^{-1})}. && (14)
\end{aligned}$$

Next, we bound cumulative regret of the first two batches. Under Assumption 1, during the first two batches, the instantaneous regret of each agent i at any round t is at most 2. Therefore

$$\sum_{t=1}^{\mathcal{T}_2} \sum_{i=1}^{N} r_t^i \leq 2N\mathcal{T}_2 = 4a\sqrt{dN}. \tag{15}$$

Note that for any $m \geq 3$, we can write T_m as

$$\begin{aligned}
T_m = aT_{m-1}^{\frac{1}{2}} = a^{\frac{3}{2}} T_{m-2}^{\frac{1}{4}} = \ldots = a^{\frac{2^{m-2}-1}{2^{m-3}}} T_2^{\frac{1}{2^{m-2}}} \\
T_m &= a^{\frac{1}{2^{m-2}}} a^{\frac{2^{m-2}-1}{2^{m-3}}} \left(\frac{T_2}{a}\right)^{\frac{1}{2^{m-2}}} \\
&= a^{\frac{2^{m-1}-1}{2^{m-2}}} \left(\sqrt{\frac{d}{N}}\right)^{\frac{1}{2^{m-2}}} \\
&= \left(\frac{a^{2^{m-1}-1} d^{\frac{1}{2}}}{N^{\frac{1}{2}}}\right)^{\frac{1}{2^{m-2}}}.
\end{aligned}$$

Our choice of a in the algorithm ensures that for any $M > 0$, $T_M = T$ and $\sum_{m=1}^{M} T_m \geq T_M = T$, and thus the choice of grid $\{\mathcal{T}_1, \ldots, \mathcal{T}_M\}$ is valid. If we let $M = 1 + \log\left(\frac{\log\left(\frac{NT}{d}\right)}{2} + 1\right)$, from (14) and (15), we conclude that, with probability at least $1 - 2\delta$, it holds that

$$\begin{aligned} R_T &\leq 4\sqrt{dNT}\left(\frac{NT}{d}\right)^{\frac{1}{2(2^{M-1}-1)}} + 8\beta M\sqrt{dNT\log d\log(NT\lambda^{-1})}\left(\frac{NT}{d}\right)^{\frac{1}{2(2^{M-1}-1)}} \\ &\leq \mathcal{O}\left(\sqrt{dNT\log d\log^2\left(\frac{KNT}{\delta\lambda}\right)\log\log\left(\frac{NT}{d}\right)}\right). \end{aligned} \tag{16}$$

B.3. Communication cost as number of bits transmitted

In this section, we consider the number of bits transmitted in a slightly modified version of DisBE-LUCB. To this end, we make the following minor modification to DisBE-LUCB. Let ϵ_0 be an additional input to the algorithm. In Line 9 of DisBE-LUCB, agent i sends vector $\tilde{\mathbf{u}}_m^i$ which is an ϵ_0-precise rounded version of $\mathbf{u}_m^i$. In particular, if it rounds each entry of $\mathbf{u}_m^i$ with precision ϵ_0, vector $\tilde{\mathbf{u}}_m^i$ will be obtained. Now, we observe how this extra rounding step affects confidence intervals in Lemma 4.4. In fact, we are interested in upper bounds on $\left|\left\langle \mathbf{x}, \tilde{\boldsymbol{\theta}}_m^i - \boldsymbol{\theta}\right\rangle\right|$, where $\tilde{\boldsymbol{\theta}}_m^i = \left(\Lambda_m^i\right)^{-1}\sum_{i=1}^N \tilde{\mathbf{u}}_m^i$.

For $\delta \in (0,1)$, let $\beta = 6\sqrt{\log\left(\frac{2KNT}{\delta}\right)} + \sqrt{\lambda}$. Then for all $\mathbf{x} \in \mathcal{X}_t^i, i \in [N], t \in [T], m \in [M]$, with probability at least $1 - \delta$, it holds that

$$\begin{aligned} \left|\left\langle \mathbf{x}, \tilde{\boldsymbol{\theta}}_m^i - \boldsymbol{\theta}\right\rangle\right| &= \left|\left\langle \mathbf{x}, \tilde{\boldsymbol{\theta}}_m^i - \boldsymbol{\theta}_m^i + \boldsymbol{\theta}_m^i - \boldsymbol{\theta}\right\rangle\right| \\ &\leq \left|\left\langle \mathbf{x}, \tilde{\boldsymbol{\theta}}_m^i - \boldsymbol{\theta}_m^i\right\rangle\right| + \left|\left\langle \mathbf{x}, \boldsymbol{\theta}_m^i - \boldsymbol{\theta}\right\rangle\right| \\ &\leq \left(\left\|\tilde{\boldsymbol{\theta}}_m^i - \boldsymbol{\theta}_m^i\right\|_{\Lambda_m^i} + \beta\right)\|\mathbf{x}\|_{\left(\Lambda_m^i\right)^{-1}} \qquad \text{(Lemma 4.4 and Cauchy Schwarz inequality)} \\ &\leq \left(\sqrt{\lambda_{\max}(\Lambda_m^i)}\left\|\tilde{\boldsymbol{\theta}}_m^i - \boldsymbol{\theta}_m^i\right\|_2 + \beta\right)\|\mathbf{x}\|_{\left(\Lambda_m^i\right)^{-1}} \\ &\leq \left(N\sqrt{dT}\epsilon_0 + \beta\right)\|\mathbf{x}\|_{\left(\Lambda_m^i\right)^{-1}}. \end{aligned} \tag{17}$$

Therefore, letting $\epsilon_0 = \frac{\beta}{N\sqrt{dT}}$, we have

$$\left|\left\langle \mathbf{x}, \tilde{\boldsymbol{\theta}}_m^i - \boldsymbol{\theta}\right\rangle\right| \leq 2\beta\|\mathbf{x}\|_{\left(\Lambda_m^i\right)^{-1}}, \tag{18}$$

which implies that replacing β in DisBE-LUCB with 2β, will result in the same order of regret as that of DisBE-LUCB for our modified algorithm. Moreover, since for transmission of each real number $\log(dNT)$ bits is used, the communication cost of our modified algorithm in terms of number of bits is same as that stated in Theorem 4.1 with an additional multiplicative factor $\log(dNT)$.

B.4. Relaxing the Assumption on Knowledge of $\mathcal{D}$

In this section, we relax this assumption and consider more realistic settings where each agent i can estimate matrix Λ_m^i in batch m up to an ϵ_m error, i.e.,

$$(1 - \epsilon_m)\Lambda_m^i \preceq \tilde{\Lambda}_m^i \preceq (1 + \epsilon_m)\Lambda_m^i, \tag{19}$$

where $\tilde{\Lambda}_m^i$ is an estimation of Λ_m^i. Given this estimation, we define

$$\tilde{\boldsymbol{\theta}}_m^i = \left(\tilde{\Lambda}_m^i\right)^{-1}\sum_{j=1}^N \mathbf{u}_m^j, \tag{20}$$

as the new estimation of $\boldsymbol{\theta}$ computed by agent i at batch m in this modified version of DisBE-LUCB.

We note that if the inequalities hold component-wise, i.e., $(1-\epsilon_m)\Lambda_m^i \leq \tilde{\Lambda}_m^i \leq (1+\epsilon_m)\Lambda_m^i$, this concludes that (19) holds. This is because for any positive semi-definite matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ such that $\mathbf{A} = \mathbf{B} + \mathbf{C}$, we have:

$$\mathbf{A} \succeq \mathbf{B},\ \mathbf{A} \succeq \mathbf{C}. \tag{21}$$

This combined with the fact that all $(1-\epsilon_m)\Lambda_m^i$, $\tilde{\Lambda}_m^i$, and $(1+\epsilon_m)\Lambda_m^i$ are positive semi-definite symmetric matrices ensures that (19) holds if $(1-\epsilon_m)\Lambda_m^i \leq \tilde{\Lambda}_m^i \leq (1+\epsilon_m)\Lambda_m^i$, and therefore, (19) is a weaker assumption than the component-wise assumption $(1-\epsilon_m)\Lambda_m^i \leq \tilde{\Lambda}_m^i \leq (1+\epsilon_m)\Lambda_m^i$.

Now, we define corresponding modified confidence intervals in the following lemma.

Lemma B.3. *Suppose* $\|\boldsymbol{\theta}\|_2 \leq 1$, $\left\|\mathbf{x}_{t,a}^i\right\|_2 \leq 1$, $\left|y_t^i\right| \leq 1$ *for all* $(a,i,t) \in [K]\times[N]\times[T]$ *and* $\epsilon_m \leq \sqrt{\frac{\lambda}{NT_m}}$ *for all* $m \in [M]$*. For* $\delta \in (0,1)$*, let* $\beta_m = 6\sqrt{\frac{\log\left(\frac{2KNT}{\delta}\right)}{1-\epsilon_m}} + 4\sqrt{\lambda}$*. Then for all* $\mathbf{x} \in \mathcal{X}_t^i, i \in [N], t \in [T], m \in [M]$*, with probability at least* $1-\delta$*, it holds that* $\left|\left\langle \mathbf{x}, \tilde{\boldsymbol{\theta}}_m^i - \boldsymbol{\theta}\right\rangle\right| \leq \beta_m \|\mathbf{x}\|_{\left(\tilde{\Lambda}_m^i\right)^{-1}}$.

Proof. The proof closely follows the steps in the proof of Lemma 4.4. For each batch $m \in [M]$, let $\mathbf{b}_m = \sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2}\sum_{i=1}^N \mathbf{x}_t^i y_t^i$ and $\mathbf{V}_m = \sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2}\sum_{i=1}^N \mathbf{x}_t^i {\mathbf{x}_t^i}^\top$. For a fixed $\mathbf{x} \in \mathcal{X}_t^i$ and $(i,t) \in [N]\times[T]$, let $z_{t,m}^{j,i} = \mathbf{x}^\top \left(\tilde{\Lambda}_m^i\right)^{-1}\left(\mathbf{x}_t^j y_t^j - \mathbb{E}_{\mathcal{X}\sim\mathcal{D}_m^i}\mathbb{E}_{\mathbf{x}\sim\pi_{m-1}^i(\mathcal{X})}[\mathbf{x}\mathbf{x}^\top]\boldsymbol{\theta}\right)$. Thus, we have

$$\begin{aligned}
\left|\left\langle \mathbf{x}, \tilde{\boldsymbol{\theta}}_m^i - \boldsymbol{\theta}\right\rangle\right| &= \left|\left\langle \mathbf{x}, \left(\tilde{\Lambda}_m^i\right)^{-1}\mathbf{b}_m - \boldsymbol{\theta}\right\rangle\right| \\
&= \left|\left\langle \mathbf{x}, \left(\tilde{\Lambda}_m^i\right)^{-1}\mathbf{b}_m\right\rangle - \left\langle \mathbf{x}, \left(\tilde{\Lambda}_m^i\right)^{-1}\tilde{\Lambda}_m^i\boldsymbol{\theta}\right\rangle\right| \\
&= \left|\left\langle \mathbf{x}, \left(\tilde{\Lambda}_m^i\right)^{-1}\mathbf{b}_m\right\rangle - \left\langle \mathbf{x}, \left(\tilde{\Lambda}_m^i\right)^{-1}\left(\Lambda_m^i - \lambda I\right)\boldsymbol{\theta}\right\rangle + \left\langle \mathbf{x}, \left(\tilde{\Lambda}_m^i\right)^{-1}\left(\Lambda_m^i - \tilde{\Lambda}_m^i - \lambda I\right)\boldsymbol{\theta}\right\rangle\right| \\
&\leq \left|\left\langle \mathbf{x}, \left(\tilde{\Lambda}_m^i\right)^{-1}\mathbf{b}_m\right\rangle - \left\langle \mathbf{x}, \left(\tilde{\Lambda}_m^i\right)^{-1}\left(\Lambda_m^i - \lambda I\right)\boldsymbol{\theta}\right\rangle\right| + \left|\left\langle \mathbf{x}, \left(\tilde{\Lambda}_m^i\right)^{-1}\left(\Lambda_m^i - \tilde{\Lambda}_m^i - \lambda I\right)\boldsymbol{\theta}\right\rangle\right| \\
&\leq \left|\mathbf{x}^\top\left(\tilde{\Lambda}_m^i\right)^{-1}\left(\mathbf{b}_m - \frac{NT_m}{2}\mathbb{E}_{\mathcal{X}\sim\mathcal{D}_m^i}\mathbb{E}_{\mathbf{x}\sim\pi_{m-1}^i(\mathcal{X})}[\mathbf{x}\mathbf{x}^\top]\boldsymbol{\theta}\right)\right| + 4\sqrt{\lambda}\|\mathbf{x}\|_{\left(\tilde{\Lambda}_m^i\right)^{-1}} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Cauchy Schwarz inequality)} \\
&= \left|\sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2}\sum_{j=1}^N z_{t,m}^{j,i}\right| + 4\sqrt{\lambda}\|\mathbf{x}\|_{\left(\tilde{\Lambda}_m^i\right)^{-1}},
\end{aligned} \tag{22}$$

where the second inequality follows from

$$
\begin{aligned}
\|\boldsymbol{\theta}\|_{\left(\tilde{\Lambda}_m^i\right)^{-1}\left(\Lambda_m^i-\tilde{\Lambda}_m^i-\lambda I\right)^2} &= \sqrt{\boldsymbol{\theta}^\top \left(\tilde{\Lambda}_m^i\right)^{-1} \left(\Lambda_m^i - \tilde{\Lambda}_m^i - \lambda I\right)^2 \boldsymbol{\theta}} \\
&\leq \|\boldsymbol{\theta}\|_2 \sqrt{\lambda_{\max}\left(\left(\tilde{\Lambda}_m^i\right)^{-1} \left(\Lambda_m^i - \tilde{\Lambda}_m^i - \lambda I\right)^2\right)} \\
&\leq \sqrt{\lambda_{\max}\left(\left(\tilde{\Lambda}_m^i\right)^{-1} \left(\Lambda_m^i - \tilde{\Lambda}_m^i\right)^2 + \lambda^2 \left(\tilde{\Lambda}_m^i\right)^{-1}\right)} && (\|\boldsymbol{\theta}\|_2 \leq 1) \\
&\leq \sqrt{\lambda_{\max}\left(\left(\tilde{\Lambda}_m^i\right)^{-1} \left(\Lambda_m^i - \tilde{\Lambda}_m^i\right)^2 + \lambda^2 \left(\tilde{\Lambda}_m^i\right)^{-1}\right)} \\
&\leq \sqrt{\lambda_{\max}\left(\left(\tilde{\Lambda}_m^i\right)^{-1} \left(\Lambda_m^i - \tilde{\Lambda}_m^i\right)^2\right)} + \sqrt{\lambda} && \text{(Cauchy Schwarz inequality)} \\
&\leq \epsilon_m \sqrt{\lambda_{\max}\left(\tilde{\Lambda}_m^i\right)} + \sqrt{\lambda} && \text{(Eqn. (19))} \\
&\leq 2\epsilon_m \sqrt{\lambda_{\max}\left(\Lambda_m^i\right)} + \sqrt{\lambda} && \text{(Eqn. (19))} \\
&\leq \epsilon_m \sqrt{NT_m} + 3\sqrt{\lambda} \\
&\leq 4\sqrt{\lambda}. && \left(\epsilon_m \leq \sqrt{\tfrac{\lambda}{NT_m}}\right)
\end{aligned}
$$

Note that

$$
\mathbb{E}\left[z_{t,m}^{j,i}\right] = \mathbb{E}\left[\mathbf{x}^\top \left(\tilde{\Lambda}_m^i\right)^{-1} \left(\mathbf{x}_t^j (\mathbf{x}_t^{j\top} \boldsymbol{\theta} + \eta_t^j) - \mathbb{E}_{\mathcal{X}\sim\mathcal{D}_m^i} \mathbb{E}_{\mathbf{x}\sim\pi_{m-1}^i(\mathcal{X})}[\mathbf{x}\mathbf{x}^\top]\boldsymbol{\theta}\right)\right] = 0,
$$

(Noise η_t^j is zero-mean and independent of $\mathbf{x}_t^j$)

By Azuma's inequality, for a fixed $\mathbf{x} \in \mathcal{X}_t^i$ and $(i,t) \in [N] \times [T]$, we have

$$
\mathbb{P}\left(\left|\sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2} \sum_{j=1}^{N} z_{t,m}^{j,i}\right| \geq \alpha \|\mathbf{x}\|_{\left(\tilde{\Lambda}_m^i\right)^{-1}}\right) \leq 2\text{exp}\left(\frac{-\alpha^2 \|\mathbf{x}\|^2_{\left(\tilde{\Lambda}_m^i\right)^{-1}}}{2c_m^i}\right), \tag{23}
$$

where

$$
\begin{aligned}
c_m^i &= \sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2} \sum_{j=1}^{N} \left| \mathbf{x}^\top \left(\tilde{\Lambda}_m^i\right)^{-1} \left(\mathbf{x}_t^j y_t^j - \mathbb{E}_{\mathcal{X}\sim\mathcal{D}_m^i} \mathbb{E}_{\mathbf{x}\sim\pi_{m-1}^i(\mathcal{X})}[\mathbf{x}\mathbf{x}^\top]\boldsymbol{\theta} \right) \right|^2 \\
&\le 2 \sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2} \sum_{j=1}^{N} \left| \mathbf{x}^\top \left(\tilde{\Lambda}_m^i\right)^{-1} \mathbf{x}_t^j y_t^j \right|^2 + NT_m \left| \mathbf{x}^\top \left(\tilde{\Lambda}_m^i\right)^{-1} \mathbb{E}_{\mathcal{X}\sim\mathcal{D}_m^i} \mathbb{E}_{\mathbf{x}\sim\pi_{m-1}^i(\mathcal{X})}[\mathbf{x}\mathbf{x}^\top]\boldsymbol{\theta} \right|^2 \\
&\le 2 \sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2} \sum_{j=1}^{N} \left| \mathbf{x}^\top \left(\tilde{\Lambda}_m^i\right)^{-1} \mathbf{x}_t^j \right|^2 + \frac{4}{NT_m} \left| \mathbf{x}^\top \left(\tilde{\Lambda}_m^i\right)^{-1} \left(\Lambda_m^i - \lambda I\right) \boldsymbol{\theta} \right|^2 \\
&= 2 \sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2} \sum_{j=1}^{N} \mathbf{x}^\top \left(\tilde{\Lambda}_m^i\right)^{-1} \mathbf{x}_t^j {\mathbf{x}_t^j}^\top \left(\tilde{\Lambda}_m^i\right)^{-1} \mathbf{x} + \frac{4}{NT_m} \left| \mathbf{x}^\top \left(\tilde{\Lambda}_m^i\right)^{-1} \left(\Lambda_m^i\right) \boldsymbol{\theta} - \lambda \mathbf{x}^\top \left(\tilde{\Lambda}_m^i\right)^{-1} \boldsymbol{\theta} \right|^2 \\
&\le 2\mathbf{x}^\top \left(\tilde{\Lambda}_m^i\right)^{-1} \mathbf{V}_m \left(\tilde{\Lambda}_m^i\right)^{-1} \mathbf{x} + \frac{1}{1-\epsilon_m} \left(4 + \frac{8\lambda}{NT_m}\right) \|\mathbf{x}\|^2_{(\tilde{\Lambda}_m^i)^{-1}} && \text{(Cauchy Schwarz inequality)} \\
&\le 4\mathbf{x}^\top \left(\tilde{\Lambda}_m^i\right)^{-1} \Lambda_m^i \left(\tilde{\Lambda}_m^i\right)^{-1} \mathbf{x} + \frac{1}{1-\epsilon_m} \left(4 + \frac{8\lambda}{NT_m}\right) \|\mathbf{x}\|^2_{(\tilde{\Lambda}_m^i)^{-1}} && \text{(Conditioned on the event in Eqn. (7))} \\
&\le \frac{4}{1-\epsilon_m} \mathbf{x}^\top \left(\tilde{\Lambda}_m^i\right)^{-1} \mathbf{x} + \frac{1}{1-\epsilon_m} \left(4 + \frac{8\lambda}{NT_m}\right) \|\mathbf{x}\|^2_{(\tilde{\Lambda}_m^i)^{-1}} && ((1-\epsilon_m)\Lambda_m^i \preceq \tilde{\Lambda}_m^i) \\
&= \frac{8}{1-\epsilon_m} \left(1 + \frac{\lambda}{NT_m}\right) \|\mathbf{x}\|^2_{(\tilde{\Lambda}_m^i)^{-1}} \\
&\le \frac{16}{1-\epsilon_m} \|\mathbf{x}\|^2_{(\tilde{\Lambda}_m^i)^{-1}}\,, && (24)
\end{aligned}
$$

where the third inequity follows from the fact that

$$
\begin{aligned}
\boldsymbol{\theta}^\top \left(\Lambda_m^i \left(\tilde{\Lambda}_m^i\right)^{-1} \Lambda_m^i \right) \boldsymbol{\theta} &\le \|\boldsymbol{\theta}\|^2 \lambda_{\max} \left(\Lambda_m^i \left(\tilde{\Lambda}_m^i\right)^{-1} \Lambda_m^i \right) \\
&\le \lambda_{\max} \left(\Lambda_m^i \left(\tilde{\Lambda}_m^i\right)^{-1} \Lambda_m^i \right) && (\|\boldsymbol{\theta}\|_2 \le 1) \\
&\le \frac{1}{1-\epsilon_m} \lambda_{\max} \left(\Lambda_m^i \right) && ((1-\epsilon_m)\Lambda_m^i \preceq \tilde{\Lambda}_m^i) \\
&\le \frac{\lambda + NT_m}{1-\epsilon_m}.
\end{aligned}
$$

Combining (22), (23) and (24), and by a union bound, we have

$$
\mathbb{P}\left(\left| \left\langle \mathbf{x}, \boldsymbol{\theta}_m^i - \boldsymbol{\theta} \right\rangle \right| \le \left(6\sqrt{\frac{\log\left(\frac{2KNT}{\delta}\right)}{1-\epsilon_m}} + 4\sqrt{\lambda} \right) \|\mathbf{x}\|_{(\tilde{\Lambda}_m^i)^{-1}}\,,\ \forall \mathbf{x} \in \mathcal{X}_t^i, i \in [N], t \in [T], m \in [M] \right) \ge 1 - \delta. \quad (25)
$$

□

Now, we state the regret bound for DisBE-LUCB with $\tilde{\Lambda}_m^i$ and $\tilde{\boldsymbol{\theta}}_m^i$.

Theorem B.4. *Fix* $M = 1 + \log\left(\log\left(NT/d\right)/2 + 1\right)$. *Under the setting of Lemma B.3, if* $T \ge \Omega\left(d^{22} \log^2(NT/\delta) \log^2 d \log^2(d\lambda^{-1})\right)$ *and* $\beta = \max_{m\in[M]} \beta_m$*, then with probability at least* $1 - 2\delta$*, it holds that* $R_T \le \mathcal{O}\left(\frac{1}{1-\max_{m\in[M]} \epsilon_m} \sqrt{dNT \log d \log^2\left(\frac{KNT}{\delta\lambda}\right) \log\log\left(\frac{NT}{d}\right)} \right)$*, where the communication cost is measured by the number of real numbers communicated by the agents.*

Proof. The proof follows similar steps to those in the proof of Theorem 4.1.

We focus on the regret of the i-th agent at m-th batch for any $m \geq 3$. Let $\mathcal{D}_m^i$ be the distribution based on which the surviving sets $\mathcal{X}_t^{i(m)}$ for all $t \in [\mathcal{T}_{m-1}+1 : \mathcal{T}_m]$ are generated when conditioned on the first $m-1$ batches. For any $t \in [\mathcal{T}_{m-1}+1 : \mathcal{T}_m]$, conditioned on the event that the confidence intervals in Lemma 4.4 hold, we have

$$
\begin{aligned}
r_t^i &= \mathbb{E}\left[\langle \boldsymbol{\theta}, \mathbf{x}_{*,t}^i\rangle - \langle \boldsymbol{\theta}, \mathbf{x}_t^i\rangle\right] \\
&\leq \mathbb{E}\left[\langle \tilde{\boldsymbol{\theta}}_{m-1}^i, \mathbf{x}_{*,t}^i\rangle - \langle \tilde{\boldsymbol{\theta}}_{m-1}^i, \mathbf{x}_t^i\rangle + \beta\left\|\mathbf{x}_{*,t}^i\right\|_{(\tilde{\Lambda}_{m-1}^i)^{-1}} + \beta\left\|\mathbf{x}_t^i\right\|_{(\tilde{\Lambda}_{m-1}^i)^{-1}}\right] && \text{(Lemma B.3)} \\
&\leq 2\beta\mathbb{E}\left[\left\|\mathbf{x}_{*,t}^i\right\|_{(\tilde{\Lambda}_{m-1}^i)^{-1}} + \left\|\mathbf{x}_t^i\right\|_{(\tilde{\Lambda}_{m-1}^i)^{-1}}\right] && (\mathbf{x}_{*,t}^i \in \mathcal{X}_t^{i(m)}) \\
&\leq 4\beta\mathbb{E}\left[\max_{\mathbf{x}\in\mathcal{X}_t^{i(m)}} \|\mathbf{x}\|_{(\tilde{\Lambda}_{m-1}^i)^{-1}}\right] \\
&\leq 4\beta\mathbb{E}_{\mathcal{X}\sim\mathcal{D}_m^i}\left[\max_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x}\|_{(\tilde{\Lambda}_{m-1}^i)^{-1}}\right] \\
&\leq 4\beta\mathbb{E}_{\mathcal{X}\sim\mathcal{D}_{m-1}^i}\left[\max_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x}\|_{(\tilde{\Lambda}_{m-1}^i)^{-1}}\right] \\
&\leq \frac{4\beta}{\sqrt{1-\epsilon_m}}\mathbb{E}_{\mathcal{X}\sim\mathcal{D}_{m-1}^i}\left[\max_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x}\|_{(\Lambda_{m-1}^i)^{-1}}\right] && ((1-\epsilon_m)\Lambda_m^i \preceq \tilde{\Lambda}_m^i) \\
&\leq \frac{8\beta}{\sqrt{NT_{m-1}(1-\epsilon_m)}}\mathbb{E}_{\mathcal{X}\sim\mathcal{D}_{m-1}^i}\left[\max_{\mathbf{x}\in\mathcal{X}} \sqrt{\mathbf{x}^\top\left(\frac{2\lambda}{NT_{m-1}}I + \mathbb{E}_{\mathcal{X}\sim\mathcal{D}_{m-1}^i}\mathbb{E}_{\mathbf{y}\sim\pi_{m-2}^i(\mathcal{X})}[\mathbf{y}\mathbf{y}^\top]\right)^{-1}\mathbf{x}}\right] \\
&= \frac{8\beta}{\sqrt{NT_{m-1}(1-\epsilon_m)}}\mathbb{V}_{\mathcal{D}_{m-1}^i}^{(\frac{2\lambda}{NT_{m-1}})}(\pi_{m-2}^i), && (26)
\end{aligned}
$$

where the third inequality follows from our established confidence intervals in Lemma B.3 guaranteeing that $\mathbf{x}_{*,t}^i \in \mathcal{X}_t^{i(m)}$ for all $(i,t,m) \in [N]\times[\mathcal{T}_{m-1}+1 : \mathcal{T}_m]\times[M]$ with probability at least $1-\delta$. The rest of the proof follows the steps as those in the proof of Theorem 4.1 with an additional $\frac{1}{\sqrt{1-\epsilon_m}}$ multiplicative factor in the bound.

Therefore, we conclude that, with probability at least $1-2\delta$, it holds that

$$
\begin{aligned}
R_T &\leq 4\sqrt{dNT}\left(\frac{NT}{d}\right)^{\frac{1}{2(2^{M-1}-1)}} + 8\beta M\sqrt{\frac{dNT\log d\log(NT\lambda^{-1})}{1-\max_{m\in[M]}\epsilon_m}}\left(\frac{NT}{d}\right)^{\frac{1}{2(2^{M-1}-1)}} \\
&\leq \mathcal{O}\left(\frac{1}{1-\max_{m\in[M]}\epsilon_m}\sqrt{dNT\log d\log^2\left(\frac{KNT}{\delta\lambda}\right)\log\log\left(\frac{NT}{d}\right)}\right). && (27)
\end{aligned}
$$

□

C. Decentralized Batch Elimination LUCB without Server

In this environment, the agents are represented by the nodes of an undirected and connected graph G. Each agent i can send and receive messages only to and from its immediate neighbors $j \in \mathcal{N}(i)$.

Definition C.1 (Communication Matrix)**.** For an undirected connected graph G with N nodes, $\mathbf{P} \in \mathbb{R}^{N\times N}$ is a symmetric communication matrix if it satisfies the following three conditions: (i) $\mathbf{P}_{i,j} = 0$ if there is no connection between nodes i and j; (ii) the sum of each row and column of $\mathbf{P}$ is 1; (iii) the eigenvalues are real and their magnitude is less than 1, i.e., $1 = |\lambda_1| > |\lambda_2| \geq \ldots |\lambda_N| \geq 0$.

We assume that $\mathbf{P}$ is known to the agents. We remark that $\mathbf{P}$ can be constructed with little global information about the graph, such as its adjacency matrix and the graph's maximal degree; For example, one can compute it as $\mathbf{P} = I_N - \frac{1}{\delta_{\max}+1}\mathbf{D}^{-1/2}\mathcal{L}\mathbf{D}^{-1/2}$, where $\delta_{\max}$ is the maximum degree of the graph, $\mathcal{L} \in \mathbb{R}^{N\times N}$ is the graph Laplacian, and $\mathbf{D} \in \mathbb{R}^{N\times N}$ is a diagonal matrix whose entries are the degrees of the nodes (see Duchi et al. (2011) for details).

Running consensus. In order to share information about agents' past actions among the network, we rely on *running consensus*, e.g., (Lynch, 1996; Xiao and Boyd, 2004). The goal of running consensus is that after enough rounds of communication, each agent has an accurate estimate of the average (over all agents) of the initial values of each agent. Precisely, let $\boldsymbol{\nu}_0 \in \mathbb{R}^N$ be a vector, where each entry $\boldsymbol{\nu}_{0,i}, i \in [N]$ represents agent's i information at some initial round. Then, running consensus aims at providing an accurate estimate of the average $\frac{1}{N}\sum_{i\in[N]} \boldsymbol{\nu}_{0,i}$ for each agent. It turns out that the communication matrix $\mathbf{P}$ defined in Definition C.1 plays a key role in reaching consensus. The details are standard in the rich related literature (Xiao and Boyd, 2004; Lynch, 1996). Here, we only give a brief explanation of the high-level principles. Roughly speaking, a consensus algorithm updates $\boldsymbol{\nu}_0$ by $\boldsymbol{\nu}_1 = \mathbf{P}\boldsymbol{\nu}_0$, $\boldsymbol{\nu}_2 = \mathbf{P}\boldsymbol{\nu}_1$ and so on. Note that this operation respects the network structure since the updated value $\boldsymbol{\nu}_{1,j}$ is a weighted average of only $\boldsymbol{\nu}_{0,j}$ itself and neighbor-only values $\boldsymbol{\nu}_{0,i}, i \in \mathcal{N}(j)$. Thus, after S rounds, agent j has access to entry j of $\boldsymbol{\nu}_S = \mathbf{P}^S\boldsymbol{\nu}_0$. We adapt *polynomial filtering* introduced in Martínez-Rubio et al. (2019); Seaman et al. (2017) to speed up the mixing of information by following an approach whose convergence rate is faster than the standard multiplication method above. Specifically, after S communication rounds, instead of $\mathbf{P}^S$, agents compute and apply to the initial vector $\boldsymbol{\nu}_0$ an appropriate re-scaled *Chebyshev polynomial* $q_S(\mathbf{P})$ of degree S of the communication matrix. Recall that Chebyshev polynomials are defined recursively. It turns out that the Chebyshev polynomial of degree ℓ for a communication matrix $\mathbf{P}$ is also given by a recursive formula as follows: $q_{\ell+1}(\mathbf{P}) = \frac{2w_\ell}{|\lambda_2|w_{\ell+1}}\mathbf{P}q_\ell(\mathbf{P}) - \frac{w_{\ell-1}}{w_{\ell+1}}q_{\ell-1}(\mathbf{P})$, where $w_0 = 0, w_1 = 1/|\lambda_2|, w_{\ell+1} = 2w_\ell/|\lambda_2| - w_{\ell-1}$, $q_0(\mathbf{P}) = I$ and $q_1(\mathbf{P}) = \mathbf{P}$. Specifically, in a Chebyshev-accelerated gossip protocol (Martínez-Rubio et al., 2019), the agents update their estimates of the average of the initial vector's $\boldsymbol{\nu}_0$ entries as follows:

$$\boldsymbol{\nu}_{\ell+1} = (2w_\ell)/(|\lambda_2|w_{\ell+1})\mathbf{P}\boldsymbol{\nu}_\ell - (w_{\ell-1}/w_{\ell+1})\boldsymbol{\nu}_{\ell-1}. \tag{28}$$

DecBE-LUCB, presented in Algorithm 2, implements the Chebyshev-accelerated gossip protocol outlined above for every entry of vectors $\mathbf{u}_m^i = \sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+T_m/2} \mathbf{x}_t^i y_t^i$ at the end of m-th batch.

The accelerated consensus algorithm, summarized in Algorithm 3, guarantees fast mixing of information thanks to the following key property stated in Lemma 3 of Martínez-Rubio et al. (2019): for $\epsilon \in (0,1)$ and any vector $\boldsymbol{\nu_0}$ in the N-dimensional simplex, it holds that

$$\|Nq_S(\mathbf{P})\boldsymbol{\nu_0} - \mathbf{1}\|_2 \leq \epsilon, \text{ if } S = \frac{\log(2N/\epsilon)}{\sqrt{2\log(1/|\lambda_2|)}}. \tag{29}$$

In view of this, DecBE-LUCB properly implements the accelerated consensus algorithm such that for every $i \in [N]$ and $m \in [M]$, the vector $\mathbf{u}_m^i$ is communicated within the network during the last S rounds of batch m. At round $\mathcal{T}_m + 1$, agent i has access to $\sum_{j=1}^N a_{i,j}\mathbf{u}_m^j$, where $a_{i,j} = N[q_S(\mathbf{P})]_{i,j}$. Thanks to (29), $a_{i,j}$ is ϵ close to 1, thus, these are good approximations of the true $\sum_{j=1}^N \mathbf{u}_m^j$. Furthermore, the choice of grid $\mathcal{T} = \{\mathcal{T}_0, \mathcal{T}_1, \ldots, \mathcal{T}_M\}$ in DecBE-LUCB is slightly different than what used in DisBE-LUCB.

C.1. Theoretical guarantees of DecBE-LUCB

As the first step in regret analysis of DecBE-LUCB, we establish the following confidence intervals.

Lemma C.2 (Confidence intervals for DecBE-LUCB)**.** *Suppose Assumption 1 holds. Fix $\delta \in (0,1)$ and let $\epsilon = \frac{\beta}{\sqrt{d}}$ and $\gamma = 2\beta$, where β is defined in Lemma 4.4. Then*

$$\mathbb{P}\left(\left|\left\langle \mathbf{x}, \hat{\boldsymbol{\theta}}_m^i - \boldsymbol{\theta}\right\rangle\right| \leq \gamma\|\mathbf{x}\|_{\left(\Lambda_m^i\right)^{-1}}, \ \forall \mathbf{x} \in \mathcal{X}_t^i, i \in [N], t \in [T], m \in [M]\right) \geq 1 - \delta. \tag{30}$$

Algorithm 2 DecBE-LUCB for agent i

```
Input: $N$, $d$, $\delta$, $T$, $M$, $\lambda$, $\epsilon$
Initialization: $S = \frac{\log(2N/\epsilon)}{\sqrt{2\log(1/|\lambda_2|)}}$, $a = \sqrt{T+S}\left(\frac{N(T+S)}{d}\right)^{\frac{1}{2(2^{M-1}-1)}}$, $T_1 = T_2 = a\sqrt{\frac{d}{N}}+S$, $T_m = \lfloor a\sqrt{T_{m-1}-S}+S\rfloor$, $\boldsymbol{\theta}_0^i = \mathbf{0}$, $\Lambda_0^i = \lambda I$, $\mathcal{T}_0 = 0$, $\mathcal{T}_m = \mathcal{T}_{m-1} + T_m$, $\lambda = 5\log\left(\frac{4dT}{\delta}\right)$, $\gamma = 12\sqrt{\log\left(\frac{2KNT}{\delta}\right)} + 2\sqrt{\lambda}$, arbitrary policy $\pi_0^i$
for $m = 1, \ldots, M$ do
    for $t = \mathcal{T}_{m-1}+1, \ldots, \min\{\mathcal{T}_m, T\}$ do
        Let $\mathcal{X}_t^{i(m)} = \cap_{k=0}^{m-1}\mathcal{E}\left(\mathcal{X}_t^i; (\Lambda_k^i, \hat{\boldsymbol{\theta}}_k^i, \gamma)\right)$
        Play arm $a_{i,t}$ associated with feature vector $\mathbf{x}_t^i \sim \pi_{m-1}\left(\mathcal{X}_t^{i(m)}\right)$ and observe $y_t^i$.
    end forSet $\mathcal{K}_0^i = \sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_{m-1}+(T_m-S)/2} \mathbf{x}_t^i y_t^i$
    for $t = \mathcal{T}_m - S + 1$ do
        Let $\mathcal{X}_t^{i(m)} = \cap_{k=0}^{m-1}\mathcal{E}\left(\mathcal{X}_t^i; (\Lambda_k^i, \hat{\boldsymbol{\theta}}_k^i, \gamma)\right)$
        Play arm $a_{i,t}$ associated with feature vector $\mathbf{x}_t^i \sim \pi_{m-1}\left(\mathcal{X}_t^{i(m)}\right)$ and observe $y_t^i$.
        Send each entry of $\mathcal{K}_0^i$, i.e., $[\mathcal{K}_0^i]_n,\ \forall n \in [d]$ to your neighbors $\mathcal{N}(j)$ and receive the corresponding values from them. For each $n \in [d]$, update $[\mathcal{K}_1^i]_n = \mathbf{P}_{i,i}[\mathcal{K}_0^i]_n + \sum_{j\in\mathcal{N}(i)} \mathbf{P}_{i,j}[\mathcal{K}_0^j]_n$
    end for
    Set $s = 1$
    for $t = \mathcal{T}_m - S + 2, \ldots, \mathcal{T}_m$ do
        Construct set $\mathcal{X}_t^{i(m)} = \cap_{k=0}^{m-1}\mathcal{E}\left(\mathcal{X}_t^i; (\Lambda_k^i, \hat{\boldsymbol{\theta}}_k^i, \gamma)\right)$.
        Play arm $a_{i,t}$ associated with feature vector $\mathbf{x}_t^i \sim \pi_{m-1}\left(\mathcal{X}_t^{i(m)}\right)$ and observe $y_t^i$. $[\mathcal{K}_{s+1}^i]_n = \mathrm{Comm}([\mathcal{K}_s^i]_n, [\mathcal{K}_{s-1}^i]_n, s+1), \forall n \in [d]$
        $s = s + 1$
    end for
    Compute/construct
        $\Lambda_m^i = \lambda I + \frac{N(T_m - S)}{2}\mathbb{E}_{\mathcal{X}\sim\mathcal{D}_m^i}\mathbb{E}_{\mathbf{x}\sim\pi_{m-1}^i(\mathcal{X})}[\mathbf{x}\mathbf{x}^\top]$,
        $\hat{\boldsymbol{\theta}}_m^i = \left(\Lambda_m^i\right)^{-1}\bar{\mathbf{u}}_{m,i}$,
        $\mathcal{S}_m^i = \left\{\mathcal{X}_t^{i(m+1)}\right\}_{t=\mathcal{T}_{m-1}+(T_m-S)/2+1}^{\mathcal{T}_m}$,
        $\pi_m^i = \mathrm{ExpPol}\left(\frac{2\lambda}{N(T_m - S)}, \mathcal{S}_m^i\right)$.
end for
```

Proof. Recall the definition of $\boldsymbol{\theta}_m^i$ in (4). For a fixed $\mathbf{x} \in \mathcal{X}_t^i$ and $(i,t) \in [N] \times [T]$, we have

$$\begin{aligned}
\left|\left\langle \mathbf{x}, \hat{\boldsymbol{\theta}}_m^i - \boldsymbol{\theta} \right\rangle\right| &\leq \left|\left\langle \mathbf{x}, \boldsymbol{\theta}_m^i - \boldsymbol{\theta} \right\rangle\right| + \left|\left\langle \mathbf{x}, \hat{\boldsymbol{\theta}}_m^i - \boldsymbol{\theta}_m^i \right\rangle\right| \\
&\leq \left|\left\langle \mathbf{x}, \boldsymbol{\theta}_m^i - \boldsymbol{\theta} \right\rangle\right| + \|\mathbf{x}\|_{(\Lambda_m^i)^{-2}} \left\| \bar{\mathbf{u}}_{m,i} - \sum_{j=1}^{N} \mathbf{u}_m^j \right\|_2 && \text{(Cauchy Schwarz inequality)} \\
&\leq \left|\left\langle \mathbf{x}, \boldsymbol{\theta}_m^i - \boldsymbol{\theta} \right\rangle\right| + \epsilon\sqrt{d}\|\mathbf{x}\|_{(\Lambda_m^i)^{-1}} && \text{(Assumption 1 and choice of } S \text{ in (29))} \\
&= \left|\left\langle \mathbf{x}, \boldsymbol{\theta}_m^i - \boldsymbol{\theta} \right\rangle\right| + \beta\|\mathbf{x}\|_{(\Lambda_m^i)^{-1}} . && (31)
\end{aligned}$$

Combining Lemma 4.4 and (31), we have

$$\mathbb{P}\left(\left|\left\langle \mathbf{x}, \hat{\boldsymbol{\theta}}_m^i - \boldsymbol{\theta} \right\rangle\right| \leq 2\beta\|\mathbf{x}\|_{(\Lambda_m^i)^{-1}} , \ \forall \mathbf{x} \in \mathcal{X}_t^i, i \in [N], t \in [T], m \in [M] \right) \geq 1 - \delta. \tag{32}$$

□

Theorem C.3. *Fix* $M = 1 + \log\left(\frac{\log\left(\frac{N(T+S)}{d}\right)}{2} + 1 \right)$, *with* S *defined in* (29) *for* $\epsilon = 6\sqrt{\frac{\log\left(\frac{2dKNT}{\delta}\right)}{d}}$ *in Algorithm 1. Suppose Assumption 1 holds. If* $T \geq \Omega\left(d^{22} \log^2(\frac{NT}{\delta}) \log^2 d \log^2(d\lambda^{-1}) \right)$, *then with probability at least* $1 - 2\delta$, *it holds that*

$$R_T \leq \mathcal{O}\left(\left(\frac{N\log(dN)}{\sqrt{1/|\lambda_2|}} + \sqrt{dN\left(T + \frac{\log(dN)}{\sqrt{1/|\lambda_2|}}\right) \log d \log^2\left(\frac{KN\left(T + \frac{\log(dN)}{\sqrt{1/|\lambda_2|}}\right)}{\delta\lambda} \right)} \right) \log\log\left(\frac{NT}{d}\right) \right), \tag{33}$$

and

$$\textit{Communication Cost} \leq \mathcal{O}\left(\frac{\delta_{\max} dN \log(dN)}{\sqrt{\log(1/|\lambda_2|)}} \right). \tag{34}$$

Proof. The proof follows similar steps as those of Theorem 4.1's proof. We focus on the regret of m-th batch for any $m \geq 3$.

For any $i \in [N]$, $t \in [\mathcal{T}_{m-1}+1 : \mathcal{T}_m]$, conditioned on the event that the confidence intervals in Lemma C.2 hold, we have

$$
\begin{aligned}
r_t^i &= \mathbb{E}\left[\langle \boldsymbol{\theta}, \mathbf{x}_{*,t}^i\rangle - \langle \boldsymbol{\theta}, \mathbf{x}_t^i\rangle\right] \\
&\leq \mathbb{E}\left[\langle \hat{\boldsymbol{\theta}}_{m-1}^i, \mathbf{x}_{*,t}^i\rangle - \langle \hat{\boldsymbol{\theta}}_{m-1}^i, \mathbf{x}_t^i\rangle + \beta\left\|\mathbf{x}_{*,t}^i\right\|_{\left(\Lambda_{m-1}^i\right)^{-1}} + \beta\left\|\mathbf{x}_t^i\right\|_{\left(\Lambda_{m-1}^i\right)^{-1}}\right] && \text{(Lemma C.2)} \\
&\leq 2\gamma\mathbb{E}\left[\left\|\mathbf{x}_{*,t}^i\right\|_{\left(\Lambda_{m-1}^i\right)^{-1}} + \left\|\mathbf{x}_t^i\right\|_{\left(\Lambda_{m-1}^i\right)^{-1}}\right] && (\mathbf{x}_{*,t}^i \in \mathcal{X}_t^{i(m)}) \\
&\leq 4\gamma\mathbb{E}\left[\max_{\mathbf{x}\in\mathcal{X}_t^{i(m)}} \|\mathbf{x}\|_{\left(\Lambda_{m-1}^i\right)^{-1}}\right] \\
&\leq 4\gamma\mathbb{E}_{\mathcal{X}\sim\mathcal{D}_m^i}\left[\max_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x}\|_{\left(\Lambda_{m-1}^i\right)^{-1}}\right] \\
&\leq 4\gamma\mathbb{E}_{\mathcal{X}\sim\mathcal{D}_{m-1}^i}\left[\max_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x}\|_{\left(\Lambda_{m-1}^i\right)^{-1}}\right] \\
&\leq \frac{8\gamma}{\sqrt{N(T_{m-1}-S)}}\mathbb{E}_{\mathcal{X}\sim\mathcal{D}_{m-1}^i}\left[\max_{\mathbf{x}\in\mathcal{X}} \sqrt{\mathbf{x}^\top\left(\frac{2\lambda}{N(T_{m-1}-S)}I + \mathbb{E}_{\mathcal{X}\sim\mathcal{D}_{m-1}^i}\mathbb{E}_{\mathbf{y}\sim\pi_{m-2}^i(\mathcal{X})}[\mathbf{y}\mathbf{y}^\top]\right)^{-1}\mathbf{x}}\right] \\
&= \frac{8\gamma}{\sqrt{N(T_{m-1}-S)}}\mathbb{V}_{\mathcal{D}_{m-1}^i}^{\left(\frac{2\lambda}{N(T_{m-1}-S)}\right)}(\pi_{m-2}^i), && (35)
\end{aligned}
$$

where the third inequality follows from our established confidence intervals in Lemma C.2 guaranteeing that $\mathbf{x}_{*,t}^i \in \mathcal{X}_t^{i(m)}$ for all $(i,t,m) \in [N] \times [\mathcal{T}_{m-1}+1 : \mathcal{T}_m] \times [M]$ with probability at least $1-\delta$. Now, continuing form (13), we bound the cumulative regret of batches $m \geq 3$, as follows:

$$
\begin{aligned}
\sum_{t=\mathcal{T}_2+1}^{T}\sum_{i=1}^{N} r_t^i &\leq 2MSN + \sum_{m=3}^{M}\sum_{t=\mathcal{T}_{m-1}+1}^{\mathcal{T}_m-S}\sum_{i=1}^{N} r_t^i \\
&\leq 2MSN + \frac{8\gamma MN(T_m-S)}{\sqrt{N(T_{m-1}-S)}}\mathbb{V}_{\mathcal{D}_{m-1}^i}^{\left(\frac{2\lambda}{N(T_{m-1}-S)}\right)}(\pi_{m-2}^i) \\
&\leq 2MSN + +8\gamma M\sqrt{dN\log d\log(NT\lambda^{-1})}\sum_{m=2}^{M}\frac{T_m-S}{\sqrt{T_{m-1}-S}} && \text{(Conditioned on the event in Eqn. (12))} \\
&= 2MSN + 8\gamma Ma\sqrt{dN\log d\log(NT\lambda^{-1})}. && (36)
\end{aligned}
$$

Next, we bound cumulative regret of the first two batches. Under Assumption 1, during the first two batches, the instantaneous regret of each agent i at any round t is at most 2. Therefore

$$
\sum_{t=1}^{\mathcal{T}_2}\sum_{i=1}^{N} r_t^i \leq 2N\mathcal{T}_2 = 4a\sqrt{dN}. \tag{37}
$$

Note that the choice of a in the algorithm ensures that for any $M > 0$, $T_M = T$ and $\sum_{m=1}^{M} T_m \geq T_M = T$, and thus the choice of grid $\{\mathcal{T}_1, \ldots, \mathcal{T}_M\}$ is valid. If we let $M = 1 + \log\left(\frac{\log\left(\frac{N(T+S)}{d}\right)}{2} + 1\right)$, from (36) and (37), we conclude that, with probability at least $1-2\delta$, it holds that

$$
\begin{aligned}
R_T &\leq 2MSN + 4\sqrt{dN(T+S)}\left(\frac{NT}{d}\right)^{\frac{1}{2(2^{M-1}-1)}} + 8\gamma M\sqrt{dNT\log d\log(NT\lambda^{-1})}\left(\frac{N(T+S)}{d}\right)^{\frac{1}{2(2^{M-1}-1)}} \\
&\leq \mathcal{O}\left(\left(\frac{N\log(dN)}{\sqrt{1/|\lambda_2|}} + \sqrt{dN\left(T+\frac{\log(dN)}{\sqrt{1/|\lambda_2|}}\right)\log d\log^2\left(\frac{KN\left(T+\frac{\log(dN)}{\sqrt{1/|\lambda_2|}}\right)}{\delta\lambda}\right)}\right)\log\log\left(\frac{NT}{d}\right)\right). \qquad (38)
\end{aligned}
$$

□

C.2. Communication Step

In this section, we summarize the accelerated Chebyshev communication step, discussed above, in Algorithm 3, which follows the same steps as those of the communication algorithm presented in Martínez-Rubio et al. (2019).

Algorithm 3 Comm for Agent i

```
Input: $x_{\text{now}}, x_{\text{prev}}, \ell$
Output: $x_{i,\text{next}}$
Initialization: $w_0 = 0, w_1 = 1/|\lambda_2|, w_r = 2w_{r-1}/|\lambda_2| - w_{r-2}, \forall 2 \leq r \leq S, x_{i,\text{now}} = x_{\text{now}}, x_{i,\text{prev}} = x_{\text{prev}}$
Send $x_{i,\text{now}}$ and receive the corresponding $x_{j,\text{now}}$ to and from $j \in \mathcal{N}(i)$    // Recall that all agents run Comm in parallel.
$x_{i,\text{next}} = \frac{2w_{\ell-1}}{|\lambda_2|w_\ell}\mathbf{P}_{i,i}x_{i,\text{now}} + \frac{2w_{\ell-1}}{|\lambda_2|w_\ell}\sum_{j\in\mathcal{N}(i)}\mathbf{P}_{i,j}x_{j,\text{now}} - \frac{w_{\ell-2}}{w_\ell}x_{i,\text{prev}}$
```

Chebyshev polynomials (Young, 2014) are defined as $T_0(x) = 1, T_1(x) = x$ and $T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x)$. Define:

$$q_\ell(\mathbf{P}) = \frac{T_\ell(\mathbf{P}/|\lambda_2|)}{T_\ell(1/|\lambda_2|)}. \qquad (39)$$

By the properties of Chebyshev polynomial (Arioli and Scott, 2014), it can be shown that:

$$q_{\ell+1}(\mathbf{P}) = \frac{2w_\ell}{|\lambda_2|w_{\ell+1}}\mathbf{P}q_\ell(\mathbf{P}) - \frac{w_{\ell-1}}{w_{\ell+1}}q_{\ell-1}(\mathbf{P}), \qquad (40)$$

where $w_0 = 1, w_1 = 1/|\lambda_2|$, $w_{\ell+1} = 2w_\ell/|\lambda_2| - w_{\ell-1}$, $q_0(\mathbf{P}) = I$ and $q_1(\mathbf{P}) = \mathbf{P}$. This implies that when agents share an specific quantity, whose initial values given by agents are denoted by vector $\boldsymbol{\nu}_0 \in \mathbb{R}^N$, by using the recursive Chebyshev-accelerated updating rule, they have:

$$\boldsymbol{\nu}_{\ell+1} = \frac{2w_\ell}{|\lambda_2|w_{\ell+1}}\mathbf{P}\boldsymbol{\nu}_\ell - \frac{w_{\ell-1}}{w_{\ell+1}}\boldsymbol{\nu}_{\ell-1}. \qquad (41)$$

In light of the above mentioned recursive procedure, the accelerated communication step is summarized in Algorithm 3 below for agent i. We denote the inputs by: 1) x_{now}, which is the quantity of interest that agent i wants to update at the current round, 2) x_{prev}, which is the estimated value for a quantity of interest that agent i updated at the previous round, and 3) ℓ which is the current round of communication. Note that inputs are scalars, however matrices and vectors also can be passed as inputs with Comm running for each of their entries.

D. Omitted Algorithms

In this section, we present a definition and necessary algorithms, that are borrowed from Ruan et al. (2021) and are used as subroutines in DisBE-LUCB and DecBE-LUCB.

Definition D.1 (Ruan et al. (2021)). Fix $\alpha = \log K$. For a given positive semi-definite matrix $\mathbf{M}$, we define the softmax policy $\pi^{\mathrm{S}}_{\mathbf{M}}(\mathcal{X})$ over a set $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$ with $k \leq K$ with

$$\pi^{\mathrm{S}}_{\mathbf{M}}(\mathbf{x}_i) = \frac{(\mathbf{x}_i^\top\mathbf{M}\mathbf{x}_i)^\alpha}{\sum_{i=1}^k(\mathbf{x}_i^\top\mathbf{M}\mathbf{x}_i)^\alpha}. \qquad (42)$$

Now, suppose we are given a set $\mathcal{M} = \{(p_i, \mathbf{M}_i)\}_{i=1}^n$ such that $p_i \geq 0$ and $\sum_{i=1}^n p_i = 1$. We define the mixed-softmax policy $\pi_{\mathcal{M}}^{\mathrm{MS}}(\mathcal{X})$ over $\mathcal{X}$ as

$$\pi_{\mathcal{M}}^{\mathrm{MS}}(\mathbf{x}_i) = \begin{cases} \pi^{\mathrm{G}}(\mathcal{X}), & \text{with probability } 1/2, \\ \pi_{\mathbf{M}_i}^{\mathrm{S}}(\mathcal{X}), & \text{with probability } p_i/2, \end{cases} \tag{43}$$

where $\pi^{\mathrm{G}}(\mathcal{X})$ is called G-optimal design and is the minimizer of $g(\pi) = \max_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x}\|^2_{\mathbf{V}(\pi)^{-1}}$, where $\mathbf{V}(\pi) = \sum_{\mathbf{x}\in\mathcal{X}} \pi(\mathbf{x})\mathbf{x}\mathbf{x}^\top$; see Section 21 in Lattimore and Szepesvári (2020) for details.

Algorithm 4 ExpPol

```
Input: $\lambda$, $\mathcal{S} = \{\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_L\}$
Output: A mixed-softmax policy $\pi$  Using Algorithm 5 find a core $\mathcal{C} \subseteq \mathcal{S}$ such that
    $$\max_{\mathcal{X}_i \in \mathcal{C}, \mathbf{x}\in\mathcal{X}_i} \mathbf{x}^\top \mathbf{A}(\mathcal{C})^{-1}\mathbf{x} > d^5 \tag{44}$$
    and
    $$\frac{|\mathcal{C}|}{L} < 1 - \mathcal{O}(d^{-2}\log\lambda^{-1}) \tag{45}$$
    where $\mathbf{A}(\mathcal{C}) := \lambda I + \frac{1}{L}\sum_{\mathcal{X}_i\in\mathcal{C}} \mathbb{E}_{\mathbf{x}\sim\pi^{\mathrm{G}}(\mathcal{X}_i)}[\mathbf{x}\mathbf{x}^\top]$, and for any set $\mathcal{X} \subset \mathbb{R}^d$, $\pi^{\mathrm{G}}(\mathcal{X})$ is called $G$-optimal design and is the maximizer of $g(\pi) = \max_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x}\|^2_{\mathbf{V}(\pi)^{-1}}$, where $\mathbf{V}(\pi) = \sum_{\mathbf{x}\in\mathcal{X}} \pi(\mathbf{x})\mathbf{x}\mathbf{x}^\top$.
Return the mixed-softmax policy $\pi$ by calling $\mathrm{MixedSoftMax}(\lambda, \mathcal{C})$.
```

Algorithm 5 CoreIdentification (Algorithm 4 in (Ruan et al., 2021))

```
Input: $\lambda$, $\mathcal{S} = \{\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_L\}$
Output: A core set $\mathcal{C} \subseteq \mathcal{S}$
Initialization: $\mathcal{C}_1 = \mathcal{S}$
for $\xi = 1, 2, \ldots$ do
    if $\max_{\mathcal{X}_i\in\mathcal{C}_\xi, \mathbf{x}\in\mathcal{X}_i} \mathbf{x}^\top \mathbf{A}(\mathcal{C}_\xi)^{-1}\mathbf{x} > d^5$ then
        Return $\mathcal{C}_\xi$.
    else
        $$\mathcal{C}_{\xi+1} = \left\{\mathcal{X}_i \in \mathcal{C}_\xi : \max_{\mathbf{x}\in\mathcal{X}_i} \mathbf{x}^\top \mathbf{A}(\mathcal{C}_\xi)^{-1}\mathbf{x} \leq \frac{1}{2}d^5\right\},$$
        where $\mathbf{A}(\mathcal{C}) := \lambda I + \frac{1}{L}\sum_{\mathcal{X}_i\in\mathcal{C}} \mathbb{E}_{\mathbf{x}\sim\pi^{\mathrm{G}}(\mathcal{X}_i)}[\mathbf{x}\mathbf{x}^\top]$, and for any set $\mathcal{X} \subset \mathbb{R}^d$, $\pi^{\mathrm{G}}(\mathcal{X})$ is called $G$-optimal design and is the maximizer of $g(\pi) = \max_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x}\|^2_{\mathbf{V}(\pi)^{-1}}$, where $\mathbf{V}(\pi) = \sum_{\mathbf{x}\in\mathcal{X}} \pi(\mathbf{x})\mathbf{x}\mathbf{x}^\top$.
    end if
end for
```

E. Auxiliary Lemmas

Lemma E.1 (Tropp (2015), Theorem 5.1.1)**.** *Consider a finite sequence $\mathbf{X}_k$ of independent, random, Hermitian matrices with common dimension d. Assume that $0 \leq \lambda_{\min}(\mathbf{X}_k)$ and $\lambda_{\max}(\mathbf{X}_k) \leq L$ for each index k. Introduce the random matrix*

$$\mathbf{Y} = \sum_{k=1}^n \mathbf{X}_k \tag{46}$$

Define the minimum eigenvalue $\mu_{\min}$ and maximum eigenvalue $\mu_{\max}$ of the expectation $\mathbb{E}[\mathbf{Y}]$:

$$\mu_{\min} = \lambda_{\min}(\mathbb{E}[\mathbf{Y}]), \quad \mu_{\max} = \lambda_{\max}(\mathbb{E}[\mathbf{Y}]). \tag{47}$$

Algorithm 6 MixedSoftMax

```
**Input:** $\lambda, \mathcal{S} = \{\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_L\}$
**Output:** A mixed-softmax policy $\pi$
**Initialization:** $Q = 2d^2 \log d$, $\mathcal{X}_{(i-1)L+j} = \mathcal{X}_j,\ \forall (i,j) \in [Q] \times L$, $\mathbf{U}_0 = \lambda QLI + \frac{Q}{2}\sum_{i=1}^{L} \mathbb{E}_{\mathbf{x} \sim \pi^{\mathrm{G}}(\mathcal{X}_i)}[\mathbf{x}\mathbf{x}^\top]$, $n = 1$, $\tau_n = \emptyset$, $\mathbf{W}_n = \mathbf{U}_0$
**for** $s = 1, \dots, QL$ **do**
    $\tau_n = \tau_n \cup \{s\}$
    $\mathbf{U}_s = \mathbf{U}_{s-1} + \mathbb{E}_{\mathbf{x} \sim \pi^{\mathrm{S}}_{\mathbf{W}_n^{-1}}(\mathcal{X}_s)}[\mathbf{x}\mathbf{x}^\top]$, where $\pi^{\mathrm{S}}_{\mathbf{W}_n^{-1}}(\mathcal{X}_s)$ is computed as in Definition D.1.
    **if** $\frac{\det \mathbf{U}_s}{\det \mathbf{W}_n} > 2$ **then**
        $n = n + 1$, $\tau_n = \emptyset$, $\mathbf{W}_n = \mathbf{U}_s$
    **end if**
**end for**
$p_i = \frac{\mathbb{I}\{|\tau_i| \geq L\}|\tau_i|}{\sum_{i=1}^{n} \mathbb{I}\{|\tau_i| \geq L\}|\tau_i|}$ and $\mathbf{M}_i = QL\mathbf{W}_i^{-1},\ \forall i \in [n]$
Return the mixed-softmax policy with parameters $\mathcal{M} = \left\{(p_i, \mathbf{M}_i)\right\}_{i=1}^{n}$ as in Definition D.1.
```

Then

$$\mathbb{P}\left(\lambda_{\min}(\mathbf{Y}) \leq (1-\varepsilon)\mu_{\min}\right) \leq d\left(\frac{\exp(-\varepsilon)}{(1-\varepsilon)^{1-\varepsilon}}\right)^{\frac{\mu_{\min}}{L}}, \quad \textit{for } \varepsilon \in [0,1) \tag{48}$$

$$\mathbb{P}\left(\lambda_{\max}(\mathbf{Y}) \geq (1+\varepsilon)\mu_{\max}\right) \leq d\left(\frac{\exp(\varepsilon)}{(1+\varepsilon)^{1+\varepsilon}}\right)^{\frac{\mu_{\max}}{L}}, \quad \textit{for } \varepsilon \geq 0. \tag{49}$$

Lemma E.2. *Suppose $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n \sim \mathcal{D}$ are d-dimensional vectors that are i.i.d. drawn from a distribution $\mathcal{D}$ and $\|\mathbf{x}_k\|_2 \leq L$ for all $k \in [n]$ almost surely. Let $\gamma = \lambda_{\min}\left(\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\mathbf{x}^\top]\right) > 0$ be the smallest eigenvalue of the co-variance matrix. We have that*

$$\mathbb{P}\left(\frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k\mathbf{x}_k^\top \preceq 2\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\mathbf{x}^\top]\right) \geq 1 - d\exp\left(\frac{-\gamma n}{3}\right). \tag{50}$$

Proof. Let $\mathbf{\Sigma} = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\mathbf{x}^\top]$ and $\mathbf{y}_k = \mathbf{\Sigma}^{\frac{-1}{2}}\mathbf{x}_k$ for all $k \in [n]$. Also, we have $\lambda_{\max}(\mathbf{y}_k\mathbf{y}_k^\top) = \|\mathbf{y}_k\|_2^2 \leq \frac{1}{\gamma}$ almost surely, and $\mathbb{E}[\mathbf{y}_k\mathbf{y}_k^\top] = I$. Therefore, plugging $\varepsilon = 1$ in (49), we have

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k\mathbf{x}_k^\top \preceq 2\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\mathbf{x}^\top]\right) &= \mathbb{P}\left(\frac{1}{n}\sum_{k=1}^{n} \mathbf{y}_k\mathbf{y}_k^\top \preceq 2\mathbf{\Sigma}^{\frac{-1}{2}}\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\mathbf{x}^\top]\mathbf{\Sigma}^{\frac{-1}{2}}\right) \\
&= \mathbb{P}\left(\frac{1}{n}\sum_{k=1}^{n} \mathbf{y}_k\mathbf{y}_k^\top \preceq 2I\right) \\
&= \mathbb{P}\left(\lambda_{\max}\left(\sum_{k=1}^{n} \mathbf{y}_k\mathbf{y}_k^\top\right) \leq 2n\right) \\
&\geq 1 - d\left(\frac{e}{4}\right)^{n\gamma} \geq 1 - d\exp\left(\frac{-\gamma n}{3}\right).
\end{aligned} \tag{51}$$

□

Lemma E.3. *Suppose $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n \sim \mathcal{D}$ are d-dimensional vectors that are i.i.d. drawn from a distribution $\mathcal{D}$ and $\|\mathbf{x}_k\|_2 \leq 1$ for all $k \in [n]$ almost surely. For any cutoff level $\gamma > 0$, we have*

$$\mathbb{P}\left(\frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k\mathbf{x}_k^\top \preceq 2\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\mathbf{x}^\top] + 6\gamma I\right) \geq 1 - 2d\exp\left(\frac{-\gamma n}{3}\right). \tag{52}$$

Proof. Suppose $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\mathbf{x}^\top] = \sum_{i=1}^d \lambda_i \boldsymbol{\nu}_i \boldsymbol{\nu}_i^\top$, where $\{\boldsymbol{\nu}_i\}_{i=1}^d$ is a set of orthonormal basis. Let $\mathbf{P}_+ = \sum_{i=1}^d \boldsymbol{\nu}_i \boldsymbol{\nu}_i^\top \mathbb{1}(\lambda_i \geq \gamma)$ and $\mathbf{P}_- = \sum_{i=1}^d \boldsymbol{\nu}_i \boldsymbol{\nu}_i^\top \mathbb{1}(\lambda_i < \gamma)$, so that $\mathbf{P}_+\mathbf{P}_- = I$. We observe that the eigenvalues of $\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{P}_+\mathbf{x}\mathbf{x}^\top\mathbf{P}_+^\top]$ are greater than or equal to γ when restricted to the space spanned by the $\mathbf{P}_+$. Therefore, by Lemmas E.2 and E.1 (Eqn. (49)), we respectively have

$$\mathbb{P}\left(\frac{1}{n}\sum_{k=1}^n \mathbf{P}_+\mathbf{x}_k\mathbf{x}_k^\top\mathbf{P}_+^\top \preceq 2\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{P}_+\mathbf{x}\mathbf{x}^\top\mathbf{P}_+^\top]\right) \geq 1 - d\exp\left(\frac{-\gamma n}{3}\right) \tag{53}$$

$$\mathbb{P}\left(\frac{1}{n}\sum_{k=1}^n \mathbf{P}_-\mathbf{x}_k\mathbf{x}_k^\top\mathbf{P}_-^\top \preceq 2\gamma I\right) \geq 1 - d\exp\left(\frac{-\gamma n}{3}\right). \tag{54}$$

Now, we observe that

$$\begin{aligned}
\frac{1}{n}\sum_{k=1}^n \mathbf{x}_k\mathbf{x}_k^\top &= \frac{1}{n}\left(\sum_{k=1}^n \mathbf{P}_+\mathbf{x}_k\mathbf{x}_k^\top\mathbf{P}_+^\top + \sum_{k=1}^n \mathbf{P}_+\mathbf{x}_k\mathbf{x}_k^\top\mathbf{P}_-^\top + \sum_{k=1}^n \mathbf{P}_-\mathbf{x}_k\mathbf{x}_k^\top\mathbf{P}_+^\top + \sum_{k=1}^n \mathbf{P}_-\mathbf{x}_k\mathbf{x}_k^\top\mathbf{P}_-^\top\right) \\
&= \frac{1}{n}\left(\sum_{k=1}^n \mathbf{P}_+\mathbf{x}_k\mathbf{x}_k^\top\mathbf{P}_+^\top + \sum_{k=1}^n \mathbf{P}_+\mathbf{P}_+\mathbf{P}_-\mathbf{x}_k\mathbf{x}_k^\top\mathbf{P}_-^\top + \sum_{k=1}^n \mathbf{P}_-\mathbf{x}_k\mathbf{x}_k^\top\mathbf{P}_-^\top\mathbf{P}_+^\top\mathbf{P}_+^\top + \sum_{k=1}^n \mathbf{P}_-\mathbf{x}_k\mathbf{x}_k^\top\mathbf{P}_-^\top\right) \\
&\preceq \frac{1}{n}\left(\sum_{k=1}^n \mathbf{P}_+\mathbf{x}_k\mathbf{x}_k^\top\mathbf{P}_+^\top + \sum_{k=1}^n \mathbf{P}_-\mathbf{x}_k\mathbf{x}_k^\top\mathbf{P}_-^\top + \sum_{k=1}^n \mathbf{P}_-\mathbf{x}_k\mathbf{x}_k^\top\mathbf{P}_-^\top + \sum_{k=1}^n \mathbf{P}_-\mathbf{x}_k\mathbf{x}_k^\top\mathbf{P}_-^\top\right) \\
&= \frac{1}{n}\sum_{k=1}^n \mathbf{P}_+\mathbf{x}_k\mathbf{x}_k^\top\mathbf{P}_+^\top + \frac{3}{n}\sum_{k=1}^n \mathbf{P}_-\mathbf{x}_k\mathbf{x}_k^\top\mathbf{P}_-^\top
\end{aligned} \tag{55}$$

Also, note that

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{P}_+\mathbf{x}\mathbf{x}^\top\mathbf{P}_+^\top] = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[\mathbf{x}\mathbf{x}^\top - \mathbf{P}_+\mathbf{x}\mathbf{x}^\top\mathbf{P}_-^\top - \mathbf{P}_-\mathbf{x}\mathbf{x}^\top\mathbf{P}_+^\top - \mathbf{P}_-\mathbf{x}\mathbf{x}^\top\mathbf{P}_-^\top\right] \preceq \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[\mathbf{x}\mathbf{x}^\top\right]. \tag{56}$$

Therefore, combining (54) and (55) and (56), we have

$$\mathbb{P}\left(\frac{1}{n}\sum_{k=1}^n \mathbf{x}_k\mathbf{x}_k^\top \preceq 2\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\mathbf{x}^\top] + 6\gamma I\right) \geq 1 - 2d\exp\left(\frac{-\gamma n}{3}\right). \tag{57}$$

□