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

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

A RELATED WORKS

The most relevant work to us is Yang et al. [2022] which considers a special case of heterogeneous rewards with known agent-specific rewards. In Section 1.2, we discussed in details how the MA2B-HR model covers the AC-MA2B model studied by Yang et al. [2022] as a special case. We provide a tighter regret lower bound and a more efficient algorithm than those given by Yang et al. [2022] as we discover the *free exploration* mechanism while Yang et al. [2022] was not aware of it. Further, Yang et al. [2022] additionally considers an asynchronous action frequencies setting, which our algorithm (with minor modifications) can address as well. We omit this extension in our paper and focus on the heterogeneous arm set setting for clearly presenting the free exploration mechanism and its importance on improving the regret.

The *free exploration* mechanism in cooperation among agents does not make sense when agent specific rewards are unknown, as discussed in Appendix B.1, and hence this setting is not at the core of this paper's interest. Nevertheless, many works study heterogeneous rewards with unknown agent specific rewards [Hossain et al., 2021, Bistritz and Leshem, 2021, Mehrabian et al., 2020, Shi et al., 2021b, Zhu et al., 2021, Shi and Shen, 2021, Shi et al., 2021a, Chen et al., 2018, Shi et al., 2021c] in the MA2B literature, which are intellectually and practically interesting under various settings and goals. Among these works, Bistritz and Leshem [2021], Mehrabian et al. [2020], Shi et al. [2021b] consider the collision model, where agents who pull the same arm at the same time collide and receive zero reward. On the other hand, Zhu et al. [2021], Shi and Shen [2021], Shi et al. [2021a] study the federated learning framework, where the central server aims to learn the global bandit model through the information agents learned from the local bandit models. It is worth noting that, the term "free exploration" is also used by Chen et al. [2018] and Shi et al. [2021c] who study the problem of incentivizing exploration in multi-armed bandit. Specifically, Chen et al. [2018], Shi et al. [2021c] consider a principal who aims to learn the global bandit model offers bonuses to agents to do explorations on the principal's behalf. Chen et al. [2018], Shi et al. [2021c] study the "free exploration" with regard to the principal's cost, while we study the *free exploration* in cooperation among agents in this work. Hence, leveraging the idea of free exploration in a cooperative multi-agent bandit setting is the unique difference of this work with the prior literature on heterogeneous multi-agent bandits. The "free exploration" also differs from another term "exploration-free" recently proposed in contextual bandits Bastani et al. [2021], where their algorithms did not need to deliberately explore arms while our algorithm explores arms without cost. Besides, Jiang and Cheng [2023] also considered a multi-agent bandits model with agent-dependent rewards. The key difference between this work and ours is that their agent-dependent reward mean was disturbed by a zero-mean Gaussian, while ours is by a non-zero-mean Gaussian. Hence, their model does not provide a chance for free exploration as ours.

Homogeneous arm rewards setting [Landgren et al., 2016, Martínez-Rubio et al., 2019, Szorenyi et al., 2013, Landgren et al., 2016, Buccapatnam et al., 2015, Martínez-Rubio et al., 2019], in which an arm generates rewards for all agents from the exact same distribution, is the most extensively studied model in the MA2B literature. It is worth noting that the models of Yang et al. [2021] and Chawla et al. [2020], though may seem close to the AC-MA2B model studied by Yang et al. [2022] at first glance, essentially fall into the category with homogeneous agent-specific reward (see Table 1). Specifically, Yang et al. [2021] considers the heterogeneity of arms in terms of their feedback rather than their rewards. Therefore, in the model of Yang et al. [2021], the reward of an arm is essentially the same for each agent, and the optimal arm is the same one for all

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

agents; hence no room for free exploration. Similarly, there exist a single optimal arm for all agents in the model of Chawla et al. [2020]; hence, Chawla et al. [2020] lets agents update their arm sets, which at the beginning contains different arms, with the goal of eventually containing this optimal arm.

Besides, stochastic rewards with heavy tails [Dubey et al., 2020] and non-stochastic rewards [Bar-On and Mansour, 2019, Cesa-Bianchi et al., 2016] have also been studied in the MA2B literature. Apart from various ways of modeling and assumptions on arm rewards or arm sets, many other variations of MA2B are also studied in the literature. For example, Kolla et al. [2018], Szorenyi et al. [2013], Chawla et al. [2020], Landgren et al. [2016], Buccapatnam et al. [2015], Martínez-Rubio et al. [2019], Bistritz and Bambos [2020], Madhushani et al. [2021], Chakraborty et al. [2017], Cesa-Bianchi et al. [2016], Hillel et al. [2013], Dubey et al. [2020], Yang et al. [2021, 2022], Sankararaman et al. [2019], Féraud et al. [2019] deal with decentralized learning scenarios where agents communicate with each other to improve their performance, while Shi et al. [2021a], Mehrabian et al. [2020], Shi et al. [2021b], Shi and Shen [2021], Wang et al. [2019, 2020], Bar-On and Mansour [2019], Chakraborty et al. [2017], Dubey et al. [2020] consider the models with central servers or leaders that can coordinate the learning process. Many different communication schemes are also considered in the literature, such as immediate broadcasting [Buccapatnam et al., 2015, Yang et al., 2021, 2022], peer-to-peer protocols [Szorenyi et al., 2013], gossip-style communication [Martínez-Rubio et al., 2019, Chawla et al., 2020], etc.

Lastly, there is a similarity between our model and meta bandits [Kveton et al., 2021, Wan et al., 2021], where we assume all agents have the same arm-specific reward distributions but have different agent-specific rewards, while meta bandits assume that all bandit instances are drawn from a common prior (function) but are different realizations. However, there is a key difference between our model and meta bandits: *The agent-specific reward means in our model are known and given, while the reward mean realizations of meta bandits are unknown and randomly drawn from the prior.* Therefore, the meta bandits algorithms—which learn the common prior via multiple instances' random realizations—does not fit in our case (because our agent-specific reward means are given and fixed) and thus cannot be applied to addressing our model.

B ADDITIONAL DISCUSSION ON MODEL

B.1 OTHER POSSIBLE SCENARIOS OF MA2B-HR

Table 1: Different possible scenarios of the MA2B-HR model based on agent-specific reward values.

$\nu^{(i)}(k)$	Homogeneous ($\nu^{(i)}(k) = \nu^{(j)}(k), \forall i, j \in \mathcal{M}$)	Heterogeneous
Unknown	(1) The majority of prior work on MA2B	(2) No useful information to cooperate
Known		(3) This work (Section 2)

While the focus of this paper is on MA2B-HR with known and heterogeneous agent-specific rewards, one can imagine other settings of this model, as outlined in Table 1. In particular, the agent-specific reward mean $\nu^{(i)}(k)$ can be (i) homogeneous or heterogeneous among different agents (i.e., $\nu^{(i)}(k) = \nu^{(j)}(k)$, $\forall i, j \in \mathcal{M}$ or not); (ii) known or unknown by the agent.

Given the above possibilities, there are three scenarios. In scenario (1) agent-specific reward means $\nu^{(i)}(k)$ are identical across all agents, and MA2B-HR reduces to the case where all agents play the same bandit game, which has been studied previously, e.g., in Landgren et al. [2016], Martínez-Rubio et al. [2019]. In scenario (2) agent-specific reward means are heterogeneous and unknown. In this case, cooperation among agents becomes impossible since each agent is essentially solving a different bandit problem and there is no useful information to share between agents. Scenario (3) with heterogeneous and known agent-specific reward means is of interest in this paper.

B.2 THE DIFFERENCE BETWEEN MA2B-HR AND CONTEXTUAL BANDITS

Although the contextual bandits model, e.g., Li et al. [2010], Slivkins [2011], can capture the reward heterogeneity of agents, contextual bandits cannot express the advantage of free exploration as clearly as MA2B-HR. One needs to associate the heterogeneous rewards (agents) with contexts to model the reward heterogeneity via contexts. However, some contexts (agents) that can be utilized to explore some arms freely may arrive rarely, and, therefore, their corresponding free arms cannot be freely explored at most times. For example, in the adversarial arrival setting, these contexts may only arrive a few times, and in the stochastic case, these contexts may arrive with a pretty small probability, e.g., $1/T$. Instead, the modeling of this paper allows agents to sample their local optimal arms and provides room for free exploration. In addition, as we

explain in the next section, several application scenarios can be captured by MA2B-HR and its special cases Yang et al. [2022], Baek and Farias [2021].

C PROOFS

C.1 PRELIMINARY LEMMAS

Lemma C.1 ([Wang et al., 2020, Lemma 3]). *Let $k \in \mathcal{K}$, and $c > 0$. Let H be a random set of rounds such that for all t, $\{t \in H\} \in \mathcal{F}_{t-1}$. Assume that there exists $(C_t)_{t \geqslant 0}$, a sequence of independent binary random variables such that for any $t \geqslant 1$, C_t is $\mathcal{F}_t$-measurable and $\mathbb{P}[C_t = 1] \geqslant c$. Further assume for any $t \in H$, arm k is selected if $C_t = 1$. Then,*

$$\sum_{t \geqslant 1} \mathbb{P}\left[\{t \in H, |\hat{\mu}_t(k) - \mu(k)| \geqslant \varepsilon\}\right] \leqslant 2c^{-1}(2c^{-1} + \varepsilon^{-2}).$$

Lemma C.2 ([Combes et al., 2015, Lemma 6]). *For any arm k, we have*

$$\sum_{t \geqslant 1} \mathbb{P}[d_t(k) \leqslant \mu(k)] \leqslant 15,$$

where $d_t(k)$ is the KL-divergence of arm k at time slot t.

Lemma C.3. *The the KL-divergence of two Gaussian random variables with means μ_1, μ_2 and same variance σ^2 has the following expression,*

$$\mathrm{kl}(\mu_1, \mu_2) = \frac{(\mu_1 - \mu_2)^2}{2\sigma^2}. \tag{1}$$

Lemma C.4 (Abel Transformation). *For two sequences $\{a_n\}$ and $\{b_n\}$ with $n \in \mathbb{N}$, we have*

$$\sum_{n=0}^{N} a_n b_n = \sum_{n=0}^{N} a_n b_N - \sum_{n=0}^{N-1} \sum_{k=0}^{n} a_k (b_{n+1} - b_n).$$

Proof of Lemma C.4. We use induction to prove this lemma. When $N = 0$, we have that $a_0 b_0 = a_0 b_0$ holds.

Suppose when $N = m\,(\in \mathbb{N})$ the above equation holds. We show when $N = m + 1$ the above equation holds as follows,

$$\begin{aligned}
\text{RHS} &= \sum_{n=0}^{m+1} a_n b_{m+1} - \sum_{n=0}^{m} \sum_{k=0}^{n} a_k (b_{n+1} - b_n) \\
&= \sum_{n=0}^{m+1} a_n b_{m+1} - \sum_{n=0}^{m-1} \sum_{k=0}^{n} a_k (b_{n+1} - b_n) - \sum_{k=0}^{m} a_k b_{m+1} + \sum_{k=0}^{m} a_k b_m \\
&= a_{m+1} b_{m+1} + \underbrace{\sum_{k=0}^{m} a_k b_m - \sum_{n=0}^{m-1} \sum_{k=0}^{n} a_k (b_{n+1} - b_n)}_{\text{use the supposition}} \\
&= \sum_{n=0}^{m+1} a_n b_n \\
&= \text{LHS}.
\end{aligned}$$

□

C.2 PROOF OF THEOREM 4.1 (REGRET LOWER BOUND)

Fix an arm k whose $\bar{\Delta}(k) > 0$. Recall that $\bar{i}(k)$ is the agent whose local optimal arm's reward mean is the closest to arm k's. For this agent $\bar{i}(k)$, we define $\theta = (P_\theta(1), P_\theta(2), \ldots, P_\theta(K))$ as an instance of the K arms' reward distributions (arm specific reward plus agent specific reward). Then, we consider another instance $\theta' = (P_{\theta'}(1), P_{\theta'}(2), \ldots, P_{\theta'}(K))$, whose

distributions are the same as instance θ's except for that arm k reward distribution $P_{\theta'}(k)$'s mean is increased by λ, i.e., $\mathbb{E}[P_{\theta'}(k)] = \mathbb{E}[P_\theta(k)] + \lambda$, where $\bar{\Delta}(k) < \lambda < \min_{i \in \mathcal{M} \setminus \{\bar{i}(k)\}} \Delta^{(i)}(k)$. Therefore, in instance θ, the arm k is suboptimal, while in instance θ', arm k becomes the optimal arm. Denote $\mathbb{P}_{\theta,\pi}$ as the probability measure of T-round action-reward histories induced by the interconnection of policy π and the environment θ. Denote $\mathbb{E}_{\theta,\pi}[\mathrm{R}_T(k)]$ as MA2B-HR's regret of pulling the suboptimal arm k under instance θ and policy π, and $N_T(k)$ as the total pulling times of this suboptimal arm k in T time slots. Then, we have

$$
\begin{aligned}
&\mathbb{E}_{\theta,\pi}[\mathrm{R}_T(k)] + \mathbb{E}_{\theta',\pi}[\mathrm{R}_T(k)] \\
&\geqslant \frac{T\bar{\Delta}(k)}{2}\mathbb{P}_{\theta,\pi}\left(N_T(k) \geqslant \frac{T}{2}\right) + \frac{T(\lambda - \bar{\Delta}(k))}{2}\mathbb{P}_{\theta',\pi}\left(N_T(k) < \frac{T}{2}\right) \\
&\geqslant \frac{T}{2}\min\left\{\bar{\Delta}(k), \lambda - \bar{\Delta}(k)\right\}\left(\mathbb{P}_{\theta,\pi}\left(N_T(k) \geqslant \frac{T}{2}\right) + \mathbb{P}_{\theta',\pi}\left(N_T(k) < \frac{T}{2}\right)\right) \\
&\overset{(a)}{\geqslant} \frac{T}{2}\min\left\{\bar{\Delta}(k), \lambda - \bar{\Delta}(k)\right\}\exp\left(-\mathrm{KL}(\mathbb{P}_{\theta,\pi}, \mathbb{P}_{\theta',\pi})\right) \\
&\geqslant \frac{T}{2}\min\left\{\bar{\Delta}(k), \lambda - \bar{\Delta}(k)\right\}\exp\left(-\mathbb{E}_{\theta,\pi}[N_T(k)]\,\mathrm{KL}(P_\theta(k), P_{\theta'}(k))\right),
\end{aligned}
$$

where the inequality (a) is due to the Bretagnolle-Huber inequality [Bretagnolle and Huber, 1978] and the KL represents the KL-divergence between two general probability distributions.

Rearranging the above inequality, we have

$$
\frac{\mathbb{E}_{\theta,\pi}[N_T(k)]}{\log T} \geqslant \frac{1}{\mathrm{KL}(P_\theta(k), P_{\theta'}(k))}\left(1 + \frac{\log\frac{\min\left\{\bar{\Delta}(k), \lambda - \bar{\Delta}(k)\right\}}{2}}{\log T} - \frac{\log(\mathbb{E}_{\theta,\pi}[\mathrm{R}_T(k)] + \mathbb{E}_{\theta',\pi}[\mathrm{R}_T(k)])}{\log T}\right).
$$

When $T \to \infty$, for RHS, the second term inside the bracket is equal to 0 and the third term becomes arbitrarily small since π is a consistent policy. So, we have

$$
\liminf_{T\to\infty} \frac{\mathbb{E}_{\theta,\pi}[N_T(k)]}{\log T} \geqslant \frac{1}{\mathrm{KL}(P_\theta(k), P_{\theta'}(k))}.
$$

Notice that the smallest cost of pulling the suboptimal arm once in instance θ is $\bar{\Delta}(k)$. We transform the pull times of arm k to regret costs of pulling arm k as follows,

$$
\liminf_{T\to\infty} \frac{\mathbb{E}_{\theta,\pi}[\mathrm{R}_T(k)]}{\log T} \geqslant \frac{\bar{\Delta}(k)}{\mathrm{KL}(P_\theta(k), P_{\theta'}(k))} \overset{(a)}{=} \frac{\bar{\Delta}(k)}{\mathrm{kl}(\bar{\omega}(k), \bar{\omega}(k) + \bar{\Delta}(k))},
$$

where kl represents the KL-divergence between two Gaussian distributions with the same variance, and the the equation (a) is by choosing the reward distribution as Gaussian and letting $\lambda \to \bar{\Delta}(k)$ and the definition of $\bar{\omega}(k)$ in (6). Lastly, we take a summation over all arms k whose suboptimality gap $\bar{\Delta}(k) > 0$ and obtain a regret lower bound as follows,

$$
\liminf_{T\to\infty} \frac{\mathbb{E}_{\theta,\pi}[\mathrm{R_T}]}{\log T} \geqslant \sum_{k:\bar{\Delta}(k)>0} \frac{\bar{\Delta}(k)}{\mathrm{kl}(\bar{\omega}(k), \bar{\omega}(k) + \bar{\Delta}(k))}.
$$

C.3 PROOF OF THEOREM 4.3 (REGRET UPPER BOUND)

Fix an agent i. Recall that $I_t^{(i)}$ is the arm index with highest reward empirical mean at time slot t for agent i, and $J_t^{(i)}$ is the arm that agent i pulls in time slot t in FreeExp. We first define two sets of time slots as follows,

$$
\begin{aligned}
\mathcal{A}^{(i)} &:= \{t \geqslant 1 : I_t^{(i)} \neq k_*^{(i)}\}, \\
\mathcal{B}^{(i)} &:= \{t \geqslant 1 : |\hat{\omega}_t^{(i)}(I_t^{(i)}) - \omega^{(i)}(I_t^{(i)})| \geqslant \delta\} = \{t \geqslant 1 : |\hat{\mu}_t(I_t^{(i)}) - \mu(I_t^{(i)})| \geqslant \delta\},
\end{aligned}
$$

where $\mathcal{A}^{(i)}$ denotes a set of time slots in which the empirical optimal arm $I_t^{(i)}$ is not the true optimal arm $k_*^{(i)}$; $\mathcal{B}^{(i)}$ denotes the set of time slots in which the empirical optimal arm $I_t^{(i)}$'s empirical mean is different from its true reward mean by at least a δ.

Denote $\mathcal{T} = \{1, 2, \ldots, T\}$ and $\tilde{\mathcal{T}} \coloneqq \cup_{i\in\mathcal{M}}(\mathcal{A}^{(i)} \cup \mathcal{B}^{(i)})$. We can decompose regret $\mathbb{E}[\mathrm{R}_{\mathrm{T}}(\mathcal{A})]$ as

$$\begin{aligned}
\mathbb{E}[\mathrm{R}_{\mathrm{T}}(\mathcal{A})] &= \mathbb{E}\left[\sum_{i\in\mathcal{M}} \mathrm{R}_T^{(i)}\right] = \mathbb{E}\left[\sum_{i\in\mathcal{M}}\sum_{t\in\mathcal{T}}\sum_{k\in\mathcal{K}^{(i)}} \Delta^{(i)}(k)\mathbb{1}\{J_t^{(i)} = k\}\right] \\
&= \mathbb{E}\left[\sum_{i\in\mathcal{M}}\sum_{t\in\tilde{\mathcal{T}}}\sum_{k\in\mathcal{K}^{(i)}} \Delta^{(i)}(k)\mathbb{1}\{J_t^{(i)} = k\} + \sum_{i\in\mathcal{M}}\sum_{t\in\mathcal{T}\setminus\tilde{\mathcal{T}}}\sum_{k\in\mathcal{K}^{(i)}} \Delta^{(i)}(k)\mathbb{1}\{J_t^{(i)} = k\}\right] \\
&\overset{(a)}{\leqslant} \underbrace{bM\mathbb{E}\left[|\tilde{\mathcal{T}}|\right]}_{(I)} + \underbrace{\mathbb{E}\left[\sum_{i\in\mathcal{M}}\sum_{t\in\mathcal{T}\setminus\tilde{\mathcal{T}}}\sum_{k\in\mathcal{K}^{(i)}\setminus\mathcal{K}^{\mathtt{fr}}} \Delta^{(i)}(k)\mathbb{1}\{J_t^{(i)} = k\}\right]}_{(II)},
\end{aligned} \tag{2}$$

where inequality (a) is due to that (1) the first term is scaled by $\Delta^{(i)}(k) < b$ for all $k \in \mathcal{K}^{(i)}$; (2) when $t \in \mathcal{T} \setminus \tilde{\mathcal{T}}$ all free exploration arms (in $\mathcal{K}^{\mathtt{fr}}$) are correctly identified, so these arms will not be explored with cost in the second term, and we can replace $k \in \mathcal{K}^{(i)}$ with $k \in \mathcal{K}^{(i)} \setminus \mathcal{K}^{\mathtt{fr}}$.

Next, we provide Lemmas C.5 and C.6 to show that the first term (I) is upper bounded by a constant, and Lemmas C.7 and C.8 to show that the second term (II) is upper bounded by a logarithmic term.

Bounding term (I) We define the following two sets,

$$\begin{aligned}
\mathcal{C}^{(i)} &\coloneqq \{t \geqslant 1 : d_t^{(i)}(k_*^{(i)}) < \omega^{(i)}(k_*^{(i)})\}, \\
\mathcal{E}^{(i)} &\coloneqq \{t \in \mathcal{A}^{(i)} \setminus (\mathcal{B}^{(i)} \cup \mathcal{C}^{(i)}), |\hat{\omega}_t^{(i)}(k_*^{(i)}) - \omega^{(i)}(k_*^{(i)})| \geqslant \delta\} \\
&= \{t \in \mathcal{A}^{(i)} \setminus (\mathcal{B}^{(i)} \cup \mathcal{C}^{(i)}), |\hat{\mu}_t(k_*^{(i)}) - \mu(k_*^{(i)})| \geqslant \delta\},
\end{aligned}$$

where $\mathcal{C}^{(i)}$ denotes the set of time slots in which the optimal arm $k_*^{(i)}$'s KL-UCB index is smaller than this optimal arm's true reward mean $\omega^{(i)}(k_*^{(i)})$; $\mathcal{E}^{(i)}$ denotes the subset of time slots of $\mathcal{A}^{(i)} \setminus (\mathcal{B}^{(i)} \cup \mathcal{C}^{(i)})$ and in which the optimal arm $k_*^{(i)}$'s empirical mean is different from its true reward mean by at least a δ.

Lemma C.5 shows the set $\mathcal{A}^{(i)} \cup \mathcal{B}^{(i)}$ can be covered by another set $\mathcal{B}^{(i)} \cup \mathcal{C}^{(i)} \cup \mathcal{E}^{(i)}$. Lemma C.6 separately upper bounds the expected set cardinality of $\mathcal{B}^{(i)}, \mathcal{C}^{(i)}, \mathcal{E}^{(i)}$. We defer both lemmas' proof to the end of this subsection.

Lemma C.5. *$\mathcal{A}^{(i)} \cup \mathcal{B}^{(i)} \subseteq \mathcal{B}^{(i)} \cup \mathcal{C}^{(i)} \cup \mathcal{E}^{(i)}$. Thus, $\mathbb{E}\left[\mathcal{A}^{(i)} \cup \mathcal{B}^{(i)}\right] \leqslant \mathbb{E}\left[\mathcal{B}^{(i)}\right] + \mathbb{E}\left[\mathcal{C}^{(i)}\right] + \mathbb{E}\left[\mathcal{E}^{(i)}\right]$.*

Lemma C.6. $\mathbb{E}\left[|\mathcal{B}^{(i)}|\right] \leqslant 4K^{(i)}(4 + \delta^{-2}), \mathbb{E}\left[|\mathcal{C}^{(i)}|\right] \leqslant 15, \mathbb{E}\left[|\mathcal{E}^{(i)}|\right] \leqslant 4(K^{(i)})^2(4K^{(i)} + \delta^{-2})$.

Given Lemmas C.5 and C.6, we have $\mathbb{E}\left[|\mathcal{A}^{(i)} \cup \mathcal{B}^{(i)}|\right] \leqslant 6(K^{(i)})^2(4K^{(i)} + \delta^{-2})$. Then, we can upper bound the term (I) as follows,

$$\begin{aligned}
(I) &= bM\mathbb{E}\left[|\tilde{\mathcal{T}}|\right] = bM\mathbb{E}\left[\left|\cup_{i\in\mathcal{M}}(\mathcal{A}^{(i)} \cup \mathcal{B}^{(i)})\right|\right] \\
&\leqslant bM\sum_{i\in\mathcal{M}}\mathbb{E}\left[\left|\mathcal{A}^{(i)} \cup \mathcal{B}^{(i)}\right|\right] \leqslant bM^2\max_{i\in\mathcal{M}}\mathbb{E}\left[\left|\mathcal{A}^{(i)} \cup \mathcal{B}^{(i)}\right|\right] \\
&\leqslant 6bM^2K^2(4K + \delta^{-2}).
\end{aligned} \tag{3}$$

Bounding term (II) One key challenge to upper bound (II) is that the pulls of an arm k is among agents with heterogeneous rewards and the costs of pulling arm k can be different among these agents. So, the common technique of bounding the pull times of arm k in current bandits literature is not applicable in bounding these agents' total pulling arm k. To address the challenge, we sort these agents according to their reward gaps $\Delta^{(i)}(k)$ of this arm k, bound the pull times of arm k among a group of incremental agent subsets where these subsets gradually include agents with higher reward gaps $\Delta^{(i)}(k)$, and apply an Abel transformation to bound these agents' total regret cost at last.

Fix an arm $k \in \mathcal{K} \setminus \mathcal{K}^{\mathtt{fr}}$ that cannot be freely explored. Denote $\mathcal{M}(k) \coloneqq \{i \in \mathcal{M} : k \in \mathcal{K}^{(i)}\}$ as the set of agents having access to arm k and $M(k) \coloneqq |\mathcal{M}(k)|$ is the number of such agents. We consider an order $\{i(k;1), i(k;2), \ldots, i(k;M(k))\}$

of those $M(k)$ agents such that $\Delta^{(i(k;1))}(k) \geqslant \Delta^{(i(k;2))}(k) \geqslant \ldots \geqslant \Delta^{(i(k;M(k)))}(k)$. With this order, we rearrange summations of (II) as follows,

$$(II) = \mathbb{E}\left[\sum_{i\in\mathcal{M}}\sum_{t\in\mathcal{T}\setminus\tilde{\mathcal{T}}}\sum_{k\in\mathcal{K}^{(i)}\setminus\mathcal{K}^{\mathtt{fr}}}\Delta^{(i)}(k)\mathbb{1}\{J_t^{(i)}=k\}\right] = \mathbb{E}\left[\sum_{i\in\mathcal{M}}\sum_{k\in\mathcal{K}^{(i)}\setminus\mathcal{K}^{\mathtt{fr}}}\Delta^{(i)}(k)\sum_{t\in\mathcal{T}\setminus\tilde{\mathcal{T}}}\mathbb{1}\{J_t^{(i)}=k\}\right]$$
$$\overset{(a)}{=} \mathbb{E}\left[\sum_{k\in\mathcal{K}\setminus\mathcal{K}^{\mathtt{fr}}}\sum_{m=1}^{M(k)}\Delta^{(i(k;m))}(k)\sum_{t\in\mathcal{T}\setminus\tilde{\mathcal{T}}}\mathbb{1}\{J_t^{(i(k;m))}=k\}\right] \overset{(b)}{=} \mathbb{E}\left[\sum_{k\in\mathcal{K}\setminus\mathcal{K}^{\mathtt{fr}}}\sum_{m=1}^{M(k)}\Delta^{(i(k;m))}(k)n_T^{(i(k;m))}(k)\right],$$

where (a) is because the arm $k\in\mathcal{K}\setminus\mathcal{K}^{\mathtt{fr}}$ is only pulled by agents $\{i(k;1),i(k;2),\ldots,i(k;M(k))\}$, and (b) is from a simplified definition $n_T^{(i(k;m))}(k) \coloneqq \sum_{t\in\mathcal{T}\setminus\tilde{\mathcal{T}}}\mathbb{1}\{J_t^{(i(k;m))}=k\}$.

In Lemma C.7, we provide an intermediate result that bounds the number of times of pulling arm k by agents $\{i(k;1),i(k;2),\ldots,i(k;m)\}$ for any $m\leqslant M(k)$. Lemma C.8 is derived via an Abel transformation and based on Lemma C.7. We defer both lemmas' proof to the end of this subsection.

Lemma C.7. $\mathbb{E}\left[\sum_{\ell=1}^{m} n_T^{(i(k;\ell))}(k)\right] \leqslant \dfrac{\log T + 4\log(\log T)}{\mathrm{kl}(\omega^{(i(k;m))}(k)+\delta, \omega^{(i(k;m))}(k_*^{(i(k;m))})-\delta)} + m(4+2\delta^{-2}).$

Lemma C.8.

$$\mathbb{E}\left[\sum_{m=1}^{M(k)} n_T^{(i(k;m))}(k)\Delta^{(i(k;m))}(k)\right] \leqslant (4+2\delta^{-2})\sum_{m=1}^{M(k)}\Delta_m + \frac{1}{2(\sigma_1^2+\sigma_2^2)}\frac{(\Delta_m-2\delta)(\log T+4\log(\log T))}{\mathrm{kl}(\bar{\omega}(k)+\delta,\bar{\omega}(k)+\bar{\Delta}(k)-\delta)}.$$

In Lemma C.8, we bound the regret cost of pulling arm $k\in\mathcal{K}\setminus\mathcal{K}^{\mathtt{fr}}$ when $t\in\mathcal{T}\setminus\tilde{\mathcal{T}}$ by agents $\{i(k;1),i(k;2),\ldots,i(k;M(k))\}$. Given Lemma C.8, we can upper bound the term (II) as follows,

$$\begin{aligned}(II) &= \sum_{k\in\mathcal{K}\setminus\mathcal{K}^{\mathtt{fr}}}\mathbb{E}\left[\sum_{m=1}^{M(k)}\Delta^{(i(k;m))}(k)n_T^{(i(k;m))}(k)\right]\\ &\leqslant \sum_{k:\bar{\Delta}(k)>0}\frac{1}{2(\sigma_1^2+\sigma_2^2)}\frac{(\Delta_m-2\delta)(\log T+4\log(\log T))}{\mathrm{kl}(\mu(k)+\delta,\mu(k)+\bar{\Delta}(k)-\delta)} + (4+2\delta^{-2})\sum_{k:\bar{\Delta}(k)>0}\sum_{m=1}^{M(k)}\Delta_m \qquad (4)\\ &\leqslant \frac{1}{2(\sigma_1^2+\sigma_2^2)}\sum_{k:\bar{\Delta}(k)>0}\frac{(\Delta_m-2\delta)(\log T+4\log(\log T))}{\mathrm{kl}(\mu(k)+\delta,\mu(k)+\bar{\Delta}(k)-\delta)} + bMK(4+2\delta^{-2}).\end{aligned}$$

Finally, we obtain the regret bound by substituting (3) and (4) into (2) as follows,

$$\mathbb{E}[\mathrm{R_T}(\mathcal{A})] \leqslant \frac{1}{2(\sigma_1^2+\sigma_2^2)}\sum_{k:\bar{\Delta}(k)>0}\frac{(\Delta_m-2\delta)(\log T+4\log(\log T))}{\mathrm{kl}(\mu(k)+\delta,\mu(k)+\bar{\Delta}(k)-\delta)} + 7bM^2K^2(4K+\delta^{-2}) \qquad (5)$$

Letting $T\to\infty,\delta\to 0$, we obtain the asymptotic regret upper bound. *The main proof of Theorem 4.3 is finished.*

In the rest of this section, we present the detailed proofs of Lemmas C.5–C.8.

Proof of Lemma C.5. Denote $t\in\mathcal{A}^{(i)}\setminus(\mathcal{B}^{(i)}\cup\mathcal{C}^{(i)})$. We only need to show that this condition leads to $t\in\mathcal{E}^{(i)}$. From the condition, we have

$$\hat{\omega}_t^{(i)}(k_*^{(i)}) \overset{(a)}{\leqslant} \hat{\omega}_t^{(i)}(I_t^{(i)}) \overset{(b)}{\leqslant} \omega^{(i)}(I_t^{(i)}) + \delta \overset{(c)}{\leqslant} \omega^{(i)}(k_*^{(i)}) - \delta,$$

where the inequality (a) is due to $t\in\mathcal{A}^{(i)}$, inequality (b) is due to $t\notin\mathcal{B}^{(i)}$, and inequality (c) is due to the definition of $\delta < \frac{1}{4}\min_{i\in\mathcal{M},k_1\neq k_2\in\mathcal{K}}|\omega^{(i)}(k_1)-\omega^{(i)}(k_2)|$.

Therefore, we have $\mu(k_*^{(i)}) - \hat{\mu}_t(k_*^{(i)}) = \omega^{(i)}(k_*^{(i)}) - \hat{\omega}_t^{(i)}(k_*^{(i)}) \geqslant \delta$. That is, $t\in\mathcal{E}^{(i)}$. □

Proof of Lemma C.6. To show $\mathbb{E}\left[|\mathcal{B}^{(i)}|\right] \leqslant 4K^{(i)}(4+\delta^{-2})$: for any arm $k \in \mathcal{K}^{(i)}$, denote $\mathcal{B}^{(i)}(k) \coloneqq \{t \geqslant 1 : I_t^{(i)} = k, |\hat{\mu}_t(k) - \mu(k)| \geqslant \delta\}$. Then, applying Lemma C.1 (in Appendix) via letting $H = \{t \geqslant 1 : I_t^{(i)} = k\}, C_t = \mathbb{1}\{J_t^{(i)} = k\}$ and $\mathbb{P}(C_t = 1|H) \geqslant \frac{1}{2}$ (because the agent has a probability of $1/2$ to pull the empirical optimal arm $I_t^{(i)}$), we have $\sum_{t>1}\mathbb{P}(t \in \mathcal{B}^{(i)}(k)) \leqslant 4(4+\delta^{-2})$, that is, $\mathbb{E}\left[|\mathcal{B}^{(i)}(k)|\right] \leqslant 4(4+\delta^{-2})$. Applying union bound over all arms in $\mathcal{K}^{(i)}$, we obtain $\mathbb{E}\left[|\mathcal{B}^{(i)}|\right] \leqslant 4K^{(i)}(4+\delta^{-2})$.

To show $\mathbb{E}\left[|\mathcal{C}^{(i)}|\right] \leqslant 15$: this is from KL-UCB's property in Lemma C.2 in Appendix.

To show $\mathbb{E}\left[|\mathcal{E}^{(i)}|\right] \leqslant 4(K^{(i)})^2(4K^{(i)}+\delta^{-2})$: for any arm $k \in \mathcal{K}^{(i)}$, denote $\mathcal{E}^{(i)}(k) \coloneqq \{t \geqslant 1 : t \in \mathcal{A}^{(i)} \setminus (\mathcal{B}^{(i)} \cup \mathcal{C}^{(i)}), I_t^{(i)} = k, |\hat{\mu}_t(k_*^{(i)}) - \mu(k_*^{(i)})| \geqslant \delta\}$. Applying Lemma C.1 via letting $H = \{t \geqslant 1 : t \in \mathcal{A}^{(i)} \setminus (\mathcal{B}^{(i)} \cup \mathcal{C}^{(i)}), I_t^{(i)} = k\}, C_t = \mathbb{1}\{J_t^{(i)} = k_*^{(i)}\}$ and thus $\mathbb{P}(C_t = 1|H) \geqslant \frac{1}{2K^{(i)}}$ (which will be proven next), we then have $\mathbb{E}\left[|\mathcal{E}^{(i)}(k)|\right] \leqslant 4K^{(i)}(4K^{(i)}+\delta^{-2})$. Then, with union bound over all arms in $\mathcal{K}^{(i)}$, we derive the result.

We now show $\mathbb{P}(C_t = 1|H) \geqslant \frac{1}{2K^{(i)}}$. From the choice of H, we have

$$d_t^{(i)}(k_*^{(i)}) \overset{(a)}{\geqslant} \omega^{(i)}(k_*^{(i)}) \overset{(b)}{\geqslant} \omega^{(i)}(I_t^{(i)}) + \delta \overset{(c)}{\geqslant} \hat{\omega}_t^{(i)}(I_t^{(i)}),$$

where inequality (a) is due to $t \notin \mathcal{C}^{(i)}$, inequality (b) is due to $t \in \mathcal{A}^{(i)}$ and the definition of δ, and inequality (c) is due to $t \notin \mathcal{B}^{(i)}$. Therefore, $d_t^{(i)}(k_*^{(i)}) > \hat{\omega}_t^{(i)}(I_t^{(i)})$, which implies that the agent may explore this arm with a probability of at least $1/2K^{(i)}$ (cf. Algorithm 1's line 11 and if this arm is removed from $\mathcal{D}_t^{(i)}$, then some other agents pull this arm with a probability of at least $1/2$). □

Proof of Lemma C.7. Fix an arm $k \in \mathcal{K} \setminus \mathcal{K}^{\mathtt{fr}}$ and an integer $m \in \{1, 2, \ldots, M(k)\}$. Denote a set of agent-time pairs (i,t) as follows,

$$\mathcal{G}(k;m) \coloneqq \{(i,t) : i \in \{i(k;1), \ldots, i(k;m)\}, t \leqslant T, t \notin \cup_{\ell=1}^{m}(\mathcal{A}^{(i(k;\ell))} \cup \mathcal{B}^{(i(k;\ell))}), J_t^{(i)} = k\}.$$

Let $n_0 \coloneqq (\log T + 4\log(\log T))/\mathrm{kl}(\omega^{(i(k;m))}(k)+\delta, \omega^{(i(k;m))}(k_*^{(i(k;m))}) - \delta)$, and denote $c_t \coloneqq \sum_{s=1}^{t}\sum_{\ell=1}^{m}\mathbb{1}\{(\ell,s) \in \mathcal{G}(k;m)\}$ as the number of times that the agent-time pair (i,t) lies inside $\mathcal{G}(k;m)$. Define two subsets of $\mathcal{G}(k;m)$ as follows,

$$\begin{aligned}\mathcal{G}_1(k;m) &= \{(i,t) \in \mathcal{G}(k;m) : |\hat{\mu}_t(k) - \mu(k)| \geqslant \delta\},\\ \mathcal{G}_2(k;m) &= \{(i,t) \in \mathcal{G}(k;m) : c_t < n_0\}.\end{aligned}$$

We first show that $\mathcal{G}(k;m) \subseteq \mathcal{G}_1(k;m) \cup \mathcal{G}_2(k;m)$. Let $t \in \mathcal{G}(k;m) \setminus (\mathcal{G}_1(k;m) \cup \mathcal{G}_2(k;m))$, from which we have

$$\begin{aligned}d_t^{(i(k;m))}(k) &\overset{(a)}{\geqslant} \hat{\omega}_t^{(i(k;m))}(I_t^{(i(k;m))}) \overset{(b)}{=} \hat{\omega}_t^{(i(k;m))}(k_*^{(i(k;m))})\\ &\overset{(c)}{>} \omega^{(i(k;m))}(k_*^{(i(k;m))}) - \delta \overset{(d)}{>} \omega^{(i(k;m))}(k) + \delta \overset{(e)}{>} \hat{\omega}_t^{(i(k;m))}(k),\end{aligned}$$

where the inequality (a) is due to the definition of KL-UCB $d_t^{(i)}(k)$, the equation (b) is due to $t \notin \mathcal{A}^{(i)}$, the inequality (c) is due to $t \notin \mathcal{B}^{(i)}$, the inequality (d) is due to the definition of δ, the inequality (e) is due to $t \notin \mathcal{G}_1(k;m)$. Recall that $n_t(k)$ is the number of times that arm k has been pulled up to time t. Then, we have

$$\begin{aligned}n_0\,\mathrm{kl}(\hat{\omega}_t^{(i(k;m))}(k), \omega^{(i(k;m))}(k_*^{(i(k;m))}) - \delta) &\overset{(a)}{\leqslant} n_t(k)\,\mathrm{kl}(\hat{\omega}_t^{(i(k;m))}(k), \omega^{(i(k;m))}(k_*^{(i(k;m))}) - \delta)\\ &\overset{(b)}{\leqslant} n_t(k)\,\mathrm{kl}(\hat{\omega}_t^{(i(k;m))}(k), d_t^{(i(k;m))}(k)) \overset{(c)}{\leqslant} \log T + 4\log(\log T),\end{aligned}$$

where the inequality (a) is due to $t \notin \mathcal{G}_2(k;m)$ and thus $n_0 \leqslant c_t \leqslant n_t(k)$, the inequality (b) is due to that $\mathrm{kl}(x,y)$ is increasing for y when $0 < x < y < 1$, the inequality (c) is due to the definition of $d_t^{(i(k;m))}(k)$.

We then substitute n_0 into the above inequality and obtain $\mathrm{kl}(\hat{\omega}_t^{(i(k;m))}(k), \omega^{(i(k;m))}(k_*^{(i(k;m))}) - \delta) \leqslant \mathrm{kl}(\omega^{(i(k;m))}(k) + \delta, \omega^{(i(k;m))}(k_*^{(i(k;m))}) - \delta)$. Noticing $\mathrm{kl}(x,y)$ is decreasing for x when $0 < x < y < 1$, the inequality further leads to $\hat{\omega}_t^{(i(k;m))}(k) \geqslant \omega^{(i(k;m))}(k) + \delta$, which contradicts $t \notin \mathcal{G}_1(k;m)$. From this contradiction, we have $\mathcal{G}(k;m) \subseteq \mathcal{G}_1(k;m) \cup \mathcal{G}_2(k;m)$.

Next, we upper bound $\mathbb{E}\left[|\mathcal{G}_1(k;m)|\right]$ and $\mathbb{E}\left[|\mathcal{G}_2(k;m)|\right]$. To bound $\mathbb{E}\left[|\mathcal{G}_1(k;m)|\right]$, we apply Lemma C.1 for a fixed agent i, let $H = \{(j,t) \in \mathcal{G}_1(k;m) : j = i\}, C_t = c = 1$, and obtain $\mathbb{E}\left[|\mathcal{G}_1(k;m)||\text{agent } i\right] \leqslant 4 + 2\delta^{-2}$. Summing up over all agent in $\{i(k;1), \ldots, i(k;m)\}$, we have $\mathbb{E}\left[|\mathcal{G}_1(k;m)|\right] \leqslant m(4 + 2\delta^{-2})$.

We bound $\mathbb{E}\left[|\mathcal{G}_2(k;m)|\right]$, via its definition as follows,

$$\mathbb{E}\left[|\mathcal{G}_2(k;m)|\right] \leqslant n_0 = \frac{\log T + 4\log(\log T)}{\mathrm{kl}(\omega^{(i(k;m))}(k) + \delta, \omega^{(i(k;m))}(k_*^{(i(k;m))}) - \delta)}.$$

Combining both bounds together, we have, for any $m \leqslant M(k)$ and arm k,

$$\mathbb{E}\left[|\mathcal{G}(k;m)|\right] \leqslant \frac{\log T + 4\log(\log T)}{\mathrm{kl}(\omega^{(i(k;m))}(k) + \delta, \omega^{(i(k;m))}(k_*^{(i(k;m))}) - \delta)} + m(4 + 2\delta^{-2}).$$

Denote $n_t^{(i)}(k)$ as the number of times that agent i pulls arm k in time slots $\{s \leqslant t : s \notin \mathcal{A}^{(i)} \cup \mathcal{B}^{(i)}\}$. The above inequality can be rewritten as

$$\mathbb{E}\left[\sum_{\ell=1}^{m} n_t^{(i(k;\ell))}(k)\right] \leqslant \frac{\log T + 4\log(\log T)}{\mathrm{kl}(\omega^{(i(k;m))}(k) + \delta, \omega^{(i(k;m))}(k_*^{(i(k;m))}) - \delta)} + m(4 + 2\delta^{-2}).$$

□

Proof of Lemma C.8. For the fixed arm k, we first simplify the notations by using $n_m = n_T^{(i^{(m)})}(k)$ and $\Delta_m = \Delta^{i(k;m)}(k)$ as follows,

$$\mathbb{E}\left[\sum_{m=1}^{M(k)} n_T^{(i(k;m))}(k)\Delta^{(i(k;m))}(k)\right] = \sum_{m=1}^{M(k)} \mathbb{E}\left[n_T^{(i(k;m))}(k)\right]\Delta^{(i(k;m))}(k) = \sum_{m=1}^{M(k)} \mathbb{E}\left[n_m\right]\Delta_m.$$

To simplify the result of Lemma C.7 as well, we denote $A_m \coloneqq \frac{\log T + 4\log(\log T)}{\mathrm{kl}(\omega^{(i(k;m))}(k)+\delta, \omega^{(i(k;m))}(k_*^{(i(k;m))})-\delta)}$ and $B_m \coloneqq m(4 + 2\delta^{-2})$. Then, Lemma C.7 becomes $\sum_{\ell}^{m} \mathbb{E}[n_l] \leqslant A_m + B_m$ for all integer $m \leqslant M(k)$.

Next, we rewrite the summation and upper bound it as follows,

$$\begin{aligned}
\sum_{m=1}^{M(k)} \mathbb{E}\left[n_m\right]\Delta_m &= (A_1 + B_1)\Delta_1 + \underbrace{((A_1 + B_1) - \mathbb{E}[n_1])}_{>0, \text{ Lemma C.7}} \underbrace{(\Delta_2 - \Delta_1)}_{<0} + ((A_2 + B_2) - (A_1 + B_1))\Delta_2 \\
&\quad + \underbrace{((A_2 + B_2) - (\mathbb{E}[n_1] + \mathbb{E}[n_2]))}_{>0, \text{ Lemma C.7}} \underbrace{(\Delta_3 - \Delta_2)}_{<0} + ((A_3 + B_3) - (A_2 + B_2))\Delta_3 \\
&\quad + \quad \vdots \\
&\quad + \underbrace{\left((A_{M(k)-1} + B_{M(k)-1}) - \sum_{m=1}^{M(k)-1} \mathbb{E}[n_m]\right)}_{>0, \text{ Lemma C.7}} \underbrace{(\Delta_{M(k)} - \Delta_{M(k)-1})}_{<0} \\
&\quad + \left(\underbrace{\sum_{m=1}^{M(k)} \mathbb{E}[n_m]}_{<A_{M(k)} + B_{M(k)}} - (A_{M(k)-1} + B_{M(k)-1})\right)\Delta_{M(k)} \\
&\leqslant (A_1 + B_1)\Delta_1 + \sum_{m=1}^{M(k)-1} ((A_{m+1} + B_{m+1}) - (A_m + B_m))\Delta_m \\
&= A_1\Delta_1 + \sum_{m=1}^{M(k)-1} (A_{m+1} - A_m)\Delta_m + B_1 \sum_{m=1}^{M(k)} \Delta_m.
\end{aligned}$$

One can bound Δ_m as follows,

$$\Delta_m \overset{(a)}{\leqslant} 2(\Delta_m - 2\delta) = 2((\omega^{(i(k;m))}(k_*^{(i(k;m))}) - \delta) - (\omega^{(i(k;m))}(k) + \delta))$$
$$\overset{(b)}{=} \sqrt{8(\sigma_1^2 + \sigma_2^2)\,\mathrm{kl}(\omega^{(i(k;m))}(k) + \delta, \omega^{(i(k;m))}(k_*^{(i(k;m))}) - \delta)} =: x_m,$$

where the inequality (a) is due to δ's definition, the equation (b) is due to the property of the KL-divergence in Lemma C.3, and we define $x_m := \sqrt{\mathrm{kl}(\omega^{(i(k;m))}(k) + \delta, \omega^{(i(k;m))}(k_*^{(i(k;m))}) - \delta)}$ for simplicity. We can substitute $\Delta_m \leqslant x_m$ into the above inequality and further scale it up as follows,

$$A_1\Delta_1 + \sum_{m=1}^{M(k)-1}(A_{m+1} - A_m)\Delta_m + B_1\sum_{m=1}^{M(k)}\Delta_m \leqslant A_1 x_1 + \sum_{m=1}^{M(k)-1}(A_{m+1} - A_m)x_m + B_1\sum_{m=1}^{M(k)}\Delta_m$$
$$\overset{(a)}{=} \sum_{m=1}^{M(k)-1} A_m(x_m - x_{m+1}) + A_{M(k)}x_{M(k)} + B_1\sum_{m=1}^{M(k)}\Delta_m$$
$$= 8(\sigma_1^2 + \sigma_2^2)\left(\sum_{m=1}^{M(k)-1}\frac{\log T + 4\log(\log T)}{x_m^2}(x_m - x_{m+1}) + \frac{\log T + 4\log(\log T)}{x_{M(k)}^2}x_{M(k)}\right) + (4 + 2\delta^{-2})\sum_{m=1}^{M(k)}\Delta_m$$
$$\leqslant 8(\sigma_1^2 + \sigma_2^2)(\log T + 4\log(\log T))\left(\int_{x_{M(k)}}^{x_1}\frac{1}{x^2}dx + \frac{1}{x_{M(k)}}\right) + (4 + 2\delta^{-2})\sum_{m=1}^{M(k)}\Delta_m$$
$$\leqslant 16(\sigma_1^2 + \sigma_2^2)\frac{\log T + 4\log(\log T)}{x_{M(k)}} + (4 + 2\delta^{-2})\sum_{m=1}^{M(k)}\Delta_m$$
$$\overset{(b)}{=} \frac{16(\sigma_1^2 + \sigma_2^2)}{\sqrt{8(\sigma_1^2 + \sigma_2^2)}}\frac{\sqrt{\mathrm{kl}(\bar{\omega}(k) + \delta, \bar{\omega}(k) + \bar{\Delta}(k) - \delta)}(\log T + 4\log(\log T))}{\mathrm{kl}(\bar{\omega}(k) + \delta, \bar{\omega}(k) + \bar{\Delta}(k) - \delta)} + (4 + 2\delta^{-2})\sum_{m=1}^{M(k)}\Delta_m$$
$$\overset{(c)}{=} \frac{4(\Delta_m - 2\delta)(\log T + 4\log(\log T))}{\mathrm{kl}(\bar{\omega}(k) + \delta, \bar{\omega}(k) + \bar{\Delta}(k) - \delta)} + (4 + 2\delta^{-2})\sum_{m=1}^{M(k)}\Delta_m,$$

where equation (a) applies the Abel transformation in Lemma C.4 in Appendix, equation (b) is by $\omega^{(i(k;M(k)))}(k_*^{(i(k;M(k)))}) = \bar{\omega}(k) + \bar{\Delta}(k)$ where $\bar{\omega}(k) = \omega^{(\bar{i}(k))}(k)$ and $\bar{i}(k) = i(k; M(k))$, and equation (c) is due to the KL-divergence's property in Lemma C.3. The above inequality upper bounds the regret cost paying for the suboptimal arm k. □

References

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

Yogev Bar-On and Yishay Mansour. Individual regret in cooperative nonstochastic multi-armed bandits. *Advances in Neural Information Processing Systems*, 32, 2019.

Hamsa Bastani, Mohsen Bayati, and Khashayar Khosravi. Mostly exploration-free algorithms for contextual bandits. *Management Science*, 67(3):1329–1349, 2021.

Ilai Bistritz and Nicholas Bambos. Cooperative multi-player bandit optimization. *Advances in Neural Information Processing Systems*, 33:2016–2027, 2020.

Ilai Bistritz and Amir Leshem. Game of thrones: Fully distributed learning for multiplayer bandits. *Mathematics of Operations Research*, 46(1):159–178, 2021.

Jean Bretagnolle and Catherine Huber. Estimation des densités: risque minimax. *Séminaire de probabilités de Strasbourg*, 12:342–363, 1978.

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

Nicol'o Cesa-Bianchi, Claudio Gentile, Yishay Mansour, and Alberto Minora. Delay and cooperation in nonstochastic bandits. In *Conference on Learning Theory*, pages 605–622. PMLR, 2016.

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

Ronshee Chawla, Abishek Sankararaman, Ayalvadi Ganesh, and Sanjay Shakkottai. The gossiping insert-eliminate algorithm for multi-agent bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3471–3481. PMLR, 2020.

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

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.

Abhimanyu Dubey et al. Cooperative multi-agent bandits with heavy tails. In *International Conference on Machine Learning*, pages 2730–2739. PMLR, 2020.

Raphaël Féraud, Réda Alami, and Romain Laroche. Decentralized exploration in multi-armed bandits. In *International Conference on Machine Learning*, pages 1901–1909. PMLR, 2019.

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.

Safwan Hossain, Evi Micha, and Nisarg Shah. Fair algorithms for multi-agent multi-armed bandits. *Advances in Neural Information Processing Systems*, 34:24005–24017, 2021.

Fan Jiang and Hui Cheng. Multi-agent bandit with agent-dependent expected rewards. *Swarm Intelligence*, pages 1–33, 2023.

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

Branislav Kveton, Mikhail Konobeev, Manzil Zaheer, Chih-wei Hsu, Martin Mladenov, Craig Boutilier, and Csaba Szepesvari. Meta-thompson sampling. In *International Conference on Machine Learning*, pages 5884–5893. PMLR, 2021.

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.

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.

Udari Madhushani, Abhimanyu Dubey, Naomi Leonard, and Alex Pentland. One more step towards reality: Cooperative bandits with imperfect communication. *Advances in Neural Information Processing Systems*, 34:7813–7824, 2021.

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.

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

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.

Aleksandrs Slivkins. Contextual bandits with similarity information. In *Proceedings of the 24th annual Conference On Learning Theory*, pages 679–702. JMLR Workshop and Conference Proceedings, 2011.

Balazs Szorenyi, Róbert Busa-Fekete, István Hegedus, Róbert Ormándi, Márk Jelasity, and Balázs Kégl. Gossip-based distributed stochastic bandit algorithms. In *International Conference on Machine Learning*, pages 19–27. PMLR, 2013.

Runzhe Wan, Lin Ge, and Rui Song. Metadata-based multi-task bandits with bayesian hierarchical models. *Advances in Neural Information Processing Systems*, 34:29655–29668, 2021.

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, 2020.

Yuanhao Wang, Jiachen Hu, Xiaoyu Chen, and Liwei Wang. Distributed bandit learning: Near-optimal regret with efficient communication. In *International Conference on Learning Representations*, 2019.

Lin Yang, Yu-Zhen Janice Chen, Stephen Pasteris, Mohammad Hajiesmaili, John Lui, and Don Towsley. Cooperative stochastic bandits with asynchronous agents and constrained feedback. *Advances in Neural Information Processing Systems*, 34:8885–8897, 2021.

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.