Multi-Agent Low-Dimensional Linear Bandits

Ronshee Chawla[*], Abishek Sankararaman[†] and Sanjay Shakkottai[‡]

Abstract

We study a multi-agent stochastic linear bandit with side information, parameterized by an unknown vector $\theta^* \in \mathbb{R}^d$. The side information consists of a finite collection of low-dimensional subspaces, one of which contains θ^*. In our setting, agents can collaborate to reduce regret by sending recommendations across a communication graph connecting them. We present a novel decentralized algorithm, where agents communicate subspace indices with each other, and each agent plays a projected variant of LinUCB on the corresponding (low-dimensional) subspace. Through a combination of collaborative best subspace identification, and per-agent learning of an unknown vector in the corresponding low-dimensional subspace, we show that the per-agent regret is much smaller than the case when agents do not communicate. By collaborating to identify the subspace containing θ^*, we show that each agent effectively solves an easier instance of the linear bandit (compared to the case of no collaboration), thus leading to the reduced per-agent regret. We finally complement these results through simulations.

1 Introduction

The Multi-Armed Bandit (MAB) model features a single decision maker making sequential decisions under uncertainty. It has found a wide range of applications: advertising [10], information retrieval [25] and operation of data-centers [15] to name a few. See also books of [19, 8]. As the scale of applications increases, several decision makers (a.k.a. agents) are involved in making repeated decisions as opposed to just a single agent. For example, in internet advertising, multiple servers are typically deployed to handle the large volume of traffic [9]. Multi-agent MAB models have emerged in recent times as a framework to design algorithms accounting for this scale. However, all the multi-agent models studied thus far only consider unstructured bandits with finitely many arms.

We introduce a multi-agent version of stochastic linear bandits. The linear bandits framework allows for a continuum of arms with a shared reward structure, thereby modeling many complex online learning scenarios [14, 1]. From a practical perspective, the linear bandit framework have shown to be more appropriate than unstructured bandits in many instances (e.g. recommendations [20], clinical studies [5]). Despite its applicability, a multi-agent version of linear bandits has not been studied thus far. The key technical challenge arises from the 'information leakage'; the reward obtained by playing an arm gives information on the reward obtained by all other arms. In a multi-agent scenario, this is further exacerbated, making design of collaborative algorithms non-trivial.

[*] Electrical and Computer Engineering, University of Texas at Austin. Email: ronsheechawla@utexas.edu
[†] Electrical Engineering and Computer Science Department, UC Berkeley. Email: abishek@berkeley.edu
[‡] Electrical and Computer Engineering, University of Texas at Austin. Email: sanjay.shakkottai@utexas.edu

We take a step in this direction by considering a collaborative multi-agent low-dimensional linear bandit problem and propose a novel decentralized algorithm. Agents in our model have side information in the form of subspaces. In our algorithm, agents collaborate by sharing these subspaces as opposed to the linear reward in our algorithm. Our main result shows that, even with minimal communications, the regret of all agents are much lower compared with the case of no collaboration.

Model Overview: Our setup consists of a single instance of a stochastic linear bandit with unknown parameter $\theta^* \in \mathbb{R}^d$, concurrently played by N agents. At each time t, each agent $i \in [N]$[1] plays an action vector $A_t^{(i)}$ from the set $\mathcal{A} = \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$ and receives a reward $\langle A_t^{(i)}, \theta^* \rangle + \eta_t^{(i)}$, where $\eta_t^{(i)}$ is zero mean sub-gaussian noise. The rewards obtained by the agents are only dependent on their actions and independent of actions of other agents. Agents have a common side information of K disjoint m-dimensional subspaces, one of which contains θ^*. The subspace containing θ^* is unknown to the agents. The agents in our model are connected through a communication graph over which they can exchange messages to collaborate. Agents have a communication budget limiting the number of times they can exchange messages. We seek decentralized algorithms for agents, i.e. the choice of action vector, communication choices and messages depend only on the observed past history (of action vectors, rewards and messages) of that agent.

Motivating Example: We first give an example of linear bandits with side information in the context of internet search advertisements and subsequently motivate the multi-agent version. In a search engine advertisement, each time a keyword is queried, an ad-server needs to make several decisions on advertising choices. The query is composed of the keyword and additional context (e.g. location of the search origin, user profile of the person). Thus, the decisions taken by the ad-server for a query can be modeled as playing an action in a high dimensional vector space.

Subspaces here arise from from historical context – different advertising choices for the categories of keywords can be clustered into distinct low-dimensional subspaces. Each subspace corresponds to decisions for a *category of keywords* (e.g. shoes, watches etc.). Thus, the advertising system can be decomposed into running several low-dimensional linear bandit problems, one for each category of keywords. Now, when a new keyword starts being seen (which has not been a-priori classified), the system needs to both *(i)* identify the correct category (subspace) it belongs to and *(ii)* make advertising choices according the rest of the query's context within the identified category.
Further, a collection of ad-servers are typically deployed to handle the large numbers of queries. Each query is routed to one of the servers which needs to make a real-time decision. Each server can be modeled as an agent and can collaborate by exchanging messages with other servers. In such a distributed system, decentralized learning algorithms with minimal communications are desirable. Our model abstracts this setup and our algorithm demonstrates that agents can indeed benefit from collaborations with minimal communications.

Our **main contributions** are as follows.

1. The SubGoss Algorithm: We propose SubGoss (Algorithm 1), where agents perform pure exploration for $O(\log T)$ time steps over a time horizon of T and play a projected version of LinUCB on the most likely subspace containing θ^*. The pure exploration is used to estimate the subspace

[1] For any positive integer $n \in \mathbb{N}$, $[n] := \{1, \cdots, n\}$

containing θ^*. Agents use pairwise communications to recommend subspaces (not samples), i.e. agents communicate the ID of the estimated best subspace and is limited by the communication budget provided as input. Our algorithm constrains agents to search for θ^* over only a small set (of cardinality $\leq \frac{K}{N} + 1$) of subspaces per agent. This set of subspaces is updated through recommendations; agents accept new recommendations and drop subspace(s) unlikely to contain θ^*, ensuring that the total number of subspaces an agent considers at all times remain small. Nevertheless, the best one *spreads* to all the agents through communications and thus all agents eventually identify the correct subspace. Thus, due to collaborations, each agent is individually solving an *easier* instance of the problem, leading to low regret.

2. Regret Guarantee and Benefit of Collaboration: Despite playing from a time-varying set of subspaces, every agent incurs a regret of at-most $\widetilde{O}(m\sqrt{T}) + O\left(\frac{d^2}{(\Delta^{(i)})^2}\log T\right)$ (Theorem 1). This scaling *does not depend* on either the gossip matrix G or communication budget[2] and we show that these communication constraints only affect the constant term in regret. The $\widetilde{O}(m\sqrt{T})$ term in regret arises from projected LinUCB on the true subspace containing θ^*, which we show every agent eventually identifies with probability 1. The logarithmic term is the regret of pure exploration in order to identify the best subspace. The term $\Delta^{(i)}$ in this logarithmic term is the minimum pairwise distance[3] between subspaces considered by agent i. We show that the set of subspaces an agent plays from is although time-varying, "freezes" after a random time. The minimum subspace gap for this frozen set is $\Delta^{(i)}$, which dictates rate of pure exploration. We observe that $\Delta^{(i)}$ is monotone in N: larger N leads to larger $\Delta^{(i)}$.

Thus, as the number of agents N increases, the individual regret for any agent is reduced, thereby demonstrating the benefit of collaboration. This is more pronounced in *high-dimensional* settings, ($d = O(T^{1/4})$ and m is a constant), when the regret from Projected LinUCB is comparable to pure exploration. In such settings, collaboration aids in reducing the leading order term in regret. The insights regarding benefit of collaboration are further confirmed numerically.

Related Work: Multi-agent MAB have become popular in recent times as the scale of applications increases. However, all of the work in this space concerns with simple unstructured bandits. The literature on multi-agent MAB can be classified into two —*competitive* where different agents compete for limited resources or *collaborative*, where agents jointly accomplish a shared objective (as in this paper). The canonical model of competitive multi-agent bandits is one wherein if multiple agents play the same arm, they all are blocked and receive no reward (colliding bandit model) for eg. [22, 3, 6, 7, 16]. Such models are motivated from applications in wireless networks [4]. The collaborative models consist of settings where if multiple agents play the same arm simultaneously, then they all receive independent rewards [23, 11, 9, 17, 21]. Such models have primarily been motivated by applications such as internet advertising [12]. However, the algorithms are all adapted to the case of unstructured bandits and cannot be applied to a linear bandit setup such as ours. Nevertheless, we adopt some of the broader principles from [23, 12] regarding the use of gossiping paradigm for communications to spread the best subspace into our algorithm design.

[2] We require G to be connected and communication budget at least logarithmic. See Appendix B

[3] Definition of distance between subspaces is given in the sequel

The stochastic linear bandit framework and the study of LinUCB algorithm was initiated by [14, 1]. From a practical perspective, the linear bandit framework has been shown to be effective for various applications: for example [2, 20] apply this framework in the context of internet advertising and [5, 24] apply in the context of clinical trials. Further, a projected version of LinUCB on low-dimensional subspaces has been recently studied in [18]. However, multi-agent version of the linear bandit framework has not been studied. We take a step in this direction by introducing a collaborative multi-agent linear bandit and proposing an algorithm that demonstrates benefit of collaboration.

2 Problem Setup

Our problem setup builds on similar settings considered for unstructured bandits in [23, 12], where a single instance of stochastic linear bandit is concurrently played by N agents. All agents play from the same set of action vectors $\mathcal{A}$, which is the unit ball with respect to Euclidean (l_2) norm, i.e. $\mathcal{A} \coloneqq \mathcal{B}^d = \{x \in \mathbb{R} : ||x||_2 \leq 1\}$. If at time t, any agent i plays an action vector $A_t^{(i)} \in \mathcal{A}$, the reward obtained is $r_t^{(i)} := \langle A_t^{(i)}, \theta^* \rangle + \eta_t^{(i)}$, where $\theta^* \in \mathbb{R}^d$ is unknown and $\eta_t^{(i)}$ is a zero mean sub-gaussian noise, conditional on the actions and rewards accumulated only by agent i (see Appendix A, (4) for details). The noise is independent across agents. The side information available to all the agents is a collection of K disjoint subspaces in $\mathbb{R}^d$ of dimension $m < d$. These subspaces are denoted by the $d \times m$ orthonormal matrices $\{U_i\}_{i=1}^K$, where $\text{span}(U_i)$ defines a m-dimensional subspace in $\mathbb{R}^d$. One of these subspaces contains θ^*, but agents are unaware of the subspace containing it. Let $P_k = U_k U_k^T$ denote the projection matrix of the subspace $\text{span}(U_k)$ for all $k \in [K]$. Agents know the parameter $L \in (0, 1]$, such that $L \leq ||\theta^*||_2 \leq 1$. Without loss of generality, we assume that K is an integral multiple of N.

Collaboration among Agents: The agents collaborate by exchanging messages over a communication network. This matrix is represented through a $N \times N$ *gossip matrix* G, with rows in this matrix being probability distributions over $[N]$. At each time step, after playing an action vector and obtaining a reward, agents can additionally choose to communicate with each other. If an agent i chooses to communicate, it will do so with another agent $j \sim G(i, \cdot)$, chosen independently of everything else. However, agents have a communication budget $(\widetilde{B}_t)_{t \geq 1}$, where, for any t, the total number of times an agent can communicate is at most $\widetilde{B}_t$. Each time an agent chooses to communicate, it can exchange only a fixed $O(\log(NK))$ number of bits.

Decentralized Algorithm: Each agent's decisions (arm play and communication decisions) in the algorithm depend only on its own history of plays and observations, along with the the recommendation messages that it has received from others.

Performance Metric: Each agent plays action vectors in order to minimize their individual cumulative regret. At any time t, the instantaneous regret for an agent i is given by $w_t^{(i)} = \langle \theta^*, A^* \rangle - \langle \theta^*, A_t^{(i)} \rangle$, where $A^* = \arg\max_{a \in \mathcal{A}} \langle \theta^*, a \rangle$. The expected cumulative regret for any agent $i \in [N]$ is given by $\mathbb{E}[R_T^{(i)}] \coloneqq \mathbb{E}[\sum_{t=1}^T w_t^{(i)}]$.

3 SubGoss Algorithm

Key Ideas and Intuition: Our setting is one where the optimal θ^* lies in one of a large number of (low-dimensional) subspaces. In our approach, agents at any time instant identify a small *active set of subspaces* (cardinality $\leq K/N + 1$) and play actions only within this set of subspaces (however, this set is time-varying). Namely, at each point of time, an agent first *identifies* amongst its current active set of subspaces the one likely to contain θ^*. It subsequently plays a projected version of LinUCB on this identified subspace. The communication medium is used for *recommendations*; whenever an agent is asked for information, it sends as message the subspace ID that it thinks most likely contains θ^*, which is then used to update the active set of the receiving agent. Thus, an agent's algorithm has two time-interleaved parts: (a) Updating active sets through collaboration, which is akin to a distributed best-arm identification problem, and (b) Determining the optimal θ^* from within its active set of subspaces, an estimation problem in low-dimensions, similar to the classical linear bandit.

SubGoss algorithm is organized in phases with the active subspaces for each agent fixed during the phase. Within a phase, all agents solve two tasks - *(i)* identify the most likely subspace amongst its active subspaces to contain θ^* and *(ii)* within this subspace, play actions optimally to minimize regret. The first point is accomplished by agents through pure exploration. The rate of exploration (different across agents) depends on the minimum distance between subspaces in their active set, which agents can explicitly compute. Thus, in order to reduce regret, agents in our algorithm only consider a *small, well separated* active set of subspaces in each phase. Otherwise, agents play action vectors within their best estimated subspace containing θ^* to minimize regret. This step is achieved by playing a projected version of the LinUCB algorithm. The second step only incurs regret in the dimension of the subspace (once the true subspace is correctly identified) as opposed to the ambient dimension, thereby keeping regret low. Due to communications, the correct subspace *spreads* to all agents, while keeping each agents minimum subspace gap large (and thus reducing the regret due to explorations).

Notation: $S^{d-1} = \{x \in \mathbb{R}^d : ||x||_2 = 1\}$ is the surface of unit sphere and $\mathbf{d}(x, \text{span}(U_l)) = \inf_{y \in \text{span}(U_l)} ||x - y||_2$, the distance between $x \in \mathbb{R}^d$ and subspace $\text{span}(U_l)$. The distance between subspaces $l_1, l_2 \in [K]$ is denoted by $\Delta_{l_1 l_2} = \inf_{x \in S^{d-1}} \{\mathbf{d}(x, \text{span}(U_{l_1})) + \mathbf{d}(x, \text{span}(U_{l_2}))\}$.

3.1 Description

Our algorithm builds on some of the ideas developed for a (non-contextual) collaborative setting for unstructured bandits in [23, 12]. For any agent $i \in [N]$, our algorithm proceeds in phases, where phase $j \in \mathbb{N}$ is from time slots $B_{j-1} + 1$ to B_j, both inclusive (with $B_0 = 0$). During each phase j, every agent $i \in [N]$ only plays from an active set $S_j^{(i)} \subset [K]$ of subspaces such that $|S_j^{(i)}| \leq (K/N) + 1$. Agents communicate at the end of the phase to update their active set. The phase lengths $(B_j)_{j \in \mathbb{N}}$ are designed so as to respect communication budget constraints.

Initialization: This step sets $S_1^{(i)}$, the initial active set also referred to as the *sticky set*[4]. Let $S_1^{(i)} := \widehat{S}^{(i)}$, which is obtained by partitioning the subspaces across agents to maximize the distance

$$(\widehat{S}^{(i)})_{i=1}^N = \arg\max\{\min_{i\in[N]} \min_{l_1\neq l_2 \in S_i} \Delta_{l_1,l_2} : (S_i)_{i=1}^N \text{ s.t. } 1 \leq |S_i| \leq (K/N), \cup_{i=1}^N S_i = [K]\}. \tag{1}$$

For numerical simulations, we provide a feasible partition that is fast and works well in practice.

Action Vectors Chosen in a Phase: If at time t in a phase j, the total number of explore samples thus far (denoted by $n_t^{(i)}$) satisfies $(32\alpha d^2/(\Delta_j^{(i)})^2 L^2)\log t \geq n_{t-1}^{(i)}$, then agent i plays a standard basis vector chosen according to the round robin method. Otherwise, the action vector is chosen according to the Projected LinUCB [18]. The subspace agent i chooses is the one closest in $S_j^{(i)}$ (denoted by $k_t^{(i)} \in S_j^{(i)}$) to the unit vector $(\widetilde{\theta}_{t-1}^{(i)})/(\|\widetilde{\theta}_{t-1}^{(i)}\|_2)$, the estimate of θ^* formed from the explore samples thus far. The precise action vector chosen is given according to the following equations [18]:

$$A_t^{(i)} = \arg\max_{a\in\mathcal{A}_t} \max_{\theta\in\mathcal{C}_t^{(i)}} \langle \theta, P_{k_t^{(i)}} a\rangle, \text{ where },$$

$$\mathcal{C}_t^{(i)} = \left\{\theta \in \mathbb{R}^d : ||\widehat{\theta}_t^{(i)} - \theta||_{\bar{V}_t(\lambda)^{(i)}} \leq \beta_t\right\},$$

and $\beta_t = S\sqrt{\lambda} + \sqrt{2\log\frac{1}{T} + m\log\left(1 + \frac{t-1}{\lambda m}\right)}$, $\bar{V}_t(\lambda)^{(i)} = P_{k_t^{(i)}}(\mathbf{A}_{t-1}^{(i)}(\mathbf{A}_{t-1}^{(i)})^T + \lambda I_d)P_{k_t^{(i)}}$ and $\widehat{\theta}_t^{(i)} = \arg\min_{\theta\in\mathbb{R}^d} ||(P_{k_t^{(i)}}\mathbf{A}_{t-1}^{(i)})^T\theta - \mathbf{r}_{t-1}^{(i)}||_2^2 + \lambda||P_{k_t^{(i)}}\theta||_2^2$. $\mathbf{A}_{t-1}^{(i)}$ is a $d \times (t-1)$ matrix whose columns are all the action vectors (explore and exploit) played up to and including time $t-1$ and $\mathbf{r}_{t-1}^{(i)}$ is a column vector of the corresponding rewards.

Communications and the Active Subspaces for the Next Phase: At the end of phase j, agent i asks for a subspace recommendation from an agent $J \sim G(i,\cdot)$ chosen independently. Denote by $\mathcal{O}_j^{(i)} \in [K]$ to be this recommendation. Agent i if asked for a recommendation at the end of phase j, recommends the subspace in $S_j^{(i)}$ closest to $(\widetilde{\theta}_{B_j}^{(i)}/\|\widetilde{\theta}_{B_j}^{(i)}\|_2)$, i.e., using only the explore samples. The next active set is given by $S_{j+1}^{(i)} := \widehat{S}^{(i)} \cup \mathcal{B}_j^{(i)}$, where $\mathcal{B}_j^{(i)} = \arg\min_{k\in\{S_j^{(i)}\cup\mathcal{O}_j^{(i)}\}} \mathbf{d}(\widetilde{\theta}_{B_j}^{(i)}/||\widetilde{\theta}_{B_j}^{(i)}||_2, \text{span}(U_k))$. Observe that $\widehat{S}^{(i)} \subseteq S_j^{(i)}$, $\forall j$, and thus $\widehat{S}^{(i)}$, is denoted sticky.

Please see Appendix A for a detailed description, and Algorithm 1 for the pseudo-code.

4 Main Result

In order to state the result, we make two mild technical assumptions: the gossip matrix P is connected and the communication budget $\widetilde{B}_t = \Omega(\log t)$ (detailed definitions in Appendix B). We define a random variable $\tau_{spr}^{(P)}$ denoting the spreading time of the following process: node i initially has a rumor; at each time, an agent without a rumor calls another chosen independently from the gossip matrix G and learns the rumor if the other person knows the rumor. The stopping time

[4] Choice of terminology will become apparent shortly

Algorithm 1 SubGoss Algorithm (at Agent i)

```
Input : $K$ subspaces $\{U_l\}_{l=1}^K$, time horizon $T$, communication budgets $(\widetilde{B}_t)_{t\in\mathbb{N}}$, explore parameter $\alpha > 0$, regularization parameter $\lambda > 0$, $L \in (0,1)$ and $\varepsilon > 0$
Initialization: $\widehat{S}^{(i)}, S_1^{(i)}$ (see Equation (1)).
$n_0^{(i)} \leftarrow 0$, $j \leftarrow 1$, $\delta \leftarrow \frac{1}{T}$
$B_j = \max\left(\min\{t \geq 0, \widetilde{B}_t \geq j\}, \lceil j^{1+\varepsilon} \rceil\right)$    ▷ Identify time of next communication
$\Delta_1^{(i)} \leftarrow \inf_{x\in S^{d-1}, l_1 \in S_1^{(i)}, l_2 \in S_1^{(i)}, l_1 \neq l_2} \{\mathbf{d}(x, \mathrm{span}(U_{l_1})) + \mathbf{d}(x, \mathrm{span}(U_{l_2}))\}$
for Time $t \in \mathbb{N}$ do
    if $\frac{32\alpha d^2}{(\Delta_j^{(i)})^2 L^2} \log t \geq n_{t-1}^{(i)}$ then    ▷ Condition for explore
        Play the standard basis vectors in $\mathbb{R}^d$ in round-robin fashion
        $n_t^{(i)} \leftarrow n_{t-1}^{(i)} + 1$
    else
        $k_t^{(i)} \leftarrow \arg\min_{k\in S_j^{(i)}} \mathbf{d}\left(\frac{\widetilde{\theta}_{t-1}^{(i)}}{||\widetilde{\theta}_{t-1}^{(i)}||_2}, \mathrm{span}(U_k)\right)$    ▷ Subspace closest to $\frac{\widetilde{\theta}_{t-1}^{(i)}}{||\widetilde{\theta}_{t-1}^{(i)}||_2}$
        Play $A_t^{(i)} \in \arg\max_{a\in\mathcal{A}_t} \max_{\theta\in\mathcal{C}_t^{(i)}} \langle \theta, P_{k_t^{(i)}} a \rangle$    ▷ Projected LinUCB
        $n_t^{(i)} \leftarrow n_{t-1}^{(i)}$
    end if
    if $t == B_j$ then    ▷ End of Phase
        $\mathcal{O}_j^{(i)} \leftarrow$ Recommend-Subspace$(i, j)$
        $\mathcal{B}_j^{(i)} \leftarrow \arg\min_{k\in\{S_j^{(i)} \cup \mathcal{O}_j^{(i)}\}} \mathbf{d}\left(\frac{\widetilde{\theta}_{B_j}^{(i)}}{||\widetilde{\theta}_{B_j}^{(i)}||_2}, \mathrm{span}(U_k)\right)$
        $S_{j+1}^{(i)} \leftarrow \widehat{S}^{(i)} \cup \mathcal{B}_j^{(i)}$    ▷ Update the set of playing subspaces
        $j \leftarrow j + 1$
        $\Delta_j^{(i)} \leftarrow \inf_{x\in S^{d-1}, l_1 \in S_j^{(i)}, l_2 \in S_j^{(i)}, l_1 \neq l_2} \{\mathbf{d}(x, \mathrm{span}(U_{l_1})) + \mathbf{d}(x, \mathrm{span}(U_{l_2}))\}$
        $B_j = \max\left(\min\{t \geq 0, \widetilde{B}_t \geq j\}, \lceil j^{1+\varepsilon} \rceil\right)$    ▷ Identify time of next communication
    end if
end for
```

$\tau_{spr}^{(G;i)}$ denotes the first time when all agents know the rumor and $\tau_{spr}^{(G)} = \max_{i\in[N]} \tau_{spr}^{(G;i)}$. For ease of exposition, we assume that $\theta^* \in \mathrm{span}(U_1)$, which the agents are unaware of. Let $\widetilde{\Delta}$ be the minimum separation between all subspaces, i.e., $\widetilde{\Delta} := \min_{l_1 \neq l_2 \in [K]} \Delta_{l_1,l_2}$ and $\Delta^{(i)} = \inf_{l_1 \neq l_2 \in \{\widehat{S}^{(i)} \cup 1\}} \Delta_{l_1,l_2}$.

Theorem 1. *Consider a system consisting of N agents connected by a gossip matrix G, all running Algorithm 1 with input parameters $\lambda \geq 1$, $\delta = \frac{1}{T}$, $\alpha > 12\left(\widetilde{\Delta}\right)^{-2}$. Furthermore, suppose the agents have a communication budget $(\tilde{B}_t)_{t\in\mathbb{N}}$ and use parameter $\varepsilon > 0$ satisfying assumption* **(A.2)** *in*

Algorithm 2 Subspace Recommendation

```
procedure RECOMMEND-SUBSPACE($(i, j)$)                          ▷ Input an agent $i$ and Phase $j$
    $p \sim G(i, \cdot)$                                        ▷ Sample an agent using the Gossip Matrix
return $\arg\min_{k \in S_j^{(p)}} \mathbf{d}\left(\frac{\widetilde{\theta}_{B_j}^{(p)}}{||\widetilde{\theta}_{B_j}^{(p)}||_2}, \mathrm{span}(U_k)\right)$   ▷ Closest subspace for agent $p$ at time slot $B_j$
end procedure
```

Appendix B. Then the regret of any agent $i \in [N]$, after any time $T \in \mathbb{N}$ is bounded by

$$\mathbb{E}[R_T^{(i)}] \leq \underbrace{\sqrt{8m\beta_T^2 \log\left(1 + \frac{T}{m\lambda}\right)} + 2 +}_{\textit{Projected LinUCB Regret}} \underbrace{2\left\lceil \frac{32\alpha d^2}{(\Delta^{(i)})^2 L^2} \log T \right\rceil}_{\textit{Cost of subspace exploration}} + \underbrace{g((B_x)_{x\in\mathbb{N}}) + 2\mathbb{E}[B_{2\tau_{spr}^{(G)}}]}_{\textit{Constant Cost of Pairwise Communications}}. \tag{2}$$

Here, $\beta_T = \sqrt{\lambda} + \sqrt{2\log T + m\log\left(1 + \frac{T-1}{\lambda m}\right)}$, $(B_x)_{x\in\mathbb{N}}$ is given in Equation (5) and $g((B_x)_{x\in\mathbb{N}}) = B_{j^*} + 4\left(\sum_{l \geq \frac{j^*}{2}} \frac{B_{2l}}{B_l^3}\right)$, where

$$j^* = 2\max\left(B^{-1}\left(\left(Nde^{\frac{1}{8d}}\right)^{\frac{1}{\frac{\alpha\widetilde{\Delta}^2}{4} - 3}}\right), \min\left\{j \in \mathbb{N} : \forall j' \geq j, B_{j'} \geq \left\lceil \frac{32\alpha d^2}{\widetilde{\Delta}^2 L^2} \log B_{j'} \right\rceil \right\}\right)$$

and $B^{-1}(x) = \sup\{y \in \mathbb{N} : B_y \leq x\}$, $\forall x \in \mathbb{R}_+$.

4.1 Discussion

In order to understand the theorem better, we consider an illustrative example and state a corollary of our main theorem. First, observe that the effect of the parameter ε used in the definition of $(B_x)_{x\in\mathbb{N}}$ in Equation (2) only affects the constant pairwise communication cost term and does not affect the scaling of regret with time horizon T. We now consider a simple example. Assume that the agents are connected by a complete graph, i.e., for any $i \neq j$, $P(i, j) = \frac{1}{N-1}$, and the communication budget $\widetilde{B}_t = \lfloor t^{1/\beta} \rfloor$, for some $\beta > 1$. Thus, from the definition of $(B_x)_{x\in\mathbb{N}}$ in Equation (5), we have $B_x = x^\beta$, where ε used in Equation (5) is sufficiently small. In this case, it is easy to see that for all N and $\beta > 0$, (for ex. Corollary 6 in [12]), $\mathbb{E}[B_{2\tau_{spr}^{(P)}}] \leq C(\log(N))^\beta$, for some universal constant $C > 0$. Thus, we have the following corollary.

Corollary 2. *Suppose the agents are connected by the complete graph, namely $P(i, j) = \frac{1}{N-1}$, for all $i \neq j$, and the communication budget and $\varepsilon > 0$ used by the agents are such that $B_x = x^\beta$ (given in Equation (5), for all large x, for some fixed $\beta > 0$. Then, when Algorithm 1 is run with the same inputs as in Theorem 1, the regret of any agent $i \in [N]$ after T times*

$$\mathbb{E}[R_T^{(i)}] \leq \underbrace{\widetilde{O}(m\sqrt{T})}_{\textit{Projected LinUCB Regret}} + \underbrace{O\left(\frac{d^2}{(\Delta^{(i)})^2}\log T\right)}_{\textit{Cost of Subspace Exploration}} + \underbrace{2^\beta\left(\frac{\pi^2}{6} + \max\left\{\left(Nde^{\frac{1}{8d}}\right)^{\frac{1}{\frac{\alpha\widetilde{\Delta}^2}{4} - 3}}, \left(\frac{32\alpha d^2}{\widetilde{\Delta}^2 L^2}\right)^2\right\}\right)}_{g((B_x)_{x\in\mathbb{N}})} + \underbrace{(2C\log(N))^\beta}_{\mathbb{E}[B_{2\tau_{spr}^{(P)}}]}, \tag{3}$$

where $C > 0$ is an universal constant.

In Equation (3), the $O(\cdot)$ notation only hides input constants and universal constants. Thus, we see from Equation (3), that the scaling of the regret of every agent $(\widetilde{O}(m\sqrt{T}))$ matches that of an oracle that knows the correct subspace in which θ^* lies in.

4.2 Proof Sketch

The proof of this theorem is given in Appendix C and we describe its salient ideas here. Similar to the phenomenon in unstructured bandits [12], we prove in Proposition 1 that in our linear bandit setup, there exists a *freezing phase* τ, such that all agents have the correct subspace containing θ^*. Consequently, for all phases $j > \tau$, all agents will play Projected LinUCB [18] from the correct subspace in the exploit time instants and recommend it at the end of the phase. Therefore, the set of subspaces every agent has does not change after phase τ and the regret after phase τ can be decomposed into regret due to pure exploration and regret due to Projected LinUCB (Proposition 2).

The technical novelty of our proof lies in bounding the regret till phase τ, i.e. $\mathbb{E}[B_\tau]$ (Proposition 3) and in particular showing it to be finite. This follows from two key observations arising from pure exploration in the explore time steps. First, for any agent $i \in [N]$, the estimate $\widetilde{\theta}_t^{(i)}$ of θ^* using the explore samples till time t, concentrates to θ^* (in l_2 norm) even with the time varying exploration rates in different phases (Lemma 3). Second, we show in Lemma 4 that if an agent has the correct subspace containing θ^* in a phase, the probability that it will not be recommended and hence dropped at the end of the phase is small.

Combining these two observations we establish that, after a random phase denoted by $\widehat{\tau}_{stab}$ in Appendix C, satisfying $\mathbb{E}[\widehat{\tau}_{stab}] < \infty$, agents never recommend incorrectly at the end of a phase and thus play the Projected LinUCB on the correct subspace in the exploit time instants. The formal argument follows by multiplying the relevant random variables by indicator random variables that the system has frozen. Subsequently, the expected value is bounded by either dropping the indicator random variable, or by considering an appropriate maximization of the other variable. To conclude, after random phase $\widehat{\tau}_{stab}$, the spreading time can be coupled with that of a standard rumor spreading [13], as once an agent learns the correct subspace, it is not dropped by the agent. This final part is similar to the one conducted for unstructured bandits in [12], giving us the desired bound on $\mathbb{E}[B_\tau]$.

5 Benefit of Collaboration and Regret-Communication Trade-off

In this section, we derive insights into the behavior of our algorithm.

1. Benefit of Collaboration: The regret guarantee in Theorem 1 implies a collaborative gain, i.e., the regret of an individual agent is strictly better than in a system without collaboration. The key reason comes from the term $\Delta^{(i)}$ in the cost of subspace exploration in the regret which increases as N increases. The case when agents do not collaborate is equivalent to saying that $N = 1$, i.e., there is only a single agent. In this case, $\widehat{S}^{(1)} = \{1, \cdots, K\}$ and $\Delta^{(1)} = \widetilde{\Delta}$. Whereas, in the presence of collaboration, the set of subspaces $S^{(i)} \subseteq [K]$, and thus, the minimum gap $\Delta^{(i)} \geq \widetilde{\Delta}$. Even though increasing N increases the cost of pairwise communications due to increased spreading time, this is only a constant and the dominant effect on regret is in reducing cost of exploration.

Hence, *collaboration reduces regret by ensuring that each agent is solving an easier instance of the problem*. This benefit of collaboration is more pronounced in 'high-dimensional' settings, when $d = O(T^{1/4})$ and m is small (a constant). In such settings, both the projected LinUCB term and the cost of subspace exploration are comparable, and therefore collaboration directly impacts in reducing the leading order term in regret. These insights are also verified numerically in simulations.

2. Regret-Communication Tradeoff: Corollary 2 shows that the per-agent regret increases as we decrease the communication budget. This follows by noticing that the right hand side of Equation (3) is monotonically increasing in β. Thus, a higher value of β (reducing communication budget) leads to higher per-agent regret. Although one would expect this to intuitively hold, the monotonicity of the bound in Equation (3) with respect to β makes it evident. Nevertheless, we see that communication cost only impacts *second order* terms in the regret as it only contributes a constant cost that does not scale with the time horizon.

6 Numerical Results

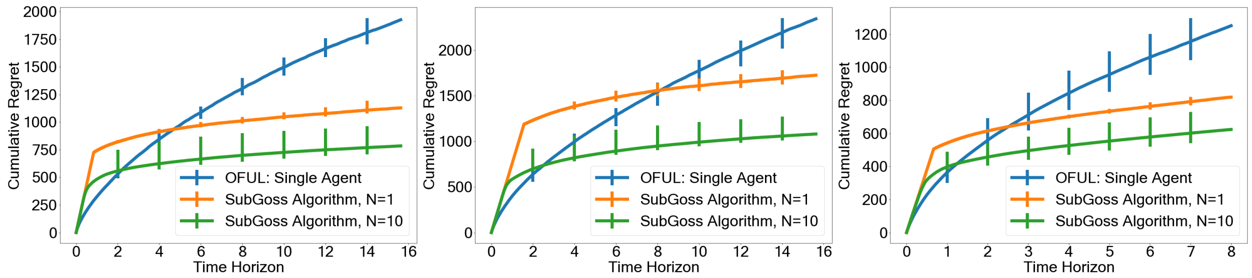

Figure 1: (d, m, K) are $(10, 2, 20)$, $(12, 3, 30)$ and $(13, 2, 10)$ respectively.

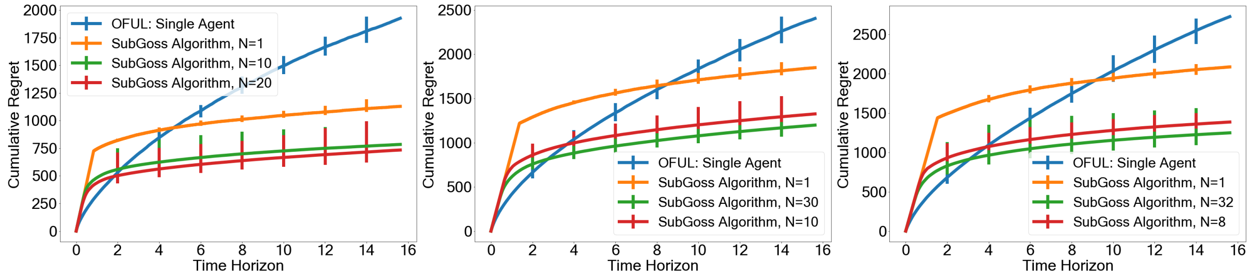

Figure 2: Illustrating benefit of collaboration. (d, m, K) are $(10, 2, 20)$, $(12, 3, 30)$ and $(13, 3, 32)$ respectively.

We evaluate the SubGoss algorithm empirically in synthetic simulations. We show the cumulative regret (after averaging across all agents) over 30 random runs of the algorithm with 95% confidence intervals. The communication budget budget $\widetilde{B}_t$ is fixed to $\widetilde{B}_t = \lfloor t^{1/3} \rfloor$ (i.e., $B_t = t^3$). The time horizon axis is scaled down by 1000 for clear representation. We compare its performance with two benchmarks: SubGoss algorithm with no collaborations (i.e. a single agent playing SubGoss algorithm) and a single agent playing the OFUL (classical LinUCB) algorithm of [1]. In this section, the condition for pure exploration is changed to $(2\alpha d/(\Delta_j^{(i)})^2 L^2) \log t \geq n_{t-1}^{(i)}$ for all $i \in [N]$. Note that as the constants in SubGoss algorithm (Algorithm 1) arise from somewhat loose tail bounds, we use different parameters here.

In our experiments, the agents are connected through a complete graph and the action set $\mathcal{A}$ is a collection of 2^d i.i.d. vectors on surface of the unit l_2 ball. Each m-dimensional subspace is the orthogonal matrix obtained by the SVD of a random $d \times m$ matrix with i.i.d. standard normal entries. The vector θ^* is the projected version of a standard Gaussian vector onto subspace 1 (the true subspace). We set $L = 1$ in simulations. For assigning the sticky sets to N agents, we re-arrange the K subspaces such that $0 = \Delta_{11} < \Delta_{12} \leq \Delta_{13} \leq \cdots \leq \Delta_{1K}$ and assign $\widehat{S}^{(i)} = \{i, N+i, \cdots, (K/N-1)\,N+i\}$, for all i. Figures 1 and 2 evaluate the performance of SubGoss algorithm for different values of problem parameters (d, m, K, N). We observe that the SubGoss algorithm effectively utilizes collaboration to reduces regret for all agents. This benefit of collaboration is demonstrated even with a sub-optimal partition for assigning sticky sets as described above. Moreover, Figure 2 demonstrates that as the number of agents increases, per-agent regret decreases. This follows as the complexity of the instance each agent solves becomes easier.

References

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

[2] Alekh Agarwal, Sarah Bird, Markus Cozowicz, Luong Hoang, John Langford, Stephen Lee, Jiaji Li, Dan Melamed, Gal Oshri, Oswaldo Ribas, et al. Making contextual decisions with low technical debt. *arXiv preprint arXiv:1606.03966*, 2016.

[3] Orly Avner and Shie Mannor. Concurrent bandits and cognitive radio networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 66–81. Springer, 2014.

[4] Orly Avner and Shie Mannor. Multi-user lax communications: a multi-armed bandit approach. In *IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications*, pages 1–9. IEEE, 2016.

[5] Hamsa Bastani and Mohsen Bayati. Online decision making with high-dimensional covariates. *Operations Research*, 68(1):276–294, 2020.

[6] Ilai Bistritz and Amir Leshem. Distributed multi-player bandits-a game of thrones approach. In *Advances in Neural Information Processing Systems*, pages 7222–7232, 2018.

[7] Etienne Boursier and Vianney Perchet. Sic-mmab: synchronisation involves communication in multiplayer multi-armed bandits. In *Advances in Neural Information Processing Systems*, pages 12048–12057, 2019.

[8] Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

[9] Swapna Buccapatnam, Jian Tan, and Li Zhang. Information sharing in distributed stochastic bandits. In *2015 IEEE Conference on Computer Communications (INFOCOM)*, pages 2605–2613. IEEE, 2015.

[10] Deepayan Chakrabarti, Ravi Kumar, Filip Radlinski, and Eli Upfal. Mortal multi-armed bandits. In *Advances in neural information processing systems*, pages 273–280, 2009.

[11] Mithun Chakraborty, Kai Yee Phoebe Chua, Sanmay Das, and Brendan Juba. Coordinated versus decentralized exploration in multi-agent multi-armed bandits. In *IJCAI*, pages 164–170, 2017.

[12] Ronshee Chawla, Abishek Sankararaman, Ayalvadi Ganesh, and Sanjay Shakkottai. The gossiping insert-eliminate algorithm for multi-agent bandits. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3471–3481, Online, 26–28 Aug 2020. PMLR.

[13] Flavio Chierichetti, Silvio Lattanzi, and Alessandro Panconesi. Almost tight bounds for rumour spreading with conductance. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 399–408, 2010.

[14] Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. *COLT*, 2008.

[15] Eshcar Hillel, Zohar S Karnin, Tomer Koren, Ronny Lempel, and Oren Somekh. Distributed exploration in multi-armed bandits. In *Advances in Neural Information Processing Systems*, pages 854–862, 2013.

[16] Dileep Kalathil, Naumaan Nayyar, and Rahul Jain. Decentralized learning for multiplayer multiarmed bandits. *IEEE Transactions on Information Theory*, 60(4):2331–2345, 2014.

[17] Ravi Kumar Kolla, Krishna Jagannathan, and Aditya Gopalan. Collaborative learning of stochastic bandits over a social network. *IEEE/ACM Transactions on Networking*, 26(4):1782–1795, 2018.

[18] Sahin Lale, Kamyar Azizzadenesheli, Anima Anandkumar, and Babak Hassibi. Stochastic linear bandits with hidden low rank structure. *CoRR*, abs/1901.09490, 2019.

[19] Tor Lattimore and Csaba Szepesvári. Bandit algorithms. *preprint*, 2019.

[20] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010.

[21] David Martínez-Rubio, Varun Kanade, and Patrick Rebeschini. Decentralized cooperative stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 4531–4542, 2019.

[22] Jonathan Rosenski, Ohad Shamir, and Liran Szlak. Multi-player bandits–a musical chairs approach. In *International Conference on Machine Learning*, pages 155–163, 2016.

[23] Abishek Sankararaman, Ayalvadi Ganesh, and Sanjay Shakkottai. Social learning in multi agent multi armed bandits. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 3(3):1–35, 2019.

[24] Ambuj Tewari and Susan A Murphy. From ads to interventions: Contextual bandits in mobile health. In *Mobile Health*, pages 495–517. Springer, 2017.

[25] Yisong Yue and Thorsten Joachims. Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1201–1208, 2009.

Appendix A Additional Details of the SubGoss Algorithm

In this section, we provide additional details about the actions (in particular the explore step) taken by an agent $i \in [N]$ at time $t \in \mathbb{N}$ (denoted by $A_t^{(i)}$) in phase j. Recall that

- the reward obtained is $r_t^{(i)} = \langle \theta^*, A_t^{(i)} \rangle + \eta_t^{(i)}$, where for all $z \in \mathbb{R}$,

$$\mathbb{E}[\exp(z\eta_t^{(i)})|\mathcal{F}_{t-1}^{(i)}] \leq \exp\left(\frac{z^2}{2}\right) \text{ a.s.}, \tag{4}$$

  and $\mathcal{F}_{t-1}^{(i)} = \sigma(A_1^{(i)}, r_1^{(i)}, \cdots, A_{t-1}^{(i)}, r_{t-1}^{(i)}, A_t^{(i)})$.

- $\Delta_j^{(i)} = \inf\{\Delta_{l_1 l_2}, l_1 \neq l_2 \in S_j^{(i)}\}$[5] is the minimum pairwise separation between the subspaces in $S_j^{(i)}$

- $n_t^{(i)}$ is the number of explore samples collected up to and including time t ($n_0^{(i)} = 0$)

The action vector $A_t^{(i)}$ is chosen as follows: (i) (explore) If $\frac{32\alpha d^2}{(\Delta_j^{(i)})^2 L^2} \log t \geq n_{t-1}^{(i)}$, agent i explores, else (ii) (exploit) agent i exploits and plays projected LinUCB. We now describe the explore step.

Explore: If agent i explores at time t, then it plays the standard basis vectors in a round robin fashion, i.e., $A_t^{(i)} = e_{n_{t-1}^{(i)} \bmod d+1}$. Subsequently, $n_t^{(i)} = n_{t-1}^{(i)} + 1$. The rewards collected during the explore time steps only is used to construct an estimate $\widetilde{\theta}_t^{(i)}$ of θ^* by solving the least squares minimization problem $\arg\min_{\theta \in \mathbb{R}^d} ||\mathbf{X}_t^{(i)T}\theta - \mathbf{Y}_t^{(i)}||_2^2$. Here, $\mathbf{X}_t^{(i)}$ is a $n_t^{(i)} \times d$ matrix whose columns are the standard basis vectors played in the explore time slots up to and including time t and $\mathbf{Y}_t^{(i)}$ is a column vector containing the corresponding obtained rewards. Solving the above least squares problem, for all $s \in [d]$, gives $\widetilde{\theta}_{t,s}^{(i)} = \frac{1}{n_{t,s}^{(i)}} \sum_{x \in [t]: s_x = s} Y_x$ when $n_{t,s}^{(i)} \neq 0$, 0 otherwise.

Appendix B Technical Assumptions for Theorem 1

Building on the communication constraints considered in [12], we make the following mild assumptions:

(A.1) [12] The gossip matrix G is irreducible: for any $i \neq j \in [N]$, there exists $2 \leq l \leq N$ and $k_1, \cdots, k_l \in [N]$, with $k_1 = i$ and $k_l = j$ such that the product $G(k_1, k_2) \cdots G(k_{l-1}, k_l) > 0$. In words, the communication graph among the agents is connected.

(A.2) [12] The communication budget $(\widetilde{B}_t)_{t \in \mathbb{N}}$ and $\varepsilon > 0$ is such that for all $D > 0$, there exists $t_0(D)$ such that for all $t \geq t_0(D)$, $\widetilde{B}_t \geq D \log t$ (i.e., $\widetilde{B}_t = \Omega(\log t)$). Furthermore, $\sum_{l \geq 2} \frac{B_{2l}}{B_l^3} < \infty$.

In order to understand assumption **A.2** better, here are some examples:

- $\widetilde{B}_t = \lceil t^{1/2} \rceil$ for all $t \geq 1$ and $\varepsilon < 1 \implies B_x = \lfloor x^2 \rfloor$ for all $x \geq 1$

[5] if there is no such feasible l_1 and l_2, then $\Delta_j^{(i)} = 2$.

- When agents have the luxury to communicate in every time slot, i.e. $\widetilde{B}_t = t$ for all $t \geq 1 \implies B_x = \lceil x^{1+\varepsilon} \rceil$ for all $x \geq 1$

One can check that the conditions in assumption **A.2** are satisfied by the above examples.

Assumption **A.1** is needed because if the communication graph among the agents is not connected, the setup becomes *degenerate*, as there exists at least a pair of agents which cannot participate in information exchange. Assumption **A.2** states that the number of times an agent performs information pulls over the time horizon T is at least $\Omega(\log T)$. Furthermore, the series in **A.2** is convergent for all *natural* examples (such as polynomial and exponential). For example, when $\tilde{B}_l = \lceil \frac{1}{D} \log^{\beta}(l) \rceil$ or $\tilde{B}_l = \lceil l^{1/(D+1)} \rceil$, for all $D > 0$ and $\beta > 1$, the series converges. Thus, the practical insights that can be obtained from our results are not affected by assumptions **A.1** and **A.2**.

Appendix C Proof of Theorem 1

In this section, we assume agents know the parameters L and S such that $L \leq \|\theta^*\|_2 \leq S$. In the paper, we set $S = 1$ for ease of exposition. Before going through the proof, we will first set some definitions and notations.

C.1 Definitions and Notations

We adapt the proof ideas developed in [12] for the unstructured bandit case. Recall that the communication budget is reparametrized as the sequence $(B_x)_{x \in \mathbb{N}}$, where

$$B_x = \max\left(\min\{t \in \mathbb{N}, \widetilde{B}_t \geq x\}, \lceil x^{1+\varepsilon} \rceil\right). \tag{5}$$

At any time $t \in \mathbb{N}$ in phase j, Let $\xi_t^{(i)} = \mathbf{1}\left(\frac{32\alpha d^2}{(\Delta_j^{(i)})^2 L^2} \log t \geq n_{t-1}^{(i)}\right)$. In words, $\xi_t^{(i)}$ is an indicator variable indicating whether the agent will explore (or exploit) in time instant t. For every agent $i \in [N]$, let $\widehat{\mathcal{O}}_j^{(i)} \in S_j^{(i)}$ be the subspace closest to $\frac{\widetilde{\theta}_{B_j}^{(i)}}{||\widetilde{\theta}_{B_j}^{(i)}||_2}$ with respect to the distance measure $\mathbf{d}$. Let $\chi_j^{(i)}$ be the indicator variable for the event $\left\{1 \in S_j^{(i)}, \widehat{\mathcal{O}}_j^{(i)} \neq 1\right\}$, i.e.

$$\chi_j^{(i)} = \mathbf{1}\left(1 \in S_j^{(i)}, \widehat{\mathcal{O}}_j^{(i)} \neq 1\right),$$

which indicates whether agent i, *if* it has the subspace span(U_1), does not recommend it at the end of phase j. We now define certain random times that will be useful in the analysis, in a similar manner as done in [12]:

$$\begin{aligned}
\widehat{\tau}_{stab}^{(i)} &= \inf\{j^{'} \geq j^* : \forall j \geq j^{'}, \chi_j^{(i)} = 0\}, \\
\widehat{\tau}_{stab} &= \max_{i \in [N]} \widehat{\tau}_{stab}^{(i)}, \\
\widehat{\tau}_{spr}^{(i)} &= \inf\{j \geq \widehat{\tau}_{stab} : 1 \in S_j^{(i)}\} - \widehat{\tau}_{stab},
\end{aligned}$$

$$\widehat{\tau}_{spr} = \max_{i \in \{1,\cdots,N\}} \widehat{\tau}_{spr}^{(i)},$$

$$\tau = \widehat{\tau}_{stab} + \widehat{\tau}_{spr}.$$

Here, $\widehat{\tau}_{stab}^{(i)}$ is the earliest phase, such that for all subsequent phases, *if* agent i has the subspace span(U_1), it will be recommend the subspace span(U_1). The term $\widehat{\tau}_{spr}^{(i)}$ denotes the number of phases it takes after $\widehat{\tau}_{stab}$ to have the subspace span(U_1) in its playing set. The following proposition shows that the system is frozen after phase τ, i.e. after phase τ, the set of subspaces of all agents remain fixed in the future.

Proposition 1. *For all agents $i \in \{1, \cdots, N\}$, we have almost-surely,*

$$\bigcap_{j \geq \tau} S_j^{(i)} = S_\tau^{(i)},$$

$$\widehat{\mathcal{O}}_l^{(i)} = 1 \ \forall l \geq \tau, \ \forall i \in \{1, \cdots, N\}.$$

Proof. Fix any agent $i \in [N]$ and any phase $j \geq \tau$. Since $\tau \geq \widehat{\tau}_{stab}^{(i)}$, we have for all $j \geq \tau$,

$$\chi_j^{(i)} = 0. \tag{6}$$

However, as $\tau \geq \widehat{\tau}_{spr}^{(i)} + \widehat{\tau}_{stab}$, we know that

$$1 \in S_j^{(i)}. \tag{7}$$

Equations (6) and (7) imply that $\widehat{\mathcal{O}}_j^{(i)} = 1$. Moreover, $\widehat{\mathcal{O}}_j^{(i)} = 1$ is true for all phases $j \geq \tau$ and all agents $i \in [N]$, as they are arbitrarily chosen. Furthermore, the update step of the algorithm along with the above reasoning tells us that none of the agents will change their subspaces after any phase $j \geq \tau$, as the agents already have the correct subspace in their respective playing sets. Thus, $\bigcap_{j \geq \tau} S_j^{(i)} = S_\tau^{(i)}$ for all agents $i \in [N]$. □

Proposition 1 also tells us that for all time instants $t > B_\tau$, in the exploit time slots, all agents will play Projected LinUCB from the subspace span(U_1), because the algorithm picks the subspace closest to $\frac{\widetilde{\theta}_t^{(i)}}{||\widetilde{\theta}_t^{(i)}||_2}$ with respect to the distance measure $\mathbf{d}$ in the exploit time slots. Combining Proposition 1 with the update step of the algorithm (line 17 of Algorithm 1) tells that $S_j^{(i)} = \{\widehat{S}^{(i)} \cup \{1\}\}$ for all $j \geq \tau$ and thus $\{\Delta_j^{(i)}\}_{j \geq \tau} = \Delta^{(i)}$ for all $i \in \mathbb{N}$.

C.2 Intermediate Results

Before stating and proving the intermediate results, we highlight the key pieces needed to prove Theorem 1. We already showed in Proposition 1 that there exists a freezing time τ, such that in the subsequent phases, all agents have the correct subspace containing θ^* and recommend it henceforth. Thus, the expected cumulative regret incurred can be decomposed into two parts: the regret up to phase τ and the regret after phase τ.

The expected cumulative regret incurred up to phase τ is a constant independent of the time horizon (Proposition 3). It is a consequence of following important observations resulting from pure exploration in the explore time steps:

- For any agent $i \in [N]$, the estimate $\widetilde{\theta}_t^{(i)}$ of θ^* concentrates to θ^* in l_2 norm, despite time-varying exploration of subspaces leading to different exploration rates in different phases (Lemma 3).
- Subsequently, we show that the probability that an agent will not recommend and thus drop the correct subspace containing θ^* is small at the end of a phase (Lemma 4).

The above observations imply that after a (random) phase, denoted by $\widehat{\tau}_{stab} \leq \tau$, agents always recommend (and never drop) the correct subspace. After phase $\widehat{\tau}_{stab}$, we stochastically dominate (in Proposition 4) the spreading time of the correct subspace with a standard rumor spreading process [13]. Hence, the expected cumulative regret up to phase τ is bounded by the total number of time steps taken to reach phase $\widehat{\tau}_{stab}$ and the additional number of phases taken to spread the correct subspace.

Post phase τ, as the active set of subspaces maintained by agents remains unchanged (as deduced in Proposition 1) and thus, the regret can be decomposed into sum of regret due to pure exploration and regret due to projected LinUCB. The regret due to projected LinUCB is adapted from the analysis of a similar algorithm conducted in [18]. This regret decomposition is further bounded (Proposition 2) by either dropping the indicator random variables corresponding to explore (exploit) time instants, or appropriately maximizing those indicator variables.

The following intermediate results will precisely characterize the intuition behind the proof of Theorem 1.

Proposition 2. *The regret of any agent $i \in \{1, \cdots, N\}$ after playing for T steps is bounded by*

$$\mathbb{E}[R_T^{(i)}] \leq 2S\mathbb{E}[B_\tau] + \mathbb{E}[R_{\text{proj},T}] + 2S\left\lceil \frac{32\alpha d^2}{(\Delta^{(i)})^2 L^2} \log T \right\rceil.$$

Proof. We will first show that for any agent i, the instantaneous regret $w_t^{(i)} \leq 2S$. In order to obtain this bound, notice that for any $a \in \mathcal{B}^d$

$$|\langle \theta^*, a\rangle| \leq ||\theta^*||_2.||a||_2 \leq S$$

by Cauchy-Schwarz inequality. Therefore, we have $w_t^{(i)} \leq 2S$ for all t. From the definition of regret $R_T^{(i)}$,

$$\begin{aligned} R_T^{(i)} &= \sum_{t=1}^{T} w_t^{(i)} \\ &= \sum_{t=1}^{B_\tau} w_t^{(i)} + \sum_{t=B_\tau+1}^{T} w_t^{(i)} \\ &\overset{(a)}{\leq} 2SB_\tau + \sum_{t=B_\tau+1}^{T} w_t^{(i)}\mathbf{1}(\xi_t^{(i)}=0) + \sum_{t=B_\tau+1}^{T} w_t^{(i)}\mathbf{1}(\xi_t^{(i)}=1) \\ &\overset{(b)}{=} 2SB_\tau + R_{\text{proj},T} + \sum_{t=B_\tau+1}^{T} w_t^{(i)}\mathbf{1}(\xi_t^{(i)}=1) \end{aligned} \tag{8}$$

In step (a), we use $w_t^{(i)} \leq 2S$ to bound the sum $\sum_{t=1}^{B_\tau} w_t^{(i)}$. In step (b), the cumulative regret in the exploit time slots after time B_τ, denoted by $R_{\text{proj},T}$ is incurred by playing Projected LinUCB on the subspace $\text{span}(U_1)$. The cumulative regret in the explore time slots after time B_τ, i.e. $\sum_{t=B_\tau+1}^{T} w_t^{(i)} \mathbf{1}(\xi_t^{(i)} = 1)$ is bounded as follows:

$$
\begin{aligned}
\sum_{t=B_\tau+1}^{T} w_t^{(i)} \mathbf{1}(\xi_t^{(i)} = 1) &\overset{(a)}{\leq} 2S \sum_{t=B_\tau+1}^{T} \mathbf{1}(\xi_t^{(i)} = 1) \\
&= 2S \sum_{t=B_\tau+1}^{T} \mathbf{1}\left(\frac{32\alpha d^2}{(\Delta^{(i)})^2 L^2} \log t \geq n_{t-1}^{(i)} \right) \\
&\overset{(b)}{\leq} 2S \sum_{t=B_\tau+1}^{T} \mathbf{1}\left(\frac{32\alpha d^2}{(\Delta^{(i)})^2 L^2} \log T \geq n_{t-1}^{(i)} \right) \\
&\leq 2S \sum_{t=1}^{T} \mathbf{1}\left(\frac{32\alpha d^2}{(\Delta^{(i)})^2 L^2} \log T \geq n_{t-1}^{(i)} \right) \\
&\overset{(c)}{\leq} 2S \left\lceil \frac{32\alpha d^2}{(\Delta^{(i)})^2 L^2} \log T \right\rceil .
\end{aligned}
\tag{9}
$$

In step (a), we use $w_t^{(i)} \leq 2S$ for all $t \in \mathbb{N}$. Step (b) is trivial as $t \in [B_\tau + 1, T]$. Step (c) is a consequence of the fact that the sequence $\{n_t^{(i)}\}_{t \in \mathbb{N}}$ is non-negative and non-decreasing in t. Substituting equation (9) into equation (8), followed by taking expectation on both sides in equation (8) completes the proof. □

The lemma below shows that $\widetilde{\theta}_t^{(i)}$ gets closer to θ^* in l_2 norm as the algorithm progresses, which will eventually help us obtain a bound on probability of picking the wrong subspace. Define $t_0 = \inf \left\{ \mathcal{T} \in \mathbb{N} : \forall t \geq \mathcal{T}, t \geq \left\lceil \frac{32\alpha d^2}{\widetilde{\Delta}^2 L^2} \log t \right\rceil \right\}$.

Lemma 3. *For any agent $i \in [N]$ and time $t \in \mathbb{N}$ such that $t \geq t_0$, we have*

$$
\mathbb{P}\left(||\widetilde{\theta}_t^{(i)} - \theta^*||_2 > \epsilon \right) \leq 2d \ \exp\left(\frac{\epsilon^2}{2d} \right) \ t^{-\frac{4\alpha\epsilon^2}{L^2}} .
$$

where $\epsilon > 0$.

Proof. We have

$$
\begin{aligned}
\widetilde{\theta}_{t,s}^{(i)} &= \frac{1}{n_{t,s}^{(i)}} \sum_{x \in [t] : s_x = s} Y_x \\
&= \theta_s^* + \frac{1}{n_{t,s}^{(i)}} \sum_{x \in [t] : s_x = s} \eta_x^{(i)}
\end{aligned}
$$

for all $s \in [d]$. From the assumption that the additive noise is conditionally 1-subgaussian, we know that

$$
\mathbb{P}\left(|\widetilde{\theta}_{T,s}^{(i)} - \theta_s^*| > \gamma \right) \leq 2e^{-\frac{\gamma^2 n_{t,s}^{(i)}}{2}}
\tag{10}
$$

for all $\gamma > 0$. If $|\widetilde{\theta}_{t,s}^{(i)} - \theta_s^*| \leq \frac{\epsilon}{\sqrt{d}}$ for all $s \in [d]$ and $\epsilon > 0$, then $||\widetilde{\theta}_t^{(i)} - \theta^*||_2 \leq \epsilon$. Hence,

$$\begin{aligned}
\mathbb{P}\left(||\widetilde{\theta}_t^{(i)} - \theta^*||_2 > \epsilon\right) &\leq \mathbb{P}\left(\exists s \in [d] : |\widetilde{\theta}_{t,s}^{(i)} - \theta_s^*| > \frac{\epsilon}{\sqrt{d}}\right) \\
&= \mathbb{P}\left(\bigcup_{s=1}^{d}\left(|\widetilde{\theta}_{t,s}^{(i)} - \theta_s^*| > \frac{\epsilon}{\sqrt{d}}\right)\right) \\
&\overset{(a)}{\leq} \sum_{s=1}^{d} \mathbb{P}\left(|\widetilde{\theta}_{t,s}^{(i)} - \theta_s^*| > \frac{\epsilon}{\sqrt{d}}\right) \\
&\overset{(b)}{\leq} 2d \exp\left(-\frac{\epsilon^2 n_{t,s}^{(i)}}{2d}\right) \\
&\overset{(c)}{\leq} 2d \exp\left(-\frac{\epsilon^2}{2d}\left\lfloor\frac{n_t^{(i)}}{d}\right\rfloor\right) \\
&\overset{(d)}{\leq} 2d \exp\left(-\frac{\epsilon^2}{2d}\left(\frac{n_t^{(i)}}{d} - 1\right)\right) \\
&\overset{(e)}{\leq} 2d \exp\left(\frac{\epsilon^2}{2d}\right) \exp\left(-\frac{\epsilon^2}{2d^2}\frac{8\alpha d^2}{L^2}\log t\right) \\
&= 2d\ \exp\left(\frac{\epsilon^2}{2d}\right)\ t^{-\frac{4\alpha\epsilon^2}{L^2}}.
\end{aligned}$$

Step (a) follows from union bound. In step (b), we use the result from equation (10). In step (c), we use the fact that $n_{t,s}^{(i)} \geq \lfloor\frac{n_t^{(i)}}{d}\rfloor$, as the basis vectors are played in a round robin fashion. Step (d) follows from the inequality $\lfloor x \rfloor \geq x - 1$ for all $x \in \mathbb{R}$. Step (e) is obtained by observing from the algorithm that $n_t^{(i)} \geq \left\lceil \frac{8\alpha d^2}{L^2}\log t\right\rceil \geq \frac{8\alpha d^2}{L^2}\log t$ as $\Delta_j^{(i)} \leq 2$ for any phase $j \geq 1$, thus concluding the proof. □

We will now obtain a bound on probability for choosing a wrong subspace. Since $\theta^* \in \text{span}(U_1)$, any subspace chosen other than $\text{span}(U_1)$ will result in an error. Equivalently, it can be expressed as

$$\left\{\mathbf{d}\left(\frac{\widetilde{\theta}_t^{(i)}}{||\widetilde{\theta}_t^{(i)}||_2}, \text{span}(U_1)\right) > \frac{\widetilde{\Delta}}{2}\right\}.$$

This is the correct choice for the error event, because the wrong subspace is chosen when $\frac{\widetilde{\theta}_t^{(i)}}{||\widetilde{\theta}_t^{(i)}||_2}$ is closer to the subspace other than $\text{span}(U_1)$ in l_2 norm, i.e.

$$\begin{aligned}
\frac{1}{2}\mathbf{d}\left(\frac{\widetilde{\theta}_t^{(i)}}{||\widetilde{\theta}_t^{(i)}||_2}, \text{span}(U_1)\right) &> \frac{1}{2}\mathbf{d}\left(\frac{\widetilde{\theta}_t^{(i)}}{||\widetilde{\theta}_t^{(i)}||_2}, \text{span}(U_k)\right) \\
\mathbf{d}\left(\frac{\widetilde{\theta}_t^{(i)}}{||\widetilde{\theta}_t^{(i)}||_2}, \text{span}(U_1)\right) &> \frac{1}{2}\mathbf{d}\left(\frac{\widetilde{\theta}_t^{(i)}}{||\widetilde{\theta}_t^{(i)}||_2}, \text{span}(U_1)\right) + \frac{1}{2}\mathbf{d}\left(\frac{\widetilde{\theta}_t^{(i)}}{||\widetilde{\theta}_t^{(i)}||_2}, \text{span}(U_k)\right) \\
&\geq \inf_{x \in S^{d-1}} \frac{1}{2}\left(\mathbf{d}(x, \text{span}(U_1)) + \mathbf{d}(x, \text{span}(U_k))\right)
\end{aligned}$$

$$= \frac{\Delta_k}{2} \geq \frac{\widetilde{\Delta}}{2},$$

where the last step follows from the definitions of Δ_k and $\widetilde{\Delta}$. From the definition of $\mathbf{d}\left(., \mathrm{span}(U_1)\right)$, it follows that

$$\mathbf{d}\left(\frac{\widetilde{\theta}_t^{(i)}}{||\widetilde{\theta}_t^{(i)}||_2}, \mathrm{span}(U_1)\right) \leq \left|\left|\frac{\widetilde{\theta}_t^{(i)}}{||\widetilde{\theta}_t^{(i)}||_2} - \frac{\theta^*}{||\widetilde{\theta}_t^{(i)}||_2}\right|\right|_2 = \frac{\left|\left|\widetilde{\theta}_t^{(i)} - \theta^*\right|\right|_2}{||\widetilde{\theta}_t^{(i)}||_2}, \tag{11}$$

as $\theta^* \in \mathrm{span}(U_1)$ and the fact that $\mathrm{span}(U_1)$ is a subspace implies any constant multiple of θ^* also lies in $\mathrm{span}(U_1)$. This results in the following lemma:

Lemma 4. *For any agent $i \in [N]$ and time $t \in \mathbb{N}$ such that $t \geq t_0$, we have*

$$\mathbb{P}\left(\mathbf{d}\left(\frac{\widetilde{\theta}_t^{(i)}}{||\widetilde{\theta}_t^{(i)}||_2}, \mathrm{span}(U_1)\right) > \frac{\widetilde{\Delta}}{2}\right) \leq 4d \ \exp\left(\frac{S^2}{8d}\right) \ t^{-\frac{\alpha\widetilde{\Delta}^2}{4}}$$

Proof. From equation (11), we have

$$\begin{aligned}
\mathbb{P}\left(\mathbf{d}\left(\frac{\widetilde{\theta}_t^{(i)}}{||\widetilde{\theta}_t^{(i)}||_2}, \mathrm{span}(U_1)\right) > \frac{\widetilde{\Delta}}{2}\right) &\leq \mathbb{P}\left(\frac{\left|\left|\widetilde{\theta}_t^{(i)} - \theta^*\right|\right|_2}{||\widetilde{\theta}_t^{(i)}||_2} > \frac{\widetilde{\Delta}}{2}\right) \\
&= \mathbb{P}\left(\left|\left|\widetilde{\theta}_t^{(i)} - \theta^*\right|\right|_2 > \frac{\widetilde{\Delta}}{2}||\widetilde{\theta}_t^{(i)}||_2\right) \\
&= \mathbb{P}\left(\left|\left|\widetilde{\theta}_t^{(i)} - \theta^*\right|\right|_2 - \frac{\widetilde{\Delta}}{4}||\theta^*||_2 > \frac{\widetilde{\Delta}}{2}\left(||\widetilde{\theta}_t^{(i)}||_2 - \frac{1}{2}||\theta^*||_2\right)\right) \\
&\leq \mathbb{P}\left(\left|\left|\widetilde{\theta}_t^{(i)} - \theta^*\right|\right|_2 > \frac{\widetilde{\Delta}}{4}||\theta^*||_2\right) + \mathbb{P}\left(||\widetilde{\theta}_t^{(i)}||_2 \leq \frac{1}{2}||\theta^*||_2\right).
\end{aligned}$$

In the last step, we apply the union bound, using the observation that for any two random variables X and Y, the event $X > Y$ is contained in the union of two events $X > 0$ and $Y \leq 0$. The first term $\mathbb{P}\left(\left|\left|\widetilde{\theta}_t^{(i)} - \theta^*\right|\right|_2 > \frac{\widetilde{\Delta}}{4}||\theta^*||_2\right)$ can be bounded using Lemma 3 as follows:

$$\begin{aligned}
\mathbb{P}\left(\left|\left|\widetilde{\theta}_t^{(i)} - \theta^*\right|\right|_2 > \frac{\widetilde{\Delta}}{4}||\theta^*||_2\right) &\leq 2d \ \exp\left(\frac{\widetilde{\Delta}^2||\theta^*||_2^2}{32d}\right) \ t^{-\frac{4\alpha}{L^2}\cdot\frac{\widetilde{\Delta}^2||\theta^*||_2^2}{16}} \\
&= 2d \ \exp\left(\frac{\widetilde{\Delta}^2||\theta^*||_2^2}{32d}\right) \ t^{-\frac{\alpha\widetilde{\Delta}^2||\theta^*||_2^2}{4L^2}} \\
&\overset{(a)}{\leq} 2d \ \exp\left(\frac{\widetilde{\Delta}^2 S^2}{32d}\right) \ t^{-\frac{\alpha\widetilde{\Delta}^2}{4}} \\
&\overset{(b)}{\leq} 2d \ \exp\left(\frac{S^2}{8d}\right) \ t^{-\frac{\alpha\widetilde{\Delta}^2}{4}},
\end{aligned}$$

where step (a) is obtained from the assumption $L \leq ||\theta^*||_2 \leq S$ and step (b) follows from the fact that $\widetilde{\Delta} \leq 2$. For bounding the term $\mathbb{P}\left(||\widetilde{\theta}_t^{(i)}||_2 \leq \frac{1}{2}||\theta^*||_2\right)$, notice that

$$\begin{aligned} ||\theta^*||_2 &= ||\theta^* - \widetilde{\theta}_t^{(i)} + \widetilde{\theta}_t^{(i)}||_2 \\ &\leq ||\theta^* - \widetilde{\theta}_t^{(i)}||_2 + ||\widetilde{\theta}_t^{(i)}||_2 \\ &= ||\widetilde{\theta}_t^{(i)} - \theta^*||_2 + ||\widetilde{\theta}_t^{(i)}||_2 \end{aligned}$$

by triangle inequality. Therefore, $||\widetilde{\theta}_t^{(i)} - \theta^*||_2 \geq ||\theta^*||_2 - ||\widetilde{\theta}_t^{(i)}||_2$ and using this observation, we can write:

$$\begin{aligned} \mathbb{P}\left(||\widetilde{\theta}_t^{(i)}||_2 \leq \frac{1}{2}||\theta^*||_2\right) &\leq \mathbb{P}\left(||\widetilde{\theta}_t^{(i)} - \theta^*||_2 \geq \frac{1}{2}||\theta^*||_2\right) \\ &\leq 2d \ \exp\left(\frac{S^2}{8d}\right) \ t^{-\alpha}, \end{aligned}$$

where the above step is obtained by a similar argument as the first term. Hence,

$$\begin{aligned} \mathbb{P}\left(\mathbf{d}\left(\frac{\widetilde{\theta}_t^{(i)}}{||\widetilde{\theta}_T^{(i)}||_2}, \operatorname{span}(U_1)\right) > \frac{\Delta_i}{2}\right) &\leq 2d \ \exp\left(\frac{S^2}{8d}\right) \ t^{-\frac{\alpha\widetilde{\Delta}^2}{4}} + 2d \ \exp\left(\frac{S^2}{8d}\right) \ t^{-\alpha} \\ &\leq 4d \ \exp\left(\frac{S^2}{8d}\right) \ t^{-\frac{\alpha\widetilde{\Delta}^2}{4}} \end{aligned}$$

where $t^{-\frac{\alpha\widetilde{\Delta}^2}{4}} \geq t^{-\alpha}$ as $\widetilde{\Delta} \leq 2$, thus concluding the proof. □

Proposition 3. *For all $N \in \mathbb{N}$ and $\alpha > \frac{12}{\widetilde{\Delta}^2}$,*

$$\mathbb{E}[B_\tau] \leq B_{j^*} + 4\sum_{l \geq \frac{j^*}{2}} \frac{B_{2l}}{B_l^3} + \mathbb{E}[B_{2\widehat{\tau}_{spr}}],$$

where j^* is defined in Theorem 1.

Proof. The proof follows in a similar way as the proof for Proposition 3 in [12]. □

Proposition 4. *The random variable $\widehat{\tau}_{spr}$ is stochastically dominated by $\tau_{spr}^{(G)}$.*

Proof. The proof follows in a similar way as the proof for Proposition 4 in [12]. □

C.3 Proof of Theorem 1

From Theorem 6 and $\delta = \frac{1}{T}$, we can bound $\mathbb{E}[R_{\text{proj},T}]$ as follows:

$$\begin{aligned} \mathbb{E}[R_{\text{proj},T}] = \mathbb{E}&\left[R_{\text{proj},T} \ \mathbf{1}\left(R_{\text{proj},T} \leq \sqrt{8mT\beta_{T,\delta}^2 \log\left(1+\frac{T}{m\lambda}\right)}\right)\right] \\ &+ \mathbb{E}\left[R_{\text{proj},T} \ \mathbf{1}\left(R_{\text{proj},T} > \sqrt{8mT\beta_{T,\delta}^2 \log\left(1+\frac{T}{m\lambda}\right)}\right)\right] \end{aligned}$$

$$\overset{(a)}{\leq} \sqrt{8mT\beta_{T,\delta}^2 \log\left(1+\frac{T}{m\lambda}\right)} \\ + 2S\mathbb{E}\left[\sum_{t=B_\tau+1}^{T} \mathbf{1}\left(R_{\text{proj},T} > \sqrt{8mT\beta_{T,\delta}^2 \log\left(1+\frac{T}{m\lambda}\right)}, \xi_t^i = 0\right)\right],$$

$$\overset{(b)}{\leq} \sqrt{8mT\beta_{T,\delta}^2 \log\left(1+\frac{T}{m\lambda}\right)} \\ + 2S\mathbb{E}\left[\sum_{t=B_\tau+1}^{T} \mathbf{1}\left(R_{\text{proj},T} > \sqrt{8mT\beta_{T,\delta}^2 \log\left(1+\frac{T}{m\lambda}\right)}\right)\right],$$

$$= \sqrt{8mT\beta_{T,\delta}^2 \log\left(1+\frac{T}{m\lambda}\right)} \\ + 2S\sum_{t=B_\tau+1}^{T} \mathbb{E}\left[\mathbf{1}\left(R_{\text{proj},T} > \sqrt{8mT\beta_{T,\delta}^2 \log\left(1+\frac{T}{m\lambda}\right)}\right)\right],$$

$$\begin{aligned} \mathbb{E}[R_{\text{proj},t}] &\overset{(c)}{\leq} \sqrt{8mT\beta_T^2 \log\left(1+\frac{T}{m\lambda}\right)} + 2S.\frac{1}{T}.\sum_{t=1}^{T} 1, \\ &= \sqrt{8mT\beta_T^2 \log\left(1+\frac{T}{m\lambda}\right)} + 2S. \qquad (12) \end{aligned}$$

In step (a), the first sum is trivially bounded by $\sqrt{8mT\beta_T^2 \log\left(1+\frac{T}{m\lambda}\right)}$ and the second sum uses the definition of $R_{\text{proj},T}$ with $w_t \leq 2S$ for all $t \in \mathbb{N}$. Step (b) is a consequence of the fact that the event $\left\{R_{\text{proj},T} > \sqrt{8mT\beta_{T,\delta}^2 \log\left(1+\frac{T}{m\lambda}\right)}, \xi_t^i = 0\right\}$ is contained in the event $\left\{R_{\text{proj},T} > \sqrt{8mT\beta_{T,\delta}^2 \log\left(1+\frac{T}{m\lambda}\right)}\right\}$. Step (c) uses Theorem 6 with $\delta = \frac{1}{T}$.

The proof of Theorem 1 is concluded by substituting the results of Propositions 3 and 4 along with equation (12) into Proposition 2.

Appendix D Analysis of Projected LinUCB

This analysis holds for all time instants $t > B_\tau$, when every agent starts playing from the correct subspace which contains θ^* in the exploit time slots, i.e. $\theta^* = P_1\theta^*$, where $P_1 = U_1U_1^T$ is the projection matrix of correct subspace. As this analysis is valid for all agents, we drop the superscript (i) from all the pertinent variables, where $i \in \mathbb{N}$. Recall that $\bar{V}_t(\lambda) = P_1(V_{t-1} + \lambda I_d)P_1 = U_1\Sigma_tU_1^T$ where $\Sigma_t = U_1^TV_{t-1}U_1 + \lambda I_m$. Define $\tilde{V}_t(\lambda) = \bar{V}_t(\lambda) - \lambda P_1 = P_1V_{t-1}P_1 = U_1\bar{\Sigma}_tU_1^T$, where $\bar{\Sigma}_t = \Sigma_t - \lambda I_m = U_1^TV_{t-1}U_1$.

D.1 Confidence Set Construction and Analysis

The construction of the confidence set is done in a similar way as in [18, 19]. The following theorem characterizes the construction of projected confidence set.

Theorem 5. *Let $\delta \in (0,1)$ and $\beta_{t,\delta} = S\sqrt{\lambda} + \sqrt{2\log\frac{1}{\delta} + m\log\left(1 + \frac{t-1}{\lambda m}\right)}$. Then,*

$$\mathbb{P}\left(\forall t > B_\tau, \theta^* \in \mathcal{C}_t^{(i)}\right) \geq 1 - \delta,$$

where

$$\mathcal{C}_t^{(i)} = \left\{\theta \in \mathbb{R}^d : ||\widehat{\theta}_t^{(i)} - \theta||_{\bar{V}_t(\lambda)} \leq \beta_{t,\delta}\right\}. \tag{13}$$

The proof of Theorem 5 is adapted from the proof of Theorem 8 in [18], which considers a different setting: the subspace in which θ^* lies is unknown and needs to be estimated. The error in estimating the correct subspace appears in the construction of projected confidence set in Theorem 8 of [18]. However, in our setting, agents are aware of the true projection matrices of the subspaces and thus do not need to account for subspace estimation error while constructing confidence sets. This necessitates a different definition of the confidence set given in Equation (13).

Thus, when agents have the correct subspace in their active set after phase τ, they don't have to pay the overhead of recovering the actual subspace from the perturbed action vectors. Hence, Theorem 5 is proved by substituting the estimated projection matrix $\hat{P}_t$ equal to the true projection matrix P_1 in the proof of Theorem 8 in [18] for all $t > B_\tau$.

D.2 Regret Analysis

Before bounding the regret, let us set some notation here. We have

$$A_t \in \arg\max_{a\in\mathcal{A}_t} \max_{\theta\in\mathcal{C}_t^{(i)}} \langle\theta, P_1 a\rangle.$$

Let $A_t^* = \arg\max_{a\in\mathcal{A}_t}\langle\theta^*, a\rangle$ and $\bar{\theta}_t \in \mathcal{C}_t^{(i)}$ such that $\bar{\theta}_t = \arg\max_{\theta\in\mathcal{C}_t^{(i)}}\langle\theta, P_1 A_t\rangle$. The following theorem characterizes the regret after every agent has the right subspace.

Theorem 6. *For any time $T > B_\tau$, with probability at least $1-\delta$, the regret of Projected LinUCB satisfies*

$$R_{\text{proj},T} \leq \sqrt{8mT\beta_{T,\delta}^2 \log\left(1 + \frac{T}{m\lambda}\right)}.$$

Proof. The proof of Theorem 6 is contingent on the following lemma. A similar lemma appears as Lemma 19.4 in [19].

Lemma 7. *Let $A_1, \cdots, A_T$ be the sequence of action vectors played up to and including time T, where $T > B_\tau$. Then,*

$$\sum_{t=B_\tau+1}^{T} \min\left\{1, ||A_t||_{V_t(\lambda)^\dagger}^2\right\} \mathbf{1}(\xi_t^{(i)} = 0) \leq 2\log\left(\frac{\det(\Sigma_{T+1})}{\det(\Sigma_{B_\tau+1})}\right).$$

Proof. For any $u \geq 0$, we have $\min\{1, u\} \leq 2\log(1+u)$. Therefore,

$$\sum_{t=B_\tau+1}^{T} \min\left\{1, ||A_t||^2_{\bar{V}_t(\lambda)^\dagger}\right\} \mathbf{1}(\xi_t^{(i)} = 0) \leq \sum_{t=B_\tau+1}^{T} \min\left\{1, ||A_t||^2_{\bar{V}_t(\lambda)^\dagger}\right\}$$
$$\leq 2\sum_{t=B_\tau+1}^{T} \log\left(1 + ||A_t||^2_{\bar{V}_t(\lambda)^\dagger}\right). \tag{14}$$

The remainder of the proof is identical to the proof of [19], Lemma 19.4, except that we use the recursive update of $\det(\Sigma_t)$ instead of $\det(V_{t-1})$. □

We are in a position to complete the proof of Theorem 6. Using the fact that $\theta^* \in \mathcal{C}_t^{(i)}$ and from the algorithm definitions, the following chain of inequalities is true for all $t > B_\tau$ whenever $\xi_t^{(i)} = 0$:

$$\langle \theta^*, A_t^* \rangle = \langle P_1\theta^*, A_t^* \rangle = \langle \theta^*, P_1 A_t^* \rangle \leq \max_{\theta \in \mathcal{C}_t^{(i)}} \langle \theta, P_1 A_t^* \rangle \leq \langle \bar{\theta}_t, P_1 A_t \rangle. \tag{15}$$

Thus, for all $t > B_\tau$, whenever $\xi_t^{(i)} = 0$,

$$w_t^{(i)} \leq 2\beta_{t,\delta}||A_t||_{\bar{V}_t(\lambda)^\dagger}, \tag{16}$$

which is shown in a similar way as bounding r_t in [19], Theorem 19.2. However, we additionally use the fact that $\theta^* = P_1\theta^*$ and $P_1^2 = P_1$.
While proving Proposition 1, we showed that $w_t^{(i)} \leq 2S$ for all $t \in \mathbb{N}$. Combining this with equation (16) and $\beta_{T,\delta} \geq S$ results in

$$w_t^{(i)}\mathbf{1}(\xi_t^{(i)} = 0) \leq 2\min\left\{S, \beta_{t,\delta}||A_t||_{\bar{V}_t(\lambda)^\dagger}\right\} \mathbf{1}(\xi_t^{(i)} = 0)$$
$$\leq 2\beta_{T,\delta}\min\left\{1, ||A_t||_{\bar{V}_t(\lambda)^\dagger}\right\} \mathbf{1}(\xi_t^{(i)} = 0). \tag{17}$$

Therefore, the cumulative regret for the projected LinUCB after time $T > B_\tau$ can be bounded as follows: let $\mathbb{I}$ denote a column vector with the $T - B_\tau$ entries $\{\mathbf{1}(\xi_t^{(i)} = 0)\}_{t=B_\tau+1}^{T}$, and $\mathbf{R}$ be a column vector containing the elements $w_{B_\tau+1}, \cdots, w_T$. Then,

$$\begin{aligned}
R_{\text{proj},T} &= \sum_{t=B_\tau+1}^{T} w_t^{(i)}\mathbf{1}(\xi_t^{(i)} = 0) \\
&= \mathbb{I}^T\mathbf{R} \\
&\overset{(a)}{\leq} \sqrt{T - B_\tau}\sqrt{\sum_{t=B_\tau+1}^{T} (w_t^{(i)})^2\mathbf{1}(\xi_t^{(i)} = 0)} \\
&\overset{(b)}{\leq} \sqrt{T}\sqrt{4\beta_{T,\delta}^2 \sum_{t=B_\tau+1}^{T} \min\left\{1, ||A_t||^2_{\bar{V}_t(\lambda)^\dagger}\right\} \mathbf{1}(\xi_t^{(i)} = 0)} \\
&\overset{(c)}{\leq} \sqrt{8T\beta_{T,\delta}^2 \log\left(\frac{\det(\Sigma_{T+1})}{\det(\Sigma_{B_\tau+1})}\right)}
\end{aligned}$$

$$
\begin{aligned}
&\overset{(d)}{\leq} \sqrt{8T\beta_{T,\delta}^2 \log\left(\frac{\det(\Sigma_{T+1})}{\det(\Sigma_1)}\right)} \\
&\overset{(e)}{\leq} \sqrt{8mT\beta_{T,\delta}^2 \log\left(1 + \frac{T}{m\lambda}\right)}.
\end{aligned}
$$

Step (a) follows from Cauchy-Schwarz inequality. In step (b), we use equation (17). Step (c) is obtained by application of Lemma 7. Step (d) uses the observation that $\det(\Sigma_t)$ is increasing in t. Step (e) results from the fact that $\det(\Sigma_1) = \det(\Lambda) = \lambda^m$ and $\det(\Sigma_t) \leq \left(\lambda + \frac{t-1}{m}\right)^m$ (which follows from Lemma 11 in [18]), thus completing the proof. □