On the Optimal Regret of Collaborative Personalized Linear Bandits

Bruce Huang
UCLA

Ruida Zhou
Amazon AGI

Lin F. Yang
UCLA

Suhas Diggavi
UCLA

Abstract

Stochastic linear bandits are a fundamental model for sequential decision making. Although well studied in the single-agent setting, many real-world scenarios involve multiple agents solving heterogeneous bandit problems, each with a different unknown parameter. This paper investigates the optimal regret achievable in collaborative personalized linear bandits. We derive an information-theoretic lower bound showing how the number of agents, the number of rounds, and the degree of heterogeneity jointly affect regret. We propose a two-stage collaborative algorithm that achieves the optimal regret. We model heterogeneity via a hierarchical Bayesian framework and introduces a novel information-theoretic technique for bounding regret. Our results offer a complete characterization of when and how collaboration helps with a optimal regret bound $\tilde{O}(d\sqrt{mn})$, $\tilde{O}(dm^{1-\gamma}\sqrt{n})$, $\tilde{O}(dm\sqrt{n})$ for the number of rounds n in the range of $o\left(\frac{d}{m\sigma^2}\right)$, $\Theta\left(\frac{d}{m^{2\gamma}\sigma^2}\right)$ and $\omega\left(\frac{d}{\sigma^2},\right)$ respectively, where σ measures the level of heterogeneity, m is the number of agents, and $\gamma \in [0, 1/2]$ is an absolute constant. In contrast, without collaboration achieves a regret bound $O(dm\sqrt{n})$ at best. Code is available at: `https://github.com/brucehuang1729/Collaborative-Personalized-Linear-Bandits`

1 Introduction

Stochastic linear bandits are a fundamental model in sequential decision-making, where in each round, an agent selects a vector-valued action and receives a stochastic reward with expectation equal to the inner product between the action and an unknown parameter vector (Lattimore and Szepesvári, 2020). This setting has been extensively studied in the single-agent case, where the goal is to maximize cumulative reward over time through efficient exploration and exploitation (Robbins, 1952; Lai and Robbins, 1985; Auer, 2002).

In this paper, we consider a generalized multi-agent setting, motivated by many practical scenarios such as distributed decision systems (Landgren et al., 2016; Zhu et al., 2020), personalized adaptive interfaces (Liu et al., 2003; Khamaj and Ali, 2024), and multi-device learning (Cho et al., 2022)—where in each round, a set of m agents must each select an action and receive feedback from a stochastic, potentially heterogeneous reward function. A natural baseline approach is to run a standard single-agent linear bandit algorithm independently for each agent. While straightforward, this approach entirely ignores potential similarities across agents. When the agents' environments are similar, such independent learning leads to redundant exploration and a suboptimal use of data. In contrast, our goal is to develop collaborative learning algorithms that adaptively share information across agents—without knowing in advance how similar they are. This raises a fundamental question: **How can collaboration reduce regret under unknown heterogeneity?**

To address the question, we model heterogeneity using a hierarchical/empirical Bayesian framework, where each agent's unknown parameter vector is drawn from an *unknown* population distribution. Then the variance of this population captures the degree of heterogeneity. Although the population distribution can be arbitrary in the general framework, in this work we focus on *unknown* Gaussian population distribution. The Gaussian population distribution captures the heterogeneity through the population variance σ^2, around the *unknown* arbitrary mean vector without any prior on it. Besides being a useful theoretical framework, effectively this population distribution has been extensively used in algorithms for personalized FL, in supervised frameworks (Ozkara et al., 2023) and references therein. We consider unknown population distributions in our problem, and we do not assume further prior distribution

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

on the population parameters since the agents will not have such information in practice.

We then propose a novel two-stage algorithm to leverage collaboration: in the first *collaborative learning stage*, agents jointly perform phased elimination to identify a promising subset of actions by pooling their information. In the second *personalized learning stage*, agents refine their individual models based on local feedback. This structure allows the algorithm to smoothly interpolate between joint learning (when agents are similar) and local adaptation (when they are not). We carefully design a threshold to guide the transition between the two stages, so that collaboration is activated only when it benefits learning. In the analysis, we show that the algorithm achieves a regret bound that can be significantly smaller than the independent-learning baseline when the number of rounds is small—highlighting the benefit of collaboration in the data-scarce regime.

To show the optimality of our algorithm, we derive an *information-theoretic lower bound* on the regret across different levels of heterogeneity. A central challenge in this analysis stems from the nature of heterogeneity: unlike the classical single-agent setting, where the worst-case parameter governs the lower bound, we must now consider the *worst-case population distribution* over agents' parameters. To overcome this, we introduce novel techniques involving *rotation and coupling* to construct hard instances and analyze the regret. Our lower bound precisely characterizes how the population variance σ^2 affects collaborative learning. Furthermore, we show that our proposed algorithm matches the lower bound, achieving regret that is *both globally and individually optimal*.

Our Contributions

- **Regret Characterization.** We provide, to the best of our knowledge, the first complete characterization of the benefit of collaboration by deriving *matching upper and lower bounds* on cumulative regret. For a fixed level of heterogeneity σ and number of users m, we characterize the regret in three regimes by varying the number of rounds n. Specifically Theorem 3.1 shows:
  - When n is small, we refer to as the *data-scarce* regime where $n = o(d/m\sigma^2)$, collaboration yields regret $\tilde{O}(d\sqrt{mn})$, matching the optimal rate for a single-agent linear bandit with mn rounds—showing that full data sharing is optimal.
  - When n is moderate, we refer to as the *intermediate* regime where $n = \Theta(d/m^{2\gamma}\sigma^2)$ for any $\gamma \in [0, 1/2]$, the regret $\tilde{O}(dm^{1-\gamma}\sqrt{n})$ smoothly interpolates between the two extremes, with sharp dependence on the data size n and heterogeneity level σ.
  - When n is large, we refer to as the *data-rich* regime where $n = \omega(d/\sigma^2)$, regret grows as $\tilde{O}(dm\sqrt{n})$, recovering the performance of independent learning per agent.
- **Information-Theoretic Lower Bound.** To establish the main result, we develop a minimax regret lower bound for heterogeneous multi-agent linear bandits. Unlike standard settings that focus on worst-case individual parameters, our bound considers the worst-case population distribution and introduces new coupling and rotation techniques to handle this multi-agent heterogeneous setting; see Theorem 4.1.
- **Optimal Algorithm.** We propose Collaborative Personalized Phased Elimination (CP-PE), see Algorithm 1 in Section 5, a two-stage collaborative algorithm that adapts to the heterogeneity and *achieves the minimax lower bound*. Notably, the algorithm is *individually optimal* for each agent and *globally optimal* for cumulative performance.

1.1 Related Work

There is extensive work in the bandit and linear bandit literature (see, e.g., the comprehensive review in (Lattimore and Szepesvári, 2020)). Most of these works focus on the single-agent, non-distributed learning setting. Another line of research explores distributed learning (Hillel et al., 2013; Wang et al., 2020; Huang et al., 2021; Zhu et al., 2021), but the majority of these studies consider the homogeneous case, where all agents collaboratively learn a shared parameter. Among studies on collaborative bandits with heterogeneous agents, several works consider clustered bandits algorithms (Gentile et al., 2014; Landgren et al., 2016; Ghosh et al., 2021; Liu et al., 2022; Blaser et al., 2024; Li et al., 2016; Korda et al., 2016; Mahadik et al., 2020). In most of the works, agents in the same cluster share a common unknown parameter. However, variations in users' parameters within the same cluster can significantly change the problem. Small variations between parameters in the same cluster are allowed in (Blaser et al., 2024). However, the variation is bounded by $O(1/m\sqrt{n})$, and applying single-agent methods to users with such variation gives optimal regret bound. This does not answer how heterogeneous data can help through collaboration. (Ghosh et al., 2021) also propose a two-stage algorithm, while the algorithm does not adapt to heterogeneity. Other works characterize similarity between the agents with different structures like networks (Buccapatnam et al., 2013; Wu et al., 2016), or consider agents sharing a global

component in their unknown parameters (Li and Wang, 2022). Instead, our work considers an unknown population model that characterizes the heterogeneity. We discuss the two most related works in the followings.

After completing our work, we recently became aware of the work (Do et al., 2023). They consider heterogeneous linear contextual bandits, and they measure heterogeneity as the maximum difference, ε, between any pair of individual parameters. They propose an algorithm and provide upper and lower bounds that depend on ε for the problem. In our work, we model heterogeneity differently through unbounded population distribution, and thus the value of ε can go to infinity even when the level of heterogeneity is low in our case. Furthermore, their upper and lower bounds do not match in all regimes with up to a $\sqrt{d}$ gap for their model, whereas our results show matching bounds in all regimes. Additionally, due to the different modeling for heterogeneity, their lower bound techniques cannot be applied to our setup. We develop novel techniques in Section 4 to resolve challenges emerging in our minimax lower bound.

Another related work is (Hong et al., 2022), which proposes a hierarchical Bayesian framework to model heterogeneity in multi-armed bandits, and introduces a hierarchical Thompson Sampling algorithm for the setting. A major limitation of their approach is the assumption of a known prior distribution over the population model. The case of fixed unknown population parameters is important since agents do not have information on such prior distribution in practice. However, the case is not captured in their setting and their result does not apply. In contrast to their pure Bayesian setting, we adopt an empirical Bayes approach without assuming any prior distribution on the unknown population parameters, which can make the analysis significantly more challenging. Furthermore, their work does not provide a regret lower bound, whereas we establish a matching lower bound that demonstrates the optimality of our proposed algorithm.

2 Problem Formulation

We now formally present the problem definitions in the paper. We first recall the conventional stochastic linear bandits, let n be the number of rounds. The user chooses an action $\boldsymbol{x}_t$ from the action set $\mathcal{A} \subset \mathbb{R}^d$ in each round $t \in [n]$, and receive a stochastic reward $y_t = \boldsymbol{x}_t^\top \boldsymbol{\theta} + \epsilon_t$ where $\boldsymbol{\theta} \in \mathbb{R}^d$ is an unknown parameter to the user and $\epsilon_t \sim \mathcal{N}(0, \sigma_0^2)$ is a Gaussian noise.

In our problem setting, there are m users, each solving a stochastic linear bandit problem with a distinct parameter $\boldsymbol{\theta}_i \in \mathbb{R}^d$ for $i = 1, 2, \ldots, m$. We model user heterogeneity using a hierarchical/empirical Bayesian

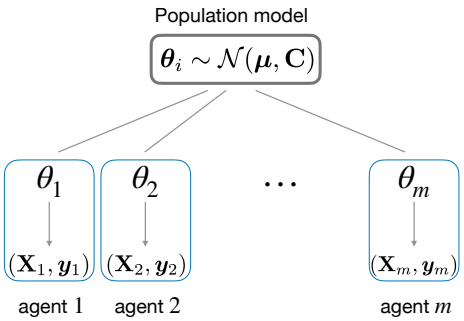

Figure 1: Hierarchical/empirical Bayesian framework

framework (Ozkara et al., 2023), where the parameters $\boldsymbol{\theta}_i$ are sampled i.i.d. from a population-level Gaussian distribution:

$$\boldsymbol{\theta}_i \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C}), \tag{1}$$

with a covariance matrix $\mathbf{C}$ capturing heterogeneity across users, and an unknown mean $\boldsymbol{\mu}$. While we assume a Gaussian population distribution in the formulation, our algorithm and regret analysis extend to sub-Gaussian population distributions. See Appendix B.6 for details.

To illustrate the role of heterogeneity, consider the special case where $\mathbf{C} = \sigma^2 \mathbf{I}$ for some $\sigma \geq 0$. In this case, a larger σ corresponds to greater heterogeneity, while $\sigma = 0$ recovers the homogeneous setting in which $\boldsymbol{\theta}_1 = \cdots = \boldsymbol{\theta}_m$. However, our results hold for general covariance matrices $\mathbf{C}$. See Appendix B.7 for details.

At time t, each user i chooses an action $\boldsymbol{x}_{i,t}$ from an action set $\mathcal{A}_i \subset \mathbb{R}^d$, and receives a reward $y_{i,t}$:

$$y_{i,t} = \boldsymbol{x}_{i,t}^\top \boldsymbol{\theta}_i + \epsilon_{i,t}, \quad \epsilon_{i,t} \sim \mathcal{N}(0, \sigma_0^2), \tag{2}$$

where $\epsilon_{i,t}$ is an unknown "local" noise and $\sigma_0 > 0$ is the variance, assumed to be the same[1] across users. A visualization of the model is provided in Figure 1. For the sake of presentation, we assume that $\mathbf{C}$ is known and local noise has a unit variance,$\sigma_0^2 = 1$ in the main paper. However, in Appendix C.3, we empirically show that for the unknown σ, case we can modify our algorithm to estimate σ throughout the algorithm and achieves a similar result.

We use the following regret for measuring the performance of an algorithm:

Definition 2.1 (Joint pseudo-regret)**.** *For any $\boldsymbol{\mu}$, the joint pseudo-regret of the users is defined as*

$$\hat{R}_{m,n}(\boldsymbol{\mu}) = \sum_{i=1}^{m} \sum_{t=1}^{n} \left(\langle \boldsymbol{x}_i^*, \boldsymbol{\theta}_i \rangle - \langle \boldsymbol{x}_{i,t}, \boldsymbol{\theta}_i \rangle \right)$$

[1]Our results naturally generalize to heterogeneous local noise levels; we assume a uniform noise level σ_0 for clarity of presentation.

where $\boldsymbol{x}_i^* = \arg\max_{\boldsymbol{x}\in\mathcal{A}_i}\langle \boldsymbol{x}, \boldsymbol{\theta}_i\rangle$ *is the optimal action for each user* $i \in [m]$*. Notice that the joint pseudo-regret is random due to the randomness of* $\boldsymbol{\theta}_i$*'s.*

Definition 2.2 (Joint Bayesian regret)**.** *For any* $\boldsymbol{\mu}$*, the joint Bayesian regret of the users is defined as*

$$R_{m,n}(\boldsymbol{\mu}) = \mathbb{E}\left[\sum_{i=1}^{m}\sum_{t=1}^{n}\left(\langle \boldsymbol{x}_i^*, \boldsymbol{\theta}_i\rangle - \langle \boldsymbol{x}_{i,t}, \boldsymbol{\theta}_i\rangle\right)\right]$$

where $n > 0$ *is the number of total rounds of interactions.*

Notations. In the rest of the paper, we represent vectors in bold lowercase letters, such as $\boldsymbol{\mu}, \boldsymbol{\theta}$. Bold capitalized letters $\mathbf{X}, \mathbf{C}$ are matrices, and $\mathbf{I}$ is the identity matrix. We denote $[m] = \{1, \ldots, m\}$. For a real symmetric matrix $\mathbf{A}$, we define $\mathbf{A}^\dagger$ to be the pseudo-inverse of $\mathbf{A}$, obtained by inverting the non-zero eigenvalues and leaving the zero eigenvalues unchanged in the eigen-decomposition of $\mathbf{A}$. We use the big-O notation $\tilde{O}$ to ignore logarithm factors. Let $\mathcal{B}_2^d = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 \le 1\}$ denote the d-dimensional ball.

3 Main Results & Technical Overview

3.1 Algorithm Overview: Collaborative Personalized Phased elimination

In this section, we briefly introduce our proposed algorithm in preparation for our main results in Section 3.2. The regret analysis and algorithm details can be found in Section 5 and Appendix B.

We propose Collaborative Personalized Phased Elimination (CP-PE) by adapting the method of phased elimination with G-optimal design exploration in Lattimore and Szepesvári (2020) (Chapter 21) to our heterogeneous setting. A phased elimination algorithm shrinks the action set in each phase by eliminating the actions that are too far from the estimated parameter with a decreasing target error. Additionally, CP-PE consists of two stages: the *collaborative learning stage* and the *personalized learning stage*. A threshold is carefully designed to guide the transition between the two stages so that collaboration is activated only when it benefits learning.

In the collaborative stage, all users perform the phased elimination algorithm together as if they have the same unknown parameter. This benefits learning when the target error is larger than the degree of heterogeneity since estimating $\boldsymbol{\mu}$ already gives a relatively good estimation to all $\boldsymbol{\theta}_i$'s. However, as target error drops below the degree of heterogeneity, collaboratively learning an estimation on $\boldsymbol{\mu}$ does not give good enough estimations on $\boldsymbol{\theta}_i$'s anymore. Thus, all users transit to the local stage, where they perform phased elimination on their own to refine their results.

3.2 Main Results: Minimax-optimal Regret Bounds

We provide our main result in the section. Let us consider the case where the covariance matrix is isotropic, i.e., $\mathbf{C} = \sigma^2\mathbf{I}$ for some $\sigma \ge 0$. To present our results on the minimax lower bound, let us define the minimax regret.

Definition 3.1. *The minimax regret is defined as*

$$\overline{R}_{m,n} = \inf_{\pi\in\Pi}\max_{\boldsymbol{\mu}\in\mathcal{B}_2^d} R_{m,n}(\boldsymbol{\mu}),$$

where Π *is the collection of all policies and* $\mathcal{B}_2^d$ *is the* d*-dimensional unit ball in Euclidean space.*

In the definition of the minimax regret, we maximize over all $\boldsymbol{\mu}$ in the unit ball, a discussion on the restriction can be found in remark 3.3.

The following Theorem 3.1 combines our results on the minimax regret lower bound in Theorem 4.1 and the regret analysis for CP-PE in Theorem B.1. This gives us a complete characterization of the benefit of collaboration.

Theorem 3.1 (Informal: A complete characterization on the benefit of collaboration)**.** *Consider the unit-ball action set* $\mathcal{A} = \mathcal{B}_2^d$*. Assume that* $\boldsymbol{\mu} \in \mathcal{B}_2^d$ *and* $\sigma \le 1$*. With assumptions*[2] *on the relation of* n *and* d*, the minimax lower bound on the Bayesian regret is achieved by CP-PE omitting the logarithm terms. The regret bounds are shown in the following table:*

	$n = o\left(\frac{d}{m\sigma^2}\right)$	$n = \Theta\left(\frac{d}{m^{2\gamma}\sigma^2}\right)$	$n = \omega\left(\frac{d}{\sigma^2}\right)$
$\overline{R}_{m,n}$	$\Omega(d\sqrt{mn})$	$\Omega(dm^{1-\gamma}\sqrt{n})$	$\Omega(dm\sqrt{n})$
$\hat{R}_{m,n}$	$\tilde{O}(d\sqrt{mn})$	$\tilde{O}(dm^{1-\gamma}\sqrt{n})$	$\tilde{O}(dm\sqrt{n})$

where $\overline{R}_{m,n}$ *is the minimax regret and* $\hat{R}_{m,n}$ *is the regret obtained by Alg. 1. The columns are separated by three different regimes, where* $\gamma \in [0, 1/2]$ *are any constants. The first row is the minimax regret lower bound on the joint Bayes regret and the second row is the joint pseudo-regret obtained by CP-PE. See details in Theorem 4.1 and Theorem B.1.*

Regret Characterization In Theorem 3.1, we identify three distinct regimes for the number of rounds n, characterized by their relationship to the heterogeneity level σ: **data-scarce regime:** $n = o(dm^{-1}\sigma^{-2})$; **intermediate regime:** $n = \Theta(dm^{-2\gamma}\sigma^{-2})$; **data-abundant regime:** $n = \omega(d\sigma^{-2})$. Equivalently, by

[2]The regret upper bounds holds for $n \ge md^2$. More details can be found in Theorem 4.1 and Theorem 5.1.

fixing n, we can interpret the regimes in terms of heterogeneity:

- **Data-scarce regime** corresponds to **low heterogeneity**, $\sigma = o(\sqrt{dm^{-1}n^{-1}})$
- **Intermediate regime** corresponds to $\sigma = \Theta(\sqrt{cdm^{-2\gamma}n^{-1}})$
- **Data-abundant regime** corresponds to **high heterogeneity**, $\sigma = \omega(\sqrt{dn^{-1}})$

The theorem shows that the regret behavior varies across these regimes, indicating whether collaboration among users is beneficial. Below, we elaborate on each regime.

In the data-scarce regime, where n is small or σ is low, we prove that no policy can achieve a regret smaller than $\Omega(d\sqrt{mn})$. This matches the minimax regret lower bound for the standard stochastic linear bandit with mn total samples. This result can be interpreted in two ways. First, it shows that collaboration is highly beneficial when heterogeneity is sufficiently low: users can collectively estimate the shared population parameter $\boldsymbol{\mu}$, achieving performance comparable to the homogeneous case. Second, when n is small, individual users lack sufficient data to accurately estimate their own parameters $\boldsymbol{\theta}_i$, and hence benefit from jointly estimating $\boldsymbol{\mu}$ using all mn samples.

In the data-abundant regime, where n is large or σ is high (i.e., the highly heterogeneous case), we show that any policy incurs regret at least $\Omega(dm\sqrt{n})$. This matches the minimax lower bound for the joint regret when each user solves their own linear bandit problem independently using n samples, and the regrets are summed across all m users. Again, this can be seen from two perspectives. First, it implies that collaboration offers little to no benefit in this regime due to the high heterogeneity. Alternatively, when n is large, each user has enough data to accurately estimate their own parameter $\boldsymbol{\theta}_i$, making collaboration unnecessary.

The most interesting behavior arises in the intermediate regime, where both the number of rounds and the heterogeneity level are moderate. We capture this regime using a smooth interpolation parameter $\gamma \in [0, 1/2]$, which governs the transition between the two extremes. Our results show that regret interpolates smoothly between the bounds in the data-scarce and data-abundant regimes, depending on γ. This reveals how the benefit of collaboration varies with both the level of heterogeneity σ and the number of samples n. As γ increases from 0 to 1/2, the problem gradually shifts from a collaboration-friendly regime to one where users are better off acting independently.

Remark 3.2 (Global and individual optimality)**.** *In Theorem 3.1, we study the joint cumulative regret. However, both our minimax lower bound and the regret upper bound achieved by CP-PE can be directly translated into individual regret bounds by dividing the results by the number of users, m. In other words, we establish not only global optimality in terms of joint regret but also individual optimality for each user's regret—a stronger result. More details can be found in Section B.3.*

Remark 3.3 (Assumptions in Theorem 3.1)**.** *In Theorem 3.1, we restrict $\boldsymbol{\mu}$ to lie within the unit ℓ_2 ball $\mathcal{B}_2^d$ and assume that $\sigma \leq 1$. These assumptions serve as analogs to the standard boundedness assumptions on the unknown parameter $\boldsymbol{\theta}$ in the classic linear bandit setting. For example, (Lai and Robbins, 1985, Chapter 19) considers the condition $\sup_{\boldsymbol{x},\boldsymbol{x}'\in\mathcal{A}}\langle\boldsymbol{\theta}, \boldsymbol{x}-\boldsymbol{x}'\rangle \leq 1$, which, when the action set is the unit ball, is equivalent to assuming $\|\boldsymbol{\theta}\|_2 \leq 2$. In our multi-agent setting, where each $\boldsymbol{\theta}_i$ is drawn from the population distribution $\mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I})$, we naturally place the boundedness assumptions on the population parameters $\boldsymbol{\mu}$ and σ instead.*

4 Information-Theoretic Lower Bound

In this section, we provide details on our minimax lower bound result and a sketch proof highlighting the challenges and novel techniques.

Theorem 4.1 (Minimax lower bound on the Bayesian regret)**.** *Consider the unit-ball action set $\mathcal{A} = \mathcal{B}_2^d$. Assume that $\boldsymbol{\mu} \in \mathcal{B}_2^d$ and $\sigma \leq 1$. Recall the minimax regret $\overline{R}_{m,n}$ in Definition 3.1. Then, we have the following results:*

- *For $n = o\left(\frac{d}{m\sigma^2}\right)$ and $mn \geq d$, we have $\overline{R}_{m,n} = \Omega\left(d\sqrt{mn}\right)$;*
- *For $n = \Theta\left(\max\left[\frac{d}{m^{2\gamma}\sigma^2}, d\right]\right)$, we have $\overline{R}_{m,n} = \Omega\left(dm^{1-\gamma}\sqrt{n}\right)$;*
- *For $n = \omega\left(\frac{d}{\sigma^2}\right)$, we have $\overline{R}_{m,n} = \Omega\left(dm\sqrt{n}\right)$.*

The interpretation of the results on the regret is discussed in Remark 3.2, and we discuss the assumptions we made in the theorem in Remark A.2.

We provide a proof sketch of Theorem 4.1 while keeping the detailed proof in Section A.1. In the data-scarce regime where $n = o(d/m\sigma^2)$, we will show that the heterogeneity is low such that the regret lower bound can not be far from the lower bound on the regret for the classic stochastic linear bandits with mn samples. Thus, we get our result with the classic lower bound $\Omega(d\sqrt{mn})$ plus an additional term due to the heterogeneity, where the term is small enough such that the final bound stays in the order of $\Omega(d\sqrt{mn})$.

A technical challenge emerges in the data-rich regime. In the classic stochastic linear bandit problem, the Bayesian regret is defined as $R_n(\boldsymbol{\theta}) = \mathbb{E}\left[\sum_{t=1}^{n}\langle \boldsymbol{x}^* - \boldsymbol{x}_t, \boldsymbol{\theta}\rangle\right]$ for any given $\boldsymbol{\theta}$. A set of "bad" $\boldsymbol{\theta}$'s is chosen to be the vertices of a d-dimensional hypercube. With the probabilistic method, it is shown that any policy must get a large regret on one of the $\boldsymbol{\theta}$'s being the true parameter, and thus the minimax lower bound is shown to be $\Omega(d\sqrt{n})$ for any policy π (Lattimore and Szepesvári (2020), Section 24.2). The proof strongly depends on the fact that the chosen $\boldsymbol{\theta}$ has the same absolute value in each coordinate and has a length in the order of $d/\sqrt{n}$.

However, in our heterogeneous setting, the parameters $\boldsymbol{\theta}_i$'s are not arbitrary but follow the population distribution $\mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I})$. Thus, our minimax bound considers the worst-case $\boldsymbol{\mu}$ instead of a worst-case $\boldsymbol{\theta}$. Especially in the data-rich regime when the heterogeneity is high, the norm of $\boldsymbol{\theta}_i$'s will be larger than the order of $d/\sqrt{n}$ since $\sigma = \omega(\sqrt{d/n})$. Additionally, each $\boldsymbol{\theta}_i$ can be in any direction following the population model. To address the challenges, we develop novel techniques to deal with $\boldsymbol{\theta}_i$'s with large norms and pointing into arbitrary directions.

Due to symmetry, we can focus on analyzing the regret from one user, say, user m in (6). To address the issue of $\boldsymbol{\theta}_m$ having a large norm, we choose the set of "bad" $\boldsymbol{\theta}_m$'s to be the vertices of a $(d-1)$-dimensional hypercube in (11) instead, which we denote them as $\{\boldsymbol{\theta}^{(\boldsymbol{z})}\}$ in the proof. Although $\boldsymbol{\theta}^{(\boldsymbol{z})}$'s has a large norm, we sacrifice one of the d coordinates to absorb most part of the norm and carefully construct them such that the distance between them is bad for the learner, meaning that the $\boldsymbol{\theta}^{(\boldsymbol{z})}$'s are hard to distinguish so the regret is large, yet separated enough such that an agent cannot get a small regret by simply keep selecting the same action.

Now we are able to construct a set of "bad" $\boldsymbol{\theta}^{(\boldsymbol{z})}$'s, we need them to marginally follow the population model as we mentioned. This is achieved by first sampling an arbitrary $\boldsymbol{\theta}^{(0)}$ as in (8) following the population model. Then, we form a set $\{\boldsymbol{\theta}^{(\boldsymbol{z})}\}$ around $\boldsymbol{\theta}^{(0)}$ as in (9) and Figure 3. Following the construction, each $\boldsymbol{\theta}^{(\boldsymbol{z})}$ marginally follows the population model due to symmetry, yet they are artificially coupled so that they form the vertices of a $(d-1)$-dimensional hypercube. Since we consider the unit ball action set, we can rotate $\boldsymbol{\theta}^{(\boldsymbol{z})}$ to align with the coordinates in (12).

Dealing with the large $\boldsymbol{\theta}_i$'s and choosing the appropriate set of "bad" $\boldsymbol{\theta}^{(\boldsymbol{z})}$'s are the most challenging parts in the proof. We achieve this by carefully choosing both the correct order of n and the suitable coefficients to form the suitable angle between the $\boldsymbol{\theta}^{(\boldsymbol{z})}$'s in (10). Unlike choosing the specific bad $\boldsymbol{\theta}$'s in the regret lower bound for classic stochastic linear bandit problem Lattimore and Szepesvári (2020) (Section 24.2), our proof also provides a way to show the lower bound when given the knowledge that the unknown parameter $\boldsymbol{\theta}$'s are large. In other words, given the information that the norm of $\boldsymbol{\theta}$ is large does not make the problem easier as larger $\boldsymbol{\theta}$ might seem to be easier to distinguish. In contrast, our construction shows that the order of the regret remains the same.

5 Algorithm: Collaborative Personalized Phased Elimination

In this section, we provide a regret analysis for Algorithm 1. The algorithm is composed of two stages: the collaborative learning stage and the personalized learning stage. We present Algorithm 1 and the Collaborative-Phased Elimination block 2 in this section. On the other hand, the Phased Elimination block 3 is rather standard and can be found in Appendix B.1.

In the collaborative learning stage (line 5 and 6 in Algorithm 1), all users start with the same action set $\mathcal{A}$ and perform phased elimination together. In each phase ℓ, we solve the G-optimal design for the current action set $\mathcal{A}_{1,\ell}$ and let $\boldsymbol{\pi}_\ell$ be the design solution, which is a probability distribution on the actions in $\mathcal{A}_{1,\ell}$. Let $n_{i,\ell}(\boldsymbol{x})$ denote the number of times for the action $\boldsymbol{x}$ to be selected by user i in the phase and choose

$$n_{i,\ell}(\boldsymbol{x}) = \left\lceil \frac{\boldsymbol{\pi}_\ell(\boldsymbol{x}) g(\boldsymbol{\pi}_\ell)}{m\varepsilon_\ell^2} \log \frac{k\ell(\ell+1)}{2\delta} \right\rceil \tag{3}$$

for all $\boldsymbol{x} \in \mathcal{A}_{i,\ell}$ and $i \in [m]$, where k is the size of the action set. Let $\bar{n}_{i,\ell} = \sum_{\boldsymbol{x}} n_{i.\ell}(\boldsymbol{x})$ denote the total number of pulls done by user i in phase ℓ. Notice that the factor m in the denominator is the benefit

Algorithm 1 Collaborative Personalized Phased Elimination (CP-PE)

1: **Input**: action set $\mathcal{A}$, dimension d, number of users m, number of rounds n, level of heterogeneity σ, $\delta \in (0,1)$
2: $\mathcal{A}_{i,1} \leftarrow \mathcal{A} \quad \forall i \in [m]$, $\ell \leftarrow 0$, calculate $h_{\sigma,\delta}$ from (5)
3: **while** total rounds for each user $\leq n$ **do**
4: $\quad \varepsilon_\ell \leftarrow 2^{-\ell}$
5: $\quad$ **if** $h_{\sigma,\delta} \leq \frac{1}{2}\varepsilon_\ell$ **then**
6: $\quad\quad \{\mathcal{A}_{i,\ell+1}\}_{i=1:m} \leftarrow$
7: $\quad\quad$ Collaborative Phased Elimination($\mathcal{A}_{1,\ell}, \varepsilon_\ell$)
8: $\quad$ **else**
9: $\quad\quad$ **for** $i \in [m]$ **do**
10: $\quad\quad\quad \mathcal{A}_{i,\ell+1} \leftarrow$ Phased Elimination($\mathcal{A}_{i,\ell}, \varepsilon_\ell$)

Algorithm 2 Collaborative Phased Elimination($\mathcal{A}_{1,\ell}$, ε_ℓ)

1: Find G-optimal design $\boldsymbol{\pi}_\ell \in \mathcal{P}(\mathcal{A}_{1,\ell})$ with $\mathrm{Supp}(\boldsymbol{\pi}_\ell) \le d(d+1)/2$
2: **for** $i \in [m]$ **do**
3: User i selects each action $\boldsymbol{x} \in \mathcal{A}_{i,\ell}$ for $n_{i,\ell}(\boldsymbol{x})$ times as defined in (3).
4: Record the actions $\mathbf{X}_i \in \mathbb{R}^{\bar{n}_{i,\ell} \times d}$ and rewards $\boldsymbol{y}_i \in \mathbb{R}^{\bar{n}_{i,\ell}}$.
5: Calculate $\hat{\boldsymbol{\mu}}_\ell = \frac{1}{m}\sum_{i=1}^m (\mathbf{X}_i{}^\top \mathbf{X}_i)^\dagger \mathbf{X}_i{}^\top \boldsymbol{y}_i$ and update

$$\mathcal{A}_{i,\ell+1} \leftarrow \left\{ x \in \mathcal{A}_{1,\ell} : \max_{\boldsymbol{x}' \in \mathcal{A}_{1,\ell}} \langle \hat{\boldsymbol{\mu}}_\ell, \boldsymbol{x}' - \boldsymbol{x} \rangle \le 2\varepsilon_\ell \right\}$$

for all $i \in [m]$.
6: **return** $\{\mathcal{A}_{i,\ell+1}\}_{i=1:m}$

of collaboration that each user only needs to afford $1/m$ of the pulls it would need if it were performing phased elimination on its own. In Lemma B.2, we see that by selecting this amount of actions, the agents collaboratively learn a good estimation on $\boldsymbol{\mu}$ with the target error ε_ℓ. And as we decrease ε_ℓ in each phase, we gradually form a better estimation on $\boldsymbol{\mu}$.

However, the ultimate goal for each user i is to find the best action with respect to $\boldsymbol{\theta}_i$ instead of $\boldsymbol{\mu}$. This motivates our second stage, the personalized learning stage. In this stage (line 7 to 9 in Algorithm 1), each user continues from the action set it gets from the end of the collaboration stage and performs phased elimination by its own. In each phase ℓ, we now have

$$n_{i,\ell}(\boldsymbol{x}) = \left\lceil \frac{\boldsymbol{\pi}_\ell(\boldsymbol{x}) g(\boldsymbol{\pi}_\ell)}{\varepsilon_\ell^2} \log \frac{km\ell(\ell+1)}{2\delta} \right\rceil \tag{4}$$

for all $\boldsymbol{x} \in \mathcal{A}_{i,\ell}$. Also in Lemma B.2, we can see that this gives each user i an estimation on $\boldsymbol{\theta}_i$ with an error ε_ℓ. Thus, each user can now refine its own result and find its optimal action.

The timing to switch from the collaborative stage to the personalized learning stage depends on the amount of heterogeneity. In the data-scarce regime, staying in the collaborative stage increases the effective number of samples for each user and benefits the learning. Also, when the heterogeneity is low, it reduces the workload for each user while providing a good estimation for each $\boldsymbol{\theta}_i$, since $\boldsymbol{\theta}_i$'s and $\boldsymbol{\mu}$. In the data-rich regime or with high heterogeneity, the users need to switch to the personalized learning stage earlier since collaboration barely helps. The timing to switch stages is explicitly determined by the term

$$h_{\sigma,\delta} = \sqrt{2\left(\max_{\boldsymbol{x}\in\mathcal{A}} \|\boldsymbol{x}\|_{\mathbf{C}}^2\right)\log\frac{4k}{\delta}} + \sqrt{\frac{2}{m}\left(\max_{\boldsymbol{x}\in\mathcal{A}} \|\boldsymbol{x}\|_{\mathbf{C}}^2\right)\log\frac{4k}{\delta}} \tag{5}$$

for any given $\delta \in (0,1)$. We can see that the term $h_{\sigma,\delta}$ depends on the covariance matrix $\mathbf{C}$, which is the covariance matrix in the population distribution reflecting the heterogeneity of the users. We can see from line 3 in Algorithm 1 that the users stay in the collaborative stage if the target error ε_ℓ is larger than $2h_{\sigma,\delta}$, and switch to the personalized learning stage once the target error goes below $2h_{\sigma,\delta}$. This switching is natural here since estimating $\boldsymbol{\mu}$ is good enough for estimating $\boldsymbol{\theta}_i$ as long as the target error ε_ℓ surpassed the heterogeneity. On the other hand, when the target error is becoming sufficiently small, estimating $\boldsymbol{\mu}$ will never give us a good estimation on $\boldsymbol{\theta}_i$, and thus we need to switch to personalized learning stage instead.

For the regret analysis on Algorithm 1, we start from the case with finite action sets in Theorem B.1, and extend the results for the unit-ball action set presented in Theorem 5.1.

Theorem 5.1. *Consider the unit ball action set $\mathcal{A} = \{\boldsymbol{x} : \|\boldsymbol{x}\|_2 \le 1\}$, and assume that $\|\boldsymbol{\mu}\|_2 \le 1$, $\sigma \le 1$, and $n \ge d^2$. For any $\delta \in (0,1)$, by running Algorithm 1, we have the following results:*

- *When $n = o\left(d/m\sigma^2\right)$, we have $\hat{R}_{m,n}(\boldsymbol{\mu}) = \tilde{O}\left(d\sqrt{mn} + md^2\right)$.*
- *When $n = \Theta\left(d/m^{2\gamma}\sigma^2\right)$ with $\gamma \in [0, \frac{1}{2}]$, we have $\hat{R}_{m,n}(\boldsymbol{\mu}) = \tilde{O}\left(dm^{1-\gamma}\sqrt{n} + md^2\right)$.*
- *When $n = \omega\left(d/\sigma^2\right)$, we have $\hat{R}_{m,n}(\boldsymbol{\mu}) = \tilde{O}\left(dm\sqrt{n} + md^2\right)$.*

with probability at least $1-\delta$.

The interpretation of the regret bounds can be found in Remark 3.2.

Remark 5.2 (General covariance matrix $\mathbf{C}$)**.** *Our regret analysis in Theorem 5.1 can be extended to the case of general known covariance matrix $\mathbf{C}$. Suppose now that $\boldsymbol{\theta}_i \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$ for some general covariance matrix $\mathbf{C}$, and let $\lambda_1(\mathbf{C})$ denote the largest eigenvalue of $\mathbf{C}$. We can simply replace σ by $\sqrt{\lambda_1(\mathbf{C})}$ in the result, and we obtain an upper bound characterized by the largest eigenvalue of $\mathbf{C}$. More details can be found in Appendix B.7.*

Remark 5.3 (Assumptions in Theorem B.1 and 5.1)**.** *We discuss the assumption on $\boldsymbol{\mu}$ and σ in Remark 3.3. The assumptions on the relations of n and d come from the phased elimination algorithm. In each phase ℓ, each action in the support of the design $\boldsymbol{\pi}_\ell$ has to be selected*

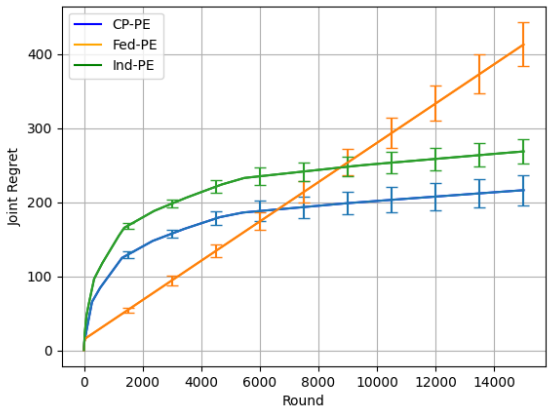

Figure 2: Joint cumulative regret obtained by CP-PE, Fed-PE, and Ind-PE. In the experiment, we have $m = 100$, $n = 15000$, $d = 2$, $\sigma = 0.3$.

at least once, and the support of the design is bounded by the size of the action set, or by $d(d+1)/2$ if the action set is compact (Lattimore and Szepesvári (2020).

Remark 5.4 (The md^2 term in the upper bound)**.** *We discuss the additional md^2 term in detail in Appendix B.5. Notice that the term is a constant with respect to the number of rounds n. As long as we have $n \geq md^2$, the second term is always dominated by the first term in the regret bound and thus we omit the term in Theorem 3.1.*

Sketch proofs of Theorem B.1 and Theorem 5.1. Following Algorithm 1, by choosing each action for the number of times indicated in (3) and (4), we show that each agent i learns a good estimation on $\boldsymbol{\theta}_i$ with the target error ε_ℓ in Lemma B.2. With the lemma, the following Lemma B.3 and B.4 are relatively standard from the proof in the single-agent case. With the lemmas, we rewrite the regret upper bound into an optimization form in (29). Then, we optimize (29) separately for the three regimes and get our final bounds. To extend the finite action set result to the unit ball action set case, we adopt the standard ϵ-net argument. The detailed proofs can be found in Section B.3 and B.4. □

6 Numerical Results

In our experiments, we compare CP-PE with Federated Phased Elimination (Fed-PE, all agents perform phased elimination as if they have the same unknown parameter) and Individual Phased Elimination (Ind-PE, each agent performs phased elimination on it own and evaluate the joint regret). The details of the experiment setting and additional results can be found in Appendix C. We pick the population parameter σ such that the heterogeneity falls in the second regime, and CP-PE usually performs one to two phases in the collaborative stage depending on the instance of $\{\boldsymbol{\theta}_i\}_{i=1:m}$ in each run.

In the experiment, there are 100 users ($m = 100$) and the level of heterogeneity $\sigma = 0.3$. Figure 2 shows the joint cumulative regret for each method. We can see CP-PE has the best final regret. Although Fed-PE has the best performance in the beginning, the cumulative regret grows linearly in n and is outperformed by CP-PE and Ind-PE eventually. The head start of Fed-PE is due to the benefit of collaboration, which significantly increases the effective sample size and eliminates the bad actions rapidly. However, without personalization, agents are forced to find a global best action that can be distinct from each agent's individual best action, and thus a linear regret is inevitable. On the other hand, CP-PE benefits from the collaborative learning stage and switch to the personalized learning stage before each agent's best action is eliminated. The transition happens in a few of the early rounds since the first few phases in phased elimination are composed of much fewer pulls of actions compared to the later phases, and the pulls are even equally divided among each user in the collaborative learning stage. By monitoring each run, we can see that agents under CP-PE rapidly shrink the action set in the collaborative learning stage and start with a much smaller action set when starting the personalized learning stage. Thus, CP-PE obtains a better regret than Ind-PE.

Experiments for unknown σ. To address the unknown σ case, we conducted experiments for unknown σ, and modified CP-PE such that σ is estimated in each phase. Details in Appendix C.3 show that the result is similar to the case of known σ.

7 Conclusion

We provided the first complete characterization of the benefit of collaboration under our hierarchical/empirical Bayesian framework for heterogeneous linear bandits. Overcoming technical challenges, we establish a minimax lower bound and propose CP-PE, an algorithm that achieves minimax-optimal regret up to polylogarithmic factors. Despite this complete characterization, several important directions remain open for future work. First, one can consider alternative heterogeneity models beyond the Gaussian/sub-Gaussian population studied in this work. Second, we empirically show that through estimating the unknown σ with modified CP-PE, we can obtain similar regrets. However, whether there is a fundamental gap in regret from the known heterogeneity case, remains an open and important question of ongoing investigations.

8 Acknowledgments

The research was supported in part by the DEVCOM Army Research Laboratory under award # W911NF1720196, and by the National Science Foundation under award #2502536. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the funding agencies.

References

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

Ethan Blaser, Chuanhao Li, and Hongning Wang. Federated linear contextual bandits with heterogeneous clients. In *International Conference on Artificial Intelligence and Statistics*, pages 631–639. PMLR, 2024.

Swapna Buccapatnam, Atilla Eryilmaz, and Ness B Shroff. Multi-armed bandits in the presence of side observations in social networks. In *52nd IEEE conference on decision and control*, pages 7309–7314. IEEE, 2013.

Hyunsung Cho, Akhil Mathur, and Fahim Kawsar. Flame: Federated learning across multi-device environments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(3): 1–29, 2022.

Anh Do, Thanh Nguyen-Tang, and Raman Arora. Multi-agent learning with heterogeneous linear contextual bandits. *Advances in Neural Information Processing Systems*, 36:78768–78790, 2023.

Claudio Gentile, Shuai Li, and Giovanni Zappella. Online clustering of bandits. In *International conference on machine learning*, pages 757–765. PMLR, 2014.

Avishek Ghosh, Abishek Sankararaman, and Kannan Ramchandran. Collaborative learning and personalization in multi-agent stochastic linear bandits. *stat*, 1050:15, 2021.

Eshcar Hillel, Zohar S Karnin, Tomer Koren, Ronny Lempel, and Oren Somekh. Distributed exploration in multi-armed bandits. *Advances in Neural Information Processing Systems*, 26, 2013.

Joey Hong, Branislav Kveton, Manzil Zaheer, and Mohammad Ghavamzadeh. Hierarchical bayesian bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 7724–7741. PMLR, 2022.

Ruiquan Huang, Weiqiang Wu, Jing Yang, and Cong Shen. Federated linear contextual bandits. *Advances in neural information processing systems*, 34:27057–27068, 2021.

Abdulrahman Khamaj and Abdulelah M Ali. Adapting user experience with reinforcement learning: Personalizing interfaces based on user behavior analysis in real-time. *Alexandria Engineering Journal*, 95: 164–173, 2024.

Nathan Korda, Balazs Szorenyi, and Shuai Li. Distributed clustering of linear bandits in peer to peer networks. In *International conference on machine learning*, pages 1301–1309. PMLR, 2016.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard. 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, 2016.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Chuanhao Li and Hongning Wang. Asynchronous upper confidence bound algorithms for federated linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 6529–6553. PMLR, 2022.

Shuai Li, Alexandros Karatzoglou, and Claudio Gentile. Collaborative filtering bandits. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 539–548, 2016.

Jiming Liu, Chi Kuen Wong, and Ka Keung Hui. An adaptive user interface based on personalized learning. *IEEE Intelligent Systems*, 18(2):52–57, 2003.

Xutong Liu, Haoru Zhao, Tong Yu, Shuai Li, and John CS Lui. Federated online clustering of bandits. In *Uncertainty in Artificial Intelligence*, pages 1221–1231. PMLR, 2022.

Kanak Mahadik, Qingyun Wu, Shuai Li, and Amit Sabne. Fast distributed bandits for online recommendation systems. In *Proceedings of the 34th ACM international conference on supercomputing*, pages 1–13, 2020.

Kaan Ozkara, Antonious M Girgis, Deepesh Data, and Suhas N Diggavi. A statistical framework for personalized federated learning and estimation: Theory, algorithms, and privacy. In *International Conference on Learning Representations (ICLR), 2023*. OpenReview. net, 2023.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Yuanhao Wang, Jiachen Hu, Xiaoyu Chen, and Liwei Wang. Distributed bandit learning: Near-optimal regret with efficient communication. In *International Conference on Learning Representations*, 2020.

Qingyun Wu, Huazheng Wang, Quanquan Gu, and Hongning Wang. Contextual bandits in a collaborative environment. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 529–538, 2016.

Jingxuan Zhu, Romeil Sandhu, and Ji Liu. A distributed algorithm for sequential decision making in multi-armed bandit with homogeneous rewards. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 3078–3083. IEEE, 2020.

Zhaowei Zhu, Jingxuan Zhu, Ji Liu, and Yang Liu. Federated bandit: A gossiping approach. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 5(1):1–29, 2021.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Instructions for Paper Submissions to AISTATS 2026: Supplementary Materials

A Information-Theoretic Lower Bound

A.1 Proof of Theorem 4.1

Let $\mathcal{A} = \mathcal{B}_d^2$ be the unit ball action set. We prove the lower bound in the three regimes separately.

A.1.1 Lower bound for large σ

We start from the third regime, the data-rich regime, in Theorem 4.1. Equivalently, we have

$$\sigma = \Theta\left(\sqrt{\frac{d}{n^{1-\alpha}}}\right)$$

for some $\alpha \in [0, 1]$. By the definition of the regret, for any given $\boldsymbol{\mu} \in \mathbb{R}^d$, we have

$$\begin{aligned} R_{m,n}(\boldsymbol{\mu}) &= \mathbb{E}\left[\sum_{i=1}^{m}\sum_{t=1}^{n}\langle \boldsymbol{x}_i^* - \boldsymbol{x}_{i,t}, \boldsymbol{\theta}_i\rangle\right] \\ &= \mathbb{E}\left[\sum_{i=1}^{m}\sum_{t=1}^{n}\|\boldsymbol{\theta}_i\|_2 - \langle \boldsymbol{x}_{i,t}, \boldsymbol{\theta}_i\rangle\right] \\ &= \sum_{i=1}^{m}\mathbb{E}\left[\sum_{t=1}^{n}\|\boldsymbol{\theta}_i\|_2 - \langle \boldsymbol{x}_{i,t}, \boldsymbol{\theta}_i\rangle\right] \\ &= m\mathbb{E}\left[\sum_{t=1}^{n}\|\boldsymbol{\theta}_m\|_2 - \langle \boldsymbol{x}_{m,t}, \boldsymbol{\theta}_m\rangle\right]. \end{aligned} \tag{6}$$

The second equality comes from the fact that $\max_{\boldsymbol{x}\in\mathcal{B}_d^2}\langle \boldsymbol{x}, \boldsymbol{\theta}_i\rangle = \|\boldsymbol{\theta}_i\|_2$ and the last equality is due to symmetry. Thus, let us focus on the term

$$\mathbb{E}\left[\sum_{t=1}^{n}\|\boldsymbol{\theta}_m\|_2 - \langle \boldsymbol{x}_{m,t}, \boldsymbol{\theta}_m\rangle\right].$$

Despite the expression, note that the actions $\{\boldsymbol{x}_{m,t}\}_{t=1:n}$ still depend on the parameters of all users, $\{\boldsymbol{\theta}_i\}_{i=1:m}$. Consider the case when

$$\sigma = c_1 d^{\frac{1}{2}} n^{-\frac{1}{2}+\alpha} \tag{7}$$

for some constant $c_1 > 0$ and $\alpha \geq 0$. Notice that the results can be easily extended to the case where $c_1 d^{\frac{1}{2}} n^{-\frac{1}{2}+\alpha} \leq \sigma \leq c_1' d^{\frac{1}{2}} n^{-\frac{1}{2}+\alpha}$.

Suppose that we have a d-dimensional random vector $\boldsymbol{\theta}^{(0)} \in \mathcal{N}(0, \sigma^2\mathbf{I})$. We have the following lemma:

Lemma A.1. *Let $z_1, \ldots, z_d$ be d independent random variables that follows the normal distribution*

$$z_i \sim \mathcal{N}(0, \sigma^2) \quad \forall i \in [d]$$

for some $\sigma > 0$. Then, we have

$$\mathbb{P}\left(\sum_{i=1}^{d} z_i^2 \in \left[\frac{d\sigma^2}{2}, \frac{3d\sigma^2}{2}\right]\right) \geq c_2$$

for some constant $c_2 \in (0, 1)$.

The lemma holds since $\sum_{i=1}^d z_i^2/\sigma^2$ follows the chi-square distribution.

Let us consider the case that $\boldsymbol{\mu} = 0$. Since $\|\boldsymbol{\theta}^{(0)}\|^2 = \sum_{\ell=1}^d (\theta_\ell^{(0)})^2$ and $\theta_\ell^{(0)} \sim \mathcal{N}(0, \sigma^2)$, by the lemma, we have

$$\mathbb{P}\left(\|\boldsymbol{\theta}^{(0)}\|^2 \in \left[\frac{d\sigma^2}{2}, \frac{3d\sigma^2}{2}\right]\right) \geq c_2$$

with some constant $c_2 \in (0, 1)$. That is,

$$\mathbb{P}\left(\|\boldsymbol{\theta}^{(0)}\| \in \left[\sigma\sqrt{\frac{d}{2}}, \sigma\sqrt{\frac{3d}{2}}\right]\right) = \mathbb{P}\left(\|\boldsymbol{\theta}^{(0)}\| \in \left[\frac{c_1\sqrt{2}}{2} dn^{-\frac{1}{2}+\alpha}, \frac{c_1\sqrt{6}}{2} dn^{-\frac{1}{2}+\alpha}\right]\right) \geq c_2. \tag{8}$$

We denote the event by $\mathcal{E}$, i.e.,

$$\mathcal{E} = \left\{\|\boldsymbol{\theta}^{(0)}\| \in \left[\frac{c_1\sqrt{2}}{2} dn^{-\frac{1}{2}+\alpha}, \frac{c_1\sqrt{6}}{2} dn^{-\frac{1}{2}+\alpha}\right]\right\}$$

and thus $\mathbb{P}(\mathcal{E}) \geq c_2$. Let $\mathcal{S}^{d-1}(\|\boldsymbol{\theta}^{(0)}\|)$ be the sphere centered at $\mathbf{0} \in \mathbb{R}^d$ with radius $\|\boldsymbol{\theta}^{(0)}\|$ in the d dimensional space. Let us define a distribution on a set of d-dimensional random vectors $\{\boldsymbol{\theta}^{(\boldsymbol{z})} : \boldsymbol{z} \in \mathcal{Z}\}$ where the index set $\mathcal{Z} = \{\pm 1\}^{d-1}$. Given the event $\mathcal{E}$, let us write

$$\|\boldsymbol{\theta}^{(0)}\| = c_3 dn^{-\frac{1}{2}+\alpha} \tag{9}$$

where $c_3 \in \left[\frac{c_1\sqrt{2}}{2}, \frac{c_1\sqrt{6}}{2}\right]$. Let us denote

$$\tilde{\mathcal{S}}_n^{d-2}(\boldsymbol{\theta}^{(0)}) = \left\{\boldsymbol{\theta} \in \mathcal{S}^{d-1}(\|\boldsymbol{\theta}^{(0)}\|) : \frac{\boldsymbol{\theta}^\top \boldsymbol{\theta}^{(0)}}{\|\boldsymbol{\theta}\|\|\boldsymbol{\theta}^{(0)}\|} = \cos\eta\right\}$$

where

$$\eta = \sin^{-1}\left(\frac{c_4\sqrt{d(d-1)}n^{-\frac{1}{2}+\frac{\alpha}{2}}}{\|\boldsymbol{\theta}^{(0)}\|}\right) = \sin^{-1}\left(\frac{c_4}{c_3}\sqrt{\frac{d-1}{dn^\alpha}}\right) \in \left(0, \frac{\pi}{2}\right), \tag{10}$$

$$c_4 = \min\left\{\frac{c_1\sqrt{2}}{2}, \sqrt{\frac{c_1}{6\sqrt{2}}}\right\} \leq c_3.$$

The set $\tilde{\mathcal{S}}_n^{d-2}(\boldsymbol{\theta}^{(0)})$ is a $(d-2)$-dimensional sphere on $\mathcal{S}^{d-1}(\|\boldsymbol{\theta}^{(0)}\|)$. Let $\tilde{\boldsymbol{\theta}}^{(0)}$ denote the center of $\tilde{\mathcal{S}}_n^{d-2}(\boldsymbol{\theta}^{(0)})$. Notice that $\tilde{\boldsymbol{\theta}}^{(0)} \neq \boldsymbol{\theta}^{(0)}$. We want to randomly pick a set $\{\boldsymbol{\theta}^{(\boldsymbol{z})} : \boldsymbol{z} \in \mathcal{Z}\} \subset \tilde{\mathcal{S}}_n^{d-2}(\boldsymbol{\theta}^{(0)})$ such that $\{\boldsymbol{\theta}^{(\boldsymbol{z})} : \boldsymbol{z} \in \mathcal{Z}\}$ is the set of vertices of a $(d-1)$-dimensional hyper-cube. To be more specific on the indices, each $(\boldsymbol{\theta}^{(\boldsymbol{z})}, \boldsymbol{\theta}^{(\boldsymbol{z}')})$ pair forms the end points of an edge if $\|\boldsymbol{z} - \boldsymbol{z}'\|_1 = 2$, i.e., $\boldsymbol{z}$ and $\boldsymbol{z}'$ only differ in one coordinate. To do so, we can simply start from uniformly choosing a $(d-1)$-dimensional basis, and find the $(d-1)$-dimensional hyper-cube on $\tilde{\mathcal{S}}_n^{d-2}(\boldsymbol{\theta}^{(0)})$ that align with the basis. Due to the symmetry of sampling $\boldsymbol{\theta}^{(0)}$ and the basis, we have

$$\boldsymbol{\theta}^{(\boldsymbol{z})} \sim \mathcal{N}(0, \sigma^2\mathbf{I}) \quad \forall \boldsymbol{z} \in \mathcal{Z}. \tag{11}$$

Now, with some rotation matrix $W \in \mathbb{R}^{d\times d}$, we can rotate $\boldsymbol{\theta}^{(0)}$ and the hyper-cube $\{\boldsymbol{\theta}^{(\boldsymbol{z})} : z \in \mathcal{Z}\}$ such that $W\boldsymbol{\theta}^{(0)}$ points toward the first coordinate and the hyper-cube aligns with the standard basis. Then, we have

$$\tilde{\boldsymbol{\theta}}^{(\boldsymbol{z})} = W\boldsymbol{\theta}^{(\boldsymbol{z})} \quad \forall z \in \mathcal{Z} \tag{12}$$

where

$$\tilde{\theta}_\ell^{(\boldsymbol{z})} = \begin{cases} \|\boldsymbol{\theta}^{(0)}\|\cos\eta & \ell = 1, \\ \frac{z_\ell}{\sqrt{d-1}}\|\boldsymbol{\theta}^{(0)}\|\sin\eta = z_\ell c_4\sqrt{d}n^{-\frac{1}{2}+\frac{\alpha}{2}} & \ell = 2, \ldots, d, \end{cases}$$

and thus

$$\|\tilde{\boldsymbol{\theta}}^{(\boldsymbol{z}_1)} - \tilde{\boldsymbol{\theta}}^{(\boldsymbol{z}_2)}\| = 2c_4 n^{-\frac{1}{2}+\frac{\alpha}{2}} \quad \forall \boldsymbol{z}_1, \boldsymbol{z}_2 \in \mathcal{Z}, \|\boldsymbol{z}_1 - \boldsymbol{z}_2\|_1 = 2.$$

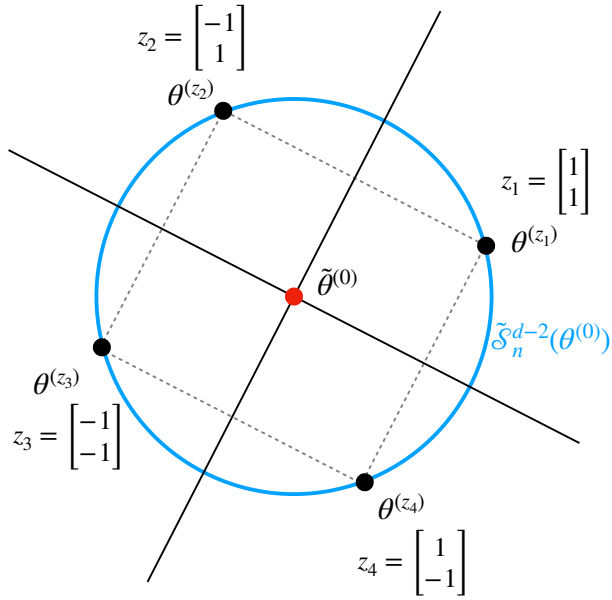

Figure 3: Constructing the $\boldsymbol{\theta}^{(\boldsymbol{z})}$'s

Thus, $\{\boldsymbol{\theta}^{(\boldsymbol{z})}\}$ are the vertices of a $(d-1)$-dimensional hypercube aligning with the standard basis, and each of the vertices has the same value in the first coordinate.

Now we want to show a lower bound for the average regret with respect to $\{\tilde{\boldsymbol{\theta}}^{(\boldsymbol{z})} : \boldsymbol{z} \in \mathcal{Z}\}$. For simplicity, let us denote $\Theta = \{\boldsymbol{\theta}^{(\boldsymbol{z})} : \boldsymbol{z} \in \mathcal{Z}\}$. Suppose that $\boldsymbol{z}$ is uniformly randomly sampled from $\{\pm 1\}^{d-1}$ and we set $\boldsymbol{\theta}_m = \boldsymbol{\theta}^{(\boldsymbol{z})}$. Since $\boldsymbol{\theta}_m$ and $\boldsymbol{\theta}^{(\boldsymbol{z})}$ follow the same distribution, we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{n} \|\boldsymbol{\theta}_m\|_2 - \langle \boldsymbol{x}_{m,t}, \boldsymbol{\theta}_m \rangle\right] &= \mathbb{E}\left[\sum_{t=1}^{n} \|\boldsymbol{\theta}^{(\boldsymbol{z})}\|_2 - \langle \boldsymbol{x}_{m,t}, \boldsymbol{\theta}^{(\boldsymbol{z})} \rangle \middle| \boldsymbol{z}\right] \\
&= \frac{1}{|\mathcal{Z}|} \sum_{\boldsymbol{z} \in \mathcal{Z}} \mathbb{E}\left[\sum_{t=1}^{n} \|\boldsymbol{\theta}^{(\boldsymbol{z})}\| - \langle \boldsymbol{x}_{m,t}, \boldsymbol{\theta}^{(\boldsymbol{z})} \rangle \middle| \boldsymbol{z}\right] \\
&= \frac{1}{|\mathcal{Z}|} \sum_{\boldsymbol{z} \in \mathcal{Z}} \mathbb{E}\left[\sum_{t=1}^{n} \|\boldsymbol{\theta}^{(\boldsymbol{z})}\| - \langle W^{-1}\boldsymbol{x}_{m,t}, W\boldsymbol{\theta}^{(\boldsymbol{z})} \rangle \middle| \boldsymbol{z}\right] \\
&= \frac{1}{|\mathcal{Z}|} \sum_{\boldsymbol{z} \in \mathcal{Z}} \mathbb{E}\left[\sum_{t=1}^{n} \|\boldsymbol{\theta}^{(\boldsymbol{z})}\| - \langle \tilde{\boldsymbol{x}}_{m,t}, \tilde{\boldsymbol{\theta}}^{(\boldsymbol{z})} \rangle \middle| \boldsymbol{z}\right],
\end{aligned} \tag{13}$$

where $\tilde{\boldsymbol{x}}_{m,t}$ still falls in the unit ball due to the rotation. For any $t \in [n]$,we have

$$\begin{aligned}
&\|\boldsymbol{\theta}^{(\boldsymbol{z})}\| - \langle \tilde{\boldsymbol{x}}_{m,t}, \tilde{\boldsymbol{\theta}}^{(\boldsymbol{z})} \rangle \\
=& \left(\|\boldsymbol{\theta}^{(\boldsymbol{z})}\| \cos^2\eta - \tilde{x}_{m,t,1}\tilde{\theta}_1^{(\boldsymbol{z})}\right) + \sum_{\ell=2}^{d} \left(\|\boldsymbol{\theta}^{(\boldsymbol{z})}\| \frac{\sin^2\eta}{d-1} - \tilde{x}_{m,t,\ell}\tilde{\theta}_\ell^{(\boldsymbol{z})}\right) \\
=& \left(\|\boldsymbol{\theta}^{(\boldsymbol{z})}\| \cos^2\eta - \tilde{x}_{m,t,1}\|\boldsymbol{\theta}^{(\boldsymbol{z})}\|(\cos\eta)\right) \\
&+ \sum_{\ell=2}^{d} \left(\|\boldsymbol{\theta}^{(\boldsymbol{z})}\| \cdot \frac{\sin^2\eta}{d-1} - \tilde{x}_{m,t,\ell}\left(\|\boldsymbol{\theta}^{(\boldsymbol{z})}\| \cdot \frac{\sin\eta}{\sqrt{d-1}}\right) z_\ell\right)
\end{aligned}$$

$$
\begin{aligned}
&= \left(\frac{\|\boldsymbol{\theta}^{(\boldsymbol{z})}\|\cos^2\eta}{2} - \tilde{x}_{m,t,1}\|\boldsymbol{\theta}^{(\boldsymbol{z})}\|\cos\eta + \frac{\|\boldsymbol{\theta}^{(\boldsymbol{z})}\|\tilde{x}_{m,t,1}^2}{2}\right) + \left(\frac{\|\boldsymbol{\theta}^{(\boldsymbol{z})}\|\cos^2\eta}{2} - \frac{\|\boldsymbol{\theta}^{(\boldsymbol{z})}\|\tilde{x}_{m,t,1}^2}{2}\right) \\
&\quad + \sum_{\ell=2}^{d}\left(\frac{\|\boldsymbol{\theta}^{(\boldsymbol{z})}\|\sin^2\eta}{2(d-1)} - \tilde{x}_{m,t,\ell}\left(\|\boldsymbol{\theta}^{(\boldsymbol{z})}\|\cdot\frac{\sin\eta}{\sqrt{d-1}}\right)z_\ell + \frac{\|\boldsymbol{\theta}^{(\boldsymbol{z})}\|\tilde{x}_{m,t,\ell}^2}{2}\right) \\
&\quad + \sum_{\ell=2}^{d}\left(\frac{\|\boldsymbol{\theta}^{(\boldsymbol{z})}\|\sin^2\eta}{2(d-1)} - \frac{\|\boldsymbol{\theta}^{(\boldsymbol{z})}\|\tilde{x}_{m,t,\ell}^2}{2}\right) \\
&= \frac{\|\boldsymbol{\theta}^{(\boldsymbol{z})}\|\cos^2\eta}{2}\left(1 - 2\frac{\tilde{x}_{m,t,1}}{\cos\eta} + \frac{\tilde{x}_{m,t,1}^2}{\cos^2\eta}\right) + \left(\frac{\|\boldsymbol{\theta}^{(\boldsymbol{z})}\|\cos^2\eta}{2} - \frac{\|\boldsymbol{\theta}^{(\boldsymbol{z})}\|\tilde{x}_{m,t,1}^2}{2}\right) \\
&\quad + \sum_{\ell=2}^{d}\frac{\|\boldsymbol{\theta}^{(\boldsymbol{z})}\|\sin^2\eta}{2}\left(\frac{1}{d-1} - \frac{2\tilde{x}_{m,t,\ell}z_\ell}{\sqrt{d-1}\sin\eta} + \frac{\tilde{x}_{m,t,\ell}^2}{\sin^2\eta}\right) \\
&\quad + \sum_{\ell=2}^{d}\left(\frac{\|\boldsymbol{\theta}^{(\boldsymbol{z})}\|\sin^2\eta}{2(d-1)} - \frac{\|\boldsymbol{\theta}^{(\boldsymbol{z})}\|\tilde{x}_{m,t,\ell}^2}{2}\right) \\
&= \frac{\|\boldsymbol{\theta}^{(\boldsymbol{z})}\|\cos^2\eta}{2}\left(1 - \frac{\tilde{x}_{m,t,1}}{\cos\eta}\right)^2 + \sum_{\ell=2}^{d}\frac{\|\boldsymbol{\theta}^{(\boldsymbol{z})}\|\sin^2\eta}{2}\left(\frac{1}{\sqrt{d-1}} - \frac{\tilde{x}_{m,t,\ell}}{\sin\eta}\right)^2 \\
&\quad + \left(\frac{\|\boldsymbol{\theta}^{(\boldsymbol{z})}\|}{2} - \frac{\|\boldsymbol{\theta}^{(\boldsymbol{z})}\|}{2}\sum_{\ell=1}^{d}\tilde{x}_{m,t,\ell}^2\right) \\
&\geq \sum_{\ell=2}^{d}\frac{\|\boldsymbol{\theta}^{(\boldsymbol{z})}\|\sin^2\eta}{2}\left(\frac{1}{\sqrt{d-1}} - \frac{\tilde{x}_{m,t,\ell}\cdot z_\ell}{\sin\eta}\right)^2.
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{n}\|\boldsymbol{\theta}^{(\boldsymbol{z})}\| - \langle\tilde{\boldsymbol{x}}_{i,t}, \tilde{\boldsymbol{\theta}}^{(\boldsymbol{z})}\rangle\middle|\boldsymbol{z}\right] \\
&\geq \mathbb{P}(\mathcal{E})\,\mathbb{E}\left[\sum_{t=1}^{n}\sum_{\ell=2}^{d}\frac{\|\boldsymbol{\theta}^{(\boldsymbol{z})}\|\sin^2\eta}{2}\left(\frac{1}{\sqrt{d-1}} - \frac{\tilde{x}_{m,t,\ell}\cdot z_\ell}{\sin\eta}\right)^2\middle|\boldsymbol{z},\mathcal{E}\right] \\
&\geq c_2\mathbb{E}\left[\sum_{\ell=2}^{d}\frac{\|\boldsymbol{\theta}^{(\boldsymbol{z})}\|\sin^2\eta}{2}\sum_{t=1}^{n'}\left(\frac{1}{\sqrt{d-1}} - \frac{\tilde{x}_{m,t,\ell}\cdot z_\ell}{\sin\eta}\right)^2\middle|\boldsymbol{z},\mathcal{E}\right] \\
&= c_2\mathbb{E}\left[\sum_{\ell=2}^{d}\frac{c_4^2 d(d-1)n^{-1+\alpha}}{2\|\boldsymbol{\theta}^{(0)}\|}\sum_{t=1}^{n'}\left(\frac{1}{\sqrt{d-1}} - \frac{\tilde{x}_{m,t,\ell}\cdot z_\ell}{\sin\eta}\right)^2\middle|\boldsymbol{z},\mathcal{E}\right] \\
&= c_2\mathbb{E}\left[\sum_{\ell=2}^{d}\frac{c_4^2 d(d-1)n^{-1+\alpha}}{2c_3 dn^{-\frac{1}{2}+\alpha}}\sum_{t=1}^{n'}\left(\frac{1}{\sqrt{d-1}} - \frac{\tilde{x}_{m,t,\ell}\cdot z_\ell}{\sin\eta}\right)^2\middle|\boldsymbol{z},\mathcal{E}\right] \\
&= \frac{c_2(d-1)}{2\sqrt{n}}\sum_{\ell=2}^{d}\mathbb{E}\left[\frac{c_4^2}{c_3}\sum_{t=1}^{n'}\left(\frac{1}{\sqrt{d-1}} - \frac{\tilde{x}_{m,t,\ell}\cdot z_\ell}{\sin\eta}\right)^2\middle|\boldsymbol{z},\mathcal{E}\right] \\
&= \frac{c_2(d-1)}{2\sqrt{n}}\sum_{\ell=2}^{d}\mathbb{E}\left[\frac{c_4^2}{c_3\sin^2\eta}\sum_{t=1}^{n'}\left(\frac{\sin\eta}{\sqrt{d-1}} - \tilde{x}_{m,t,\ell}\cdot z_\ell\right)^2\middle|\boldsymbol{z},\mathcal{E}\right] \\
&= \frac{c_2(d-1)}{2\sqrt{n}}\sum_{\ell=2}^{d}\mathbb{E}\left[\frac{c_3 dn^{\alpha}}{d-1}\sum_{t=1}^{n'}\left(\frac{\sin\eta}{\sqrt{d-1}} - \tilde{x}_{m,t,\ell}\cdot z_\ell\right)^2\middle|\boldsymbol{z},\mathcal{E}\right]
\end{aligned}
$$

$$
\begin{aligned}
&\geq \frac{c_2 d n^{-\frac{1}{2}+\alpha}}{2} \sum_{\ell=2}^{d} \mathbb{E}\left[\frac{\sqrt{2}c_1}{2} \sum_{t=1}^{n'} \left(\frac{\sin\eta}{\sqrt{d-1}} - \tilde{x}_{m,t,\ell} \cdot z_\ell\right)^2 \middle| \boldsymbol{z}, \mathcal{E}\right] \\
&= c_5 d n^{-\frac{1}{2}+\alpha} \sum_{\ell=2}^{d} \boldsymbol{z}, \mathbb{E}\left[\sum_{t=1}^{n'} \left(\frac{\sin\eta}{\sqrt{d-1}} - \tilde{x}_{m,t,\ell} \cdot z_\ell\right)^2 \middle| \mathcal{E}\right] \qquad (14)
\end{aligned}
$$

where $c_5 = (\sqrt{2}c_1c_2)/4$ and

$$
n' = \min\left\{n, \min\left\{\tau : \sum_{t=1}^{\tau} \tilde{x}_{m,t,\ell}^2 \geq \frac{n\sin^2\eta}{d-1}\right\}\right\}.
$$

Now, we want to show that give a $(\boldsymbol{z}, \boldsymbol{z}')$ pair such that $\boldsymbol{z}$ and $\boldsymbol{z}'$ only differ in the ℓ-th coordinate, the regret

$$
\mathbb{E}\left[\sum_{t=1}^{n'} \left(\frac{\sin\eta}{\sqrt{d-1}} - \tilde{x}_{m,t,\ell} \cdot z_\ell\right)^2 \middle| \boldsymbol{z}, \mathcal{E}\right] + \mathbb{E}\left[\sum_{t=1}^{n'} \left(\frac{\sin\eta}{\sqrt{d-1}} - \tilde{x}_{m,t,\ell} \cdot z'_\ell\right)^2 \middle| \boldsymbol{z}', \mathcal{E}\right]
$$

is bounded by some lower bound. For any $w \in \{1, -1\}$, we have

$$
\begin{aligned}
\sum_{t=1}^{n'} \left(\frac{\sin\eta}{\sqrt{d-1}} - \tilde{x}_{m,t,\ell} \cdot w\right)^2 &\leq 2\sum_{t=1}^{n'} \frac{\sin^2\eta}{d-1} + 2\sum_{t=1}^{n'} \tilde{x}_{m,t,\ell}^2 \\
&\leq \frac{2n\sin^2\eta}{d-1} + 2\left(\frac{n\sin^2\eta}{d-1} + 1\right) \\
&= \frac{4n}{d-1} \cdot \frac{c_4^2}{c_3^2} \cdot \frac{d-1}{dn^\alpha} + 2 \\
&\leq \frac{4c_4^2 n^{1-\alpha}}{c_3^2 d} + 2 \\
&\leq \frac{4(c_4^2/c_3^2)n^{1-\alpha} + 2d}{d} \\
&\leq \frac{6(c_4^2/c_3^2)n^{1-\alpha}}{d}.
\end{aligned}
$$

Following the proof in Section 24.2 of Lattimore and Szepesvári (2020), with Pinsker's inequality and Exercise 14.4 in Lattimore and Szepesvári (2020), we have

$$
\begin{aligned}
&\left|\mathbb{E}\left[\sum_{t=1}^{n'} \left(\frac{\sin\eta}{\sqrt{d-1}} - \tilde{x}_{m,t,\ell}\right)^2 \middle| \boldsymbol{z}', \mathcal{E}\right] - \mathbb{E}\left[\sum_{t=1}^{n'} \left(\frac{\sin\eta}{\sqrt{d-1}} - \tilde{x}_{m,t,\ell}\right)^2 \middle| \boldsymbol{z}', \mathcal{E}\right]\right| \\
\leq &\left(\frac{8(c_4^2/c_3^2)n^{1-\alpha}}{d}\right) \sqrt{\frac{1}{2}\mathrm{D}\left(\mathbb{P}_{\boldsymbol{z}} \| \mathbb{P}_{\boldsymbol{z}'}\right)},
\end{aligned}
$$

where $\mathbb{P}_{\boldsymbol{z}}$ is the joint distribution of $\{(\boldsymbol{x}_{i,t}, y_{i,t})\}_{i=1:m,t=1:n}$ given that $\boldsymbol{\theta}_m = \boldsymbol{\theta}^{(\boldsymbol{z})}$ and event $\mathcal{E}$ holds. With Section 24.2 and Exercise 15.7 in Lattimore and Szepesvári (2020), we know that

$$
\begin{aligned}
\mathrm{D}\left(\mathbb{P}_{\boldsymbol{z}} \| \mathbb{P}_{\boldsymbol{z}'}\right) &= \frac{1}{2}\mathbb{E}_{\boldsymbol{z}}\left[\sum_{t=1}^{n'} \left\langle \tilde{\boldsymbol{x}}_{m,t}, \tilde{\boldsymbol{\theta}}^{(\boldsymbol{z})} - \tilde{\boldsymbol{\theta}}^{(\boldsymbol{z}')} \right\rangle^2\right] \\
&= \frac{1}{2}\mathbb{E}_{\boldsymbol{z}}\left[\sum_{t=1}^{n'} \left(x_{m,t,\ell} \cdot \left(\theta_\ell^{(\boldsymbol{z})} - \theta_\ell^{(\boldsymbol{z}')}\right)\right)^2\right] \\
&\leq \frac{1}{2}\left(2c_4\sqrt{d}n^{-\frac{1}{2}+\frac{\alpha}{2}}\right)^2 \mathbb{E}_{\boldsymbol{z}}\left[\sum_{t=1}^{n'} x_{m,t,\ell}^2\right]
\end{aligned}
$$

$$
\begin{aligned}
&\leq \frac{1}{2}\left(2c_4\sqrt{d}n^{-\frac{1}{2}+\frac{\alpha}{2}}\right)^2\left(\frac{n\sin^2\eta}{d-1}+1\right)\\
&= 2c_4^2 dn^{-1+\alpha}\left(\frac{n}{d-1}\frac{c_4^2}{c_3^2}\frac{d-1}{dn^\alpha}+1\right)\\
&\leq \frac{2c_4^4}{c_3^2}+2c_4^2dn^{-1+\alpha}\\
&\leq \frac{2c_4^4}{c_3^2}+\frac{2c_4^4}{c_3^2}\\
&\leq \frac{4c_4^4}{c_3^2}.
\end{aligned}
$$

With the bound on the KL-divergence term, we have

$$
\begin{aligned}
&\left|\mathbb{E}_{\boldsymbol{z}}\left[\sum_{t=1}^{n'}\left(\frac{\sin\eta}{\sqrt{d-1}}-\tilde{x}_{m,t,\ell}\right)^2\right]-\mathbb{E}_{\boldsymbol{z}'}\left[\sum_{t=1}^{n'}\left(\frac{\sin\eta}{\sqrt{d-1}}-\tilde{x}_{m,t,\ell}\right)^2\right]\right|\\
&\leq \left(\frac{6(c_4^2/c_3^2)n^{1-\alpha}}{d}\right)\sqrt{\frac{4c_4^4}{c_3^2}}\\
&= \frac{12(c_4^4/c_3^3)n^{1-\alpha}}{d}.
\end{aligned}
$$

Thus, we have

$$
\mathbb{E}_{\boldsymbol{z}}\left[\sum_{t=1}^{n'}\left(\frac{\sin\eta}{\sqrt{d-1}}-\tilde{x}_{m,t,\ell}\right)^2\right]\geq \mathbb{E}_{\boldsymbol{z}'}\left[\sum_{t=1}^{n'}\left(\frac{\sin\eta}{\sqrt{d-1}}-\tilde{x}_{m,t,\ell}\right)^2\right]-\frac{12(c_4^4/c_3^3)n^{1-\alpha}}{d}.
$$

Adding an extra term on the both side gives us

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{z}}\left[\sum_{t=1}^{n'}\left(\frac{\sin\eta}{\sqrt{d-1}}-\tilde{x}_{m,t,\ell}\right)^2\right]+\mathbb{E}_{\boldsymbol{z}'}\left[\sum_{t=1}^{n'}\left(\frac{\sin\eta}{\sqrt{d-1}}+\tilde{x}_{m,t,\ell}\right)^2\right]\\
&\geq \mathbb{E}_{\boldsymbol{z}'}\left[\sum_{t=1}^{n'}\left(\frac{\sin\eta}{\sqrt{d-1}}-\tilde{x}_{m,t,\ell}\right)^2\right]+\mathbb{E}_{\boldsymbol{z}'}\left[\sum_{t=1}^{n'}\left(\frac{\sin\eta}{\sqrt{d-1}}+\tilde{x}_{m,t,\ell}\right)^2\right]-\frac{12n^{1-\alpha}}{d}\\
&= 2\sum_{t=1}^{n'}\frac{\sin^2\eta}{d-1}+2\mathbb{E}_{\boldsymbol{z}'}\left[\sum_{t-1}^{n'}\tilde{x}_{m,t,\ell}\right]-\frac{12n^{1-\alpha}}{d}\\
&\geq \frac{2n\sin^2\eta}{d-1}-\frac{12n^{1-\alpha}}{d}\\
&= \frac{2n}{d-1}\frac{c_4^2}{c_3^2}\frac{d-1}{dn^\alpha}-\frac{12n^{1-\alpha}}{d}\\
&= \frac{2c_4^2n^{1-\alpha}}{c_3^2d}\left(1-\frac{6c_4^2}{c_3}\right).
\end{aligned}
$$

Since we have

$$
c_4^2=\frac{c_1}{6\sqrt{2}}\leq\frac{(\sqrt{2}c_3/2)}{6\sqrt{2}}\leq\frac{c_3}{12},
$$

we have

$$
\mathbb{E}_{\boldsymbol{z}}\left[\sum_{t=1}^{n'}\left(\frac{\sin\eta}{\sqrt{d-1}}-\tilde{x}_{m,t,\ell}\right)^2\right]+\mathbb{E}_{\boldsymbol{z}'}\left[\sum_{t=1}^{n'}\left(\frac{\sin\eta}{\sqrt{d-1}}+\tilde{x}_{m,t,\ell}\right)^2\right]
$$

$$
\begin{aligned}
&\geq \frac{c_4^2 n^{1-\alpha}}{c_3^2 d} \\
&\geq \min\left\{\frac{c_1^2}{2}, \frac{c_1}{6\sqrt{2}}\right\} \cdot \left(\frac{\sqrt{6}}{2}c_1\right)^{-2} \frac{n^{1-\alpha}}{d} \\
&= \frac{c_6 n^{1-\alpha}}{d}
\end{aligned}
$$

where $c_6 = \min\left\{\frac{1}{3}, \frac{1}{9\sqrt{2}c_1}\right\}$. Similarly, we also have

$$
\mathbb{E}_{\boldsymbol{z}}\left[\sum_{t=1}^{n'}\left(\frac{\sin\eta}{\sqrt{d-1}} + \tilde{x}_{m,t,\ell}\right)^2\right] + \mathbb{E}_{\boldsymbol{z}'}\left[\sum_{t=1}^{n'}\left(\frac{\sin\eta}{\sqrt{d-1}} - \tilde{x}_{m,t,\ell}\right)^2\right] \geq \frac{c_6 n^{1-\alpha}}{d}.
$$

These two give us

$$
\mathbb{E}_{\boldsymbol{z}}\left[\sum_{t=1}^{n'}\left(\frac{\sin\eta}{\sqrt{d-1}} - \tilde{x}_{m,t,\ell}\cdot z_\ell\right)^2\right] + \mathbb{E}_{\boldsymbol{z}'}\left[\sum_{t=1}^{n'}\left(\frac{\sin\eta}{\sqrt{d-1}} - \tilde{x}_{m,t,\ell}\cdot z'_\ell\right)^2\right] \geq \frac{c_6 n^{1-\alpha}}{d}.
$$

Plug this to (13) and (14), we have

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{n}\|\boldsymbol{\theta}_m\|_2 - \langle \boldsymbol{x}_{i,t}, \boldsymbol{\theta}_m\rangle\right] \\
&= \frac{1}{|\mathcal{Z}|}\sum_{\boldsymbol{z}\in\mathcal{Z}}\mathbb{E}\left[\sum_{t=1}^{n}\|\boldsymbol{\theta}^{(\boldsymbol{z})}\| - \langle \tilde{\boldsymbol{x}}_{i,t}, \tilde{\boldsymbol{\theta}}^{(\boldsymbol{z})}\rangle\right] \\
&\geq \frac{1}{|\mathcal{Z}|}\sum_{\boldsymbol{z}\in\mathcal{Z}} c_5 d n^{-\frac{1}{2}+\alpha}\sum_{\ell=2}^{d}\mathbb{E}\left[\sum_{t=1}^{n'}\left(\frac{\sin\eta}{\sqrt{d-1}} - \tilde{x}_{m,t,\ell}\cdot z_\ell\right)^2\Bigg|\mathcal{E}\right] \\
&= c_5 d n^{-\frac{1}{2}+\alpha}\sum_{\ell=2}^{d}\frac{1}{|\mathcal{Z}|}\sum_{\boldsymbol{z}\in\mathcal{Z}}\mathbb{E}\left[\sum_{t=1}^{n'}\left(\frac{\sin\eta}{\sqrt{d-1}} - \tilde{x}_{m,t,\ell}\cdot z_\ell\right)^2\Bigg|\mathcal{E}\right] \\
&= c_5 d n^{-\frac{1}{2}+\alpha}\sum_{\ell=2}^{d}\frac{1}{2|\mathcal{Z}|}\sum_{\boldsymbol{z}\in\mathcal{Z}}\mathbb{E}\left[\sum_{t=1}^{n'}\left(\frac{\sin\eta}{\sqrt{d-1}} - \tilde{x}_{m,t,\ell}\cdot z_\ell\right)^2\Bigg|\mathcal{E}\right] + \mathbb{E}\left[\sum_{t=1}^{n'}\left(\frac{\sin\eta}{\sqrt{d-1}} - \tilde{x}_{m,t,\ell}\cdot z'_\ell\right)^2\Bigg|\mathcal{E}\right] \\
&\geq c_5 d n^{-\frac{1}{2}+\alpha}\sum_{\ell=2}^{d}\frac{1}{2|\mathcal{Z}|}\sum_{\boldsymbol{z}\in\mathcal{Z}}\frac{c_6 n^{1-\alpha}}{d} \\
&= \frac{c_5 c_6}{2}(d-1)\sqrt{n} \\
&\geq c_7 d\sqrt{n},
\end{aligned}
$$

where $c_7 = (c_5 c_6)/4$.

The assumption we use is that

$$
d \leq n^{1-\alpha}\cdot\min\left\{\frac{1}{3}, \frac{1}{9\sqrt{2}c_1}\right\}
$$

A.1.2 Middle regime

For the second regime in Theorem 4.1, the proof is very similar to the proof for large σ. We just need an extra $m^{-\gamma}$ term and most parts of the proof stay the same. In the middle regime, instead of (7), we now have

$$
\sigma = c_1 d^{\frac{1}{2}} m^{-\gamma} n^{-\frac{1}{2}}
$$

for $\gamma \in \left[0, \frac{1}{2}\right]$ and some constant $c_1 > 0$. Then, we have

$$\eta = \sin^{-1}\left(\frac{c_4\sqrt{d(d-1)}m^{-\gamma}n^{-\frac{1}{2}}}{\|\boldsymbol{\theta}^{(0)}\|}\right) = \sin^{-1}\left(\frac{c_4}{c_3}\sqrt{\frac{d-1}{d}}\right) \in \left(0, \frac{\pi}{2}\right).$$

compared to (10). Then, in (14), we have the extra $m^{-\gamma}$ term in the front,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{n}\|\boldsymbol{\theta}^{(\boldsymbol{z})}\| - \langle\tilde{\boldsymbol{x}}_{i,t}, \tilde{\boldsymbol{\theta}}^{(\boldsymbol{z})}\rangle\middle|\boldsymbol{z}\right] \\
&\geq \mathbb{P}(\mathcal{E})\,\mathbb{E}\left[\sum_{t=1}^{n}\sum_{\ell=2}^{d}\frac{\|\boldsymbol{\theta}^{(\boldsymbol{z})}\|\sin^2\eta}{2}\left(\frac{1}{\sqrt{d-1}} - \frac{\tilde{x}_{m,t,\ell}\cdot z_\ell}{\sin\eta}\right)^2\middle|\boldsymbol{z}, \mathcal{E}\right] \\
&\geq c_2\mathbb{E}\left[\sum_{\ell=2}^{d}\frac{\|\boldsymbol{\theta}^{(\boldsymbol{z})}\|\sin^2\eta}{2}\sum_{t=1}^{n'}\left(\frac{1}{\sqrt{d-1}} - \frac{\tilde{x}_{m,t,\ell}\cdot z_\ell}{\sin\eta}\right)^2\middle|\boldsymbol{z}, \mathcal{E}\right] \\
&= c_2\mathbb{E}\left[\sum_{\ell=2}^{d}\frac{c_4^2 d(d-1)m^{-2\gamma}n^{-1+\alpha}}{2\|\boldsymbol{\theta}^{(0)}\|}\sum_{t=1}^{n'}\left(\frac{1}{\sqrt{d-1}} - \frac{\tilde{x}_{m,t,\ell}\cdot z_\ell}{\sin\eta}\right)^2\middle|\boldsymbol{z}, \mathcal{E}\right] \\
&= c_2\mathbb{E}\left[\sum_{\ell=2}^{d}\frac{c_4^2 d(d-1)m^{-2\gamma}n^{-1+\alpha}}{2c_3 d m^{-\gamma}n^{-\frac{1}{2}+\alpha}}\sum_{t=1}^{n'}\left(\frac{1}{\sqrt{d-1}} - \frac{\tilde{x}_{m,t,\ell}\cdot z_\ell}{\sin\eta}\right)^2\middle|\boldsymbol{z}, \mathcal{E}\right] \\
&= \frac{c_2(d-1)}{2m^\gamma\sqrt{n}}\sum_{\ell=2}^{d}\mathbb{E}\left[\frac{c_4^2}{c_3}\sum_{t=1}^{n'}\left(\frac{1}{\sqrt{d-1}} - \frac{\tilde{x}_{m,t,\ell}\cdot z_\ell}{\sin\eta}\right)^2\middle|\boldsymbol{z}, \mathcal{E}\right] \\
&= \frac{c_2(d-1)}{2m^\gamma\sqrt{n}}\sum_{\ell=2}^{d}\mathbb{E}\left[\frac{c_4^2}{c_3\sin^2\eta}\sum_{t=1}^{n'}\left(\frac{\sin\eta}{\sqrt{d-1}} - \tilde{x}_{m,t,\ell}\cdot z_\ell\right)^2\middle|\boldsymbol{z}, \mathcal{E}\right] \\
&= \frac{c_2(d-1)}{2m^\gamma\sqrt{n}}\sum_{\ell=2}^{d}\mathbb{E}\left[\frac{c_3 d}{d-1}\sum_{t=1}^{n'}\left(\frac{\sin\eta}{\sqrt{d-1}} - \tilde{x}_{m,t,\ell}\cdot z_\ell\right)^2\middle|\boldsymbol{z}, \mathcal{E}\right] \\
&\geq \frac{c_2 d n^{-\frac{1}{2}}}{2m^\gamma}\sum_{\ell=2}^{d}\mathbb{E}\left[\frac{\sqrt{2}c_1}{2}\sum_{t=1}^{n'}\left(\frac{\sin\eta}{\sqrt{d-1}} - \tilde{x}_{m,t,\ell}\cdot z_\ell\right)^2\middle|\boldsymbol{z}, \mathcal{E}\right] \\
&= c_5 d m^{-\gamma}n^{-\frac{1}{2}}\sum_{\ell=2}^{d}\mathbb{E}\left[\sum_{t=1}^{n'}\left(\frac{\sin\eta}{\sqrt{d-1}} - \tilde{x}_{m,t,\ell}\cdot z_\ell\right)^2\middle|\boldsymbol{z}, \mathcal{E}\right].
\end{aligned}$$

The rest of the proof are the same. Due to the $m^{-\gamma}$ term in the front, we have

$$\mathbb{E}\left[\sum_{i=1}^{m}\sum_{t=1}^{n}\|\boldsymbol{\theta}_m\|_2 - \langle\boldsymbol{x}_{i,t}, \boldsymbol{\theta}_m\rangle\right] = \Omega\left(dm^{1-\gamma}\sqrt{n}\right).$$

A.2 Regret lower bound for the first regime

Equivalently we have

$$\sigma = o\left(\sqrt{\frac{d}{mn}}\right).$$

Let Π be the set of all policies

$$\Pi = \left\{\pi : \pi = (\pi_{i,t})_{i=1:m,t=1:n}, \pi_{i,t} : \{(\boldsymbol{x}_{i,\tau}, y_{i,\tau})\}_{i=1:m,\tau=1:t-1} \mapsto \boldsymbol{x}_t\right\}$$

and

$$\boldsymbol{x}_i^* = \arg\max_{\boldsymbol{x}\in\mathcal{A}} \langle \boldsymbol{x}, \boldsymbol{\theta}_i \rangle,$$
$$\boldsymbol{x}^* = \arg\max_{\boldsymbol{x}\in\mathcal{A}} \langle \boldsymbol{x}, \boldsymbol{\mu} \rangle.$$

Then,

$$\begin{aligned}
R_{m,n}(\boldsymbol{\mu}) &= \min_{\pi\in\Pi} \mathbb{E}\left[\sum_{i=1}^m \sum_{t=1}^n \langle \boldsymbol{x}_i^* - \boldsymbol{x}_{i,t}, \boldsymbol{\theta}_i \rangle\right] \\
&\geq \min_{\pi\in\Pi} \mathbb{E}\left[\sum_{i=1}^m \sum_{t=1}^n \langle \boldsymbol{x}^* - \boldsymbol{x}_{i,t}, \boldsymbol{\theta}_i \rangle\right] \\
&= \min_{\pi\in\Pi} \mathbb{E}\left[\sum_{i=1}^m \sum_{t=1}^n \langle \boldsymbol{x}^* - \boldsymbol{x}_{i,t}, \boldsymbol{\mu} \rangle + \sum_{i=1}^m \sum_{t=1}^n \langle \boldsymbol{x}^* - \boldsymbol{x}_{i,t}, \boldsymbol{\theta}_i - \boldsymbol{\mu} \rangle\right] \\
&\geq \min_{\pi\in\Pi} \mathbb{E}\left[\sum_{i=1}^m \sum_{t=1}^n \langle \boldsymbol{x}^* - \boldsymbol{x}_{i,t}, \boldsymbol{\mu} \rangle\right] + \min_{\pi\in\Pi} \mathbb{E}\left[\sum_{i=1}^m \sum_{t=1}^n \langle \boldsymbol{x}^* - \boldsymbol{x}_{i,t}, \boldsymbol{\theta}_i - \boldsymbol{\mu} \rangle\right] \\
&\geq \min_{\pi\in\Pi} \mathbb{E}\left[\sum_{i=1}^m \sum_{t=1}^n \langle \boldsymbol{x}^* - \boldsymbol{x}_{i,t}, \boldsymbol{\mu} \rangle\right] - \left|\min_{\pi\in\Pi} \mathbb{E}\left[\sum_{i=1}^m \sum_{t=1}^n \langle \boldsymbol{x}^* - \boldsymbol{x}_{i,t}, \boldsymbol{\theta}_i - \boldsymbol{\mu} \rangle\right]\right|. \qquad (15)
\end{aligned}$$

Since $\sigma = o\left(\sqrt{d/mn}\right)$, we have

$$\begin{aligned}
\left|\mathbb{E}\left[\sum_{i=1}^m \sum_{t=1}^n \langle \boldsymbol{x}^* - \boldsymbol{x}_{i,t}, \boldsymbol{\theta}_i - \boldsymbol{\mu} \rangle\right]\right| &\leq \mathbb{E}\left[\sum_{i=1}^m \sum_{t=1}^n \|\boldsymbol{x}^* - \boldsymbol{x}_{i,t}\|_2 \|\boldsymbol{\theta}_i - \boldsymbol{\mu}\|_2\right] \\
&\leq 2mn\mathbb{E}\left[\|\boldsymbol{\theta}_i - \boldsymbol{\mu}\|_2\right] \\
&\leq 2mn\sqrt{\mathbb{E}\left[\|\boldsymbol{\theta}_i - \boldsymbol{\mu}\|_2^2\right]} \\
&= 2mn\sqrt{d\sigma^2} \\
&= o\left(d\sqrt{mn}\right). \qquad (16)
\end{aligned}$$

Now, let Π' be the set of policies that knows $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$, that is,

$$\Pi = \left\{\pi : \pi = (\pi_{i,t})_{i=1:m,t=1:n}, \pi_{i,t} : (\{\boldsymbol{u}_i\}_{i=1:m}, \{(\boldsymbol{x}_{i,\tau}, y_{i,\tau})\}_{i=1:m,\tau=1:t-1}) \mapsto \boldsymbol{x}_t\right\}.$$

Since $\Pi \subset \Pi'$, we have

$$R_{m,n}(\boldsymbol{\mu}) \geq \min_{\pi\in\Pi'} \mathbb{E}\left[\sum_{i=1}^m \sum_{t=1}^n \langle \boldsymbol{x}^* - \boldsymbol{x}_{i,t}, \boldsymbol{\mu} \rangle\right] - 2mn\sqrt{d\sigma^2}.$$

Since $\pi \in \Pi'$ knows $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$, equivalently it knows

$$y'_{i,t} = \langle \boldsymbol{x}_{i,t}, \boldsymbol{\mu} \rangle + \epsilon_{i,t} = \langle \boldsymbol{x}_{i,t}, \boldsymbol{\theta}_i \rangle + \langle \boldsymbol{x}_{i,t}, \boldsymbol{\mu} - \boldsymbol{\theta}_i \rangle + \epsilon_{i,t} = y_{i,t} - \langle \boldsymbol{x}_{i,t}, \boldsymbol{u}_i \rangle.$$

Thus, we must have

$$\min_{\pi\in\Pi'} \mathbb{E}\left[\sum_{i=1}^m \sum_{t=1}^n \langle \boldsymbol{x}^* - \boldsymbol{x}_{i,t}, \boldsymbol{\mu} \rangle\right]$$

obeying the lower bound for the conventional bandits problem where there are mn samples and $\boldsymbol{\mu}$ is the true parameter. From Section 24.2 in Lattimore and Szepesvári (2020), we have that

$$\min_{\pi\in\Pi'} \mathbb{E}\left[\sum_{i=1}^m \sum_{t=1}^n \langle \boldsymbol{x}^* - \boldsymbol{x}_{i,t}, \boldsymbol{\mu} \rangle\right] \geq \frac{1}{32} d\sqrt{mn} \qquad (17)$$

for some carefully chosen $\boldsymbol{\mu} \in \mathcal{B}_2^d$. Plug (17) and (16) back to (15) and we have

$$R_{m,n}(\boldsymbol{\mu}) = \Omega\left(d\sqrt{mn}\right).$$

Remark A.2 (Assumptions made in Theorem 4.1)**.** *In Theorem 4.1, we made minor assumptions on the relation between the total number of rounds, n, and the dimension, d. This naturally follows the assumptions on $\boldsymbol{\mu}$ and σ, which are discussed in remark 3.3. In the first regime, we assume that $d \leq mn$. This corresponds to the classic minimax bound result Lattimore and Szepesvári (2020) (Theorem 24.4), where we assume that the dimension is smaller than the number of samples. In the second regimes, we the assumptions follow the identical order as in the classic result that $d \leq n$. In the third regime, we assume that $n = \Omega(d^{-(1-\gamma)})$. This naturally follows the assumption we made on $\boldsymbol{\mu}$ and σ being at most constant. In this case, we must have $d = O(n^{1-\gamma})$ in order to get $\sigma = O(1)$. The constants is the assumption are adjusted to deal with the technical proofs.*

B Collaborative Personalized Phased Elimination and Regret Analysis

B.1 Details for Algorithm 1

We provide the details of Algorithm 1 and the Phased Elimination block

Algorithm 3 Phased Elimination($\mathcal{A}_{i,\ell}$, ε_ℓ)

1: Find G-optimal design $\boldsymbol{\pi}_{i,\ell} \in \mathcal{P}(\mathcal{A}_{i,\ell})$ with $\mathrm{Supp}(\boldsymbol{\pi}_{i,\ell}) \leq d(d+1)/2$
2: Select each action $\boldsymbol{x} \in \mathcal{A}_{i,\ell}$ exactly $n_{i,\ell}(x)$ times as defined in (4).
3: Record the actions $\mathbf{X}_i \in \mathbb{R}^{\bar{n}_{i,\ell} \times d}$ and rewards $\boldsymbol{y}_i \in \mathbb{R}^{\bar{n}_{i,\ell}}$.
4: Calculate $\hat{\boldsymbol{\theta}}_{i,\ell} \leftarrow (\mathbf{X}_i^\top \mathbf{X}_i)^\dagger \mathbf{X}_i^\top \boldsymbol{y}_i$ and update

$$\mathcal{A}_{i,\ell+1} \leftarrow \left\{ \boldsymbol{x} \in \mathcal{A}_{i,\ell} : \max_{\boldsymbol{x}' \in \mathcal{A}_{i,\ell}} \langle \hat{\boldsymbol{\theta}}_{i,\ell}, \boldsymbol{x}' - \boldsymbol{x} \rangle \leq 2\varepsilon_\ell \right\}$$

5: **return** $\mathcal{A}_{i,\ell+1}$

B.2 Regret analysis of Algorithm 1 for finite action sets

Theorem B.1. *Consider any finite action set $\mathcal{A}$ with $|\mathcal{A}| = k$, $k \in \mathbb{N}$, and assume that $\|\boldsymbol{\mu}\|_2 \leq 1$, $\sigma \leq 1$, and $n \geq \min\{k, d^2\}$. For any $\delta \in (0,1)$, by running Algorithm 1, we have the following results:*

- *When $n = o\left(d/m\sigma^2\right)$, we have $\hat{R}_{m,n}(\boldsymbol{\mu}) = \tilde{O}\left(\sqrt{dmn} + md^2\right)$.*
- *When $n = \Theta\left(d/m^{2\gamma}\sigma^2\right)$ with $\gamma \in [0, \frac{1}{2}]$, we have $\hat{R}_{m,n}(\boldsymbol{\mu}) = \tilde{O}\left(m^{1-\gamma}\sqrt{dn} + md^2\right)$.*
- *When $n = \omega\left(d/\sigma^2\right)$, we have $\hat{R}_{m,n}(\boldsymbol{\mu}) = \tilde{O}\left(m\sqrt{dn} + md^2\right)$.*

with probability at least $1 - \delta$.

B.3 Proof of Theorem B.1

For notational convenience, denote by $\hat{\boldsymbol{\theta}}_{1,\ell} = \cdots = \hat{\boldsymbol{\theta}}_{1,\ell} = \hat{\boldsymbol{\mu}}_\ell$ during the collaborative learning stage, i.e., when $h_{\sigma,\delta} \leq \frac{1}{2}\varepsilon_\ell$.

Lemma B.2. *Following Algorithm 1, we have*

$$\mathbb{P}\left(\forall \ell \in \mathbb{N},\ \forall i \in [m],\ \forall \boldsymbol{x} \in \mathcal{A}_{i,\ell} : |\langle \hat{\boldsymbol{\theta}}_{i,\ell} - \boldsymbol{\theta}_i, \boldsymbol{x} \rangle| \leq \varepsilon_\ell\right) \geq 1 - \delta.$$

Proof of Lemma B.2. Suppose that the algorithm has $L_c \in \mathbb{N}$ collaboration phases. That is,

$$L_c = \max\left\{0, \left\lfloor \log_2 \frac{1}{2h_{\sigma,\delta}} \right\rfloor\right\}.$$

In the collaboration phases, for any $\ell \in [L_c]$ and action $\boldsymbol{x} \in \mathcal{A}_\ell$, we have

$$\langle \hat{\boldsymbol{\mu}}_\ell - \boldsymbol{\theta}_i, \boldsymbol{x} \rangle = \langle \hat{\boldsymbol{\mu}}_\ell - \boldsymbol{\mu}, \boldsymbol{x} \rangle + \langle \boldsymbol{\mu} - \boldsymbol{\theta}_i, \boldsymbol{x} \rangle. \tag{18}$$

Let $\mathbf{X}_i$ be a $\bar{n}_{i,\ell}$ by d matrix such that each row of $\mathbf{X}_i$ is the actions chosen by user i at phase ℓ, and $\boldsymbol{y}_i$ is the vector with corresponding rewards. For simplicity, we omit the index ℓ in the notation. For the first term in (18), we have

$$\begin{aligned}
\langle \hat{\boldsymbol{\mu}}_\ell - \boldsymbol{\mu}, \boldsymbol{x} \rangle &= \left\langle \frac{1}{m} \sum_{i=1}^m \left(\mathbf{X}_i^\top \mathbf{X}_i\right)^\dagger \mathbf{X}_i^\top \boldsymbol{y}_i - \boldsymbol{\mu}, \boldsymbol{x} \right\rangle \\
&= \left\langle \frac{1}{m} \sum_{i=1}^m \left(\mathbf{X}_i^\top \mathbf{X}_i\right)^\dagger \mathbf{X}_i^\top \left(\mathbf{X}_i(\boldsymbol{\mu} + \boldsymbol{u}_i) + \boldsymbol{\epsilon}_i\right) - \boldsymbol{\mu}, \boldsymbol{x} \right\rangle \\
&= \left\langle \frac{1}{m} \sum_{i=1}^m \left(\mathbf{X}_i^\top \mathbf{X}_i\right)^\dagger \mathbf{X}_i^\top \mathbf{X}_i \boldsymbol{\mu} - \boldsymbol{\mu}, \boldsymbol{x} \right\rangle + \left\langle \frac{1}{m} \sum_{i=1}^m \left(\mathbf{X}_i^\top \mathbf{X}_i\right)^\dagger \mathbf{X}_i^\top \mathbf{X}_i \boldsymbol{u}_i, \boldsymbol{x} \right\rangle && (19) \\
&\quad + \left\langle \frac{1}{m} \sum_{i=1}^m \left(\mathbf{X}_i^\top \mathbf{X}_i\right)^\dagger \mathbf{X}_i^\top \boldsymbol{\epsilon}_i, \boldsymbol{x} \right\rangle && (20)
\end{aligned}$$

For the first term in (20), we have

$$\left\langle \left(\mathbf{X}_i^\top \mathbf{X}_i\right)^\dagger \mathbf{X}_i^\top \mathbf{X}_i \boldsymbol{\mu}, \boldsymbol{x} \right\rangle = \left\langle \boldsymbol{\mu}, \mathbf{X}_i^\top \mathbf{X}_i \left(\mathbf{X}_i^\top \mathbf{X}_i\right)^\dagger \boldsymbol{x} \right\rangle = \langle \boldsymbol{\mu}, \boldsymbol{x} \rangle$$

and thus

$$\left\langle \frac{1}{m} \sum_{i=1}^m \left(\mathbf{X}_i^\top \mathbf{X}_i\right)^\dagger \mathbf{X}_i^\top \mathbf{X}_i \boldsymbol{\mu} - \boldsymbol{\mu}, \boldsymbol{x} \right\rangle = 0.$$

For the second term in (20), we have

$$\left\langle \frac{1}{m} \sum_{i=1}^m \left(\mathbf{X}_i^\top \mathbf{X}_i\right)^\dagger \mathbf{X}_i^\top \mathbf{X}_i \boldsymbol{u}_i, \boldsymbol{x} \right\rangle = \frac{1}{m} \sum_{i=1}^m \left\langle \boldsymbol{u}_i, \mathbf{X}_i^\top \mathbf{X}_i \left(\mathbf{X}_i^\top \mathbf{X}_i\right)^\dagger \boldsymbol{x} \right\rangle = \frac{1}{m} \sum_{i=1}^m \langle \boldsymbol{u}_i, \boldsymbol{x} \rangle.$$

Since each $\boldsymbol{u}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$, we have

$$\left\langle \frac{1}{m} \sum_{i=1}^m \boldsymbol{u}_i, \boldsymbol{x} \right\rangle \sim \mathcal{N}\left(0, \frac{1}{m} \boldsymbol{x}^\top \mathbf{C} \boldsymbol{x}\right), \tag{21}$$

and thus

$$\mathbb{P}\left(\left\langle \frac{1}{m} \sum_{i=1}^m \left(\mathbf{X}_i^\top \mathbf{X}_i\right)^\dagger \mathbf{X}_i^\top \mathbf{X}_i \boldsymbol{u}_i, \boldsymbol{x} \right\rangle \geq s \right) \leq \frac{1}{2} \exp\left\{ -\frac{ms^2}{2\boldsymbol{x}^\top \mathbf{C} \boldsymbol{x}} \right\} \tag{22}$$

for any $s \geq 0$. Thus, we have

$$\mathbb{P}\left(\exists \ell \in [L_c],\ \exists \boldsymbol{x} \in \mathcal{A}_\ell : \left| \left\langle \frac{1}{m} \sum_{i=1}^m \left(\mathbf{X}_i^\top \mathbf{X}_i\right)^\dagger \mathbf{X}_i^\top \mathbf{X}_i \boldsymbol{u}_i, \boldsymbol{x} \right\rangle \right| \geq \sqrt{\frac{2}{m} \left(\max_{\boldsymbol{x} \in \mathcal{A}} \|\boldsymbol{x}\|_{\mathbf{C}}^2 \right) \log \frac{4k}{\delta}} \right) \leq \frac{\delta}{4} \tag{23}$$

for any $\delta > 0$. For the third term in (20), since $\mathbf{X}_i$ and $\boldsymbol{\epsilon}_i$ are independent, we have

$$\frac{1}{m} \sum_{i=1}^m \left(\mathbf{X}_i^\top \mathbf{X}_i\right)^\dagger \mathbf{X}_i^\top \boldsymbol{\epsilon}_i \sim \mathcal{N}\left(0, \frac{1}{m^2} \sum_{i=1}^m \left(\mathbf{X}_i^\top \mathbf{X}_i\right)^\dagger \right).$$

Then, we have

$$\mathbb{P}\left(\exists \boldsymbol{x} \in \mathcal{A}_\ell : \left| \left\langle \frac{1}{m} \sum_{i=1}^m \left(\mathbf{X}_i^\top \mathbf{X}_i\right)^\dagger \mathbf{X}_i^\top \boldsymbol{\epsilon}_i, \boldsymbol{x} \right\rangle \right| \geq \sqrt{\frac{2\sigma^2}{m^2} \|\boldsymbol{x}\|^2_{\sum_{i=1}^m \left(\mathbf{X}_i^\top \mathbf{X}_i\right)^\dagger} \log \frac{2k\ell(\ell+1)}{\delta}} \right) \leq \frac{\delta}{2\ell(\ell+1)}$$

for any $\delta > 0$. From the algorithm, we choose the actions $\mathbf{X}_i$ such that

$$\begin{aligned}
&\frac{2}{m^2}\|\boldsymbol{x}\|^2_{\sum_{i=1}^m (\mathbf{X}_i^\top \mathbf{X}_i)^\dagger} \log \frac{2k\ell(\ell+1)}{\delta} \\
&\le \frac{2}{m^2}\boldsymbol{x}^\top \sum_{i=1}^m \left(\left\lceil \frac{8g(\pi_\ell)}{m\varepsilon_\ell^2} \log \frac{2k\ell(\ell+1)}{\delta} \right\rceil \sum_{\boldsymbol{x}' \in \mathcal{A}_{i,\ell}} \pi_\ell(\boldsymbol{x}') \boldsymbol{x}'^\top \boldsymbol{x}' \right)^\dagger \boldsymbol{x} \log \frac{2k\ell(\ell+1)}{\delta} \\
&\le \frac{\varepsilon_\ell^2}{4}.
\end{aligned}$$

Thus, we have

$$\forall \ell \in [L_c]: \quad \mathbb{P}\left(\exists \boldsymbol{x} \in \mathcal{A}_\ell : \left| \left\langle \frac{1}{m} \sum_{i=1}^m \left(\mathbf{X}_i^\top \mathbf{X}_i\right)^\dagger \mathbf{X}_i^\top \boldsymbol{\epsilon}_i, \boldsymbol{x} \right\rangle \right| \ge \frac{\boldsymbol{\epsilon}_\ell}{2} \right) \le \frac{\delta}{2\ell(\ell+1)}. \tag{24}$$

For the second term in (18), we have

$$\langle \boldsymbol{\mu} - \boldsymbol{\theta}_i, \boldsymbol{x} \rangle = \boldsymbol{x}^\top \boldsymbol{u}_i \sim \mathcal{N}(0, \boldsymbol{x}^\top \mathbf{C} \boldsymbol{x}).$$

Thus, for any $s \ge 0$,

$$\mathbb{P}\left(\langle \boldsymbol{\mu} - \boldsymbol{\theta}_i, \boldsymbol{x} \rangle \ge s \right) \le \frac{1}{2} \exp\left\{ -\frac{s^2}{2\boldsymbol{x}^\top \mathbf{C} \boldsymbol{x}} \right\}. \tag{25}$$

Therefore, we have

$$\mathbb{P}\left(\exists \ell \in [L_c],\ \exists i \in [m],\ \exists \boldsymbol{x} \in \mathcal{A}_\ell : |\langle \boldsymbol{\mu} - \boldsymbol{\theta}_i, \boldsymbol{x} \rangle| \ge \sqrt{2 \left(\max_{\boldsymbol{x} \in \mathcal{A}} \|\boldsymbol{x}\|^2_{\mathbf{C}} \right) \log \frac{4km}{\delta}} \right) \le \frac{\delta}{4} \tag{26}$$

for any $\delta > 0$. In the collaborations, we know that

$$h_{\sigma,\delta} = \sqrt{\frac{2}{m} \left(\max_{\boldsymbol{x} \in \mathcal{A}} \|\boldsymbol{x}\|^2_{\mathbf{C}} \right) \log \frac{4k}{\delta}} + \sqrt{2 \left(\max_{\boldsymbol{x} \in \mathcal{A}} \|\boldsymbol{x}\|^2_{\mathbf{C}} \right) \log \frac{4k}{\delta}} \le \frac{\boldsymbol{\epsilon}_\ell}{2}.$$

Plug (23), (24) and (26) back to (20) and we have

$$\mathbb{P}\left(\exists \ell \in [L_c],\ \exists i \in [m],\ \exists \boldsymbol{x} \in \mathcal{A}_\ell : \langle \hat{\boldsymbol{\mu}}_\ell - \boldsymbol{\theta}_i, \boldsymbol{x} \rangle \ge \boldsymbol{\epsilon}_\ell \right) \le \left(1 - \frac{1}{L_c + 1} \right) \delta. \tag{27}$$

Now, for the local training phases, we know that for any $\boldsymbol{x} \in \mathcal{A}_{i,\ell}$

$$\hat{\boldsymbol{\theta}}_{i,\ell} - \boldsymbol{\theta}_i = \left(\mathbf{X}_i^\top \mathbf{X}_i\right)^\dagger \mathbf{X}_i^\top \boldsymbol{\epsilon}_i \sim \mathcal{N}\left(0, \left(\mathbf{X}_i^\top \mathbf{X}_i\right)^\dagger\right)$$

Therefore, we have

$$\mathbb{P}\left(\exists i \in [m],\ \exists \boldsymbol{x} \in \mathcal{A}_{i,\ell} : \left| \langle \hat{\boldsymbol{\theta}}_{i,\ell} - \boldsymbol{\theta}_i, \boldsymbol{x} \rangle \right| \ge \sqrt{2\|\boldsymbol{x}\|^2_{(\mathbf{X}_i^\top \mathbf{X}_i)^\dagger} \log \frac{2km\ell(\ell+1)}{\delta}} \right) \le \frac{\delta}{2\ell(\ell+1)}$$

for any $\delta > 0$. From the algorithm, we choose $\mathbf{X}_i$'s such that

$$\begin{aligned}
&2\|\boldsymbol{x}\|^2_{(\mathbf{X}_i^\top \mathbf{X}_i)^\dagger} \log \frac{2km\ell(\ell+1)}{\delta} \\
&\le 2\boldsymbol{x}^\top \left(\left\lceil \frac{2g(\pi_{i,\ell})}{\varepsilon_\ell^2} \log \frac{2km\ell(\ell+1)}{\delta} \right\rceil \sum_{\boldsymbol{x}' \in \mathcal{A}_{i,\ell}} \pi_\ell(\boldsymbol{x}') \boldsymbol{x}'^\top \boldsymbol{x}' \right)^{-1} \boldsymbol{x} \log \frac{2km\ell(\ell+1)}{\delta} \\
&\le \boldsymbol{\epsilon}_\ell^2.
\end{aligned}$$

Thus, we have

$$\mathbb{P}\left(\exists \ell > L_c,\ \exists i \in [m],\ \exists \boldsymbol{x} \in \mathcal{A}_{i,\ell} : \langle \hat{\boldsymbol{\theta}}_{i,\ell} - \boldsymbol{\theta}_i, \boldsymbol{x} \rangle \geq \boldsymbol{\epsilon}_\ell\right) \leq \left(\frac{1}{L_c + 1}\right)\delta. \tag{28}$$

With (27) and (28), we have that

$$\mathbb{P}\left(\exists \ell \in \mathbb{N},\ \exists i \in [m],\ \exists \boldsymbol{x} \in \mathcal{A}_{i,\ell} : \langle \hat{\boldsymbol{\theta}}_{i,\ell} - \boldsymbol{\theta}_i, \boldsymbol{x} \rangle \geq \boldsymbol{\epsilon}_\ell\right) \leq \delta.$$

□

Lemma B.3. *Let $\boldsymbol{x}_i^* = \max_{\boldsymbol{x} \in \mathcal{A}} \langle \boldsymbol{\theta}_i, \boldsymbol{x} \rangle$. Then, under the good event in Lemma B.2 we have*

$$\forall \ell \in \mathbb{N} : \boldsymbol{x}_i^* \notin \mathcal{A}_{i,\ell}.$$

Proof of Lemma B.3. For any $\ell \in \mathbb{N},\ i \in [m]$, suppose that $\boldsymbol{x}_i^* \in \mathcal{A}_{i,\ell}$. Then, for any $\boldsymbol{x} \in \mathcal{A}_{i,\ell}$,

$$\begin{aligned}
\langle \hat{\boldsymbol{\theta}}_{i,\ell}, \boldsymbol{x} - \boldsymbol{x}_i^* \rangle &= \langle \hat{\boldsymbol{\theta}}_{i,\ell} - \boldsymbol{\theta}_i, \boldsymbol{x} \rangle - \langle \hat{\boldsymbol{\mu}}_\ell - \boldsymbol{\theta}_i, \boldsymbol{x}_i^* \rangle - \langle \theta_i, \boldsymbol{x}_i^* - \boldsymbol{x} \rangle \\
&\leq \langle \hat{\boldsymbol{\theta}}_{i,\ell} - \boldsymbol{\theta}_i, \boldsymbol{x} \rangle - \langle \hat{\boldsymbol{\theta}}_{i,\ell} - \boldsymbol{\theta}_i, \boldsymbol{x}_i^* \rangle \\
&\leq 2\varepsilon_\ell
\end{aligned}$$

from Lemma B.2. Since $\boldsymbol{x}_i^* \in \mathcal{A}_1 = \mathcal{A}$, the lemma is proved. □

Lemma B.4. *Let $\ell_{i,\boldsymbol{x}} = \min\{\ell \in \mathbb{N} : 4\varepsilon_\ell < \Delta_{i,\boldsymbol{x}}\}$ for any $\boldsymbol{x} \in \mathcal{A}$, where $\Delta_{i,\boldsymbol{x}} = \langle \boldsymbol{x}_i^* - \boldsymbol{x}, \boldsymbol{\theta}_i \rangle$. Then, under the good event in Lemma B.2 we have*

$$\forall \boldsymbol{x} \in \mathcal{A} : \boldsymbol{x} \notin \mathcal{A}_{i,\ell_{i,\boldsymbol{x}}+1}.$$

Proof of Lemma B.4. The proof is straightforward from Algorithm 1. □

With Lemma B.2, B.3, and B.4, we can focus on the good event in Lemma B.2 and it's corresponding regret. We have

$$\begin{aligned}
\hat{R}_{m,n}(\boldsymbol{\mu}) &= \sum_{t=1}^{n}\sum_{i=1}^{m} \langle \boldsymbol{\theta}_i, \boldsymbol{x}_i^* \rangle - \langle \boldsymbol{\theta}_i, \boldsymbol{x}_{i,t} \rangle \\
&= \sum_{i=1}^{m}\sum_{\boldsymbol{x} \in \mathcal{A}} \Delta_{i,\boldsymbol{x}} n_{i,\boldsymbol{x}} \\
&= \sum_{i=1}^{m}\left(\sum_{\boldsymbol{x}:\Delta_{i,\boldsymbol{x}} \leq \eta} \Delta_{i,\boldsymbol{x}} n_{i,\boldsymbol{x}} + \sum_{\boldsymbol{x}:\Delta_{i,\boldsymbol{x}} > \eta} \Delta_{i,\boldsymbol{x}} n_{i,\boldsymbol{x}}\right)
\end{aligned}$$

For any $\boldsymbol{x} \in \mathcal{A}$ such that $\Delta_{i,\boldsymbol{x}} > \eta$, we have

$$\ell_{i,\boldsymbol{x}} \leq \left\lfloor \log \frac{4}{\Delta_{i,\boldsymbol{x}}} \right\rfloor + 1 < \left\lfloor \log \frac{4}{\eta} \right\rfloor + 1.$$

Thus,

$$\begin{aligned}
\hat{R}_{m,n}(\boldsymbol{\mu}) &\leq mn\eta + \sum_{i=1}^{m}\sum_{\boldsymbol{x}:\Delta_{i,\boldsymbol{x}} > \eta}\sum_{\ell=1}^{\lceil \log_2 \frac{4}{\eta} \rceil} \Delta_{i,\boldsymbol{x}} n_{i,\ell}(\boldsymbol{x}) \\
&= mn\eta + \sum_{i=1}^{m}\sum_{\boldsymbol{x}:\Delta_{i,\boldsymbol{x}} > \eta}\left(\sum_{\ell=1}^{L_c'} \Delta_{i,\boldsymbol{x}} n_{i,\ell}(\boldsymbol{x}) + \sum_{\ell=L_c'+1}^{\lceil \log_2 \frac{4}{\eta} \rceil} \Delta_{i,\boldsymbol{x}} n_{i,\ell}(\boldsymbol{x})\right),
\end{aligned}$$

where $L_c' = \min\left\{L_c, \left\lceil \log \frac{4}{\eta} \right\rceil\right\}$ is the number of collaboration phases.

Suppose that we have $L_c \geq 1$. Let r_1 denote the regret in the first collaboration phase. From (26), we know that

$$|\langle \boldsymbol{\mu} - \boldsymbol{\theta}_i, \boldsymbol{x} \rangle| \leq \sqrt{2\left(\max_{\boldsymbol{x}\in\mathcal{A}} \|\boldsymbol{x}\|_{\mathbf{C}}^2\right) \log \frac{4km}{\delta}}.$$

Thus, in the first phase we have

$$\begin{aligned}
&\sum_{i=1}^{m} \sum_{\boldsymbol{x}:\Delta_{i,\boldsymbol{x}}>\eta} \Delta_{i,\boldsymbol{x}} n_{i,1}(\boldsymbol{x}) \\
&\leq \sum_{i=1}^{m} \sum_{\boldsymbol{x}:\Delta_{i,\boldsymbol{x}}>\eta} \left(2\max_{\boldsymbol{x}\in\mathcal{A}} \langle \boldsymbol{\mu}, \boldsymbol{x}\rangle + 2\sqrt{2\left(\max_{\boldsymbol{x}\in\mathcal{A}} \|\boldsymbol{x}\|_{\mathbf{C}}^2\right) \log \frac{4km}{\delta}}\right) \left\lceil \frac{8\pi_\ell(\boldsymbol{x}) g(\pi_\ell)}{m(1/2)^2} \log \frac{2k\ell(\ell+1)}{\delta} \right\rceil \\
&\leq \sum_{i=1}^{m} \sum_{\boldsymbol{x}:\Delta_{i,\boldsymbol{x}}>\eta} \left(2\max_{\boldsymbol{x}\in\mathcal{A}} \langle \boldsymbol{\mu}, \boldsymbol{x}\rangle + 2\sqrt{2\left(\max_{\boldsymbol{x}\in\mathcal{A}} \|\boldsymbol{x}\|_{\mathbf{C}}^2\right) \log \frac{4km}{\delta}}\right) \left(\frac{32\pi_\ell(\boldsymbol{x}) d}{m} \log \frac{2k\ell(\ell+1)}{\delta} + 1\right) \\
&\leq \left(2\max_{\boldsymbol{x}\in\mathcal{A}} \langle \boldsymbol{\mu}, \boldsymbol{x}\rangle + 2\sqrt{2\left(\max_{\boldsymbol{x}\in\mathcal{A}} \|\boldsymbol{x}\|_{\mathbf{C}}^2\right) \log \frac{4km}{\delta}}\right) \left(md(d+1) + 32d \log \frac{2k\ell(\ell+1)}{\delta}\right).
\end{aligned}$$

Let us denote this by

$$\begin{aligned}
r_1 &= \left(2\max_{\boldsymbol{x}\in\mathcal{A}} \langle \boldsymbol{\mu}, \boldsymbol{x}\rangle + 2\sqrt{2\left(\max_{\boldsymbol{x}\in\mathcal{A}} \|\boldsymbol{x}\|_{\mathbf{C}}^2\right) \log \frac{4km}{\delta}}\right) \left(md(d+1) + 32d \log \frac{2k\ell(\ell+1)}{\delta}\right) \\
&= O\left(md^2\right).
\end{aligned}$$

Notice that r_1 does not depend on n. For the rest of the collaboration phases, we have

$$\begin{aligned}
\sum_{i=1}^{m} \sum_{\boldsymbol{x}:\Delta_{i,\boldsymbol{x}}>\eta} \sum_{\ell=2}^{L_c'} \Delta_{i,\boldsymbol{x}} n_{i,\ell}(\boldsymbol{x}) &\leq \sum_{i=1}^{m} \sum_{\ell=2}^{L_c'} \sum_{\boldsymbol{x}:\Delta_{i,\boldsymbol{x}}>\eta} 2\varepsilon_\ell \left\lceil \frac{8\pi_\ell(\boldsymbol{x}) g(\pi_\ell)}{m\varepsilon_\ell^2} \log \frac{2k\ell(\ell+1)}{\delta} \right\rceil \\
&\leq m \sum_{\ell=2}^{L_c'} 2\varepsilon_\ell \left(|\mathrm{Supp}(\pi_\ell)| + \frac{8d}{m\varepsilon_\ell^2} \log \frac{2k\ell(\ell+1)}{\delta}\right) \\
&\leq m \sum_{\ell=2}^{L_c'} 2\cdot 2^{-\ell} \left(\frac{d(d+1)}{2} + \frac{8d}{m2^{-2\ell}} \log \frac{2k\ell(\ell+1)}{\delta}\right) \\
&\leq md(d+1)\left(\frac{1}{2} - 2^{-L_c'-1}\right) + 16d \left(\sum_{\ell=2}^{L_c'} 2^\ell \log \frac{2k\ell(\ell+1)}{\delta}\right) \\
&\leq md(d+1)\left(\frac{1}{2} - 2^{-L_c'-1}\right) + 2^{L_c'+1}(16d) \log \frac{4k{L_c'}^2}{\delta}.
\end{aligned}$$

For the local training phase, we have

$$\begin{aligned}
\sum_{i=1}^{m} \sum_{\boldsymbol{x}:\Delta_{i,\boldsymbol{x}}>\eta} \sum_{\ell=L_c'+1}^{\lceil \log_2 \frac{4}{\eta} \rceil} \Delta_{i,\boldsymbol{x}} n_{i,\ell}(\boldsymbol{x}) &\leq \sum_{i=1}^{m} \sum_{\ell=L_c'+1}^{\lceil \log_2 \frac{4}{\eta} \rceil} \sum_{\boldsymbol{x}:\Delta_{i,\boldsymbol{x}}>\eta} 2\varepsilon_\ell \left\lceil \frac{2\pi_{i,\ell}(\boldsymbol{x}) g_i(\pi_{i,\ell})}{\varepsilon_\ell^2} \log \frac{2km\ell(\ell+1)}{\delta} \right\rceil \\
&\leq \sum_{i=1}^{m} \sum_{\ell=L_c'+1}^{\lceil \log_2 \frac{4}{\eta} \rceil} 2\varepsilon_\ell \left(|\mathrm{Supp}(\pi_{i,\ell})| + \frac{2d}{\varepsilon_\ell^2} \log \frac{2km\ell(\ell+1)}{\delta}\right) \\
&\leq m \sum_{\ell=L_c'+1}^{\lceil \log_2 \frac{4}{\eta} \rceil} 2\cdot 2^{-\ell} \left(\frac{d(d+1)}{2} + \frac{2d}{2^{-2\ell}} \log \frac{2km\ell(\ell+1)}{\delta}\right)
\end{aligned}$$

$$\leq md(d+1)2^{-L_c'-1} + 4dm\left(\sum_{\ell=L_c'+1}^{\lceil \log_2 \frac{4}{\eta}\rceil} 2^{\ell}\log\frac{2km\ell(\ell+1)}{\delta}\right)$$
$$\leq md(d+1)2^{-L_c'-1} + \left(2^{\lceil \log_2 \frac{4}{\eta}\rceil+1} - 2^{L_c'+1}\right)4dm\log\frac{4km\left(L_c'\right)^2}{\delta}.$$

Thus, the over all regret

$$\begin{aligned}\hat{R}_{m,n}(\boldsymbol{\mu}) &\leq r_1 + mn\eta + \frac{1}{2}md(d+1) + 2^{L_c'+1}\left(16d\log\frac{4k(L_c')^2}{\delta}\right)\\ &\quad + \left(2^{\lceil \log_2 \frac{4}{\eta}\rceil+1} - 2^{L_c'+1}\right)\left(4dm\log\frac{4km\left(L_c'\right)^2}{\delta}\right)\\ &\leq r_1 + mn\eta + \frac{1}{2}md(d+1) + \min\left\{\frac{1}{h_{\sigma,\delta}}, \frac{16}{\eta}\right\}\left(16d\log\frac{4k\left(\log_2\frac{8}{\eta}\right)^2}{\delta}\right)\\ &\quad + \left(\frac{16}{\eta} - \min\left\{\frac{1}{h_{\sigma,\delta}}, \frac{16}{\eta}\right\}\right)\left(4dm\log\frac{4km\left(\log_2\frac{8}{\eta}\right)^2}{\delta}\right)\end{aligned}$$

for any $\eta > 0$.

When $L_c = 0$, it means that $h_{\sigma,\delta}$ is large and there is no collaboration phase. Thus, in the first local training phase, we have

$$\begin{aligned}&\sum_{i=1}^{m}\sum_{\boldsymbol{x}:\Delta_{i,\boldsymbol{x}}>\eta}\Delta_{i,\boldsymbol{x}}n_{i,1}(\boldsymbol{x})\\ &\leq \sum_{i=1}^{m}\sum_{\boldsymbol{x}:\Delta_{i,\boldsymbol{x}}>\eta}\left(2\max_{\boldsymbol{x}\in\mathcal{A}}\langle\boldsymbol{\mu},\boldsymbol{x}\rangle + 2\sqrt{2\left(\max_{\boldsymbol{x}\in\mathcal{A}}\|\boldsymbol{x}\|_{\mathbf{C}}^2\right)\log\frac{4km}{\delta}}\right)\left\lceil\frac{8\pi_{i,\ell}(\boldsymbol{x})g(\pi_{i,\ell})}{m(1/2)^2}\log\frac{2k\ell(\ell+1)}{\delta}\right\rceil\\ &\leq \sum_{i=1}^{m}\sum_{\boldsymbol{x}:\Delta_{i,\boldsymbol{x}}>\eta}\left(2\max_{\boldsymbol{x}\in\mathcal{A}}\langle\boldsymbol{\mu},\boldsymbol{x}\rangle + 2\sqrt{2\left(\max_{\boldsymbol{x}\in\mathcal{A}}\|\boldsymbol{x}\|_{\mathbf{C}}^2\right)\log\frac{4km}{\delta}}\right)\left(\frac{32\pi_{i,\ell}(\boldsymbol{x})d}{m}\log\frac{2k\ell(\ell+1)}{\delta}+1\right)\\ &\leq md(d+1) + \left(2\max_{\boldsymbol{x}\in\mathcal{A}}\langle\boldsymbol{\mu},\boldsymbol{x}\rangle + 2\sqrt{2\left(\max_{\boldsymbol{x}\in\mathcal{A}}\|\boldsymbol{x}\|_{\mathbf{C}}^2\right)\log\frac{4km}{\delta}}\right)\left(32d\log\frac{2k\ell(\ell+1)}{\delta}\right)\\ &= r_1\end{aligned}$$

where we defined r_1 earlier for the case when $L_c \geq 1$. Thus, we can see that the same upper bound on the regret holds for this case. Now, for the over all regret bound

$$\begin{aligned}\hat{R}_{m,n}(\boldsymbol{\mu}) &\leq r_1 + mn\eta + \frac{1}{2}md(d+1) + \min\left\{\frac{1}{h_{\sigma,\delta}}, \frac{16}{\eta}\right\}\left(16d\log\frac{4k\left(\log_2\frac{8}{\eta}\right)^2}{\delta}\right)\\ &\quad + \left(\frac{16}{\eta} - \min\left\{\frac{1}{h_{\sigma,\delta}}, \frac{16}{\eta}\right\}\right)\left(4dm\log\frac{4km\left(\log_2\frac{8}{\eta}\right)^2}{\delta}\right)\end{aligned} \tag{29}$$

for any $\eta > 0$, we look into different cases. Instead of $h_{\sigma,\delta}$, we look into

$$\tilde{h}_{\sigma,\delta} = \sqrt{\frac{n}{d}}h_{\sigma,\delta} = \left(\sqrt{\frac{2n}{dm}\log\frac{2k}{\delta}} + \sqrt{\frac{2n}{d}\log\frac{2k}{\delta}}\right)\sqrt{\max_{\boldsymbol{x}\in\mathcal{A}}\boldsymbol{x}^\top\mathbf{C}\boldsymbol{x}}$$

First, if $\tilde{h}_{\sigma,\delta} = O\left(m^{-\frac{1}{2}}\right)$, say, $\tilde{h}_{\sigma,\delta} \le c_1 m^{-1/2}$ for some constant $c_1 > 0$. Then, by choosing

$$\eta = 16c_1\sqrt{\frac{d}{mn}} \ge 16\sqrt{\frac{d}{n}}\tilde{h}_{\sigma,\delta} = 16h_{\sigma,\delta}, \tag{30}$$

we have

$$\begin{aligned}
\hat{R}_{m,n}(\boldsymbol{\mu}) &\le r_1 + mn\eta + \frac{1}{2}d(d+1) + \frac{16}{\eta}\left(16d\log\frac{4k\left(\log_2\frac{8}{\eta}\right)^2}{\delta}\right) \\
&\le r_1 + 16c_1\sqrt{dmn} + \frac{1}{2}d(d+1) + \frac{16}{c_1}\sqrt{dmn}\log\left(\frac{4k\left(\log_2\sqrt{mn/4c_1^2d}\right)^2}{\delta}\right) \\
&= \tilde{O}\left(\sqrt{dmn} + md^2\right)
\end{aligned}$$

Secondly, if we have $\tilde{h}_{\sigma,\delta} = \Theta(m^{-\gamma})$ for some $\gamma \in [0, -\frac{1}{2}]$, say,

$$c_2 m^{-\gamma} \le \tilde{h}_{\sigma,\delta} \le c_3 m^{-\gamma},$$

for some constant $c_2, c_3 > 0$. Then, by choosing

$$\eta = 16\tilde{h}_{\sigma,\delta}\sqrt{\frac{d}{n}} = 16h_{\sigma,\delta}, \tag{31}$$

we have

$$\begin{aligned}
\hat{R}_{m,n}(\boldsymbol{\mu}) &\le r_1 + c_3 m^{1-\gamma}\sqrt{dn} + \frac{1}{2}d(d+1) + \frac{1}{c_2 m^{-\gamma}}\sqrt{\frac{n}{d}}\left(16d\log\frac{4k\left(\log_2 m^\gamma\sqrt{n/4c_2^2d}\right)^2}{\delta}\right) \\
&\le r_1 + c_3 m^{1-\gamma}\sqrt{dn} + \frac{1}{2}d(d+1) + \frac{16}{c_2}m^\gamma\sqrt{dn}\log\frac{4k\left(\log_2 m^\gamma\sqrt{n/4c_2^2d}\right)^2}{\delta} \\
&= \tilde{O}\left(m^{1-\gamma}\sqrt{dn} + md^2\right).
\end{aligned}$$

Finally, if we have $\tilde{h}_{\sigma,\delta} = \Omega(1)$, by choosing $\eta = \min\left\{16\tilde{h}_{\sigma,\delta}\sqrt{d/n}, \sqrt{d/n}\right\}$, we can ontain

$$\hat{R}_{m,n}(\boldsymbol{\mu}) = \tilde{O}\left(m\sqrt{dn} + md^2\right).$$

B.4 Proof of Theorem 5.1

We apply the typical ε-net argument to cover the unit-ball action set with a finite action set. By Corollary 4.2.13 in Vershynin (2018), the covering number of the d-dimensional unit Euclidean ball satisfies

$$\mathcal{N}(\mathcal{B}_d, \varepsilon) \le \left(1 + \frac{2}{\varepsilon}\right)^d.$$

Let $\mathcal{A}_c$ be the covering and use $\mathcal{A}_c$ as the action set for Algorithm 1. Plug

$$k = |\mathcal{A}_c| \le \left(1 + \frac{2}{\varepsilon}\right)^d$$

into (29), we have

$$\hat{R}_{m,n}(\boldsymbol{\mu}) \leq r_1 + mn\eta + \frac{1}{2}md(d+1) + \min\left\{\frac{1}{h_{\sigma,\delta}}, \frac{16}{\eta}\right\}\left(16d^2 \log \frac{4\left(1+\frac{2}{\varepsilon}\right)\left(\log_2 \frac{8}{\eta}\right)^2}{\delta}\right)$$
$$+ \left(\frac{16}{\eta} - \min\left\{\frac{1}{h_{\sigma,\delta}}, \frac{16}{\eta}\right\}\right)\left(4d^2 m \log \frac{4\left(1+\frac{2}{\varepsilon}\right) m \left(\log_2 \frac{8}{\eta}\right)^2}{\delta}\right) + \varepsilon mn$$

for any $\varepsilon > 0$. Thus, if we choose $\varepsilon = O\left(\frac{1}{\sqrt{mn}}\right)$ and optimize over η like we did for Theorem B.1, we will get the results in Theorem 5.1 with optimal η shifted by the order of $\sqrt{d}$ comparing to the proof of Theorem B.1.

B.5 The additional term in the upper bound

The extra md^2 term in the regret upper bound comes from (3) in our collaborative phased elimination algorithm. As we will see in the details of Algorithm 1 in Section 5, in each phase of the collaborative learning stage, we first decide how many times each action $\boldsymbol{x}$ will be selected to reach the target error in the phase. Let us denote it by $n_\ell(\boldsymbol{x})$ for now. Then, we split the workload so that each user is responsible to select each action $\boldsymbol{x}$ for $n_{i,\ell}(\boldsymbol{x}) = \lceil n_\ell(\boldsymbol{x})/m \rceil$ times. Since the ceiling function can increase the value by almost 1 for each user and action in the worst-case, and the number of users is m and the number of the actions to be pulled (the cardinality of support of the design) can be upper bounded by $d(d+1)/2$ (Lattimore and Szepesvári (2020), Theorem 21.1), we will have at most extra md^2 pulls in each phase. By the way we choose the exponentially decreasing target error, this leads to an md^2 term in the regret. A more efficient way to split the workload might be first calculating the total rounds needed for all action combined, $n_\ell = \sum_{\boldsymbol{x}} n_\ell(\boldsymbol{x})$, then split it among the users. In this case, the users might be selecting different actions in the phase, and our proof need to be modified since we assume that each user selects the same actions during the collaborative learning stage. After the change, we will have an extra $(m + d^2)$ pull instead. The m term remains due to the fact that our problem formulation is synchronized. For instance, if the number of users m is much larger than the workload n_ℓ in some phase ℓ, our current formulation does not allow users to remain idle while a each of the n_ℓ selected user perform a pull. And thus since all users are forced to select some action in the phase, there are at least m pulls instead of n_ℓ pulls in each stage. This cause the m term in the regret. To resolve this, we have to remove the synchronized assumption on the formulation. For example, if the users are sequentially selecting actions, there can be only a part of the users that have to select actions in every phase. This prevent from the extra m term in the regret. However, the regret analysis will be different from our current one.

B.6 Sub-Gaussian population distributions

Our results on the regret analysis for CP-PE can be translated to the case where $(\boldsymbol{\theta}_i - \boldsymbol{\mu})$'s are sub-Gaussian random vectors. This is because sub-Gaussian random variables have similar concentration results to Gaussian random variables. Let us formally define the term first:

Definition B.1 (Sub-Gaussian random variables)**.** *A random variable $\boldsymbol{X}$ is sub-Gaussian with variance proxy σ^2 if*

$$\mathbb{E}\left[e^{\lambda \boldsymbol{X}}\right] \leq e^{\frac{\lambda^2\sigma^2}{2}}$$

for all $\lambda \in \mathbb{R}$.

With the definition of a sub-Gaussian random variable, we can extend the idea to a random vector.

Definition B.2 (Sub-Gaussian random vectors)**.** *A random vector $\boldsymbol{X}$ in the d-dimension space is said to be sub-Gaussian with variance proxy σ^2 if it is centered and for any $\boldsymbol{u} \in \mathbb{R}^d$ such that $\|\boldsymbol{u}\|_2 = 1$, the real random variable $\boldsymbol{u}^\top \boldsymbol{X}$ is sub-Gaussian with variance proxy σ^2. We write $\boldsymbol{X} \sim \mathrm{subG}(\sigma^2)$.*

Let $\boldsymbol{u}_i = \boldsymbol{\theta}_i - \boldsymbol{\mu}$ follow a sub-Gaussian distribution with variance proxy σ^2 for each $i \in [m]$. Then, for any $\boldsymbol{x} \in \mathcal{B}_d^2$,

$\boldsymbol{u}_i^\top \boldsymbol{x}$ is subG(σ^2) and thus

$$\left\langle \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{u}_i, \boldsymbol{x}_i \right\rangle \sim \text{subG}(\sigma^2/m). \tag{32}$$

From the concentration property of sub-Gaussian random variables (Wainwright (2019), Chapter 2.1), we have

$$\mathbb{P}\left\{ \left\langle \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{u}_i, \boldsymbol{x}_i \right\rangle \geq s \right\} \leq \exp\left\{ -\frac{ms^2}{2\sigma^2} \right\}. \tag{33}$$

Recall that in Section B.3, the regret analysis of CP-PE, the Gaussianity of $\boldsymbol{\theta}_i$ is only used for the additivity of Gaussianity and its concentration properties. Since similar properties are also established for sub-Gaussian random variables in (33), we can simply replace (21) with sub-Gaussian $\boldsymbol{u}_i$'s, the replace (22) and (25) with the concentration inequality of sub-Gaussian random variables. The rest of the proof is essentially the same. Additionally, CP-PE is not using the fact that $\boldsymbol{\theta}_i$ follows a Gaussian distribution. Thus, our algorithm and the regret analysis also work for any sub-Gaussian population distribution.

B.7 Extend the results to general covariance matrix C

To extend our regret analysis of Algorithm 1 to the case of general known $\mathbf{C}$, we can directly modify the proof in Section B.3. Notice that we kept the covariance $\mathbf{C}$ in the beginning of the proof instead of replacing it by $\sigma\mathbf{I}$. This allows us to simply apply the largest eigenvalue of $\mathbf{C}$ as an upper bound on the term $\max_{\boldsymbol{x}\in\mathcal{A}}\|\boldsymbol{x}\|_{\mathbf{C}}^2$. Then, we obtain an upper bound characterized by the largest eigenvalue of the covariance matrix $\mathbf{C}$.

C Experiments

C.1 Experiment setting

We compare CP-PE, Fed-PE, and Ind-PE on synthetic datasets. In the first experiment, we set the number of users $m = 100$, number of rounds $n = 15000$, and dimension $d = 2$. We set $\boldsymbol{\mu} = [1\ 0]^\top$, $\sigma = 0.3$. We let the action set $\mathcal{A}$ consist of 10 actions ($k = 10$) uniformly distributed on the unit ball. We set the variance of noise to be $\sigma_0^2 = 1$ and $\delta = 0.01$.

In each run, we randomly sample the unknown parameter $\boldsymbol{\theta}_i \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I})$ for $i \in [m]$. Then, we apply all CP-PE, Fed-PE, and Ind-PE on the instance of $\{\boldsymbol{\theta}_i\}_{i=1:m}$. We evaluate the joint cumulative regret for each algorithm and normalized it by the number of users m. Our result is averaged over 80 independent runs and the 1-sigma error bar is plotted in the figure.

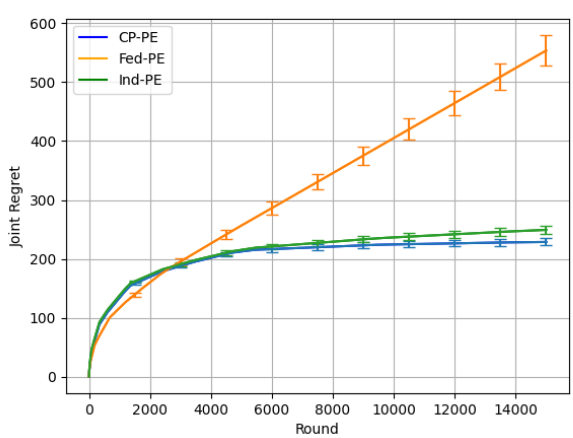

Figure 4: Joint cumulative regret obtained by CP-PE, Fed-PE, and Ind-PE. In the experiment, we have $m = 2$, $n = 15000$, $d = 2$, $\sigma = 0.3$.

C.2 Additional results

In this experiment, we decrease the number of users to $m = 2$ and the result is averaged over 350 independent runs. The joint cumulative regrets obtained by CP-PE, Fed-PE, and Ind-PE are shown in Figure 4. We can see that the plot looks similar to Figure 2 since we also consider the regime where CP-PE usually perform one to two phases in the collaborative learning stage and switch to the personalized learning stage. The difference shown in the plot is that since the number of users m decreases, the performance gap between CP-PE and Ind-PE becomes smaller. Comparing the result to Figure 2, we can see the benefit of collaboration scale with the number of users.

C.3 Experiments for unknown covariance matrix C

We conducted additional experiments to investigate the case where σ is unknown. We modified CP-PE so that during each collaborative phase, users use the data not only to estimate their parameters $\hat{\mu}$ but also to infer the heterogeneity level $\hat{\sigma}$ and decide whether they should switch to the personalization stage. Empirically, we observed that this adaptive version of CP-PE experiences a modest increase in cumulative regret (about 6%) compared to experiment conducted in the main paper. Nonetheless, CP-PE continues to outperform both Ind-PE and Fed-PE, preserving the relative performance. Figure 5 shows the joint cumulative regret obtained by modified CP-PE, Fed-PE, and Ind-PE. The modified CP-PE does not require prior knowledge on σ. Instead, an estimation of σ is estimated during each phase of the algorithm. On the other hand, Fed-Pe and Ind-Pe remain full knowledge of σ.

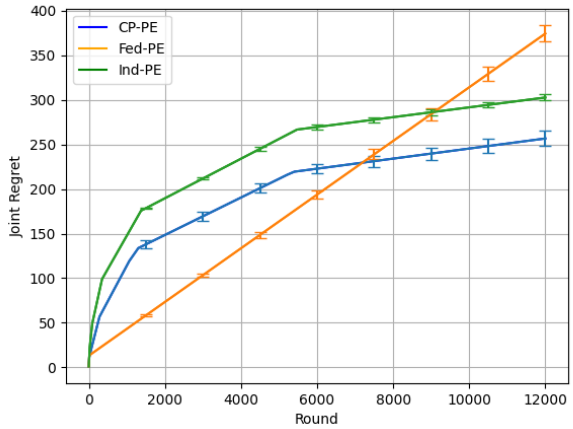

Figure 5: Joint cumulative regret obtained by modified CP-PE (for unknown σ), Fed-PE, and Ind-PE. In the experiment, we have $m = 200$, $n = 12000$, $d = 2$, $\sigma = 0.3$.