Exploration for Free: How Does Reward Heterogeneity Improve Regret in Cooperative Multi-agent Bandits?

Xuchuang Wang[1] **Lin Yang**[2] **Yu-Zhen Janice Chen**[3] **Xutong Liu**[1] **Mohammad Hajiesmaili**[3] **Don Towsley**[3] **John C.S. Lui**[1]

[1]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong
[2]School of Intelligence Science and Technology, Nanjing University, Jiangsu, China
[3]College of Information and Computer Sciences, University of Massachusetts Amherst, Massachusetts, USA

Abstract

This paper studies a cooperative multi-agent bandit scenario in which the rewards observed by agents are heterogeneous—one agent's meat can be another agent's poison. Specifically, the total reward observed by each agent is the sum of two values: an arm-specific reward, capturing the intrinsic value of the arm, and a privately-known agent-specific reward, which captures the personal preference/limitations of the agent. This heterogeneity in total reward leads to different local optimal arms for agents but creates an opportunity for *free exploration* in a cooperative setting—an agent can freely explore its local optimal arm with no regret and share this free observation with some other agents who would suffer regrets if they pull this arm since the arm is not optimal for them. We first characterize a regret lower bound that captures free exploration, i.e., arms that can be freely explored have no contribution to the regret lower bound. Then, we present a cooperative bandit algorithm that takes advantage of free exploration and achieves a near-optimal regret upper bound which tightly matches the regret lower bound up to a constant factor. Lastly, we run numerical simulations to compare our algorithm with various baselines without free exploration.

1 INTRODUCTION

Multi-armed bandit (MAB) [Lai et al., 1985, Bubeck et al., 2012] is a classic sequential decision making problem. In the stochastic MAB, an agent faces a set $\mathcal{K} := \{1, 2, \dots, K\}$ ($K \in \mathbb{N}^+$) of arms, where each arm k is associated with a reward random variable with unknown mean $\mu(k)$. The agent sequentially pulls arms from $\mathcal{K}$ in $T \in \mathbb{N}^+$ decision rounds and observes the pulled arm rewards. The goal of the agent is to maximize its total reward over all decision rounds, which is equivalent to minimizing the total *regret*, i.e., the cumulative reward difference between the aggregate reward of the optimal arm k_* with the highest mean and the agent's sequential choices. To achieve this goal, the agent needs to balance between exploration and exploitation, i.e., either optimistically choose the arm with high uncertainty in reward (exploration), or myopically pull the one with high empirical mean reward (exploitation).

Multi-agent MAB (MA2B) is an extension of the basic MAB, where a group of $M \in \mathbb{N}^+$ agents (denoted as $\mathcal{M} := \{1, 2, \dots, M\}$) pulls arms from the same arm set $\mathcal{K}$. This model has been studied in various settings, e.g., federated bandits [Shi and Shen, 2021, Shi et al., 2021a, Zhu et al., 2021, Huang et al., 2021], cooperative pure exploration [Hillel et al., 2013, Tao et al., 2019, Karpov et al., 2020], multi-agent MAB with collision [Boursier and Perchet, 2019, Mehrabian et al., 2020, Shi et al., 2021b], and cooperative multi-agent MAB [Landgren et al., 2016, Martínez-Rubio et al., 2019, Wang et al., 2020a,b].

The majority of prior works on MA2B, with a few exceptions (see Appendix A), study a homogeneous reward setting, where the reward distribution of an arm is the same for all agents. The homogeneous reward setting, however, fails to capture agent-specific preferences/limitations. In many real-world applications, the agents represent different clusters of users with specific preferences, or users in different geographical locations with different costs/limits to access the arm set. In such settings, the reward of each arm might be different for different agents. We refer to Section 2.3 for a detailed explanation of various application scenarios.

This paper introduces a multi-agent multi-armed bandits problem with heterogeneous reward (MA2B-HR). In MA2B-HR, the reward observed by an agent consists of two components representing arm- and agent-specific terms. Specifically, when agent $i \in \mathcal{M}$ pulls arm $k \in \mathcal{K}$, the observed reward is $X_t^{(i)}(k) = X_{t,\text{arm}}(k) + X_{t,\text{agent}}^{(i)}(k)$, where $X_{t,\text{arm}}(k)$ is the arm-specific reward with bounded mean

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

$\mu(k) \in (0, b)$ (where b is a positive constant) and $X_{t,\text{agent}}^{(i)}(k)$ is the agent-specific reward with mean $\nu^{(i)}(k)$. We denote $\omega^{(i)}(k) := \mu(k) + \nu^{(i)}(k)$ as the reward mean of this pull. In MA2B-HR, we assume both $X_{t,\text{arm}}(k)$ and $X_{t,\text{agent}}^{(i)}(k)$ are stochastic and independent. The arm-specific reward mean $\mu(k)$ is not known to agents, and each agent i only privately knows its own agent-specific mean values $\nu^{(i)}(k), \forall k \in \mathcal{K}$. Further, in the MA2B-HR setting, the agents can broadcast the observed values of the arm-specific term in the total reward (by subtracting the agent-specific reward mean from the observed reward, i.e., $X_t^{(i)}(k) - \nu^{(i)}(k)$) at no cost. We note that one may consider other settings for MA2B-HR, e.g., known vs. unknown and homogeneous vs. heterogeneous assumptions for the agent-specific reward. We refer to Appendix B.1 for a detailed discussion and the connection of each setting to the prior literature.

In MA2B-HR, the reward heterogeneity of agents creates a counterintuitive opportunity for *free exploration* of a subset of arms. With heterogeneous rewards among agents, there might be no global optimal arm(s). In other words, agents may have different *local* optimal arms, i.e., the arms with the largest reward mean are different among agents, so the characterization of the regret of agents becomes more complicated. However, the existence of multiple local optimal arms poses a surprising opportunity to develop a cooperative learning algorithm to explore local optimal arms for free (without cost), share the free observations with others, and significantly improve the total regret among all agents.

While the idea of free exploration is intuitive, designing a cooperative bandit algorithm that effectively implements this idea is nontrivial. The main challenge is that the local optimal arms are unknown in advance to the bandit agents. Hence, an algorithm should be designed to economically identify the local optimal arms and assign them to agents that can freely explore them and prevent other agents from pulling these arms (with cost).

We note that MA2B-HR could be considered as an extended version of two recent models in the bandits' literature: action-constrained multi-agent multi-armed bandits (AC-MA2B) Yang et al. [2022] and grouped K-armed bandits Baek and Farias [2021]. The idea of free exploration is applicable to both Yang et al. [2022], Baek and Farias [2021], however, they did not explicitly utilize free exploration in algorithm design, so they fail to achieve optimal performance that takes into account the free exploration. A detailed discussion on both models and their connection to MA2B-HR, and the significance of our results with respect to both models are given in Section 1.2.

It is worth noting that the high-level idea of free exploration has been leveraged in some other bandit settings in the literature [Chen et al., 2018, Shi et al., 2021c]. However, these works considered the problem of incentivizing exploration; specifically, they considered a principal, aiming to learn the global bandit model, offering bonuses to agents to do explorations on the principal's behalf. In these settings, Chen et al. [2018], Shi et al. [2021c] studied free exploration in the sense that the principal pays no cost rather than free exploration in cooperation among agents. Hence, these works are in clear contrast to the idea of free exploration in MA2B-HR introduced in this paper. A comprehensive comparison to related works are presented in Appendix A.

1.1 CONTRIBUTIONS

In this paper, we first present the MA2B-HR model and highlight its real-world applications. Then, we propose FreeExp, a cooperative algorithm designed to enable free exploration in the learning process. Finally, we characterize a regret lower bound that explicitly captures the impact of free exploration on MA2B-HR, and show that the regret of FreeExp matches the regret lower bound up to a constant factor. The contributions of this work are:

Modeling and practical relevance of MA2B-HR: We present the MA2B-HR model in Section 2 and justify its practical relevance by highlighting several application scenarios in online advertising, wireless networks, and cloud and edge resource allocation. We also introduce a new definition for the suboptimality gap in MA2B-HR as a key parameter to explicitly characterize the impact of free exploration in the regret analysis.

Algorithm design: In Section 3, we present FreeExp, a cooperative learning algorithm that tackles MA2B-HR and implements the idea of free exploration. The high level idea of FreeExp is that agents judiciously reduce the selection of arms that are likely to be local optimal for other agents. Instead, by cooperation, those agents can still get the observations on those arms from others without regret cost. In doing so, free exploration of some arms becomes possible and the cooperative bandit algorithm achieves significant improvement in regret. A key technique in FreeExp is to perform periodic pulls of the empirical local optimal arms (i.e., the arm with the highest empirical mean) while balancing between exploration and exploitation, which guarantees that the empirical optimal arm is indeed the ground truth local optimal arm in most time slots.

Regret analysis: In contrast to the common regret analysis in multi-agent bandits where only the pulled arm matters regardless of the agent who pull the arm, in MA2B-HR, we have to address a unique technical challenge since the regret cost of pulling an arm depends not only on which arm is pulled, but also on which agent pulls it. In Section 4, we tackle this challenge and derive a regret lower bound for MA2B-HR that echos the importance of recognizing free explorations: arms that can be freely explored only cause constant regret, instead of the usual logarithmic regret in MA2B. We derive the regret upper bound of the FreeExp

Table 1: A simple example with three agents and three arms ($b > \mu(1) > \mu(2) > \mu(3) > 0$). The entries of the table show the total reward of each arm for each agent, e.g., $\omega^{(1)}(1) = \mu(1)$ or $\omega^{(3)}(2) = \mu(2) - b < 0$. Arms 1, 2, and 3 are the local optimal arms of agents 1, 2, and 3, respectively. On the right-hand side, denoting $\Delta(i, j) = \mu(i) - \mu(j)$, the regret of our work is compared with a classic non-cooperative algorithm [Auer, 2002] and the works of Yang et al. [2022] and Baek and Farias [2021] as two special cases of `MA2B-HR`.

	Arm 1	Arm 2	Arm 3
Agent 1	$\mu(1)$	$\mu(2)$	$\mu(3)$
Agent 2	< 0	$\mu(2)$	$\mu(3)$
Agent 3	< 0	< 0	$\mu(3)$

`UCB` [Auer, 2002]	$O\left(\left(\frac{1}{\Delta(1,2)} + \frac{1}{\Delta(1,3)} + \frac{1}{\Delta(2,3)}\right)\log T\right)$
`CO-UCB` [Yang et al., 2022]	$O\left(\left(\frac{1}{\Delta(1,2)} + \frac{1}{\Delta(2,3)}\right)\log T\right)$
`KL-UCB` [Baek and Farias, 2021]	$O\left(\log\log T\right)$
`FreeExp` (our work)	$O(1)$

algorithm which matches the regret lower bound up to a constant factor. Deriving this result requires new analysis techniques (see Theorem 4.3's proof sketch in Section 4 for detail). The tightness of both regret upper and lower bounds reflects the intrinsic property of `MA2B-HR` where free exploration plays a key role, and that `FreeExp` is near-optimal. A surprising observation is that in the special cases where every arm is local optimal for at least one agent (reasonable when $M \geqslant K$), `FreeExp` achieves an $O(1)$ regret.

Numerical results: In Section 5, we report numerical experiments of comparing our algorithm to several baselines.

1.2 TECHNICAL COMPARISON TO THE PRIOR WORK

In this section, we highlight our contribution in leveraging free exploration by applying our algorithm to the action-constrained `MA2B` problem (`AC-MA2B`) which was recently studied by Yang et al. [2022]. In `AC-MA2B`, each agent $i \in \mathcal{M}$ only pulls from a subset of arms $\mathcal{K}^{(i)} \subset \mathcal{K}$ and its goal is to find the local optimal arm in $\mathcal{K}^{(i)}$. `AC-MA2B` can be regarded as a special case of `MA2B-HR` when agent i's specific reward $\nu^{(i)}(k)$ for arm k is 0 if $k \in \mathcal{K}^{(i)}$, and $-b$ if $k \notin \mathcal{K}^{(i)}$, where $b > 0$ and $\mu(k) \in (0, b)$ for all arm k (see Remark 2.1 for a formal definition). Since agent i knows its agent-specific reward means, she would never pull arms with $\nu^{(i)}(k) = -b$ and thus is equivalent to only having access to arms in the constrained arm set $\mathcal{K}^{(i)}$. We provide a simple example in Table 1 to illustrate the benefit of free exploration which substantially improves regret as compared to the classic non-cooperative algorithms and the cooperative approach in Yang et al. [2022] as a special case.

Next, we present the theoretical improvement. Recall that the non-cooperative optimal total regret of classic MAB [Lai et al., 1985] for all agents in $\mathcal{M}$ is

$$O\left(\sum_{i\in\mathcal{M}}\sum_{k\in\mathcal{K}^{(i)}\setminus\{k_*^{(i)}\}}\frac{\Delta^{(i)}(k)\log T}{\mathrm{kl}(\mu(k),\mu(k)+\Delta^{(i)}(k))}\right),$$

where the suboptimality gap $\Delta^{(i)}(k) := \mu(k_*^{(i)}) - \mu(k)$ is the difference of reward means between agent i's optimal arm $k_*^{(i)}$ and arm k, and $\mathrm{kl}(a, b)$ is the KL-divergence between two Gaussian distributions with means a and b and the same variance (defined later). To improve total regret through cooperation, Yang et al. [2022] proposed cooperative extensions to classic learning algorithms, e.g., `UCB` [Auer, 2002], which improved the total regret to

$$O\left(\sum_{k\in\cup_i(\mathcal{K}^{(i)}\setminus\{k_*^{(i)}\})}\frac{\bar{\Delta}(k)\log T}{\mathrm{kl}(\mu(k),\mu(k)+\bar{\Delta}(k))}\right), \tag{1}$$

where $\bar{\Delta}(k)$ denotes the smallest reward mean gap of arm k compared to the local optimal arms *(excluding arm k)* among agents having access to arm k.

The regret of applying `FreeExp` to `AC-MA2B` is

$$O\left(\sum_{k\in\cup_i\mathcal{K}^{(i)}\setminus\cup_i\{k_*^{(i)}\}}\frac{\bar{\Delta}(k)\log T}{\mathrm{kl}(\mu(k),\mu(k)+\bar{\Delta}(k))}\right). \tag{2}$$

The improvement of our result lies in the summation range. Specifically, the summation range $\cup_i\mathcal{K}^{(i)}\setminus\cup_i\{k_*^{(i)}\}$ in (2) is a *subset* of (1)'s $\cup_i(\mathcal{K}^{(i)}\setminus\{k_*^{(i)}\})$. The summation range in (2) excludes the regret impact of arms in $\cup_i\{k_*^{(i)}\}$, i.e., arms that are optimal to at least one agent; these arms are freely explored. In contrast, the regret of Yang et al. [2022] in (1) is over $\cup_i(\mathcal{K}^{(i)}\setminus\{k_*^{(i)}\})$, which counts some arms that are optimal for some agents (and can be freely explored). We note that this improvement can be substantial. Especially, when all arms in $\mathcal{K}$ are locally optimal for some agents, the regret upper in (2) is $O(1)$, e.g., the simple example in Table 1. This implies that capturing the benefit of free exploration requires the development of a completely new cooperative algorithm as explained in Section 3.

The grouped K-armed bandits model proposed by Baek and Farias [2021] is almost equivalent to `AC-MA2B` Yang et al. [2022] except for minor differences in how their actions are constrained—the grouped bandits' action constraint depends on the arrived group while `AC-MA2B`'s is associates to the agents. Therefore, the grouped bandits model can also be regarded as a special case of our `MA2B-HR` model. Baek and Farias [2021] proved that the KL-UCB algorithm Cappé et al. [2013] can address their grouped bandits model with the regret performance as follows,

$$\limsup_{T\to\infty} \frac{\mathbb{E}[\mathrm{R}_\mathrm{T}(\mathcal{A})]}{\log T} \leqslant \sum_{k\in\cup_i\mathcal{K}^{(i)}\setminus\cup_i\{k_*^{(i)}\}} \frac{\bar{\Delta}(k)}{\mathrm{kl}(\mu(k),\mu(k)+\bar{\Delta}(k))}.$$

We emphasize that the above bound of Baek and Farias [2021] was in an asymptotic form (i.e., for $T\to\infty$), while FreeExp's regret bound is in a non-asymptotic form (i.e., for any time T, see Eq.(10) of Theorem 4.3), which differs a lot in handling the regret of free arms (see Remark 4.7 for detail). Here, we pick the toy example in Table 1 to illustrate the difference; this can be generalized to any case that all arms are free arms. In this example, FreeExp attains the $O(1)$ regret, while KL-UCB's regret was $o(\log T)$ (or, $O(\log\log T)$ specifically) [Baek and Farias, 2021]. In Section 5, we conduct numerical comparisons to corroborate the advantage of FreeExp over KL-UCB. Also, we emphasize that our regret upper bound is proved for the MA2B-HR model which is more general than Baek and Farias [2021]'s grouped bandits model.

2 MODEL AND NOTATIONS

We first present the multi-agent multi-armed bandits with heterogeneous rewards problem (MA2B-HR) in Section 2.1 and its performance metric in Section 2.2. In Section 2.4, we introduce notations related to free exploration to facilitate our algorithm design and analysis.

2.1 MA2B-HR: THE MULTI-AGENT MULTI-ARMED BANDITS WITH HETEROGENEOUS REWARDS

In MA2B-HR, there are $K\in\mathbb{N}^+$ arms and $M\in\mathbb{N}^+$ agents. Each arm $k\in\mathcal{K}(:=\{1,2,\dots,K\})$ is associated with a Gaussian reward random variable with unknown mean $\mu(k)\in(0,b)$ and variance σ_1^2, where b is positive and known.[1] This is the *arm-specific reward* representing the intrinsic value of the arm and it is independent of the preference of the agents. In addition, each agent has its own private *agent-specific reward* for each arm to capture its private preference for different arms. The agent-specific reward of agent i for arm k is modelled by a Gaussian random variable with mean $\nu^{(i)}(k)$ and variance σ_2^2. The variances σ_1^2 and σ_2^2 are common for all arms and agents. The agent- and arm-specific rewards are independent, and both are also independent across arms $\mathcal{K}$ and time $t=1,2,\dots$.

By pulling an arm k at time t, agent i observes a Gaussian reward $X_t^{(i)}(k)$ with mean $\omega^{(i)}(k):=\mu(k)+\nu^{(i)}(k)$ and variance $\sigma_1^2+\sigma_2^2$. In this paper, we assume that the value of $\nu^{(i)}(k)$ is only known to agent i, but unknown to other agents, for all agent $i\in\mathcal{M}$. Similar to the basic setting of stochastic bandits, the arm-specific reward means $\mu(k)$ are unknown to all agents. We also assume, for each agent i, that all mean rewards $\omega^{(i)}(k)$ $(\forall k\in\mathcal{K})$ are different; hence each agent has a unique optimal arm.

[1] If b is unknown, we can set it as an arbitrarily large constant.

Remark 2.1 (Agent's local arm set)**.** Observe that $\mu(k)\in(0,b)$. Consequently, if there exist two arms k_1,k_2 such that $\nu^{(i)}(k_1)\geqslant\nu^{(i)}(k_2)+b$ for agent $i\in\mathcal{M}$, then

$$\begin{aligned}\omega^{(i)}(k_1)-\omega^{(i)}(k_2)&\\ &=(\mu(k_1)+\nu^{(i)}(k_1))-(\mu(k_2)+\nu^{(i)}(k_2))\\ &>\mu(k_1)-\mu(k_2)+b>0,\end{aligned}$$

that is, for agent i, the reward mean of arm k_1 is higher than that of arm k_2. Therefore, there is no need for agent i to pull arm k_2. More generally, we define agent i's *local arm set* as follows, Therefore, agent i's local arm set is

$$\mathcal{K}^{(i)}:=\left\{k\in\mathcal{K}:\nu^{(i)}(k)+b>\max\nolimits_{\ell\in\mathcal{K}}\nu^{(i)}(\ell)\right\},$$

and agent i only needs to explore arms in its local arm set.

Another relevant model for reward heterogeneity is contextual bandits [Li et al., 2010]. We discuss it in Appendix B.2. The MA2B-HR model finds applications in diverse domains, e.g., online advertising, online shortest path routing, online cloud and edge resources allocation, and personalized clinical trial, cf., the detail application scenarios in Appendix 2.3.

2.2 PERFORMANCE METRICS

Since rewards are heterogeneous across agents, agents may have different optimal arms. The goal of each agent is to find its *local* optimal arm, the one with the largest total reward, which is the sum of arm- and agent-specific rewards. Let $k_*^{(i)}$ be the local optimal arm of agent i, i.e., $k_*^{(i)}:=\arg\max_{k\in\mathcal{K}^{(i)}}\omega^{(i)}(k)$. For an algorithm $\mathcal{A}$, let $J_t^{(i)}(\mathcal{A})$ be the arm pulled by agent i at time t. The expected regret of agent i under algorithm $\mathcal{A}$ is the difference between the aggregate reward of pulling its local optimal arm and the aggregate reward of pulling arms in an online manner according to a bandit algorithm, i.e.,

$$\mathbb{E}[\mathrm{R}_T^{(i)}(\mathcal{A})]:=T\omega^{(i)}(k_*^{(i)})-\mathbb{E}\left[\sum\nolimits_{t=1}^{T}\omega^{(i)}(J_t^{(i)}(\mathcal{A}))\right],$$

where the expectation is taken over the randomness of action sequence $\{J_1^{(i)}(\mathcal{A}),J_2^{(i)}(\mathcal{A}),\dots\}$.

In the MA2B-HR model, agents can cooperate and share information to accelerate bandit learning. In particular, we assume that each agent can broadcast the arm-specific reward term (the observed rewards minus the agent-specific reward mean, $X_t^{(i)}(k)-\nu^{(i)}(k)$) at no cost to all other agents, and other agents immediately receives the broadcast observations. Note that this basic system model can be extended to include the communication costs, or an underlying topology to govern communication between agents, or agent privacy, etc. We leave these extensions to future works and focus on presenting the key idea of free exploration in this paper. The learning environment is a cooperative one, hence, we consider *aggregate regret* as the performance metric, which is simply the aggregate regret over M agents, i.e.,

$$\mathbb{E}[\mathrm{R_T}(\mathcal{A})] := \sum_{i=1}^{M}\left(T\omega^{(i)}(k_*^{(i)}) - \mathbb{E}\left[\sum_{t=1}^{T}\omega^{(i)}(k_t^{(i)})\right]\right). \quad (3)$$

2.3 APPLICATION SCENARIOS

The heterogeneous and known agent-specific reward means for MA2B-HR is a practically relevant setting and can find applications in diverse domains. The applications mentioned in Yang et al. [2022] and Baek and Farias [2021] can also be handled by MA2B-HR since their models are special cases of MA2B-HR. In the following, we present four motivating application scenarios that MA2B-HR could model. We note that we focus on motivating the arm- and agent-specific rewards. Detailed modeling of each application may require additional effort, which is beyond the scope of this paper.

Online Advertising in Social Networks: Online advertising is a classic example of the MAB problem [Tang et al., 2014, Mahadik et al., 2020]. Consider a scenario where there are multiple bandit agents that select ads to be placed on a social platform. Each agent is responsible for a cluster of users with similar interests. The cluster may be constructed based on different criteria, e.g., location, age, etc. Indeed, the popularity of products can differ across different locations or age groups. But the ads (arms) could be selected from a shared pool of available ads. In this scenario, the agent is aware of the personal preferences of users in its cluster, i.e., the agent-specific reward is known. However, the agents need to learn the potential value of ads as well; hence, arm-specific rewards are unknown. Since the learning agents all belong to the same social platform advertising engine, they can cooperate to share arm-specific observations and improve learning performance.

Online Shortest Path Routing in Wireless Networks: Another example is the problem of finding shortest paths in a multi-hop wireless network. Consider a scenario in which multiple learning agents try to learn the shortest paths for different communication sessions. In this scenario, bandit algorithms can be implemented to learn the shortest routing paths [He et al., 2013, Zou et al., 2014, Talebi et al., 2017]. The cost (or latency) of a certain path (arm) depends on the physical condition of the path itself, representing an arm-specific cost unknown to the learning agents. Further, the session of each agent might have its local physical conditions, e.g., distance and the hardware spec of the mobile device, which is known only to the agent and impacts the overall cost of each path. In this scenario, the former is an arm-specific cost, which is homogeneous and unknown among all agents, while the latter varies across agents and whose mean is privately known to each agent only.

Online Cloud and Edge Resource Allocation: In prior literature, the MAB framework has been used for workload allocation into a pool of cloud/edge servers [Talebi and Proutiere, 2018, Johari et al., 2017, Lattimore et al., 2014, Dagan and Koby, 2018]. In this scenario, the cloud provider may categorize the compute jobs into multiple types, e.g., ML training workload, video processing, financial analytics, etc., and create a learning agent for finding the best server type for them. In this scenario, the arm-specific reward captures the hardware spec of the servers, and the agent-specific reward captures the job-specific hardware requirement of the workload, e.g., video processing is memory-intensive, while finance workload is compute-intensive. In edge scenarios where the workload could be run in multiple locations, the agent-specific reward could be represented as the cost of moving the workload to different locations as well, which is known and heterogeneous for different agents.

Personalized Medicine and Clinical Trial: A classic MAB application is clinical trial Lai [1987], Villar et al. [2015], Aziz et al. [2021]. Consider a scenario where patients have different covariates, e.g., age, gender, genomic features, and medical history, and, therefore, should be categorized to several heterogeneous groups, and the doctor should create personalized agents (drug application policies) for every group. In this scenario, the effectiveness of a treatment for a certain patient group depends not only on the treatment itself but also on the patient group's covariates. For example, the effectiveness of a treatment that disturbs patients' blood glucose concentrations may be discounted on diabetics. In this scenario, the arm-specific reward captures treatments' or medicines' basic effectiveness on a diseases, and the agent-specific reward (or cost) captures the discounted or additional effectiveness due to the patient group features. The latter is known to (or can be well evaluated by) an expert.

2.4 NOTATIONS RELATED TO FREE EXPLORATION

To ease the presentation of FreeExp and its analysis, we introduce some key notations relevant to free exploration. In MA2B-HR, arms that are local optimal for at least one agent can be freely explored. Then, in a cooperative environment, other agents who take these arms as their suboptimal choices can enjoy the freely explored observations of these arms.

Definition 2.2 (Set of free arms)**.** We define the set of free arms $\mathcal{K}^{fr}$ as

$$\mathcal{K}^{fr} := \{k \in \mathcal{K} : \mathcal{M}_*(k) \neq \emptyset\}, \tag{4}$$

where $\mathcal{M}_*(k) := \{i \in \mathcal{M} : k \in \mathcal{K}^{(i)}, k = k_*^{(i)}\}$ is a subset of agents with arm k as their local optimal arm. Any arm $k \in \mathcal{K}^{fr}$ can be freely explored without incurring regret by any agent in $\mathcal{M}_*(k)$. In the rest of this paper, we refer to the arms in $\mathcal{K}^{fr}$ as free arms.

Recall that in the classic MAB, the difficulty of distinguishing a suboptimal arm k from the optimal arm depends on $\Delta(k)$—the reward mean gap between arm k and the optimal arm k^*. In `MA2B-HR`, the notion of optimality gap needs to be redefined since agents may have different local optimal arms. In the following, we formally define the suboptimality gap of each arm k as the smallest gap between arm k and any local optimal arms. A formal definition is given below.

Definition 2.3 (Suboptimality gap)**.** The suboptimality gap of arm k is defined as

$$\bar{\Delta}(k) := \min_{i \in \mathcal{M}} \Delta^{(i)}(k), \tag{5}$$

where $\Delta^{(i)}(k) := \omega^{(i)}(k_*^{(i)}) - \omega^{(i)}(k)$ is the gap between the mean rewards of arm k and $k_*^{(i)}$—the local optimal arm of agent i.

All free arms have zero suboptimality gaps, i.e., $\bar{\Delta}(k) = 0, \forall k \in \mathcal{K}^{fr}$. Denote $\bar{i}(k) \in \arg\min_{i \in \mathcal{M}(k)} \Delta^{(i)}(k)$ to be an agent with the smallest reward gap of arm k (one can break ties arbitrarily). Then, $\bar{\Delta}(k)$ can be rewritten as $\bar{\Delta}(k) = \omega^{(\bar{i}(k))}(k_*^{(\bar{i}(k))}) - \omega^{(\bar{i}(k))}(k)$, where for simplicity, we denote $\omega^{(\bar{i}(k))}(k)$ as $\bar{\omega}(k)$, i.e.,

$$\bar{\omega}(k) := \omega^{(\bar{i}(k))}(k) = \mu(k) + \nu^{(\bar{i}(k))}(k). \tag{6}$$

3 THE FREEEXP ALGORITHM

In this section, we present the `FreeExp` algorithm, which solves a multi-agent bandit problem in the `MA2B-HR` model. Each agent runs its own `FreeExp` algorithm and cooperates with each other. In Section 4, we demonstrate that with `FreeExp`, the reward heterogeneity not only does no harm, but in fact benefits the cooperative learning by the unique opportunity of free exploration.

High-level idea of `FreeExp`: We now explain how `FreeExp` implements the idea of free exploration to reduce regret. The pivot of `FreeExp` is the local optimal (free) arm of each agent, which is unknown in advance. To address that for an agent i, `FreeExp` maintains an local optimal arm estimate $I_t^{(i)}$ of the agent i and an *exploration arm set* $\mathcal{D}_t^{(i)}$ containing arms that might be the ground truth local optimal arm and thus need further explorations. To utilize free exploration, agent i periodically announces her estimated optimal arm $I_t^{(i)}$ to others to discourage other agents exploring this arm.

Algorithm 1 The `FreeExp` Algorithm (for Agent i)

Initialize: $d_t(k) = 0, \hat{\mu}_t(k) = 0, \hat{\omega}_t^{(i)}(k) := \hat{\mu}_t(k) + \nu^{(i)}(k)$.
for each time slot t **do**
  $I_t^{(i)} \leftarrow \arg\max_{k \in \mathcal{K}^{(i)}} \hat{\omega}_t^{(i)}(k)$ `{identify the empirical optimal arm}`
  Send $I_t^{(i)}$ to other agents and collect their $I_t^{(j)}$
  $\mathcal{D}_t^{(i)} \leftarrow \{k \in \mathcal{K}^{(i)} \setminus \{I_t^{(i)}\} : d_t^{(i)}(k) > \hat{\omega}_t^{(i)}(I_t^{(i)})\}$ `{choose arms with high KL-UCB}`
  $\mathcal{D}_t^{(i)} \leftarrow \mathcal{D}_t^{(i)} \setminus \{I_t^{(j)} : \forall j \in \mathcal{M}\}$ `{take advantage of free exploration}`
  if $\mathcal{D}_t^{(i)} = \emptyset$ **then**
    $J_t^{(i)} \leftarrow I_t^{(i)}$
  else
    w.p., $\frac{1}{2}$, $J_t^{(i)} \leftarrow I_t^{(i)}$
    w.p., $\frac{1}{2}$, $J_t^{(i)} \leftarrow$ uniformly pick an arm from $\mathcal{D}_t^{(i)}$
  end if
  Pull arm $J_t^{(i)}$ and receive observations $X_t^{(i)}(J_t^{(i)})$
  Send observations $X_t^{(i)}(J_t^{(i)}) - \nu^{(i)}(J_t^{(i)})$ to other agents and also collect theirs
  Update $\hat{\omega}_t^{(i)}(k)$ and $d_t^{(i)}(k)$ for arm k and agent i
end for

Remark 3.1. We note that some prior works [Combes and Proutiere, 2014, Combes et al., 2015, Wang et al., 2020a], such as the `DPE2` algorithm in cooperative `MA2B` [Wang et al., 2020a], also involved a pivot arm and an exploration arm set in the algorithm design. However, the technical usage of both components in those works is very different from ours. For example, `DPE2` estimates the pivot arm to gather all exploration responsibility to a single leader agent, while our usage is relegating/dispersing the free arms to the agents for which they are locally optimal.

Local optimal arm estimate and construction of exploration arm set: Let $n_t(k)$ and $\hat{\mu}_t(k)$ denote the total number of times arm k is pulled up to time t and the empirical mean of these $n_t(k)$ reward observations of arm k among all M agents. Denote $\hat{\omega}_t^{(i)}(k) := \hat{\mu}_t(k) + \nu^{(i)}(k)$ as the empirical reward mean of agent i pulling arm k and it is based on all agents' observations of arm k. `FreeExp` uses agent i's *empirical local optimal arm* $I_t^{(i)}$ (the arm with the largest empirical reward mean $\hat{\omega}_t^{(i)}(k)$ of agent i at time t) as an estimate of the pivot. Given this empirical optimal arm as the pivot, the agent either pulls its own empirical optimal arm $I_t^{(i)}$ for free exploration, or explores other arms in $\mathcal{D}_t^{(i)}$ to guarantee the correctness of this estimated pivot. To improve the efficiency of exploring other arms, we construct the *exploration arm set* $\mathcal{D}_t^{(i)}$ for each agent i using the KL-UCB index [Cappé et al., 2013]. The index of arm k at time slot t is

$$\begin{aligned} d_t^{(i)}(k) \coloneqq \sup\{q \geqslant 0 : \\ n_t(k)\,\mathrm{kl}(\hat{\omega}_t^{(i)}(k), q) \leqslant \log t + 4\log(\log t)\}, \end{aligned} \tag{7}$$

where $\mathrm{kl}(a,b)$ is the KL-divergence between two Gaussian distributions with means a and b and same variance $\sigma_1^2 + \sigma_2^2$. The exploration arm set $\mathcal{D}_t^{(i)}$ includes arms whose KL-UCB indexes $d_t^{(i)}(k)$ are greater than the agent's highest empirical mean $\hat{\omega}_t^{(i)}(I_t^{(i)})$ (Line 5) and excludes arms that are empirically optimal for at least one agent (Line 6)—discourage agent i exploring others' local optimal arms. Note that the agents only share the arm-specific reward to other, i.e., the agent subtracts the agent-specific reward from the observed compound reward before sharing (Line 14).

Arm pulling policy: To guarantee the accuracy of the pivot estimation (i.e., the empirical optimal arm is correct with high probability), each agent needs to have enough observations for her empirically optimal arm. To accomplish this, `FreeExp` implements an arm pulling policy (Lines 7-11) as follows: if exploration arm set $\mathcal{D}_t^{(i)}$ is empty, the agent i pulls the empirical optimal arm $I_t^{(i)}$; if exploration arm set $\mathcal{D}_t^{(i)}$ is not empty, with probability $1/2$, the agent, uniformly at random picks an arm from $\mathcal{D}_t^{(i)}$ to explore; and with probability $1/2$, pulls her empirical optimal arm—encourage free explorations of the agent's empirical optimal arm. This policy produces sufficient observations of this arm to guarantee fast correction if the current empirical optimal arm is not the correct one. Let $J_t^{(i)}$ denote the arm selected by agent i in time slot t under `FreeExp`. We present pseudocode for `FreeExp` in Algorithm 1.

Remark 3.2 (`NoFreeExp` Algorithm)**.** There is a counterpart algorithm of `FreeExp`, which does not utilize free exploration, i.e., Algorithm 1 without Line 6. We name it as `NoFreeExp`. Even without making use of free exploration, `NoFreeExp` should have a better regret performance than known baselines, e.g., `CO-UCB`, because `NoFreeExp` is based on the KL-UCB algorithm, which is theoretically better than UCB-like algorithms [Cappé et al., 2013].

4 THEORETICAL RESULTS

We present our theoretical results and their significance discussions in this section. The rigorous proofs of these results are deferred to Appendix C. We first derive a regret lower bound in Theorem 4.1 which reflects the impact of free exploration.

Theorem 4.1 (Regret lower bound)**.** *For any consistent policy π (i.e., for any bandit instance $\boldsymbol{\nu}$ and any $\alpha > 0$, the policy π always guarantees $\mathbb{E}_{\boldsymbol{\nu},\pi}[\mathrm{R}_T] = O(T^\alpha)$), the regret cost of addressing the* `MA2B-HR` *model in T time slots is lower bounded by*

$$\liminf_{T\to\infty} \frac{\mathbb{E}[\mathrm{R_T}(\mathcal{A})]}{\log T} \geqslant \sum_{k:\bar{\Delta}(k)>0} \frac{\bar{\Delta}(k)}{\mathrm{kl}(\bar{\omega}(k), \bar{\omega}(k)+\bar{\Delta}(k))}, \tag{8}$$

where $\bar{\Delta}(k)$ defined in (5) is the smallest reward gap of pulling arm k and $\bar{\omega}(k)$ defined in (6) is the reward mean of pulling arm k by the agent who enjoys the smallest gap.

Theorem 4.1's proof leverages similar techniques of the classic stochastic bandits [Lai et al., 1985]. Since $\bar{\Delta}(k) = 0$ for all free arms $k \in \mathcal{K}^{\mathtt{fr}}$ and *vice versa*, the regret lower bound can be rewritten as

$$\liminf_{T\to\infty} \frac{\mathbb{E}[\mathrm{R_T}(\mathcal{A})]}{\log T} \geqslant \sum_{k\in\mathcal{K}\setminus\mathcal{K}^{\mathtt{fr}}} \frac{\bar{\Delta}(k)}{\mathrm{kl}(\bar{\omega}(k), \bar{\omega}(k)+\bar{\Delta}(k))}. \tag{9}$$

Remark 4.2 (Free arms have no contribution to the asymptotic regret lower bound)**.** Free arms in $\mathcal{K}^{\mathtt{fr}}$ contribute at most sub-logarithmic costs to the regret lower bound. In fact, given our finite regret upper bound of `FreeExp` next, free arms only contribute finite regret.

Theorem 4.3 (Regret upper bound for `FreeExp` (Algorithm 1))**.** *The* `FreeExp` *algorithm's regret is upper bounded as follows,*

$$\begin{aligned} &\mathbb{E}[\mathrm{R_T}(\mathcal{A})] \leqslant 7bM^2K^2(4K + \delta^{-2}) \\ &+ \sum_{k:\bar{\Delta}(k)>0} \frac{4(\bar{\Delta}(k) - 2\delta)(\log T + 4\log(\log T))}{\mathrm{kl}(\bar{\omega}(k)+\delta, \bar{\omega}(k)+\bar{\Delta}(k)-\delta)} \end{aligned} \tag{10}$$

where $0 < \delta < \frac{1}{4}\min_{i\in\mathcal{M}, k_1\neq k_2\in\mathcal{K}} |\omega^{(i)}(k_1) - \omega^{(i)}(k_2)|$, and that σ_1^2 and σ_2^2 are the variance of arm- and agent-specific Gaussian rewards respectively, and b is an upper bound of arm-specific reward mean $\mu(k)$ for all $k \in \mathcal{K}$.[2]

If we let $T \to \infty$ and $\delta \to 0$ (e.g., $\delta = (\log(\log T))^{-1}$), the above finite-time regret upper bound has the following asymptotical form,

$$\limsup_{T\to\infty} \frac{\mathbb{E}[\mathrm{R_T}(\mathcal{A})]}{\log T} \leqslant O\left(\sum_{k:\bar{\Delta}(k)>0} \frac{\bar{\Delta}(k)}{\mathrm{kl}(\bar{\omega}(k), \bar{\omega}(k)+\bar{\Delta}(k))}\right). \tag{11}$$

Proof sketch and technical challenges. The proof of the regret upper bound in Theorem 4.3 consists of two steps: (i) bound the regret cost of pulling free arms in $\mathcal{K}^{\mathtt{fr}}$, and (ii) other arms outside $\mathcal{K}^{\mathtt{fr}}$. To bound (i), notice that for any free arm k in $\mathcal{K}^{\mathtt{fr}}$, there exists "corresponding" agent(s) that takes arm k as its local optimal and can explore it with no cost. Hence, we only need to count the number of

[2] One can also obtain a near-optimal regret upper bound if the arm- and agent-specific rewards follow Bernoulli distributions.

times that arm k is pulled by agents other than "corresponding" one(s), which only happens when the "corresponding" agent's empirical optimal arm $I_t^{(i)}$ is not its true local optimal arm $k_t^{(i)}$. Such events only occur with finite number of times even with a very large value of T. The proof of (i) shares the similar logical flow to that of [Wang et al., 2020b, Theorem 1]. To proof (ii), however, we need to develop new techniques for addressing the heterogeneous rewards in `MA2B-HR`. Note that in `MA2B-HR` the suboptimality reward gaps of pulling the same arm depend on the agents and thus are different. Hence, one cannot bound the cost of pulling a suboptimal arm k via multiplying the number of times of pulling the suboptimal arm k by one suboptimality reward gap as the usual bandits literature did. To address the challenge, we introduce two new techniques. First, we respectively count the number of times of the suboptimal arm pulls by agents (see Lemma C.7 and its proof), and secondly, we apply an Abel transformation to summing up the regret costs of all agents on pulling the arm k according to the order of magnitude of the arm's reward gaps $\Delta^{(i)}(k)$ for these agents (see Lemma C.8 and its proof).

Similar to the regret lower bound's another expression in (9), this regret upper bound's summation range can also be expressed according to the free arms,

$$\limsup_{T\to\infty} \frac{\mathbb{E}[\mathrm{R_T}(\mathcal{A})]}{\log T} \leqslant O\left(\sum_{k\in\mathcal{K}\setminus\mathcal{K}^{\mathtt{fr}}} \bar{\Delta}(k)/\mathrm{kl}(\bar{\omega}(k), \bar{\omega}(k)+\bar{\Delta}(k)) \right). \tag{12}$$

Remark 4.4 (Regret optimality of the `FreeExp` algorithm)**.** This regret upper bound in (11) matches the regret lower bound in (8) up to a constant factor, which implies that both bounds are near-optimal, and therefore the `FreeExp` algorithm is near-optimal as well.

Remark 4.5 (Comparison to Yang et al. [2022]'s regret bounds)**.** Yang et al. [2022] proposed algorithms achieving regret upper bounds [Yang et al., 2022, Theorems 2 and 4] for `AC-MA2B` as follows (adapted to our notations),[3]

$$\limsup_{T\to\infty} \frac{\mathbb{E}[\mathrm{R_T}(\mathcal{A})]}{\log T} \leqslant O\left(\sum_{k\in\cup_{i\in\mathcal{M}}(\mathcal{K}^{(i)}\setminus\{k_*^{(i)}\})} \frac{\bar{\Delta}(k)}{\mathrm{kl}(\bar{\omega}(k), \bar{\omega}(k)+\bar{\Delta}(k))} \right).$$

Note that $\mathcal{K} = \cup_{i\in\mathcal{M}}\mathcal{K}^{(i)}$ and $\mathcal{K}^{\mathtt{fr}} = \cup_{i\in\mathcal{M}}\{k_*^{(i)}\}$. So, we have $\mathcal{K}\setminus\mathcal{K}^{\mathtt{fr}} \subset \cup_{i\in\mathcal{M}}(\mathcal{K}^{(i)}\setminus\{k_*^{(i)}\})$. For example, if an arm $k\in\mathcal{K}^{\mathtt{fr}}$ is also a suboptimal arm for another agent, then $k\in\cup_{i\in\mathcal{M}}(\mathcal{K}^{(i)}\setminus\{k_*^{(i)}\})$ but $k\notin\mathcal{K}\setminus\mathcal{K}^{\mathtt{fr}}$. In other words, the arm k contributes logarithmic regret costs to their upper bound but only contributes finite costs in ours. Therefore, their regret upper bound *failed to capture the advantage of free exploration* and their algorithms did not utilize this appealing mechanism.

[3] To express Yang et al. [2022]'s result, we abuse $\bar{\Delta}(k)$ notation *once*, where $\bar{\Delta}(k) := \min_{i\in\mathcal{M}\setminus\mathcal{M}_*(k)} \Delta^{(i)}(k)$—the smallest reward mean gap of arm k compared to the local optimal arms *(excluding arm k)* among agents having access to k. The difference between this definition and the original one in (5) is that for arm k in $\mathcal{K}^{\mathtt{fr}}$ this $\bar{\Delta}(k)$ is positive while the original one is zero.

Remark 4.6 (Special cases with $O(1)$ finite regret in `MA2B-HR`)**.** The regret upper bound in (12) echos the regret lower bound's Remark 4.2 that arms in $\mathcal{K}^{\mathtt{fr}}$ only cause finite $O(1)$ costs in regret. Therefore, if all arms are local optimal for some agents, $\mathcal{K}\setminus\mathcal{K}^{\mathtt{fr}} = \emptyset$ (e.g., the example in Table 1), then the regret upper bound in (11) becomes $O(1)$, i.e., a time horizon independent finite regret.

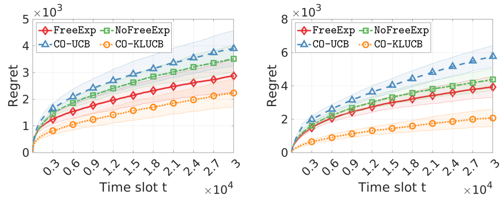
Remark 4.7 (Comparsion to Baek and Farias [2021])**.** Recall that the set of *free arms* $\mathcal{K}^{\mathtt{fr}}$ defined in our Eq.(4) contains arms that can be freely explored. In our regret upper bound, we show that `FreeExp`'s regret cost due to pulling arms in $\mathcal{K}^{\mathtt{fr}}$ is $O(1)$, while Baek and Farias [2021]'s regret bound was asymptotic with respect to $\log T$, implying that `KL-UCB`'s regret due to pulling arms in $\mathcal{K}^{\mathtt{fr}}$ was $o(\log T)$ (the analysis in Baek and Farias [2021] upper bounds the cost for arm set $\mathcal{K}^{\mathtt{fr}}$ by $O(\log\log T)$).

Remark 4.8 (Generalization to the homogeneous reward setting)**.** If all agents' local arm sets are the same, then only one unique optimal arm can be freely explored (i.e., $|\mathcal{K}^{\mathtt{fr}}| = 1$) and all other arms would appear in the summation range in regret bounds (8) and (11). Then, both the regret upper and lower bounds reduce to the ones in classic MABs in Lai et al. [1985] (also the same as the optimal bounds of cooperative `MA2B`). This observation highlights the *"generality"* of our regret bounds and shows that `FreeExp` also works for the homogeneous reward setting.

5 NUMERICAL SIMULATIONS

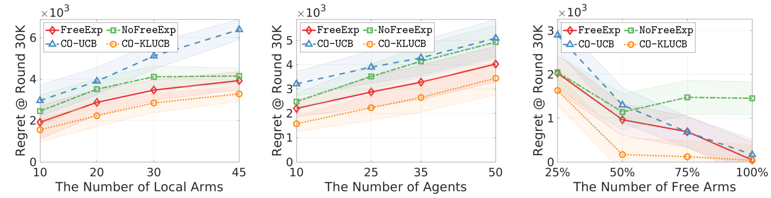
Baselines: We report results of numerical experiments that compare `FreeExp` to three known cooperative algorithms that do not leverage free exploration: (1) `CO-UCB` and (2) `CO-KLUCB`, extensions of `UCB` and `KLUCB` algorithms to cooperative multi-agent scenarios proposed by Yang et al. [2022] and Baek and Farias [2021] respectively; and (3) `NoFreeExp`, a variant of `FreeExp` that does not make use of free exploration (see Remark 3.2).

Experimental setup: Unless otherwise specified, we consider a `MA2B-HR` model with $M = 25$ agents and $K = 50$ arms. Each arm is associated with a Gaussian distribution whose arm-specific mean $\mu(k) \in (0,1)$ is chosen uniformly at random from the click-through-rates of Kaggle's *Ad-Click* dataset [Avito, 2015] and with variance $1/2$. We consider two special cases of agent-specific reward means: Case (1) $\nu^{(i)}(k)$ is either 0 or -1 $\forall k\in\mathcal{K}, i\in\mathcal{M}$ (i.e., `AC-MA2B` [Yang et al., 2022, Baek and Farias, 2021]



(a) Case (1) (b) Case (2)

Figure 1: `FreeExp` *vs.* baselines



(a) Vary # local arms (b) Vary # agents (c) Vary % of free arms

Figure 2: Vary parameters of `MA2B-HR`

where agents have different local arm sets) and Case (2) $\nu^{(i)}(k) \in (-1/2, 1/2)\ \forall k \in \mathcal{K}, i \in \mathcal{M}$ (i.e., all agents have the same local arm sets) as the more general heterogeneous reward scenario. The variances of all agent-specific rewards are set to $1/2$. In the `AC-MA2B` setting (Case (1)), for each agent, we randomly select 20 of these 50 arms and set their agent-specific rewards $\nu^{(i)}(k) = 0$, i.e., as local arms. The remaining arms' agent-specific rewards is set to $\nu^{(i)}(k) = -1$. In the heterogeneous reward setting (Case (2)), all agents have the same 50 arms but different agent-specific rewards whose means are uniformly and randomly generated between $(-1/2, 1/2)$ for each arm and agent. All simulations are averaged over 50 runs and their standard deviations are plotted as shadow regions.

Experimental results: In Figures 1a and 1b, we compare the cumulative regret of all algorithms in Cases (1) and (2). The notable observations are: (1) Comparison of `FreeExp` to `NoFreeExp` shows that utilizing the free exploration mechanism can further improve an algorithm's performance. (2) The `KLUCB` algorithm outperform our `FreeExp` algorithm. This is because `FreeExp` needs to explicitly exclude arms likely to be local optimal (Line 6) and thus suffers a high time-independent cost at the beginning, while `KLUCB` does not; and the additional cost of `FreeExp` cannot be compensated by the advantage of `FreeExp` in saving cost on free arms in these two scenarios. Especially, we note that when the number of free arms are large (e.g., see Figure 2c's 100% free arm case below), the advantage of `FreeExp` in saving cost on free arms becomes significant and, therefore, `FreeExp` has similar performance to `KLUCB`.

We report the results of varying the number of parameters of `MA2B-HR` (Case (1)) in Figure 2. In Figure 2a, we vary the number of local arms between 10 and 45 and report their cumulative regret at round 30K. All algorithm regrets increase linearly with respect to the number of local arms. Figure 2b shows the impact of the number of agents M (from 10 agents to 50) on the regrets. Their regrets also have linear increasing rate in M, which is due to the fixed per-agent costs (independent of T). Lastly, we consider an `MA2B-HR` consisting of $M = 20$ agents and $K = 20$ arms, and devise fours cases containing $\{5, 10, 15, 20\}$ free arms respectively (i.e., $25\%, 50\%, 75\%, 100\%$ of all arms are free arms). We report their regret performance in Figure 2c. The notable observations are: (1) The regret of `FreeExp` decreases as the percentage of free arms increases which corroborates that `FreeExp` saves the costs due to pulling free arms. (2) when all (100%) arms are free, `FreeExp` has similar performance to `KLUCB` and outperforms other algorithms.

6 CONCLUSION

This paper introduced a multi-agent multi-armed bandit problem with heterogeneous rewards among agents. The heterogeneous scenario creates a unique opportunity to explore a subset of arms for free and share the observation by cooperation, and hence, improve the aggregate regret significantly. We proposed a cooperative learning algorithm which would benefit from the free exploration and its regret is tight up to a constant factor. As a notable special case, when each arm is a local optimal arm in at least one agent, the proposed algorithm achieves an $O(1)$ regret.

This problem of multi-agent bandits with heterogeneous reward calls for several interesting follow-up questions, i.e., an interesting question is to extend the `FreeExp` algorithm with an effective communication protocol. In a distributed multi-agent setting, cooperation may come with a cost of communication, and hence the goal is to enhance the cooperative algorithms with a communication policies that only needs sublinear communication times w.r.t. decision rounds T, while directly extend current algorithm requires $O(T)$ communication times.

Acknowledgements

The work of Mohammad Hajiesmaili is supported by NSF CAREER-2045641, CPS-2136199, CNS-2106299, and CNS-2102963. The work of Don Towsley is supported by U.S. Army Research Laboratory under Cooperative Agreement W911NF-17-2-0196. The work of John C.S. Lui is supported in part by the RGC GRF 14215722. Lin Yang is the corresponding author (`linyang@nju.edu.cn`).

References

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

Avito. *Avito Context Ad Clicks*, 2015. https://www.kaggle.com/c/avito-context-ad-clicks.

Maryam Aziz, Emilie Kaufmann, and Marie-Karelle Riviere. On multi-armed bandit designs for dose-finding clinical trials. *Journal of Machine Learning Research*, 22(1-38): 4, 2021.

Jackie Baek and Vivek Farias. Fair exploration via axiomatic bargaining. *Advances in Neural Information Processing Systems*, 34:22034–22045, 2021.

Etienne Boursier and Vianney Perchet. Sic-mmab: synchronisation involves communication in multiplayer multi-armed bandits. *Advances in Neural Information Processing Systems*, 32, 2019.

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.

Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Kullback-leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, pages 1516–1541, 2013.

Bangrui Chen, Peter Frazier, and David Kempe. Incentivizing exploration by heterogeneous users. In *Conference On Learning Theory*, pages 798–818. PMLR, 2018.

Richard Combes and Alexandre Proutiere. Unimodal bandits: Regret lower bounds and optimal algorithms. In *International Conference on Machine Learning*, pages 521–529. PMLR, 2014.

Richard Combes, Stefan Magureanu, Alexandre Proutiere, and Cyrille Laroche. Learning to rank: Regret lower bounds and efficient algorithms. In *Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, pages 231–244, 2015.

Yuval Dagan and Crammer Koby. A better resource allocation algorithm with semi-bandit feedback. In *Algorithmic Learning Theory*, pages 268–320. PMLR, 2018.

Ting He, Dennis Goeckel, Ramya Raghavendra, and Don Towsley. Endhost-based shortest path routing in dynamic networks. In *Proc. IEEE INFOCOM*, pages 2202–2210, 2013.

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

Ruiquan Huang, Weiqiang Wu, Jing Yang, and Cong Shen. Federated linear contextual bandits. *Advances in Neural Information Processing Systems*, 34, 2021.

Ramesh Johari, Vijay Kamble, and Yash Kanoria. Matching while learning. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 119–119, 2017.

Nikolai Karpov, Qin Zhang, and Yuan Zhou. Collaborative top distribution identifications with limited interaction. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 160–171. IEEE, 2020.

Tze Leung Lai. Adaptive treatment allocation and the multi-armed bandit problem. *The Annals of Statistics*, pages 1091–1114, 1987.

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

Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard. Distributed cooperative decision-making in multiarmed bandits: Frequentist and bayesian algorithms. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 167–172. IEEE, 2016.

Tor Lattimore, Koby Crammer, and Csaba Szepesvári. Optimal resource allocation with semi-bandit feedback. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, pages 477–486, 2014.

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, 2010.

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

David Martínez-Rubio, Varun Kanade, and Patrick Rebeschini. Decentralized cooperative stochastic bandits. *Advances in Neural Information Processing Systems*, 32, 2019.

Abbas Mehrabian, Etienne Boursier, Emilie Kaufmann, and Vianney Perchet. A practical algorithm for multiplayer bandits when arm means vary among players. In *International Conference on Artificial Intelligence and Statistics*, pages 1211–1221. PMLR, 2020.

Chengshuai Shi and Cong Shen. Federated multi-armed bandits. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

Chengshuai Shi, Cong Shen, and Jing Yang. Federated multi-armed bandits with personalization. In *International Conference on Artificial Intelligence and Statistics*, pages 2917–2925. PMLR, 2021a.

Chengshuai Shi, Wei Xiong, Cong Shen, and Jing Yang. Heterogeneous multi-player multi-armed bandits: Closing the gap and generalization. *Advances in Neural Information Processing Systems*, 34, 2021b.

Chengshuai Shi, Haifeng Xu, Wei Xiong, and Cong Shen. (almost) free incentivized exploration from decentralized learning agents. *Advances in Neural Information Processing Systems*, 34:560–571, 2021c.

Mohammad Sadegh Talebi and Alexandre Proutiere. Learning proportionally fair allocations with low regret. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 2(2):1–31, 2018.

Mohammad Sadegh Talebi, Zhenhua Zou, Richard Combes, Alexandre Proutiere, and Mikael Johansson. Stochastic online shortest path routing: The value of feedback. *IEEE Transactions on Automatic Control*, 63(4):915–930, 2017.

Liang Tang, Yexi Jiang, Lei Li, and Tao Li. Ensemble contextual bandits for personalized recommendation. In *Proceedings of the 8th ACM Conference on Recommender Systems*, pages 73–80, 2014.

Chao Tao, Qin Zhang, and Yuan Zhou. Collaborative learning with limited interaction: Tight bounds for distributed exploration in multi-armed bandits. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 126–146. IEEE, 2019.

Sofía S Villar, Jack Bowden, and James Wason. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2): 199, 2015.

Po-An Wang, Alexandre Proutiere, Kaito Ariu, Yassir Jedra, and Alessio Russo. Optimal algorithms for multiplayer multi-armed bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 4120–4129. PMLR, 2020a.

Yuanhao Wang, Jiachen Hu, Xiaoyu Chen, and Liwei Wang. Distributed bandit learning: Near-optimal regret with efficient communication. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020b.

Lin Yang, Yu-Zhen Janice Chen, Mohammad Hajiesmaili, John C.S. Lui, and Don Towsley. Distributed bandits with heterogeneous agents. In *In Proceedings of The IEEE International Conference on Computer Communications 2022*, 2022.

Zhaowei Zhu, Jingxuan Zhu, Ji Liu, and Yang Liu. Federated bandit: A gossiping approach. In *Abstract Proceedings of the 2021 ACM SIGMETRICS/International Conference on Measurement and Modeling of Computer Systems*, pages 3–4, 2021.

Zhenhua Zou, Alexandre Proutiere, and Mikael Johansson. Online shortest path routing: The value of information. In *2014 American Control Conference*, pages 2142–2147. IEEE, 2014.