Optimal Regret Bounds for Collaborative Learning in Bandits

Amitis Shidani   AMITIS.SHIDANI@STATS.OX.AC.UK
University of Oxford

Sattar Vakili   SV388@CORNELL.EDU
MediaTek Research

Editors: Claire Verdane and Daniel Hsu

Abstract

We consider regret minimization in a general collaborative multi-agent multi-armed bandit model, in which each agent faces a finite set of arms and may communicate with other agents through a central controller. The optimal arm for each agent in this model is the arm with the largest expected *mixed* reward, where the mixed reward of each arm is a weighted average of its rewards across all agents, making communication among agents crucial. While near-optimal sample complexities for best arm identification are known under this collaborative model, the question of optimal regret remains open. In this work, we address this problem and propose the first algorithm with order optimal regret bounds under this collaborative bandit model. Furthermore, we show that only a small constant number of expected communication rounds is needed.

Keywords: Multi-agent multi-armed bandit, Collaborative learning

1. Introduction

In the classic multi-armed bandit problem (Lai and Robbins, 1985; Auer et al., 2002) an agent sequentially interacts with a set of arms. At each time, the agent selects one of the arms and receives the corresponding reward drawn from an unknown distribution. The goal of the agent is to minimize *regret* defined as the total loss in the rewards compared to always selecting the arm with the highest expected reward. The trade-off between *exploration* —selecting each arm several times to learn the unknown distributions— and *exploitation* —capitalizing on the information gathered so far to select the arm with the highest reward— is central to designing efficient bandit algorithms that minimize the regret. The bandit problem has also been studied under a different setting, referred to as the best arm identification, where the goal of the agent is to identify the optimal arm with high probability with as a small number of samples as possible (Even-Dar et al., 2006; Jamieson et al., 2014).

There is an increasing interest in collaborative learning due to modern applications in traffic routing, communication systems, social dynamics, federated learning, and the Internet of Things (IoT). We consider a statistical learning point of view to collaborative learning through a general multi-agent multi-armed bandit model. This collaborative bandit model was proposed in Réda et al. (2022), which itself was a generalization of the federated learning with the personalization model introduced in Shi et al. (2021). In this model, M agents interact with K arms. An agent $m \in [M]$ observes a *local* reward $X_{k,m}(t)$ for playing arm $k \in [K]$ at time $t \in [T]$. The actual reward that the agent receives, however, is a *mixed* reward $X'_{k,m}(t)$ that is a weighted average of local rewards over all M agents for the same arm k. A special case of this collaborative bandit model limited to particular weights was introduced in Shi et al. (2021) for a federated learning setting. A detailed problem formulation highlighting the special case of Shi et al. (2021) is given in Section 2.

The interdependence among agents due to mixed rewards naturally makes *communication* among agents crucial. In fact, it is easy to design problem instances where without communication a trivial linear in T regret is inevitable. The agents in the collaborative bandit model are allowed to communicate through a central controller or server. Specifically, the agents can send information (*e.g.*, empirical means of past local observations) to the server, which will broadcast this information to all other agents. An ideal algorithm obtains a good performance in terms of regret with few communication rounds.

The realistic scenarios of collaborative bandit include applications to recommendation systems, for which one may want to go beyond uniformly personalized learning; favoring the local rewards of agent m over other agents' observations in the identification of the optimal arm (Shi et al., 2021). The setting is also suitable for adaptive clinical trials on K therapies, run by M teams who have access to different sub-populations of patients. In this context, each sub-population typically aims to find its best treatment (see details in Réda et al., 2022). Similarly, this setting has applications in routing and communication systems (Shi et al., 2021).

1.1. Related Work and Contribution

The authors in Shi et al. (2021) provided a conjecture on the lower bound on regret under the special case of federated learning with personalization[1]. They also introduced *personalized federated upper confidence bound* (PF-UCB) algorithm that achieves an $\mathcal{O}(\log(T))$ regret where the implied constant in the $\mathcal{O}$ notation is suboptimal in terms of problem parameters. The authors in Réda et al. (2022) primarily studied best arm identification under the collaborative bandit model. They proved an $\Omega(s^* \log(\frac{1}{\delta}))$ information-theoretic lower bound on the sample complexity for the event that, with probability at least $1 - \delta$, each agent identifies their own best arm. In this lower bound, s^* is the solution to a constrained optimization problem involving the weights and the mixed reward gaps. In addition, they used the knowledge of the lower bound to propose *weighted collaborative phased elimination for best arm identification* (W-CPE-BAI) algorithm, achieving near-optimal performance $\mathcal{O}(s^* \log(\frac{1}{\Delta'_{\min}}) \log(\frac{1}{\delta}))$, where $\Delta'_{\min}$ is the smallest gap in the expected mixed rewards of the best arm and a suboptimal arm across all agents.

In the context of our work, Réda et al. (2022) established an $\Omega(c^* \log(T))$ lower bound on regret that also confirmed the conjecture in Shi et al. (2021). Similar to the previous case, in this lower bound, c^* is the solution to a constrained optimization problem involving the weights and the mixed reward gaps (see Lemma 1 for details). However, designing an algorithm achieving near-optimal regret was proven to be a challenging task and remained as an open problem. The difficulty arises from the structure of the constrained optimization problem associated with c^*, which is essentially different from the one associated with s^*. As noted in Réda et al. (2022), a standard analysis of phased elimination algorithms would lead to an $\mathcal{O}(\frac{c^*}{\Delta'_{\min}} \log(T))$ regret bound. In this case, the $\frac{1}{\Delta'_{\min}}$ gap is exponentially worse than the $\log(\frac{1}{\Delta'_{\min}})$ gap in the previous case, and can result in poor performance guarantees.

In this work, we present a novel regret minimization algorithm called *collaborative double exploration* (CExp[2]) and demonstrate that it achieves an $\mathcal{O}(c^* \log(T))$ regret, with only absolute constants hidden in the $\mathcal{O}$ notation. Our regret guarantees for this algorithm match the problem-

---

1. The term personalization here emphasizes the property of this collaborative learning model that the mixed reward distributions are agent specific, in contrast to settings where all agents share the same global reward model (see, *e.g.*, Tao et al., 2019; Shi and Shen, 2021).

specific regret lower bound given in Réda et al. (2022). CExp^2 comprises a sub-logarithmic initial round-robin exploration phase, followed by a guided exploration phase that is designed based on the constrained optimization problem associated with c^* (hence, the name double exploration). We prove that by utilizing the initial exploration phase, the guided exploration phase can gather enough information to efficiently exploit the best arm, resulting in an optimally bounded total regret. Moreover, we show that CExp^2 achieves the order optimal regret bound with only a constant number of expected communication rounds. Specifically, the number of communication rounds is $\mathcal{O}(\log(\frac{1}{\Delta_{\min}}))$. Furthermore, with probability at least $1 - \frac{1}{\log(T)}$, there are only 2 communication rounds in CExp^2, implying that the expected number of communication rounds is only 2 as T grows large.

1.2. Paper Structure

In Section 2, we introduce the notation and provide the details of the collaborative bandit model. We also overview the background on regret lower bound and confidence intervals in this setting. In Section 3, we describe our collaborative bandit algorithm CExp^2. The analysis of the algorithms is given in Section 4, while some details are deferred to the appendix. Further related work is discussed in Appendix A.

2. Problem formulation

In the collaborative bandit model, there are K arms and M agents. At each sequential time $t \in [T]$, $T \in \mathbb{N}$, each agent $m \in [M]$ is allowed to select an arm $k \in [K]$ and observe the corresponding local reward $X_{k,m}(t)$ independent across k, m and independent and identically distributed across t. We use the notation $[N] = \{1, 2, \ldots, N\}$, for $N \in \mathbb{N}$, throughout the paper. The random local reward $X_{k,m}(t)$ is distributed according to a σ-sub-Gaussian distribution $\nu_{k,m}$. We say that the distribution of a random variable X is σ-sub-Gaussian when for all $\eta \in \mathbb{R}$, $\mathbb{E}[\exp(\eta(X - \mathbb{E}[X]))] \leq \exp(\frac{\eta^2 \sigma^2}{2})$, for some $\sigma > 0$. This property allows us to use statistical confidence bounds on the mean rewards in algorithm design. We use $\mu_{k,m} = \mathbb{E}[X_{k,m}(t)]$ to denote the expected value of the rewards. The agents are interested in the mixed rewards of the arms defined using a fixed and known weight matrix $W = (w_{n,m})_{n,m} \in [0,1]^{M \times M}$, satisfying $\sum_{n \in [M]} w_{n,m} = 1$ for all $m \in [M]$, i.e. the columns of W add up to 1. The weight $w_{n,m}$ represents the importance weight of agent n for agent m. Specifically, the mixed reward of agent m from arm k at time t is defined as a weighted average of the local rewards over all agents

$$X'_{k,m}(t) = \sum_{n=1}^{M} w_{n,m} X_{k,n}(t), \tag{1}$$

and its expectation, referred to as expected mixed reward, is denoted by $\mu'_{k,m} = \sum_{n=1}^{M} w_{n,m} \mu_{k,n}$. We use $k_m^* = \arg\max_{k \in [K]} \mu'_{k,m}$ to denote the arm with the largest expected mixed reward for agent m, assumed unique. Also, for all m and k, we use $\Delta'_{k,m} = \mu'_{k_m^*,m} - \mu'_{k,m}$ to denote the gap between the expected mixed reward of arm k and the arm with the highest mean mixed reward, for agent m. In addition, for all m and $k \neq k_m^*$, we define $\tilde{\Delta}'_{k,m} = \Delta'_{k,m}$. In the case of the best arm k_m^* of each agent m, we define $\tilde{\Delta}'_{k_m^*,m} = \min_{k \neq k_m^*} \Delta'_{k,m}$, equal to the gap for the second best arm. We introduce the notation $\tilde{\Delta}'$ that is different from Δ' only in the case of the best arm, for the

convenience of presenting our results and proofs. We also use the notations $\Delta'_{\min} = \min_{k,m} \tilde{\Delta}'_{k,m}$ and $\Delta'_{\max} = \max_{k,m} \tilde{\Delta}'_{k,m}$ to denote the smallest and largest gaps for agents, respectively.

At each time t, the agents are allowed to communicate information to a server, in the form of empirical mean rewards of local observations (similar to Réda et al., 2022; Shi et al., 2021). The information is then broadcast to all other agents. We aim at designing efficient algorithms with as few as possible communication rounds. A collaborative learning bandit algorithm $\pi = \{\pi_m\}_{m \in [M]}$ is a specification of the arm $\pi_m(t) \in [K]$ selected by agent m at time t. The performance of π is measured in terms of its collaborative regret defined as follows:

$$\mathcal{R}(T) = \mathbb{E}\left[\sum_{m=1}^{M} \sum_{t=1}^{T} (\mu'_{k_m^*,m} - \mu'_{\pi_m(t),m})\right]. \tag{2}$$

Shi et al. (2021) considered a special case of the collaborative bandit model with $w_{m,m} = \alpha + \frac{1-\alpha}{M}$ and $w_{n,m} = \frac{1-\alpha}{M}$ for $n \neq m$, for some $\alpha \in [0, 1]$. We note that when $w_{m,m} = 1$ for all $m \in [M]$ the problem reduces to the degenerated case of M independent single-agent bandit problems. Moreover, when $w_{n,m} = \frac{1}{M}$ for all $n, m \in [M]$, the problem reduces to another special case where all agents are interested in a global average of the rewards across all agents considered in Shi and Shen (2021).

We use the following notations for the sample means. The sample mean of local observations of arm k by agent m is denoted by

$$\hat{\mu}_{k,m}(t) = \frac{1}{\tau_{k,m}(t)} \sum_{s=1}^{t} X_{k,m}(s)\mathbf{1}\{\pi_m(s) = k\}, \tag{3}$$

where $\tau_{k,m}(t)$ denotes the number of times arm k was selected by agent m by the end of round t, and $\mathbf{1}$ is the indicator function, i.e. for an event $\mathcal{F}$, $\mathbf{1}\{\mathcal{F}\} = 1$ if $\mathcal{F}$ holds true, and $\mathbf{1}\{\mathcal{F}\} = 0$, otherwise. The sample mean of the mixed reward is denoted by

$$\hat{\mu}'_{k,m}(t) = \sum_{n=1}^{K} w_{n,m}\hat{\mu}_{k,n}(t). \tag{4}$$

2.1. Regret Lower Bound

Réda et al. (2022) established a regret lower bound for the collaborative bandit model that also confirmed the conjecture in Shi et al. (2021) for their special case of weights. This result is formally stated in the following lemma. The lower bound holds for normal distributions over the rewards. Furthermore, the agents are allowed to communicate after each round.

Lemma 1 *Any uniformly efficient collaborative bandit algorithm in which all agents communicate after each round satisfies* $\liminf_{T \to \infty} \frac{\mathcal{R}(T)}{\log(T)} \geq c^*$, *where*

$$c^* = \min_{q \in (\mathbb{R}^+)^{K \times M}} \sum_{m \in [M]} \sum_{k \in [K], k \neq k_m^*} q_{k,m} \tilde{\Delta}'_{k,m}, \tag{5}$$

$$\text{subject to: } \forall m \in [M], \forall k \in [K], \sum_{\{n \in [M]: k_n^* \neq k\}} \frac{w_{n,m}^2}{q_{k,n}} \leq \frac{(\tilde{\Delta}'_{k,m})^2}{2}. \tag{6}$$

The proof is based on the change of measure technique used in proving regret lower bounds in the single-agent bandit problem (see, Lai and Robbins, 1985; Salomon et al., 2011). A uniformly efficient algorithm satisfies $\mathcal{R}(T) = o(T^\gamma)$ for any $\gamma \in (0, 1)$ and for all possible distribution instances $\nu = \{\nu_{k,m}\}_{k \in [K], m \in [M]}$. The lower bound complexity term c^* can be intuitively understood as the value of regret divided by $\log(T)$ for the optimum allocation of the arm plays, guaranteeing sufficiently small confidence intervals for all mean mixed rewards. The intuition is based on the observation that the term $\sum_{\{n \in [M] : k_n^* \neq k\}} \frac{w_{n,m}^2}{q_{k,n}}$ in the constraints in Lemma 1 appears in the confidence interval width of the mean mixed rewards (see Equation 11).

A relaxed complexity term was given in Réda et al. (2022), which does not require the knowledge of the best arm and is easier to work with in algorithm design. Let us define an oracle for the following optimization problem related to the lower bound, which will also be used in the algorithm.

Definition 2 *For any $\Delta \in (\mathbb{R}^+)^{K \times M}$, the oracle $\mathcal{P}(\Delta)$ is the solution to the following constrained optimization problem.*

$$\underset{q \in (\mathbb{R}^+)^{K \times M}}{\arg\min} \sum_{k \in [K], m \in [M]} q_{k,m} \Delta_{k,m}, \tag{7}$$

$$\textit{subject to:} \ \forall m \in [M], \forall k \in [K], \ \sum_{n \in [M]} \frac{w_{n,m}^2}{q_{k,n}} \leq \frac{\Delta_{k,m}^2}{2}. \tag{8}$$

Réda et al. (2022) proved that the complexity term c^* in the regret lower bound in Lemma 1 and a complexity term $\tilde{c}^*$ obtained from the relaxed oracle $\mathcal{P}$ are the same up to small constant factors.

Lemma 3 *Let $(q_{k,m}^*)_{k,m}$ and $\tilde{c}^*$ be as $(q_{k,m}^*)_{k,m} = \mathcal{P}(\tilde{\Delta}')$ and $\tilde{c}^* = \sum_{k \in [K], m \in [M]} q_{k,m}^* \tilde{\Delta}'_{k,m}$. Then it holds that*

$$c^* \leq \tilde{c}^* \leq 4c^*. \tag{9}$$

We thus have $\frac{\tilde{c}^*}{4} \leq c^*$, which implies, under the setting of Lemma 1,

$$\liminf_{T \to \infty} \frac{\mathcal{R}(T)}{\log(T)} \geq \frac{\tilde{c}^*}{4}. \tag{10}$$

2.2. Confidence Intervals

Confidence intervals serve as essential building blocks in the design of efficient bandit algorithms. For any $k \in [K], m \in [M], \delta \in (0, 1)$, at time t, we define:

$$\Omega_{k,m}^\delta(t) := \sqrt{\beta_\delta(\tau_{k,\cdot}(t)) \sum_{n=1}^{M} \frac{w_{n,m}^2}{\tau_{k,n}(t)}} \tag{11}$$

where $\beta_\delta : (\mathbb{R}^+)^M \to \mathbb{R}^+$ is a threshold function defined as:

$$\beta_\delta(N) := 2\left(g_M\left(\frac{\delta}{KM}\right) + 2\sum_{m=1}^{M} \log(4 + \log(N_m))\right) \tag{12}$$

5

where g_M is some non-explicit function, defined in Kaufmann and Koolen (2021) that satisfies:

$$g_M(\delta) \simeq \log(\frac{1}{\delta}) + M \log\log(\frac{1}{\delta}) \tag{13}$$

We then have the following, uniform in time, confidence intervals for mixed mean rewards.

Lemma 4 *(Kaufmann and Koolen, 2021; Réda et al., 2022) For any $\delta \in (0,1)$, the event $\mathcal{F}$ defined as follows holds with probability larger than $1 - \delta$.*

$$\mathcal{F} \coloneqq \left\{ \forall t \in \mathbb{N}, \forall k \in [K], m \in [M]; \ |\hat{\mu}'_{k,m}(t) - \mu'_{k,m}| \le \Omega^{\delta}_{k,m}(t) \right\} \tag{14}$$

Lemma 4 is proven in (Kaufmann and Koolen, 2021, Proposition 24) by constructing mixture martingales resulting in tight confidence intervals for certain functions of the means of the arms. We note that, in our setting, the standard Chernoff-Hoeffding type inequalities cannot be directly applied, as we need to create confidence intervals for $\mu'_{k,m}$ utilizing observations of $X_{k,n}$, $n \in [M]$, in contrast to directly observing samples form $X'_{k,m}$.

3. Collaborative Learning in Bandits: A Regret Minimization Algorithm

In this section, we present CExp^2, a collaborative bandit algorithm for regret minimization. The existing work has primarily focused on phased elimination algorithms, which seems a natural choice in settings like collaborative bandits. These algorithm designs are motivated by the lower bound. During each phase, the lower bound oracle is employed to obtain a good allocation of the arm plays based on the current estimate of the problem parameters (here the gaps $\tilde{\Delta}'_{k,m}$).

The allocation obtained by the oracle is then used for a guided exploration of arms in the next phase. Suboptimal arms are identified and removed from the set of candidate optimal arms at the end of each phase. In the best arm identification setting, Réda et al. (2022) introduced W-CPE-BAI, a phased elimination algorithm that achieves near-optimal sample complexity. The performance of a similar algorithm in the regret minimization setting, however, does not match the regret lower bound, as discussed in Réda et al. (2021).

A similar shortcoming is also observed for PF-UCB introduced in Shi et al. (2021) for regret minimization in the special weights setting. The reason for this drastic difference between best arm identification and regret minimization settings is the subtle difference in the constrained optimization problems associated with the lower bound complexities in each setting. In particular, we note that the unknown gaps $\tilde{\Delta}'_{k,m}$ appear both in the objective as well as in the constraints in the Oracle $\mathcal{P}$ given in Definition 2.

This is in contrast to the constrained optimization problem associated with the lower bound on the sample complexity where the unknown gap parameters appear only in the constraints (see, Réda et al., 2022, Theorem 1). This subtle difference breaks the analysis of standard phased elimination algorithms for regret minimization in the collaborative bandit model and leaves a significant gap in the performance (Réda et al., 2022). To address this problem, we introduce CExp^2(see Algorithm 1).

3.1. Description of CExp^2 Algorithm

CExp^2 comprises two exploration phases followed by a third phase in which all agents either exploit —by playing the arm with the highest sample mean of mixed reward— or switch to any policy π' with an $\mathcal{O}(\log(T))$ regret.

An example of policy π' with logarithmic regret is a variation of W-CPE-BAI adopted to the regret minimization setting. We formally give this variation, W-CPE-Reg, and the proof that it has an $\mathcal{O}(\frac{c^*}{\Delta'_{\min}}\log(T))$ regret using an $\mathcal{O}(\log(\frac{1}{\Delta'_{\min}}))$ communication rounds, in Section 3.2.

Initial Exploration: In the initial sub-logarithmic exploration phase of CExp^2, each agent $m \in [M]$ plays each arm $k \in [K]$ for $\tau_1 = \lceil\sqrt{\log(T)}\rceil$ times. The length of the initial exploration phase is denoted by

$$T_{\text{ie}} = MK\tau_1. \tag{15}$$

At the end of the initial exploration phase, each agent m computes their local sample mean reward $\hat{\mu}_{k,m}(T_{\text{ie}})$ for each arm k. The sample means of local rewards are shared through the server with other agents (the first round of communication), which are used to compute the sample means of mixed rewards $\hat{\mu}'_{k,m}(T_{\text{ie}})$, and estimates of gaps in mixed rewards $\hat{\Delta}'_{k,m}(T_{\text{ie}})$.

Guided Exploration: CExp^2 utilizes the gap estimates in mixed rewards and Oracle $\mathcal{P}$ to obtain an allocation of the arm plays $(\tau^{\text{ge}}_{k,m})_{k,m}$ for the guided exploration phase, and ensures each agent m plays each arm k until reaching $\tau^{\text{ge}}_{k,m}$. To avoid extreme cases of allocation, which may lead to large regret, the gap estimates are first projected onto the $(\frac{1}{\log\log(T)}, \log\log(T))$ interval. Specifically, we introduce, $\forall k, m$,

$$\hat{\Delta}^{\text{ge}}_{k,m} = \min\left\{\max\left\{\hat{\Delta}'_{k,m}(T_{\text{ie}}), \frac{1}{\log\log(T)}\right\}, \ \log\log(T)\right\}. \tag{16}$$

Then, the allocation design $q = \mathcal{P}(\hat{\Delta}^{\text{ge}})$ is obtained using $\hat{\Delta}^{\text{ge}}$. The allocation of the arm plays in the guided exploration phase is then set to $\lceil 18q^{\text{ge}}_{k,m}B(T)\rceil$, where $B(T) = \beta_{\delta'}(\tau_{k,\cdot})$, with $\delta' = \frac{1}{T}$ and $\tau_{k,n} = T, \forall k, n$, is an upper bound on $\beta_{\delta'}(\tau_{k,\cdot})$. As given in Section 2.2,

$$B(T) \simeq 2\log(KMT) + 2M\log\log(KMT) + 4M\log(4 + \log(T)). \tag{17}$$

Thus, we have $B(T) = \mathcal{O}(\log(KMT))$ hiding a small absolute constant 2 for large T. In the guided exploration phase, each agent m plays each arm k to reach a total number of $\tau^{\text{ge}}_{k,m} = \max\{\tau_1, \lceil 18q^{\text{ge}}_{k,m}B(T)\rceil\}$. Let T_{ge} denote the end of guided exploration phase. At this time, the sample means of local rewards are shared through the server with other agents (the second round of communication). Those are used to compute $\hat{\mu}'_{k,m}(T_{\text{ge}})$ and $\hat{\Delta}'_{k,m}(T_{\text{ge}})$ similar to the end of initial exploration phase.

Exploitation/ Switch to π': At the end of the guided exploration phase, CExp^2 either exploits —by playing the arm with the highest sample mean of mixed reward— or switches to π', based on the following condition, with $\delta' = \frac{1}{T}$,

$$\mathcal{C} = \left\{\forall k \in [K], m \in [M], \ \Omega^{\delta'}_{k,m}(T) < \frac{\hat{\Delta}'_{k,m}(T_{\text{ge}})}{2}\right\}. \tag{18}$$

If $\mathcal{C}$ does not hold, CExp^2 switches to π'. Otherwise, at the end of the guided exploration phase, each agent m identifies the arm with the largest empirical mean mixed reward

$$\hat{k}_m = \arg\max_{k\in[K]} \hat{\mu}'_{k,m}(T_{\text{ge}}), \tag{19}$$

7

and plays this arm for the remaining time. A detailed pseudo-code for CExp^2 is given in Algorithm 1. The regret performance and the number of communication rounds of CExp^2 are presented in the following theorem and lemma.

Algorithm 1 CExp^2, regret minimization in collaborative bandit model.

Input: M agents, K arms, weight matrix W, π'
Initial Exploration:
Each agent m, plays each arm k for $\tau_1 = \lceil \sqrt{\log(T)} \rceil$ times.
Agents communicate the sample mean of their local observations.
Agents obtain $(\hat{\Delta}'_{k,m}(T_{\text{ie}}))_{k,m}$.
Guided Exploration:
$\hat{\Delta}^{\text{ge}}_{k,m} \leftarrow \min\left\{\max\{\hat{\Delta}'_{k,m}(T_{\text{ie}}), \frac{1}{\log\log(T)}\}, \log\log(T)\right\}, \forall k \in [K], m \in [M]$
Let $(q^{\text{ge}}_{k,m})_{k,m} = \mathcal{P}(\hat{\Delta}^{\text{ge}})$
Each agent $m \in [M]$ plays each action $k \in [K]$ until it reaches to a total number of

$$\tau^{\text{ge}}_{k,m} = \max\{\tau_1, \lceil 18 q^{\text{ge}}_{k,m} B(T) \rceil\}$$

Agents communicate the sample mean of their local observations
Agents update $(\hat{\Delta}'_{k,m}(T_{\text{ge}}))_{k,m}$
Exploitation/Switch to π':
if $\forall k \in [K], m \in [M], \Omega^{\delta'}_{k,m}(T) < \frac{\hat{\Delta}'_{k,m}(T_{\text{ge}})}{2}$ **then**
    Each agent m plays $\arg\max_{k \in [K]} \hat{\mu}'_{k,m}(t)$ until T
else
    Reset all data and play based on π' until T
end if

Theorem 5 *Consider the collaborative bandit model described in Section 2. Consider CExp^2 given in Algorithm 1. Let π' be W-CPE-Reg. The regret performance of CExp^2 satisfies, for some $T_0 \in \mathbb{N}$, and for all $T \geq T_0$,*

$$\mathcal{R}(T) = \mathcal{O}\left(c^* \log(T) + \frac{(\Delta'_{\max})^2}{\Delta'_{\min}} (\log\log(T))^4\right) \tag{20}$$

The $\mathcal{O}$ notation hides only absolute constants which are specified in the analysis section. The value of the constant T_0 is also given in the analysis section. The $\mathcal{O}(c^* \log(T))$ regret bound given in Theorem 5 proves that the performance of CExp^2 matches the lower bound up to an absolute constant factor. As mentioned, CExp^2 also enjoys very few expected number of communication rounds, as given in the following lemma.

Lemma 6 *Under the setting of Theorem 5, the expected number of communication rounds in CExp^2 is bounded by*

$$2 + \lceil \log_2(\frac{8}{\Delta'_{\min}}) \rceil / \log(T). \tag{21}$$

In addition, with probability at least $1 - \frac{1}{\log(T)}$ the number of communication rounds in CExp^2 is exactly 2.

According to Lemma 6, as T grows large, CExp^2 requires only 2 communication rounds. In comparison, W-CPE-BAI required $\lceil \log_2(\frac{8}{\Delta'_{\min}}) \rceil$ communication rounds. The proof is provided in Appendix B.

3.2. An Example of Algorithms with Logarithmic Regret (π')

For completeness and correctness of our results, we provide a concrete example of a collaborative bandit algorithm with logarithmic regret (not necessarily with optimal constant in front of log term), that can be used as π' in CExp^2. This algorithm, referred to as W-CPE-Reg, is similar to W-CPE-BAI (Réda et al., 2022), but adopted to the regret minimization setting. W-CPE-BAI was originally introduced for best arm identification, where the goal of each agent m is to return their candidate best arm $\hat{k}_m$, with as few samples from arms as possible, satisfying, for a given $\delta \in (0, 1)$,

$$\mathbb{P}\left[\forall m \in [M], \hat{k}_m = k_m^* \right] \geq 1 - \delta. \tag{22}$$

We run W-CPE-BAI with $\delta = \frac{1}{T}$ until it stops. Then, each agent m plays $\hat{k}_m$ for the remaining time. Referring to this procedure as W-CPE-Reg, we have the following result on its regret performance and communication rounds. The proof is provided in Appendix C.

Lemma 7 *Consider the collaborative bandit model described in Section 2. The number of communication rounds is at most $\lceil \log_2(\frac{8}{\Delta'_{\min}}) \rceil$. Moreover, the regret performance of W-CPE-Reg described above satisfies*

$$\mathcal{R}(T) \leq \frac{128c^*}{\Delta'_{\min}} \log_2(\frac{8}{\Delta'_{\min}}) \log(T) + o(\log(T)). \tag{23}$$

4. Analysis

In this section, we overview the main steps and challenges in the proof of Theorem 5, while deferring some details and proof of lemmas to the appendix. The regret of CExp^2 is partitioned into three disjoint parts: the regret $\mathcal{R}_{\text{ie}}$ during the initial exploration phase $t \in \mathcal{T}_{\text{ie}}$, the regret $\mathcal{R}_{\text{ge}}$ during the guided exploration phase $t \in \mathcal{T}_{\text{ge}}$, and the regret $\mathcal{R}_{\text{ex}}$ during the third phase $t \in \mathcal{T}_{\text{ex}}$ (exploitation or switch to π'), which we refer to as exploitation for brevity. We note that

$$\mathcal{R}(T) = \mathcal{R}_{\text{ie}} + \mathcal{R}_{\text{ge}} + \mathcal{R}_{\text{ex}}. \tag{24}$$

Let us first introduce two high probability events, $\mathcal{E}_T$ and $\mathcal{B}_T$, on the sample means of mixed rewards with $\delta = \frac{1}{\log(T)}$ and $\delta' = \frac{1}{T}$.

$$\mathcal{E}_T = \left\{ \forall t \leq T, \forall k \in [K], m \in [M]; \|\hat{\mu}'_{k,m}(t) - \mu'_{k,m}\| \leq \Omega^{\delta}_{k,m}(t) \right\}, \tag{25}$$

$$\mathcal{B}_T = \left\{ \forall t \leq T, \forall k \in [K], m \in [M]; \|\hat{\mu}'_{k,m}(t) - \mu'_{k,m}\| \leq \Omega^{\delta'}_{k,m}(t) \right\}. \tag{26}$$

Based on Lemma 4, we have $\mathbb{P}[\mathcal{E}_T] \geq 1 - \frac{1}{\log(T)}$ and $\mathbb{P}[\mathcal{B}_T] \geq 1 - \frac{1}{T}$. These two events are used in the analysis of $\mathcal{R}_{\mathrm{ge}}$ and $\mathcal{R}_{ex}$. Also, for the sake of brevity, we use the notations $\delta = \frac{1}{\log(T)}$ and $\delta' = \frac{1}{T}$ throughout the proof. We now bound $\mathcal{R}_{\mathrm{ie}}$, $\mathcal{R}_{\mathrm{ge}}$ and $\mathcal{R}_{\mathrm{ex}}$ to prove the theorem.

Regret in Initial Exploration: In the initial exploration phase each agent $m \in [M]$ plays each arm $k \in [K]$ for $\tau_1 = \lceil \sqrt{\log(T)} \rceil$ times. Thus, $\mathcal{R}_{ie}$ is simply bounded as follows

$$\mathcal{R}_{ie} \leq KM\lceil\sqrt{\log(T)}\rceil\Delta'_{\max}. \tag{27}$$

Regret in Guided Exploration: By law of total expectation and using the notation "not $\mathcal{F}$" for complement of an event $\mathcal{F}$, we have, $\mathcal{R}_{\mathrm{ge}}$ is equal to

$$\mathbb{E}\left[\sum_{t \in \mathcal{T}_{\mathrm{ge}}}\sum_{m=1}^{M}\Delta'_{\pi_m(t),m}\right] = \underbrace{\mathbb{E}\left[\mathbf{1}\{\mathcal{E}_T\}\sum_{t \in \mathcal{T}_{\mathrm{ge}}}\sum_{m=1}^{M}\Delta'_{\pi_m(t),m}\right]}_{\text{Term 1}}$$
$$+$$
$$\underbrace{\mathbb{E}\left[\mathbf{1}\{\text{not }\mathcal{E}_T\}\sum_{t \in \mathcal{T}_{\mathrm{ge}}}\sum_{m=1}^{M}\Delta'_{\pi_m(t),m}\right]}_{\text{Term 2}}. \tag{28}$$

We proceed by bounding $\mathcal{R}_{\mathrm{ge}}$ when $\mathcal{E}_T$ holds true and when it does not.

$\mathcal{E}_T$ **holds true.** We first show that $\hat{\Delta}^{\mathrm{ge}}$ used in designing the allocation in the guided exploration phase are within constant factors of the gaps $\tilde{\Delta}'$. Recall $\Omega^{\delta}_{k,m}(t)$ given in (11). At the end of the initial exploration phase, we have

$$\Omega^{\delta}_{k,m}(T_{\mathrm{ie}}) = \sqrt{2\left(g_M(\frac{\delta}{KM}) + 2M\log(4 + \log(\tau_1))\right)\sum_{n=1}^{M}\frac{w_{n,m}^2}{\tau_1}} \tag{29}$$

$$= \mathcal{O}(\sqrt{\frac{\log(KM) + \log\log(T) + \log\log(KM\log(T))}{\sqrt{\log(T)}}}), \tag{30}$$

where $\mathcal{O}$ notation hides only small absolute constants. Note that

$$\sum_{n=1}^{M}w_{n,m}^2 \leq \sum_{n=1}^{M}w_{n,m} = 1. \tag{31}$$

Therefore, $\Omega^{\delta}_{k,m}(T_{\mathrm{ie}})$ is decreasing in T as given above. Thus, for a sufficiently large T_1, for all $T \geq T_1$ and $\forall k, m$, we have

$$\Omega^{\delta}_{k,m}(T_{\mathrm{ie}}) \leq \frac{\Delta'_{k,m}}{4}. \tag{32}$$

From the confidence bounds on the sample means of mixed rewards, conditioned on $T \geq T_1$ and $\mathcal{E}_T$, we have that

$$\|\hat{\Delta}'_{k,m}(T_{\text{ie}}) - \tilde{\Delta}'_{k,m}\| \leq \frac{\tilde{\Delta}'_{k,m}}{2}. \tag{33}$$

Thus $\hat{\Delta}'$ are the same as $\tilde{\Delta}'$ up to absolute constants. *I.e.* $\frac{1}{2}\Delta'_{k,m} \leq \hat{\Delta}'_{k,m} \leq \frac{3}{2}\Delta'_{k,m}$, for all k, m.

In addition, let T_2 and T_3 be the smallest numbers for which $\frac{1}{\log\log(T_2)} \leq \frac{1}{2}\Delta'_{k,m}$, for all k, m, and $\frac{3}{2}\Delta'_{k,m} \leq \log\log(T_3)$, for all k, m, respectively. Let $T_0 = \max\{T_1, T_2, T_3\}$.

Recall the projections $\hat{\Delta}^{\text{ge}}_{k,m}$ of $\hat{\Delta}'_{k,m}(T_{\text{ie}})$ onto $(\frac{1}{\log\log(T)}, \log\log(T))$ interval. Under event $\mathcal{E}_T$, when $T \geq T_0$, we have $\hat{\Delta}_{k,m}(T_{\text{ie}}) = \hat{\Delta}'^{\text{ge}}_{k,m}$. Thus,

$$\frac{1}{2}\Delta'_{k,m} \leq \hat{\Delta}^{\text{ge}}_{k,m} \leq \frac{3}{2}\Delta'_{k,m} \tag{34}$$

The following Lemma shows that the complexity term in the oracle $\mathcal{P}$ remains within constant factors, when the input is scaled by up to constant factors. The proof is provided in Appendix D.

Lemma 8 *Recall Oracle $\mathcal{P}$ from Definition 2. Consider two sets of gaps $\Delta, \hat{\Delta} \in (\mathbb{R}^+)^{K \times M}$. Let $q^* = \mathcal{P}(\Delta)$, $\hat{q}^* = \mathcal{P}(\hat{\Delta})$, and $c = \sum_{k \in [K], m \in [M]} q_{k,m}\Delta_{k,m}, \hat{c} = \sum_{k \in [K], m \in [M]} \hat{q}_{k,m}\Delta_{k,m}$. Assume $\forall k \in [K], m \in [M], a\Delta_{k,m} \leq \hat{\Delta}_{k,m} \leq b\Delta_{k,m}$, for some $0 < a < b$. Then, $\hat{c} \leq \frac{b}{a^3}c$.*

Using Lemma 8, the regret in the guided exploration phase, under event $\mathcal{E}_T$ and assuming that $T \geq T_0$, can be bounded as

$$\text{Term 1} \leq \sum_{m \in [M], k \in [K]} \lceil 18 q^{\text{ge}}_{k,m} B(T) \rceil \Delta'_{k,m} \tag{35}$$

$$\leq 18 \sum_{m \in [M], k \in [K]} (q^{\text{ge}}_{k,m} \tilde{\Delta}'_{k,m}) B(T) + \sum_{m \in [M], k \in [K]} \Delta'_{k,m} \tag{36}$$

$$\leq 216 \tilde{c}^* B(T) + KM\Delta'_{\max}, \tag{37}$$

where the first inequality comes from $\mathbb{P}[\mathcal{E}_T] \leq 1$ and the third inequality holds by Lemma 8 and (34). We note that $\frac{B(T)}{\log(T)} \simeq 2$ for large T by definition. Term 1 is the dominant term in the regret bound.

not $\mathcal{E}_T$ holds true. To upper bound Term 2, let $a = \frac{1}{\Delta'_{\max} \log\log(T)}$ and $b = \frac{\log\log(T)}{\Delta'_{\min}}$. For $T \geq T_0$, (34) implies

$$a\Delta'_{k,m} \leq \hat{\Delta}'_{k,m} \leq b\Delta'_{k,m} \quad \forall k \in [K], m \in [M]. \tag{38}$$

Using Lemma 8.

$$\text{Term 2} = \mathbb{P}(\text{not } \mathcal{E}_T) \sum_{m \in [M], k \in [K]} \lceil 18 q^{\text{ge}}_{k,m} B(T) \rceil \Delta'_{k,m} \tag{39}$$

$$\leq \mathbb{P}(\text{not } \mathcal{E}_T) \left(18 \sum_{m \in [M], k \in [K]} (q^{\text{ge}}_{k,m} \tilde{\Delta}'_{k,m}) B(T) + \sum_{m \in [M], k \in [K]} \Delta'_{k,m} \right) \tag{40}$$

$$\leq \frac{1}{\log(T)} \left(18 (\log\log(T))^4 \frac{(\Delta'_{\max})^3}{\Delta'_{\min}} \tilde{c}^* B(T) + KM\Delta'_{\max} \right), \tag{41}$$

where the last inequality comes from Lemma 8 and $\mathbb{P}[\text{not } \mathcal{E}_T] \leq \frac{1}{\log(T)}$. Since $\frac{B(T)}{\log(T)} \simeq 2$ for large T, we have

$$\text{Term 2} = \mathcal{O}\left((\log\log(T))^4 \frac{(\Delta'_{\max})^3}{\Delta'_{\min}} \tilde{c}^* \right), \tag{42}$$

up to an absolute constant factor of approximately 36.

Regret in Exploitation/Switch to π': We first consider the exploitation case where the condition $\mathcal{C}$ in (18) is satisfied, and the algorithm proceeds with choosing $\hat{k}_m = \arg\max_{k \in [K]} \hat{\mu}'_{k,m}(T_{\text{ge}})$. When $\mathcal{B}_T$ holds true, the best arm will be chosen. Because, $\forall m, k \neq k_m^*$

$$\mu'_{\hat{k}_m,m} \geq \hat{\mu}'_{\hat{k}_m,m}(T_{\text{ge}}) - \Omega^{\delta'}_{\hat{k}_m,m} \tag{43}$$

$$> \hat{\mu}'_{\hat{k}_m,m}(T_{\text{ge}}) + \Omega^{\delta'}_{k,m} \tag{44}$$

$$\geq \mu'_{k,m}. \tag{45}$$

The second inequality holds by $\mathcal{C}$. Thus, there is no regret in the exploitation phase. When $\mathcal{B}_T$ does not hold true, the regret can be simply bounded by $KM\Delta'_{\max}$, since $\mathbb{P}(\text{not } \mathcal{B}_T) \leq \frac{1}{T}$. Therefore, the expected regret under the exploitation case $\mathcal{R}_{\text{ex}} \leq KM\Delta'_{\max}$.

Now, for the case where $\mathcal{C}$ fails and the algorithm switches to π', we have the following lemma.

Lemma 9 *Recall $\mathcal{E}_T$ and $\mathcal{B}_T$ and condition $\mathcal{C}$. We have*

$$(\mathcal{E}_T \text{ and } \mathcal{B}_T) \implies \mathcal{C}. \tag{46}$$

Moreover, we note that "not $\mathcal{E}_T$" implies "not $\mathcal{B}_T$", since $\Omega^{\delta}_{k,m} \leq \Omega^{\delta'}_{k,m}$. Thus, the probability of switching is bounded by the probability of "not $\mathcal{E}_T$", which is at most $\frac{1}{\log(T)}$. Hence, the regret in this case is bounded by

$$\mathcal{R}_{\text{ex}} \leq \frac{\mathcal{R}^{\pi'}(T)}{\log(T)} \leq \frac{128c^*}{\Delta'_{\min}} \log_2(\frac{8}{\Delta'_{\min}}) + o(1). \tag{47}$$

Adding up the bounds on $\mathcal{R}_{\text{ie}}$, $\mathcal{R}_{\text{ge}}$ and $\mathcal{R}_{\text{ex}}$ completes the proof.

5. Conclusion

In this work, we considered the collaborative bandit model given in Réda et al. (2022). While near-optimal sample complexities were known under this model, the problem of optimal regret bounds was left as an open question. We addressed this problem and proposed a novel algorithm, CExp^2, achieving order optimal regret bounds by running two rounds of exploration. We also showed that CExp^2 needs only a small constant expected number of communication rounds.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, 2011.

Peter Auer and Ronald Ortner. Ucb revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 2010.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 2002.

Etienne Boursier, Emilie Kaufmann, Abbas Mehrabian, and Vianney Perchet. A practical algorithm for multiplayer bandits when arm means vary among players. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Lijie Chen, Jian Li, and Mingda Qiao. Nearly instance optimal sample complexity bounds for top-k arm selection. In *International Conference on Artificial Intelligence and Statistics*, 2017.

Yu-Zhen Janice Chen, Stephen Pasteris, Mohammad Hajiesmaili, John Lui, Don Towsley, et al. Cooperative stochastic bandits with asynchronous agents and constrained feedback. *Advances in Neural Information Processing Systems*, 2021.

Yihan Du, Wei Chen, Yuko Yuroki, and Longbo Huang. Collaborative pure exploration in kernel bandit. *arXiv preprint arXiv:2110.15771*, 2021.

E. Even-Dar, S. Mannor, and Y. Mansour. Action Elimination and Stopping Conditions for the Multi-Armed Bandit and Reinforcement Learning Problems. *Journal of Machine Learning Research*, 2006.

Tanner Fiez, Lalit Jain, Kevin G Jamieson, and Lillian Ratliff. Sequential experimental design for transductive linear bandits. *Advances in Neural Information Processing Systems*, 2019.

Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances in Neural Information Processing Systems*, 2012.

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

K. Jamieson, M. Malloy, R. Nowak, and S. Bubeck. lil'UCB: an Optimal Exploration Algorithm for Multi-Armed Bandits. In *Conference on Learning Theory*, 2014.

Tianyuan Jin, Pan Xu, Xiaokui Xiao, and Quanquan Gu. Double explore-then-commit: Asymptotic optimality and beyond. In *Conference on Learning Theory*, pages 2584–2633. PMLR, 2021.

Nikolai Karpov, Qin Zhang, and Yuan Zhou. Collaborative top distribution identifications with limited interaction. In *IEEE Annual Symposium on Foundations of Computer Science*, 2020.

Emilie Kaufmann and Wouter Koolen. Mixture martingales revisited with applications to sequential tests and confidence intervals. *Journal of Machine Learning Research*, 2021.

T.L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 1985.

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

Clémence Réda, Andrea Tirinzoni, and Rémy Degenne. Dealing with misspecification in fixed-confidence linear top-m identification. In *Advances in Neural Information Processing Systems*, 2021.

Clémence Réda, Sattar Vakili, and Emilie Kaufmann. Near-optimal collaborative learning in bandits. In *Advances in Neural Information Processing Systems*, 2022.

Sudeep Salgia, Sattar Vakili, and Qing Zhao. Collaborative learning in kernel-based bandits for distributed users. *IEEE Transactions on Signal Processing*, 2023.

Antoine Salomon, Jean-Yves Audibert, and Issam El Alaoui. Regret lower bounds and extended upper confidence bounds policies in stochastic multi-armed bandit problem. *arXiv preprint arXiv:1112.3827*, 2011.

Chengshuai Shi and Cong Shen. Federated multi-armed bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

Chengshuai Shi, Cong Shen, and Jing Yang. Federated multi-armed bandits with personalization. In *International Conference on Artificial Intelligence and Statistics*, 2021.

Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *International Conference on Machine Learning*, 2010.

Chao Tao, Qin Zhang, and Yuan Zhou. Collaborative learning with limited interaction: Tight bounds for distributed exploration in multi-armed bandits. In *IEEE Annual Symposium on Foundations of Computer Science*, 2019.

Sattar Vakili, Keqin Liu, and Qing Zhao. Deterministic sequencing of exploration and exploitation for multi-armed bandit problems. *IEEE Journal of Selected Topics in Signal Processing*, 7(5): 759–767, 2013.

Jingxuan Zhu, Ethan Mulle, Christopher Salomon Smith, and Ji Liu. Decentralized multi-armed bandit can outperform classic upper confidence bound. *CoRR*, 2021.

Appendix A. Other related work

There is a vast literature on single-agent stochastic bandits under various statistical models and assumptions (see, Lai and Robbins, 1985; Auer et al., 2002; Vakili et al., 2013; Abbasi-Yadkori et al., 2011; Srinivas et al., 2010; Lattimore and Szepesvári, 2020, as some representative works). In the classic setting, the performance of algorithms was primarily measured in terms of regret (Lai and Robbins, 1985; Auer et al., 2002), while a best arm identification setting was also later considered in several works (Even-Dar et al., 2006; Gabillon et al., 2012; Jamieson et al., 2014). The collaborative bandit model formalized in Shi et al. (2021); Réda et al. (2022) is a generalization of the classic single-agent bandit to a multi-agent setting.

Collaborative bandit under a pure exploration setting, where all agents face the same bandit problem, has been considered in Hillel et al. (2013); Tao et al. (2019); Karpov et al. (2020); Chen et al. (2021); Zhu et al. (2021). Some of these works consider a communication matrix where agents are allowed to communicate only with their neighbors (see, Zhu et al., 2021). The goal of collaborative pure exploration is to reduce sample complexity at the cost of communication rounds. Du et al. (2021) considered a collaborative bandit setting where a kernel is used to model the reward means. They provided a lower bound on the sample complexity of the best arm identification and also an algorithm with order optimal performance. Salgia et al. (2023) considered regret minimization under a collaborative kernel-based setting. The kernel model, however, is inherently different from the multi-armed bandit considered in the collaborative learning model of Réda et al. (2022).

The phased elimination algorithms are inspired by a simple phased elimination algorithm introduced in Auer and Ortner (2010) for subtle improvements in the regret bounds of the classic single-agent bandit. These algorithms maintain a set of potential optimal arms which are pruned at the end of each phase based on the observations gathered so far. This technique has been used in several bandit algorithms in various settings for both regret minimization and best arm identification (Hillel et al., 2013; Chen et al., 2017; Fiez et al., 2019; Boursier et al., 2020; Shi et al., 2021; Réda et al., 2022; Du et al., 2021). Recently, Jin et al. (2021) studied the standard K-armed bandits using a double-exploration algorithm. However, the algorithm in Jin et al. (2021) follows an "explore-commit-explore-commit" pattern, while none of the exploration phases are guided in the sense of our guided exploration phase. Our algorithm follows an "explore-guided explore-exploit/switch" phases, where in addition to the other difference mentioned before, the last phase includes switching to a policy which could be a UCB-based policy instead of always committing to one arm, which is suboptimal. Moreover, our multi-agent setting includes a constrained optimisation problem governing the lower bound which makes the problem significantly different than the setting in Jin et al. (2021). In addition, our model due to its collaborative nature requires a more complex confidence bound.

Appendix B. Proof of Lemma 6

In CExp^2, agents communicate with the server at the end of both exploration phases, i.e. initial exploration and guided exploration. If the condition (18) holds, the algorithm enters the exploitation phase where no communication happens. Otherwise, the algorithm switches to π', where the agents communicate according to π'. Therefore, the expected number of communication rounds is bounded by

$$2 + \lceil \log_2(\frac{8}{\Delta'_{\min}}) \rceil \mathbb{P}(\text{not } \mathcal{C}), \tag{48}$$

where $\lceil \log_2(\frac{8}{\Delta'_{\min}}) \rceil$ is the number of communication rounds in π' introduced is Section 3.2.

Appendix C. Proof of Lemma 7

Réda et al. (2022) considered the best arm identification setting. They introduced the following relaxed complexity term for the sample complexity lower bound.

$$s^* = \min_{q \in (\mathbb{R}^+)^{K \times M}} \sum_{m \in [M]} \sum_{k \in [K]} q_{k,m},$$

$$\text{subject to: } \forall m \in [M], \forall k \in [K], \qquad \sum_{\{n \in [M]: k_n^* \neq k\}} \frac{w_{n,m}^2}{q_{k,n}} \leq \frac{(\Delta'_{k,m})^2}{2}.$$

They proved that the sample complexity of W-CPE-BAI is at most

$$32 s^* \log_2(\frac{1}{\Delta_{\min}}) \log(\frac{1}{\delta}) + o(\log(\frac{1}{\delta})). \tag{49}$$

Choosing $\delta = \frac{1}{T}$, we have that the regret of W-CPE-BAI satisfies

$$\mathcal{R}(T) \leq 32 s^* \log_2(\frac{1}{\Delta_{\min}}) \log(T) + o(\log(T)) + \frac{TM\Delta_{\max}}{T}. \tag{50}$$

In addition, we prove that $s^* \leq \frac{4c^*}{\Delta'_{\min}}$ that completes the proof. Let $\hat{q}^*$ be the $\arg\min$ in the above optimization problem. Also, let q^* be the $\arg\min$ in the optimisation problem of Oracle $\mathcal{P}$ given in Definition 2 and $\tilde{c}^*$ be as in Lemma 3. By definition

$$
\begin{aligned}
s^* &= \sum_{m \in [M], k \in [K]} \hat{q}_{k,m}^* \\
&\leq \sum_{m \in [M], k \in [K]} q_{k,m}^* \\
&\leq \frac{1}{\Delta'_{\min}} \sum_{m \in [M], k \in [K]} q_{k,m}^* \Delta'_{k,m} \\
&= \frac{\tilde{c}^*}{\Delta'_{\min}} \\
&\leq \frac{4c^*}{\Delta'_{\min}}.
\end{aligned}
$$

The last inequality holds by Lemma 3. Also, note that the number of communication rounds follows the best arm identification setting, which is proved in Réda et al. (2022) to be at most $\lceil \log_2(\frac{8}{\Delta'_{\min}}) \rceil$.

Appendix D. Proof of Lemma 8

We note that, $\forall m \in [M]$, $\sum_{n \in [M]} \frac{a^2 w_{n,m}^2}{q_{k,n}} \leq \frac{a^2 \Delta_{k,m}^2}{2} \leq \frac{\hat{\Delta}_{k,m}^2}{2}$. This shows that $\{\frac{q_{k,m}}{a^2}\}_{k,m}$ satisfy the constraints in $\mathcal{P}(\hat{\Delta})$. Thus, by definition

$$\sum_{k \in [K], m \in [M]} \hat{q}_{k,m} \hat{\Delta}_{k,m} \leq \sum_{k \in [K], m \in [M]} \frac{q_{k,m}}{a^2} \hat{\Delta}_{k,m}. \tag{51}$$

Then, we have

$$
\begin{aligned}
\hat{c} &= \sum_{k \in [K], m \in [M]} \hat{q}_{k,m} \Delta_{k,m} \\
&\leq \frac{1}{a} \sum_{k \in [K], m \in [M]} \hat{q}_{k,m} \hat{\Delta}_{k,m} \\
&\leq \frac{1}{a^3} \sum_{k \in [K], m \in [M]} q_{k,m} \hat{\Delta}_{k,m} \\
&\leq \frac{b}{a^3} \sum_{k \in [K], m \in [M]} q_{k,m} \Delta_{k,m} \\
&= \frac{b}{a^3} c
\end{aligned}
$$

The first and third inequalities follows from $a\Delta_{k,m} \leq \hat{\Delta}_{k,m} \leq b\Delta_{k,m}$. The second inequality was shown in (51).

Appendix E. Proof of Lemma 9

Assume $\mathcal{E}_T$ and $\mathcal{B}_T$ hold true. Recall the Oracle $\mathcal{P}$ from Definition (2). For τ^{ge}, the number of samples in the guided exploration phase, we have

$$
\begin{aligned}
\Omega_{k,m}^{\delta'}(T) &\leq \sqrt{\beta_{\delta'}(\tau_{k,\cdot}(T)) \sum_{n=1}^{M} \frac{w_{n,m}^2}{\tau_{k,n}^{\mathrm{ge}}}} \\
&\leq \sqrt{\beta_{\delta'}(\tau_{k,\cdot}(T)) \sum_{n=1}^{M} \frac{w_{n,m}^2}{\lceil 18 q_{k,n}^{ge} B(T) \rceil}} \\
&\leq \sqrt{\frac{(\hat{\Delta}_{k,m}^{\mathrm{ge}})^2}{36}} \\
&\leq \frac{\Delta_{k,m}'}{4}.
\end{aligned}
\tag{52}
$$

The third inequality holds by definition of q^{ge} and the constraints in Oracle $\mathcal{P}$. The last inequality holds by (34). Under $\mathcal{B}_T$, we also have, $\forall k, m$

$$
\|\hat{\mu}_{k,m}(T_{\mathrm{ge}}) - \mu_{k,m}'\| \leq \Omega_{k,m}^{\delta'}(T).
$$

That implies, $\forall k, m$,

$$
\hat{\Delta}_{k,m}'(T_{\mathrm{ge}}) \geq \Delta_{k,m}' - 2\Omega_{k,m}^{\delta'}(T) \geq \frac{\Delta_{k,m}'}{2}.
\tag{53}
$$

From (52) and (53), we conclude, $\forall k, m$,

$$
\Omega_{k,m}^{\delta'}(T) \leq \frac{\hat{\Delta}_{k,m}'}{2}.
\tag{54}
$$

We thus proved that if both $\mathcal{E}_T$ and $\mathcal{B}_T$ hold true, then condition $\mathcal{C}$ holds true.